О сходимости методов типа стохастического градиентного спуска для выпуклых и невыпуклых задач оптимизации

Общероссийский семинар по оптимизации

Эдуард Горбунов

Московский Физико-Технический Институт

В конце презентации добавлены некоторые ссылки, которые упоминались в докладе

8 июля 2020

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(\mathbf{x}, \xi_i)]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(\mathbf{x}, \xi_i)]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(\mathbf{x}, \xi_i)]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(\mathbf{x}, \xi_i)]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[f(\mathbf{x}, \xi)\right]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[f_i(\mathbf{x}, \xi_i)\right]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(\mathbf{x}, \xi_i)]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[f(\mathbf{x}, \xi)\right]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[f_i(\mathbf{x}, \xi_i)\right]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[f(\mathbf{x}, \xi)\right]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[f_i(\mathbf{x}, \xi_i)\right]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Задачи стохастической оптимизации и задачи минимизации суммы функций

Рассмотрим задачу с регуляризацией

$$F(\mathbf{x}) = f(\mathbf{x}) + R(\mathbf{x}) \to \min_{\mathbf{x} \in \mathbb{R}^n}. \tag{1}$$

- Функция f является L-гладкой и может иметь вид:
  - $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}\left[f(\mathbf{x}, \xi)\right]$,
  - $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$ и $f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[f_i(\mathbf{x}, \xi_i)\right]$ или $f_i(\mathbf{x}) = \frac{1}{t}\sum\limits_{j=1}^{t} f_{ij}(\mathbf{x})$
- $R(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ — правильная выпуклая замкнутая функция. Здесь
  - правильная функция = функция, которая не всюду равна $+\infty$,
  - замкнутая функция = функция, у которой множества уровня замкнуты, т.е. для всех $\alpha \in \mathbb{R}$ множество $\{\mathbf{x} \in \mathbb{R}^n \mid R(\mathbf{x}) \leq \alpha\}$ замкнуто.

Стохастические градиентные методы — естественный и очень популярный выбор для решения таких задач.

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x}, \xi)$: $\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m} \sum\limits_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1, 2, \ldots, m\}$: $\mathbb{E}_j[\nabla f_j(\mathbf{x})] = \sum\limits_{i=1}^{m} \nabla f_i(\mathbf{x}) \mathbb{P}\{j = i\} = \frac{1}{m} \sum\limits_{i=1}^{m} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$.
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle \nabla f(\mathbf{x}), e_j \rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1, 2, \ldots, n\}$ и $\{e_1, \ldots, e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j[\langle \nabla f(\mathbf{x}), e_j \rangle e_j] = \frac{1}{n} \sum\limits_{i=1}^{n} \langle \nabla f(\mathbf{x}), e_i \rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \mathcal{D}}[f(\mathbf{x}, \xi)]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x}, \xi)$: $\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1, 2, \ldots, m\}$: $\mathbb{E}_j[\nabla f_j(\mathbf{x})] = \sum_{i=1}^{m} \nabla f_i(\mathbf{x}) \mathbb{P}\{j = i\} = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$.
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle \nabla f(\mathbf{x}), e_j \rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1, 2, \ldots, n\}$ и $\{e_1, \ldots, e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j[\langle \nabla f(\mathbf{x}), e_j \rangle e_j] = \frac{1}{n} \sum_{i=1}^{n} \langle \nabla f(\mathbf{x}), e_i \rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(\mathbf{x},\xi)]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x},\xi)$: $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(\mathbf{x},\xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,m\}$:
$$\mathbb{E}_j\left[\nabla f_j(\mathbf{x})\right] = \sum_{i=1}^{m}\nabla f_i(\mathbf{x})\mathbb{P}\{j=i\} = \frac{1}{m}\sum_{i=1}^{m}\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x}).$$
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle\nabla f(\mathbf{x}), e_j\rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,n\}$ и $\{e_1,\ldots,e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j\left[\langle\nabla f(\mathbf{x}), e_j\rangle e_j\right] = \frac{1}{n}\sum\limits_{i=1}^{n}\langle\nabla f(\mathbf{x}), e_i\rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(\mathbf{x},\xi)\right]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x},\xi)$: $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(\mathbf{x},\xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,m\}$: $\mathbb{E}_j\left[\nabla f_j(\mathbf{x})\right] = \sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x})\mathbb{P}\left\{j=i\right\} = \frac{1}{m}\sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$.
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle\nabla f(\mathbf{x}), e_j\rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,n\}$ и $\{e_1,\ldots,e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j\left[\langle\nabla f(\mathbf{x}), e_j\rangle e_j\right] = \frac{1}{n}\sum\limits_{i=1}^{n}\langle\nabla f(\mathbf{x}), e_i\rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(\mathbf{x},\xi)\right]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x},\xi)$: $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(\mathbf{x},\xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,m\}$: $\mathbb{E}_j\left[\nabla f_j(\mathbf{x})\right] = \sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x})\mathbb{P}\left\{j=i\right\} = \frac{1}{m}\sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$.
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle\nabla f(\mathbf{x}), e_j\rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,n\}$ и $\{e_1,\ldots,e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j\left[\langle\nabla f(\mathbf{x}), e_j\rangle e_j\right] = \frac{1}{n}\sum\limits_{i=1}^{n}\langle\nabla f(\mathbf{x}), e_i\rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(\mathbf{x},\xi)]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x},\xi)$: $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(\mathbf{x},\xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,m\}$: $\mathbb{E}_j[\nabla f_j(\mathbf{x})] = \sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x})\mathbb{P}\{j=i\} = \frac{1}{m}\sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$.
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle\nabla f(\mathbf{x}), e_j\rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,n\}$ и $\{e_1,\ldots,e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j[\langle\nabla f(\mathbf{x}), e_j\rangle e_j] = \frac{1}{n}\sum\limits_{i=1}^{n}\langle\nabla f(\mathbf{x}), e_i\rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

- Если $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f(\mathbf{x},\xi)\right]$, то в качестве стох. градиента можно выбрать $\nabla f(\mathbf{x},\xi)$: $\mathbb{E}_{\xi\sim\mathcal{D}}[\nabla f(\mathbf{x},\xi)] = \nabla f(\mathbf{x})$.
- Если $f(\mathbf{x}) = \frac{1}{m}\sum\limits_{i=1}^{m} f_i(\mathbf{x})$, то в качестве стохастического градиента можно выбрать $\nabla f_j(\mathbf{x})$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,m\}$: $\mathbb{E}_j\left[\nabla f_j(\mathbf{x})\right] = \sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x})\mathbb{P}\{j=i\} = \frac{1}{m}\sum\limits_{i=1}^{m}\nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x})$.
- Если размерность задачи большая и подсчёт частной производной обходится гораздо «дешевле», чем подсчёт полного градиента, то имеет смысл использовать покоординатные методы. Например, в качестве стох. градиента можно выбирать $\langle\nabla f(\mathbf{x}), e_j\rangle e_j$, где j выбирается случайно и равновероятно из множества $\{1,2,\ldots,n\}$ и $\{e_1,\ldots,e_n\}$ — стандартный базис в $\mathbb{R}^n$: $\mathbb{E}_j\left[\langle\nabla f(\mathbf{x}), e_j\rangle e_j\right] = \frac{1}{n}\sum\limits_{i=1}^{n}\langle\nabla f(\mathbf{x}), e_i\rangle e_i = \nabla f(\mathbf{x})$

Стохастический градиент

Во всех рассмотренных примерах предлагается использовать некоторый случайный вектор $\mathbf{g}(\mathbf{x})$, который является несмещённой оценкой градиента в точке $\mathbf{x}$:

$$\mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] = \nabla f(\mathbf{x}).$$

Существует бесконечно много способов сформировать такой вектор $\mathbf{g}(\mathbf{x})$.

Стохастический градиент

Во всех рассмотренных примерах предлагается использовать некоторый случайный вектор $\mathbf{g}(\mathbf{x})$, который является несмещённой оценкой градиента в точке $\mathbf{x}$:

$$\mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] = \nabla f(\mathbf{x}).$$

Существует бесконечно много способов сформировать такой вектор $\mathbf{g}(\mathbf{x})$.

Стохастический градиент

Во всех рассмотренных примерах предлагается использовать некоторый случайный вектор $\mathbf{g}(\mathbf{x})$, который является несмещённой оценкой градиента в точке $\mathbf{x}$:

$$\mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] = \nabla f(\mathbf{x}).$$

Существует бесконечно много способов сформировать такой вектор $\mathbf{g}(\mathbf{x})$.

Как выбирать стохастический градиент?

Ответ зависит от того, что мы хотим:

- сделать итерации метода «быстрыми»?
- быстрое время достижения не очень высокой точности решения?
- быстрое время достижения высокой точности решения?
- возможность эффективного распараллеливания вычислений?

Как выбирать стохастический градиент?

Ответ зависит от того, что мы хотим:

- сделать итерации метода «быстрыми»?
- быстрое время достижения не очень высокой точности решения?
- быстрое время достижения высокой точности решения?
- возможность эффективного распараллеливания вычислений?

Как выбирать стохастический градиент?

Ответ зависит от того, что мы хотим:

- сделать итерации метода «быстрыми»?
- быстрое время достижения не очень высокой точности решения?
- быстрое время достижения высокой точности решения?
- возможность эффективного распараллеливания вычислений?

Как выбирать стохастический градиент?

Ответ зависит от того, что мы хотим:

- сделать итерации метода «быстрыми»?
- быстрое время достижения не очень высокой точности решения?
- быстрое время достижения высокой точности решения?
- возможность эффективного распараллеливания вычислений?

Как выбирать стохастический градиент?

Ответ зависит от того, что мы хотим:

- сделать итерации метода «быстрыми»?
- быстрое время достижения не очень высокой точности решения?
- быстрое время достижения высокой точности решения?
- возможность эффективного распараллеливания вычислений?

Проксимальный стохастический градиентный спуск: prox-SGD

На шаге k:

$$\text{Вычислить } \mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),$$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma^k R}\left(\mathbf{x}^k - \gamma^k \mathbf{g}^k\right), \tag{2}$$

где

$$\operatorname{prox}_R(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n}\left\{R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\right\} \text{ — проксимальный оператор.} \tag{3}$$

Некоторые свойства проксимального оператора:

1. $\mathbf{x}^* = \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$, где $\mathbf{x}^*$ — решение задачи (1)
2. $\|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2 \le \|\mathbf{x}-\mathbf{y}\|_2$
3. $\langle \mathbf{x}-\mathbf{y}, \operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\rangle \ge \|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2^2$.

Проксимальный стохастический градиентный спуск: prox-SGD

На шаге k:

$$\text{Вычислить } \mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),$$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma^k R}\left(\mathbf{x}^k - \gamma^k \mathbf{g}^k\right), \tag{2}$$

где

$$\operatorname{prox}_R(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n}\left\{R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\right\} \text{ — проксимальный оператор.} \tag{3}$$

Некоторые свойства проксимального оператора:

1. $\mathbf{x}^* = \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$, где $\mathbf{x}^*$ — решение задачи (1)
2. $\|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2 \le \|\mathbf{x}-\mathbf{y}\|_2$
3. $\langle \mathbf{x}-\mathbf{y}, \operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\rangle \ge \|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2^2$.

Проксимальный стохастический градиентный спуск: prox-SGD

На шаге k:

$$\text{Вычислить } \mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),$$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma^k R}\left(\mathbf{x}^k - \gamma^k \mathbf{g}^k\right), \tag{2}$$

где

$$\operatorname{prox}_R(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n}\left\{R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\right\} \text{ — проксимальный оператор.} \tag{3}$$

Некоторые свойства проксимального оператора:

1. $\mathbf{x}^* = \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$, где $\mathbf{x}^*$ — решение задачи (1)
2. $\|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2 \le \|\mathbf{x}-\mathbf{y}\|_2$
3. $\langle \mathbf{x}-\mathbf{y}, \operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\rangle \ge \|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2^2$.

Проксимальный стохастический градиентный спуск: prox-SGD

На шаге k:

$$\text{Вычислить } \mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),$$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma^k R}\left(\mathbf{x}^k - \gamma^k \mathbf{g}^k\right), \tag{2}$$

где

$$\operatorname{prox}_R(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n}\left\{R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\right\} \text{ — проксимальный оператор.} \tag{3}$$

Некоторые свойства проксимального оператора:

1. $\mathbf{x}^* = \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$, где $\mathbf{x}^*$ — решение задачи (1)
2. $\|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2 \leq \|\mathbf{x}-\mathbf{y}\|_2$
3. $\langle \mathbf{x}-\mathbf{y}, \operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\rangle \geq \|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2^2$.

Проксимальный стохастический градиентный спуск: prox-SGD

На шаге k:

$$\text{Вычислить } \mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),$$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma^k R}\left(\mathbf{x}^k - \gamma^k \mathbf{g}^k\right), \tag{2}$$

где

$$\operatorname{prox}_R(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n}\left\{R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\right\} \text{ — проксимальный оператор.} \tag{3}$$

Некоторые свойства проксимального оператора:

1. $\mathbf{x}^* = \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$, где $\mathbf{x}^*$ — решение задачи (1)
2. $\|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2 \le \|\mathbf{x}-\mathbf{y}\|_2$
3. $\langle \mathbf{x}-\mathbf{y}, \operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\rangle \ge \|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2^2$.

Проксимальный стохастический градиентный спуск: prox-SGD

На шаге k:

$$\text{Вычислить } \mathbf{g}^k = \mathbf{g}(\mathbf{x}^k),$$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma^k R}\left(\mathbf{x}^k - \gamma^k \mathbf{g}^k\right), \tag{2}$$

где

$$\operatorname{prox}_R(\mathbf{x}) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^n}\left\{R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2\right\} \text{ — проксимальный оператор.} \tag{3}$$

Некоторые свойства проксимального оператора:

1. $\mathbf{x}^* = \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma\nabla f(\mathbf{x}^*))$, где $\mathbf{x}^*$ — решение задачи (1)
2. $\|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2 \le \|\mathbf{x}-\mathbf{y}\|_2$
3. $\langle \mathbf{x}-\mathbf{y}, \operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\rangle \ge \|\operatorname{prox}_R(\mathbf{x}) - \operatorname{prox}_R(\mathbf{y})\|_2^2$.

План лекции

- Методы для решения (сильно) выпуклых задач
- Методы для решения невыпуклых задач
- Over-parameterized модели машинного обучения

План лекции

- Методы для решения (сильно) выпуклых задач
- Методы для решения невыпуклых задач
- Over-parameterized модели машинного обучения

План лекции

- Методы для решения (сильно) выпуклых задач
- Методы для решения невыпуклых задач
- Over-parameterized модели машинного обучения

Предположения и обозначения

Помимо выпуклости и дифференцируемости функции f мы используем следующее предположение.

$(\mu, \mathbf{x}^*)$-квази сильная выпуклость

Мы предполагаем, что функция f является $(\mu, \mathbf{x}^*)$-квази сильно выпуклой, т.е. для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется следующее неравенство:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2$$

Дивергенция Брегмана

Дивергенцией Брегмана относительно дифференцируемой функции f будем называть

$$V_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

Предположения и обозначения

Помимо выпуклости и дифференцируемости функции f мы используем следующее предположение.

$(\mu, \mathbf{x}^*)$-квази сильная выпуклость

Мы предполагаем, что функция f является $(\mu, \mathbf{x}^*)$-квази сильно выпуклой, т.е. для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется следующее неравенство:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2$$

Дивергенция Брегмана

Дивергенцией Брегмана относительно дифференцируемой функции f будем называть

$$V_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

Предположения и обозначения

Помимо выпуклости и дифференцируемости функции f мы используем следующее предположение.

$(\mu, \mathbf{x}^*)$-квази сильная выпуклость

Мы предполагаем, что функция f является $(\mu, \mathbf{x}^*)$-квази сильно выпуклой, т.е. для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется следующее неравенство:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$$

Дивергенция Брегмана

Дивергенцией Брегмана относительно дифференцируемой функции f будем называть

$$V_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

Предположения и обозначения

Помимо выпуклости и дифференцируемости функции f мы используем следующее предположение.

$(\mu, \mathbf{x}^*)$-квази сильная выпуклость

Мы предполагаем, что функция f является $(\mu, \mathbf{x}^*)$-квази сильно выпуклой, т.е. для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется следующее неравенство:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2} \|\mathbf{x} - \mathbf{x}^*\|_2^2$$

Дивергенция Брегмана

Дивергенцией Брегмана относительно дифференцируемой функции f будем называть

$$V_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

Предположения и обозначения

Помимо выпуклости и дифференцируемости функции f мы используем следующее предположение.

$(\mu, \mathbf{x}^*)$-квази сильная выпуклость

Мы предполагаем, что функция f является $(\mu, \mathbf{x}^*)$-квази сильно выпуклой, т.е. для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется следующее неравенство:

$$f(\mathbf{x}^*) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$$

Дивергенция Брегмана

Дивергенцией Брегмана относительно дифференцируемой функции f будем называть

$$V_f(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$$

Попытка проанализировать prox-SGD

$$\begin{aligned}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(\mathbf{x}^k - \gamma \mathbf{g}^k) - \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))\|_2^2 \\ &\leq \|\mathbf{x}^k - \mathbf{x}^* - \gamma\left(\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right)\|_2^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{g}^k - \nabla f(\mathbf{x}^*)\rangle + \gamma^2\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2.\end{aligned}$$

Возьмём теперь условное математическое ожидание $\mathbb{E}\left[\cdot \mid \mathbf{x}^k\right]$ от левой и правой частей:

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

Попытка проанализировать prox-SGD

$$\begin{aligned}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(\mathbf{x}^k - \gamma \mathbf{g}^k) - \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))\|_2^2 \\ &\leq \|\mathbf{x}^k - \mathbf{x}^* - \gamma\left(\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right)\|_2^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{g}^k - \nabla f(\mathbf{x}^*)\rangle + \gamma^2\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2.\end{aligned}$$

Возьмём теперь условное математическое ожидание $\mathbb{E}\left[\cdot \mid \mathbf{x}^k\right]$ от левой и правой частей:

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

Попытка проанализировать prox-SGD

$$\begin{aligned}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(\mathbf{x}^k-\gamma\mathbf{g}^k)-\operatorname{prox}_{\gamma R}(\mathbf{x}^*-\gamma\nabla f(\mathbf{x}^*))\|_2^2\\ &\leq \|\mathbf{x}^k-\mathbf{x}^*-\gamma\left(\mathbf{g}^k-\nabla f(\mathbf{x}^*)\right)\|_2^2\\ &= \|\mathbf{x}^k-\mathbf{x}^*\|_2^2-2\gamma\langle\mathbf{x}^k-\mathbf{x}^*,\mathbf{g}^k-\nabla f(\mathbf{x}^*)\rangle+\gamma^2\|\mathbf{g}^k-\nabla f(\mathbf{x}^*)\|_2^2.\end{aligned}$$

Возьмём теперь условное математическое ожидание $\mathbb{E}\left[\cdot \mid \mathbf{x}^k\right]$ от левой и правой частей:

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

Попытка проанализировать prox-SGD

$$\begin{aligned}\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(\mathbf{x}^k-\gamma\mathbf{g}^k)-\operatorname{prox}_{\gamma R}(\mathbf{x}^*-\gamma\nabla f(\mathbf{x}^*))\|_2^2\\ &\le \|\mathbf{x}^k-\mathbf{x}^*-\gamma\left(\mathbf{g}^k-\nabla f(\mathbf{x}^*)\right)\|_2^2\\ &= \|\mathbf{x}^k-\mathbf{x}^*\|_2^2-2\gamma\langle\mathbf{x}^k-\mathbf{x}^*,\mathbf{g}^k-\nabla f(\mathbf{x}^*)\rangle+\gamma^2\|\mathbf{g}^k-\nabla f(\mathbf{x}^*)\|_2^2.\end{aligned}$$

Возьмём теперь условное математическое ожидание $\mathbb{E}\left[\cdot \mid \mathbf{x}^k\right]$ от левой и правой частей:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] &\le \|\mathbf{x}^k-\mathbf{x}^*\|_2^2-2\gamma\langle\mathbf{x}^k-\mathbf{x}^*,\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right]-\nabla f(\mathbf{x}^*)\rangle+\gamma^2\mathbb{E}\left[\|\mathbf{g}^k-\nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right]\\ &\le \|\mathbf{x}^k-\mathbf{x}^*\|_2^2-2\gamma\langle\mathbf{x}^k-\mathbf{x}^*,\nabla f(\mathbf{x}^k)-\nabla f(\mathbf{x}^*)\rangle+\gamma^2\mathbb{E}\left[\|\mathbf{g}^k-\nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].\end{aligned}$$

Попытка проанализировать prox-SGD

$$\begin{aligned}\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 &= \|\operatorname{prox}_{\gamma R}(\mathbf{x}^k - \gamma \mathbf{g}^k) - \operatorname{prox}_{\gamma R}(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))\|_2^2 \\ &\leq \|\mathbf{x}^k - \mathbf{x}^* - \gamma\left(\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right)\|_2^2 \\ &= \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma\langle \mathbf{x}^k - \mathbf{x}^*, \mathbf{g}^k - \nabla f(\mathbf{x}^*)\rangle + \gamma^2\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2.\end{aligned}$$

Возьмём теперь условное математическое ожидание $\mathbb{E}\left[\cdot \mid \mathbf{x}^k\right]$ от левой и правой частей:

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

Попытка проанализировать prox-SGD

Используя $(\mu, \mathbf{x}^*)$-квази сильную выпуклость, оценим второе слагаемое:

$$-\langle \mathbf{x}^k - \mathbf{x}^*, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*) \rangle \leq -\frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - V_f(\mathbf{x}^k, \mathbf{x}^*).$$

Отсюда мы получаем, что

$$\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad (1 - \gamma\mu)\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma V_f(\mathbf{x}^k, \mathbf{x}^*) + \gamma^2 \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].$$

Как оценить последнее слагаемое? Из сделанных нами предположений о $\mathbf{g}^k$ никаких оценок на это слагаемое не следует.

Попытка проанализировать prox-SGD

Используя $(\mu, \mathbf{x}^*)$-квази сильную выпуклость, оценим второе слагаемое:

$$-\langle \mathbf{x}^k - \mathbf{x}^*, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*) \rangle \leq -\frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - V_f(\mathbf{x}^k, \mathbf{x}^*).$$

Отсюда мы получаем, что

$$\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad (1 - \gamma\mu)\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma V_f(\mathbf{x}^k, \mathbf{x}^*) + \gamma^2 \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].$$

Как оценить последнее слагаемое? Из сделанных нами предположений о $\mathbf{g}^k$ никаких оценок на это слагаемое не следует.

Попытка проанализировать prox-SGD

Используя $(\mu, \mathbf{x}^*)$-квази сильную выпуклость, оценим второе слагаемое:

$$-\langle \mathbf{x}^k - \mathbf{x}^*, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*) \rangle \leq -\frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - V_f(\mathbf{x}^k, \mathbf{x}^*).$$

Отсюда мы получаем, что

$$\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad (1 - \gamma\mu)\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma V_f(\mathbf{x}^k, \mathbf{x}^*) + \gamma^2 \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].$$

Как оценить последнее слагаемое? Из сделанных нами предположений о $\mathbf{g}^k$ никаких оценок на это слагаемое не следует.

Попытка проанализировать prox-SGD

Используя $(\mu, \mathbf{x}^*)$-квази сильную выпуклость, оценим второе слагаемое:

$$-\langle \mathbf{x}^k - \mathbf{x}^*, \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*) \rangle \leq -\frac{\mu}{2}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - V_f(\mathbf{x}^k, \mathbf{x}^*).$$

Отсюда мы получаем, что

$$\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad (1 - \gamma\mu)\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma V_f(\mathbf{x}^k, \mathbf{x}^*) + \gamma^2 \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].$$

Как оценить последнее слагаемое? Из сделанных нами предположений о $\mathbf{g}^k$ никаких оценок на это слагаемое не следует.

Основное параметрическое предположение

Gorbunov, E., Hanzely, F., & Richtárik, P. (2020, June). **A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent**. *In International Conference on Artificial Intelligence and Statistics*, pp. 680-690 *(AISTATS 2020)*.

Предположение 1

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая `prox-SGD`.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы A, B, C, D_1, D_2, ρ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2AV_f(\mathbf{x}^k, \mathbf{x}^*) + B\sigma_k^2 + D_1, \tag{4}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k, \mathbf{x}^*) + D_2. \tag{5}$$

Возникает вопрос: удовлетворяет ли этому предположению хотя бы один алгоритм?

Основное параметрическое предположение

Gorbunov, E., Hanzely, F., & Richtárik, P. (2020, June). **A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent**. *In International Conference on Artificial Intelligence and Statistics*, pp. 680-690 *(AISTATS 2020)*.

Предположение 1

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая `prox-SGD`.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы A, B, C, D_1, D_2, ρ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2AV_f(\mathbf{x}^k, \mathbf{x}^*) + B\sigma_k^2 + D_1, \tag{4}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k, \mathbf{x}^*) + D_2. \tag{5}$$

Возникает вопрос: удовлетворяет ли этому предположению хотя бы один алгоритм?

Основное параметрическое предположение

Gorbunov, E., Hanzely, F., & Richtárik, P. (2020, June). **A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent**. *In International Conference on Artificial Intelligence and Statistics*, pp. 680-690 *(AISTATS 2020)*.

Предположение 1

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая `prox-SGD`.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы A, B, C, D_1, D_2, ρ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2AV_f(\mathbf{x}^k, \mathbf{x}^*) + B\sigma_k^2 + D_1, \tag{4}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k, \mathbf{x}^*) + D_2. \tag{5}$$

Возникает вопрос: удовлетворяет ли этому предположению хотя бы один алгоритм?

Основное параметрическое предположение

Gorbunov, E., Hanzely, F., & Richtárik, P. (2020, June). **A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent**. *In International Conference on Artificial Intelligence and Statistics*, pp. 680-690 *(AISTATS 2020)*.

Предположение 1

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая `prox-SGD`.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы A, B, C, D_1, D_2, ρ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2AV_f(\mathbf{x}^k, \mathbf{x}^*) + B\sigma_k^2 + D_1, \tag{4}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k, \mathbf{x}^*) + D_2. \tag{5}$$

Возникает вопрос: удовлетворяет ли этому предположению хотя бы один алгоритм?

Основное параметрическое предположение

Gorbunov, E., Hanzely, F., & Richtárik, P. (2020, June). **A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent**. *In International Conference on Artificial Intelligence and Statistics*, pp. 680-690 *(AISTATS 2020)*.

Предположение 1

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая `prox-SGD`.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы A, B, C, D_1, D_2, ρ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2AV_f(\mathbf{x}^k, \mathbf{x}^*) + B\sigma_k^2 + D_1, \tag{4}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k, \mathbf{x}^*) + D_2. \tag{5}$$

Возникает вопрос: удовлетворяет ли этому предположению хотя бы один алгоритм?

Санитарная проверка Предположения 1: prox-GD

Самый тривиальный пример стох. градиента — это сам градиент: $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ для всех $k \geq 0$. Действительно,

$$\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \mathbb{E}\left[\nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$$

и prox-SGD в таком случае совпадает с prox-GD. Более того, если функция f выпуклая и L-гладкая, то, как мы знаем, выполняется следующее неравенство:

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] = \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*).$$

Таким образом, мы показали, что Предположение 1 выполняется для prox-GD с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: prox-GD

Самый тривиальный пример стох. градиента — это сам градиент: $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ для всех $k \geq 0$. Действительно,

$$\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \mathbb{E}\left[\nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$$

и prox-SGD в таком случае совпадает с prox-GD. Более того, если функция f выпуклая и L-гладкая, то, как мы знаем, выполняется следующее неравенство:

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] = \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*).$$

Таким образом, мы показали, что Предположение 1 выполняется для prox-GD с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: prox-GD

Самый тривиальный пример стох. градиента — это сам градиент: $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ для всех $k \geq 0$. Действительно,

$$\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \mathbb{E}\left[\nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$$

и prox-SGD в таком случае совпадает с prox-GD. Более того, если функция f выпуклая и L-гладкая, то, как мы знаем, выполняется следующее неравенство:

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] = \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*).$$

Таким образом, мы показали, что Предположение 1 выполняется для prox-GD с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: prox-GD

Самый тривиальный пример стох. градиента — это сам градиент: $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ для всех $k \geq 0$. Действительно,

$$\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \mathbb{E}\left[\nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$$

и prox-SGD в таком случае совпадает с prox-GD. Более того, если функция f выпуклая и L-гладкая, то, как мы знаем, выполняется следующее неравенство:

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] = \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*).$$

Таким образом, мы показали, что Предположение 1 выполняется для prox-GD с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: prox-GD

Самый тривиальный пример стох. градиента — это сам градиент: $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ для всех $k \geq 0$. Действительно,

$$\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \mathbb{E}\left[\nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$$

и prox-SGD в таком случае совпадает с prox-GD. Более того, если функция f выпуклая и L-гладкая, то, как мы знаем, выполняется следующее неравенство:

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] = \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*).$$

Таким образом, мы показали, что Предположение 1 выполняется для prox-GD с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: стох. градиент с равномерно ограниченной дисперсией

Предположим, что существует такое число $\sigma \geq 0$, что $\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$. В этом случае

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] \\ &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] + 2\left\langle \mathbb{E}\left[\mathbf{g}^k - \nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right], \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\right\rangle \\ &\quad + \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \\ &\leq \sigma^2 + 2LV_f(\mathbf{x}^k, \mathbf{x}^*),\end{aligned}$$

то есть Предположение 1 выполняется с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: стох. градиент с равномерно ограниченной дисперсией

Предположим, что существует такое число $\sigma \geq 0$, что $\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$. В этом случае

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] \\ &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] + 2\left\langle \mathbb{E}\left[\mathbf{g}^k - \nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right], \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\right\rangle \\ &\quad + \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \\ &\leq \sigma^2 + 2LV_f(\mathbf{x}^k, \mathbf{x}^*),\end{aligned}$$

то есть Предположение 1 выполняется с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: стох. градиент с равномерно ограниченной дисперсией

Предположим, что существует такое число $\sigma \geq 0$, что $\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$. В этом случае

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] \\ &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] + 2\left\langle \mathbb{E}\left[\mathbf{g}^k - \nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right], \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\right\rangle \\ &\quad + \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \\ &\leq \sigma^2 + 2LV_f(\mathbf{x}^k, \mathbf{x}^*),\end{aligned}$$

то есть Предположение 1 выполняется с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: стох. градиент с равномерно ограниченной дисперсией

Предположим, что существует такое число $\sigma \geq 0$, что $\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$. В этом случае

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] \\
&= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] + 2\left\langle \mathbb{E}\left[\mathbf{g}^k - \nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right], \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\right\rangle \\
&\quad + \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \\
&\leq \sigma^2 + 2LV_f(\mathbf{x}^k, \mathbf{x}^*),
\end{aligned}$$

то есть Предположение 1 выполняется с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Санитарная проверка Предположения 1: стох. градиент с равномерно ограниченной дисперсией

Предположим, что существует такое число $\sigma \geq 0$, что $\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$. В этом случае

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] &= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right] \\
&= \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] + 2\left\langle \mathbb{E}\left[\mathbf{g}^k - \nabla f(\mathbf{x}^k) \mid \mathbf{x}^k\right], \nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\right\rangle \\
&\quad + \|\nabla f(\mathbf{x}^k) - \nabla f(\mathbf{x}^*)\|_2^2 \\
&\leq \sigma^2 + 2LV_f(\mathbf{x}^k, \mathbf{x}^*),
\end{aligned}$$

то есть Предположение 1 выполняется с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Анализ prox-SGD в Предположении 1

Итак, ранее мы доказали неравенство:

$$\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad (1-\gamma\mu)\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 - 2\gamma V_f(\mathbf{x}^k, \mathbf{x}^*) + \gamma^2 \mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].$$

Кроме того, из Предположения 1 имеем:

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] &\leq 2AV_f(\mathbf{x}^k, \mathbf{x}^*) + B\sigma_k^2 + D_1, \\
\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] &\leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k, \mathbf{x}^*) + D_2.
\end{aligned}$$

Анализ prox-SGD в Предположении 1

Итак, ранее мы доказали неравенство:

$$\mathbb{E}\left[\|\mathbf{x}^{k+1}-\mathbf{x}^*\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad (1-\gamma\mu)\|\mathbf{x}^k-\mathbf{x}^*\|_2^2 - 2\gamma V_f(\mathbf{x}^k,\mathbf{x}^*) + \gamma^2\mathbb{E}\left[\|\mathbf{g}^k-\nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k\right].$$

Кроме того, из Предположения 1 имеем:

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{g}^k-\nabla f(\mathbf{x}^*)\right\|_2^2 \mid \mathbf{x}^k\right] &\leq 2AV_f(\mathbf{x}^k,\mathbf{x}^*) + B\sigma_k^2 + D_1,\\ \mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] &\leq (1-\rho)\sigma_k^2 + 2CV_f(\mathbf{x}^k,\mathbf{x}^*) + D_2.\end{aligned}$$

Анализ prox-SGD в Предположении 1

Комбинируя полученные неравенства, можно доказать следующий результат.

Теорема 1

Пусть $f(\mathbf{x})$ — $(\mu, \mathbf{x}^*)$-квази сильно выпуклая L-гладкая функция, $R(\mathbf{x})$ — правильная выпуклая замкнутая функция и выполняется Предположение 1. Пусть

$$M \geq \frac{B}{\rho} \quad \text{и} \quad 0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}.$$

Тогда для любого $N > 0$ выход prox-SGD удовлетворяет неравенству:

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^N T^0 + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \quad \text{где } T^k = \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 + M\gamma^2\sigma_k^2. \quad (6)$$

Если $B = 0$, то считаем, что $M = 0$ и $\frac{B}{M} = 0$ в Теореме 1.

Анализ prox-SGD в Предположении 1

Комбинируя полученные неравенства, можно доказать следующий результат.

Теорема 1

Пусть $f(\mathbf{x})$ — $(\mu, \mathbf{x}^*)$-квази сильно выпуклая L-гладкая функция, $R(\mathbf{x})$ — правильная выпуклая замкнутая функция и выполняется Предположение 1. Пусть

$$M \geq \frac{B}{\rho} \quad \text{и} \quad 0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}.$$

Тогда для любого $N > 0$ выход prox-SGD удовлетворяет неравенству:

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^N T^0 + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \quad \text{где } T^k = \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 + M\gamma^2\sigma_k^2. \quad (6)$$

Если $B = 0$, то считаем, что $M = 0$ и $\frac{B}{M} = 0$ в Теореме 1.

Анализ prox-SGD в Предположении 1

Комбинируя полученные неравенства, можно доказать следующий результат.

Теорема 1

Пусть $f(\mathbf{x})$ — $(\mu, \mathbf{x}^*)$-квази сильно выпуклая L-гладкая функция, $R(\mathbf{x})$ — правильная выпуклая замкнутая функция и выполняется Предположение 1. Пусть

$$M \geq \frac{B}{\rho} \quad \text{и} \quad 0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}.$$

Тогда для любого $N > 0$ выход prox-SGD удовлетворяет неравенству:

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^N T^0 + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \quad \text{где } T^k = \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 + M\gamma^2\sigma_k^2. \quad (6)$$

Если $B = 0$, то считаем, что $M = 0$ и $\frac{B}{M} = 0$ в Теореме 1.

Анализ prox-SGD в Предположении 1

Комбинируя полученные неравенства, можно доказать следующий результат.

Теорема 1

Пусть $f(\mathbf{x})$ — $(\mu, \mathbf{x}^*)$-квази сильно выпуклая L-гладкая функция, $R(\mathbf{x})$ — правильная выпуклая замкнутая функция и выполняется Предположение 1. Пусть

$$M \geq \frac{B}{\rho} \quad \text{и} \quad 0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}.$$

Тогда для любого $N > 0$ выход prox-SGD удовлетворяет неравенству:

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^N T^0 + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \quad \text{где } T^k = \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 + M\gamma^2\sigma_k^2. \quad (6)$$

Если $B = 0$, то считаем, что $M = 0$ и $\frac{B}{M} = 0$ в Теореме 1.

Анализ prox-SGD в Предположении 1

Комбинируя полученные неравенства, можно доказать следующий результат.

Теорема 1

Пусть $f(\mathbf{x})$ — $(\mu, \mathbf{x}^*)$-квази сильно выпуклая L-гладкая функция, $R(\mathbf{x})$ — правильная выпуклая замкнутая функция и выполняется Предположение 1. Пусть

$$M \geq \frac{B}{\rho} \quad \text{и} \quad 0 < \gamma \leq \min\left\{\frac{1}{\mu}, \frac{1}{A + CM}\right\}.$$

Тогда для любого $N > 0$ выход prox-SGD удовлетворяет неравенству:

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}\right)^N T^0 + \frac{\gamma^2(D_1 + MD_2)}{\min\left\{\gamma\mu, \rho - \frac{B}{M}\right\}}, \quad \text{где } T^k = \|\mathbf{x}^k - \mathbf{x}^*\|_2^2 + M\gamma^2\sigma_k^2. \quad (6)$$

Если $B = 0$, то считаем, что $M = 0$ и $\frac{B}{M} = 0$ в Теореме 1.

prox-GD

Ранее мы доказали, что prox-GD удовлетворяет Предположению 1 в случае, если функция f выпуклая и L-гладкая с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2,$$

Иными словами, мы доказали оценку $O\left(\frac{L}{\mu} \ln \frac{R^2}{\varepsilon}\right)$ на число итераций для достижения $\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq \varepsilon$ при $\gamma = \frac{1}{L}$.

prox-GD

Ранее мы доказали, что prox-GD удовлетворяет Предположению 1 в случае, если функция f выпуклая и L-гладкая с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2,$$

Иными словами, мы доказали оценку $O\left(\frac{L}{\mu} \ln \frac{R^2}{\varepsilon}\right)$ на число итераций для достижения $\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq \varepsilon$ при $\gamma = \frac{1}{L}$.

prox-GD

Ранее мы доказали, что prox-GD удовлетворяет Предположению 1 в случае, если функция f выпуклая и L-гладкая с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2,$$

Иными словами, мы доказали оценку $O\left(\frac{L}{\mu} \ln \frac{R^2}{\varepsilon}\right)$ на число итераций для достижения $\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq \varepsilon$ при $\gamma = \frac{1}{L}$.

prox-GD

Ранее мы доказали, что prox-GD удовлетворяет Предположению 1 в случае, если функция f выпуклая и L-гладкая с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2,$$

Иными словами, мы доказали оценку $O\left(\frac{L}{\mu} \ln \frac{R^2}{\varepsilon}\right)$ на число итераций для достижения $\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq \varepsilon$ при $\gamma = \frac{1}{L}$.

prox-GD

Ранее мы доказали, что prox-GD удовлетворяет Предположению 1 в случае, если функция f выпуклая и L-гладкая с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2,$$

Иными словами, мы доказали оценку $O\left(\frac{L}{\mu} \ln \frac{R^2}{\varepsilon}\right)$ на число итераций для достижения $\|\mathbf{x}^N - \mathbf{x}^*\|_2^2 \leq \varepsilon$ при $\gamma = \frac{1}{L}$.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Мы показали, что в этом случае prox-SGD удовлетворяет Предположению 1 с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}.$$

Это означает, что prox-SGD с постоянным шагом γ сходится с линейной скоростью к некоторой окрестности решения радиуса $\sqrt{\frac{\gamma\sigma^2}{\mu}}$.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Мы показали, что в этом случае prox-SGD удовлетворяет Предположению 1 с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}.$$

Это означает, что prox-SGD с постоянным шагом γ сходится с линейной скоростью к некоторой окрестности решения радиуса $\sqrt{\frac{\gamma\sigma^2}{\mu}}$.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Мы показали, что в этом случае prox-SGD удовлетворяет Предположению 1 с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}.$$

Это означает, что prox-SGD с постоянным шагом γ сходится с линейной скоростью к некоторой окрестности решения радиуса $\sqrt{\frac{\gamma\sigma^2}{\mu}}$.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Мы показали, что в этом случае prox-SGD удовлетворяет Предположению 1 с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}.$$

Это означает, что prox-SGD с постоянным шагом γ сходится с линейной скоростью к некоторой окрестности решения радиуса $\sqrt{\frac{\gamma\sigma^2}{\mu}}$.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Мы показали, что в этом случае prox-SGD удовлетворяет Предположению 1 с параметрами

$$A = L, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Теорема 1 даёт следующий результат: если

$$\gamma \leq \frac{1}{L},$$

то для всех целых $N \geq 0$ выполняется

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\gamma\sigma^2}{\mu}.$$

Это означает, что prox-SGD с постоянным шагом γ сходится с линейной скоростью к некоторой окрестности решения радиуса $\sqrt{\frac{\gamma\sigma^2}{\mu}}$.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

При доказательстве Теоремы 1 возникает рекуррентное неравенство

$$\mathbb{E}[T^{k+1}] \leq \left(1 - \min\left\{\gamma_k \mu, \rho - \frac{B}{M}\right\}\right) \mathbb{E}[T^k] + \gamma_k^2 (D_1 + M D_2),$$

которое в случае prox-SGD записывается в виде:

$$\mathbb{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2] \leq (1 - \gamma_k \mu)\, \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^*\|_2^2] + \gamma_k^2 \sigma^2.$$

В недавнем препринте

Stich, S.U., 2019. **Unified optimal analysis of the (stochastic) gradient method.** *arXiv preprint* arXiv:1907.04232.

подробно исследуется вопрос, как анализировать рекурренты такого вида, и рассматриваются разные способы выбора шага γ_k.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

При доказательстве Теоремы 1 возникает рекуррентное неравенство

$$\mathbb{E}[T^{k+1}] \leq \left(1 - \min\left\{\gamma_k \mu, \rho - \frac{B}{M}\right\}\right) \mathbb{E}[T^k] + \gamma_k^2 (D_1 + M D_2),$$

которое в случае prox-SGD записывается в виде:

$$\mathbb{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2] \leq (1 - \gamma_k \mu)\, \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^*\|_2^2] + \gamma_k^2 \sigma^2.$$

В недавнем препринте

Stich, S.U., 2019. **Unified optimal analysis of the (stochastic) gradient method.** *arXiv preprint* arXiv:1907.04232.

подробно исследуется вопрос, как анализировать рекурренты такого вида, и рассматриваются разные способы выбора шага γ_k.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

При доказательстве Теоремы 1 возникает рекуррентное неравенство

$$\mathbb{E}[T^{k+1}] \leq \left(1 - \min\left\{\gamma_k \mu, \rho - \frac{B}{M}\right\}\right) \mathbb{E}[T^k] + \gamma_k^2 (D_1 + M D_2),$$

которое в случае prox-SGD записывается в виде:

$$\mathbb{E}[\|\mathbf{x}^{k+1} - \mathbf{x}^*\|_2^2] \leq (1 - \gamma_k \mu)\, \mathbb{E}[\|\mathbf{x}^k - \mathbf{x}^*\|_2^2] + \gamma_k^2 \sigma^2.$$

В недавнем препринте

Stich, S.U., 2019. **Unified optimal analysis of the (stochastic) gradient method.** *arXiv preprint* arXiv:1907.04232.

подробно исследуется вопрос, как анализировать рекурренты такого вида, и рассматриваются разные способы выбора шага γ_k.

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Рассмотрим некоторое $N > 0$ и следующую политику выбора шага γ_k:

$$\text{Если } N \leq \frac{L}{\mu}, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k < k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k \geq k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{2}{L + \mu(k - k_0)}.$$

Тогда

$$\mu \mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq 64 L R_0^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{36\sigma^2}{\mu N}, \quad \text{где } R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|_2.$$

Иными словами, для достижения $\mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq \varepsilon$ необходимо

$$O\left(\kappa \ln\left(\frac{\kappa R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon}\right) \quad \text{итераций.}$$

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Рассмотрим некоторое $N > 0$ и следующую политику выбора шага γ_k:

$$\text{Если } N \leq \frac{L}{\mu}, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k < k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k \geq k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{2}{L + \mu(k - k_0)}.$$

Тогда

$$\mu \mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq 64 L R_0^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{36\sigma^2}{\mu N}, \quad \text{где } R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|_2.$$

Иными словами, для достижения $\mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq \varepsilon$ необходимо

$$O\left(\kappa \ln\left(\frac{\kappa R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon}\right) \quad \text{итераций.}$$

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Рассмотрим некоторое $N > 0$ и следующую политику выбора шага γ_k:

$$\text{Если } N \leq \frac{L}{\mu}, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k < k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k \geq k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{2}{L + \mu(k - k_0)}.$$

Тогда

$$\mu \mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq 64 L R_0^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{36\sigma^2}{\mu N}, \quad \text{где } R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|_2.$$

Иными словами, для достижения $\mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq \varepsilon$ необходимо

$$O\left(\kappa \ln\left(\frac{\kappa R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon}\right) \quad \text{итераций.}$$

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Рассмотрим некоторое $N > 0$ и следующую политику выбора шага γ_k:

$$\text{Если } N \leq \frac{L}{\mu}, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k < k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k \geq k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{2}{L + \mu(k - k_0)}.$$

Тогда

$$\mu \mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq 64 L R_0^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{36\sigma^2}{\mu N}, \quad \text{где } R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|_2.$$

Иными словами, для достижения $\mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq \varepsilon$ необходимо

$$O\left(\kappa \ln\left(\frac{\kappa R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon}\right) \quad \text{итераций.}$$

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Рассмотрим некоторое $N > 0$ и следующую политику выбора шага γ_k:

$$\text{Если } N \leq \frac{L}{\mu}, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k < k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k \geq k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{2}{L + \mu(k - k_0)}.$$

Тогда

$$\mu \mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq 64 L R_0^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{36\sigma^2}{\mu N}, \quad \text{где } R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|_2.$$

Иными словами, для достижения $\mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq \varepsilon$ необходимо

$$O\left(\kappa \ln\left(\frac{\kappa R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon}\right) \quad \text{итераций.}$$

prox-SGD: равномерно ограниченная дисперсия стох. градиента

Рассмотрим некоторое $N > 0$ и следующую политику выбора шага γ_k:

$$\text{Если } N \leq \frac{L}{\mu}, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k < k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{1}{L},$$

$$\text{Если } N > \frac{L}{\mu} \text{ и } k \geq k_0 = \left\lceil \frac{N}{2} \right\rceil, \quad \text{то} \quad \gamma_k = \frac{2}{L + \mu(k - k_0)}.$$

Тогда

$$\mu \mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq 64 L R_0^2 \exp\left(-\frac{\mu N}{4L}\right) + \frac{36\sigma^2}{\mu N}, \quad \text{где } R_0 = \|\mathbf{x}^0 - \mathbf{x}^*\|_2.$$

Иными словами, для достижения $\mathbb{E}\left[\|\mathbf{x}^{N+1} - \mathbf{x}^*\|_2^2\right] \leq \varepsilon$ необходимо

$$O\left(\kappa \ln\left(\frac{\kappa R_0^2}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon}\right) \quad \text{итераций.}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E. and Richtárik, P. **SGD: General Analysis and Improved Rates.** *In International Conference on Machine Learning*, pp. 5200-5209 *(ICML 2019)*.

Стохастическая переформулировка

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \quad f(\mathbf{x}) = \mathbf{E}_{\mathcal{D}}\left[f_\xi(\mathbf{x})\right], \quad f_\xi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \xi_i f_i(\mathbf{x}) \tag{7}$$

1. $\xi \sim \mathcal{D}$: $\mathbf{E}_{\mathcal{D}}\left[\xi_i\right] = 1$ для всех i
2. f_i — гладкие функции

На шаге k:

Просэмплировать $\xi \sim \mathcal{D}$, вычислить $\mathbf{g}^k = \nabla f_\xi(\mathbf{x}^k)$

$$\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right), \tag{8}$$

SGD-SR: стохастические переформулировки задач минимизации суммы

Усреднённая гладкость (expected smoothness)

Будем говорить, что функция f является $\mathcal{L}$-гладкой в среднем относительно распределения $\mathcal{D}$, если существует такая константа $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f_\xi(\mathbf{x}^*)\|^2\right] \leq 2\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*), \tag{9}$$

для всех $x \in \mathbb{R}^d$, и будем обозначать это, как $(f, \mathcal{D}) \sim ES(\mathcal{L})$.

- Если все функции f_i являются L_i-гладкими и выпуклыми, а $\mathcal{D}$ — равномерное распределение на базисных векторах в $\mathbb{R}^m$, умноженных на m, то $\mathcal{L} = \max\limits_{i=1,\ldots,m} L_i$.
- Если все функции f_i являются L_i-гладкими и выпуклыми, а распределение $\mathcal{D}$ таково, что $\xi = (0, \ldots, 0, {}^{m\sum_{i=1}^m L_i}/_{L_j}, 0, \ldots, 0)^\top$ (все компоненты нулевые, кроме j-й) с вероятностью $p_j = {}^{L_j}/_{\sum_{i=1}^m L_i}$, то $\mathcal{L} = \overline{L} = \frac{1}{m}\sum_{i=1}^m L_i$.

SGD-SR: стохастические переформулировки задач минимизации суммы

Усреднённая гладкость (expected smoothness)

Будем говорить, что функция f является $\mathcal{L}$-гладкой в среднем относительно распределения $\mathcal{D}$, если существует такая константа $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f_\xi(\mathbf{x}^*)\|^2\right] \leq 2\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*), \tag{9}$$

для всех $x \in \mathbb{R}^d$, и будем обозначать это, как $(f, \mathcal{D}) \sim ES(\mathcal{L})$.

- Если все функции f_i являются L_i-гладкими и выпуклыми, а $\mathcal{D}$ — равномерное распределение на базисных векторах в $\mathbb{R}^m$, умноженных на m, то $\mathcal{L} = \max\limits_{i=1,\ldots,m} L_i$.
- Если все функции f_i являются L_i-гладкими и выпуклыми, а распределение $\mathcal{D}$ таково, что $\xi = (0, \ldots, 0, {}^{m\sum_{i=1}^m L_i}/_{L_j}, 0, \ldots, 0)^\top$ (все компоненты нулевые, кроме j-й) с вероятностью $p_j = {}^{L_j}/_{\sum_{i=1}^m L_i}$, то $\mathcal{L} = \overline{L} = \frac{1}{m}\sum_{i=1}^m L_i$.

SGD-SR: стохастические переформулировки задач минимизации суммы

Усреднённая гладкость (expected smoothness)

Будем говорить, что функция f является $\mathcal{L}$-гладкой в среднем относительно распределения $\mathcal{D}$, если существует такая константа $\mathcal{L} = \mathcal{L}(f, \mathcal{D}) > 0$, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f_\xi(\mathbf{x}^*)\|^2\right] \leq 2\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*), \tag{9}$$

для всех $x \in \mathbb{R}^d$, и будем обозначать это, как $(f, \mathcal{D}) \sim ES(\mathcal{L})$.

- Если все функции f_i являются L_i-гладкими и выпуклыми, а $\mathcal{D}$ — равномерное распределение на базисных векторах в $\mathbb{R}^m$, умноженных на m, то $\mathcal{L} = \max\limits_{i=1,\ldots,m} L_i$.
- Если все функции f_i являются L_i-гладкими и выпуклыми, а распределение $\mathcal{D}$ таково, что $\xi = (0, \ldots, 0, {}^{m\sum_{i=1}^m L_i}/_{L_j}, 0, \ldots, 0)^\top$ (все компоненты нулевые, кроме j-й) с вероятностью $p_j = {}^{L_j}/_{\sum_{i=1}^m L_i}$, то $\mathcal{L} = \overline{L} = \frac{1}{m}\sum_{i=1}^m L_i$.

SGD-SR: стохастические переформулировки задач минимизации суммы

При сделанных предположениях можно доказать, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2\right] \leq 4\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*) + 2\sigma^2, \quad \text{где } \sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)\|^2\right], \tag{10}$$

то есть Предположение 1 выполняется с параметрами

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

Применяя Теорему 1, получаем, что при $\gamma \leq \frac{1}{2\mathcal{L}}$ для всех $N \geq 0$ выполняется неравенство:

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1-\gamma\mu)^N\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

Для prox-SGD-SR с постоянным шагом мы наблюдаем линейную сходимость лишь к окрестности решения, в то время как prox-GD сходится линейно к точному решению, но его итерации гораздо дороже. В попытке взять «лучшее» от SGD и GD мы приходим к методам редукции дисперсии.

SGD-SR: стохастические переформулировки задач минимизации суммы

При сделанных предположениях можно доказать, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2\right] \leq 4\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*) + 2\sigma^2, \quad \text{где } \sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)\|^2\right], \tag{10}$$

то есть Предположение 1 выполняется с параметрами

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

Применяя Теорему 1, получаем, что при $\gamma \leq \frac{1}{2\mathcal{L}}$ для всех $N \geq 0$ выполняется неравенство:

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1 - \gamma\mu)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

Для prox-SGD-SR с постоянным шагом мы наблюдаем линейную сходимость лишь к окрестности решения, в то время как prox-GD сходится линейно к точному решению, но его итерации гораздо дороже. В попытке взять «лучшее» от SGD и GD мы приходим к методам редукции дисперсии.

SGD-SR: стохастические переформулировки задач минимизации суммы

При сделанных предположениях можно доказать, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2\right] \leq 4\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*) + 2\sigma^2, \quad \text{где } \sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)\|^2\right], \tag{10}$$

то есть Предположение 1 выполняется с параметрами

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

Применяя Теорему 1, получаем, что при $\gamma \leq \frac{1}{2\mathcal{L}}$ для всех $N \geq 0$ выполняется неравенство:

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1-\gamma\mu)^N\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

Для prox-SGD-SR с постоянным шагом мы наблюдаем линейную сходимость лишь к окрестности решения, в то время как prox-GD сходится линейно к точному решению, но его итерации гораздо дороже. В попытке взять «лучшее» от SGD и GD мы приходим к методам редукции дисперсии.

SGD-SR: стохастические переформулировки задач минимизации суммы

При сделанных предположениях можно доказать, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2\right] \leq 4\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*) + 2\sigma^2, \quad \text{где } \sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)\|^2\right], \tag{10}$$

то есть Предположение 1 выполняется с параметрами

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

Применяя Теорему 1, получаем, что при $\gamma \leq \frac{1}{2\mathcal{L}}$ для всех $N \geq 0$ выполняется неравенство:

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1-\gamma\mu)^N\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

Для prox-SGD-SR с постоянным шагом мы наблюдаем линейную сходимость лишь к окрестности решения, в то время как prox-GD сходится линейно к точному решению, но его итерации гораздо дороже. В попытке взять «лучшее» от SGD и GD мы приходим к методам редукции дисперсии.

SGD-SR: стохастические переформулировки задач минимизации суммы

При сделанных предположениях можно доказать, что

$$\mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2\right] \leq 4\mathcal{L}V_f(\mathbf{x}, \mathbf{x}^*) + 2\sigma^2, \quad \text{где } \sigma^2 \stackrel{\text{def}}{=} \mathbf{E}_{\mathcal{D}}\left[\|\nabla f_\xi(\mathbf{x}^*) - \nabla f(\mathbf{x}^*)\|^2\right], \tag{10}$$

то есть Предположение 1 выполняется с параметрами

$$A = 2\mathcal{L}, B = 0, D_1 = 2\sigma^2, \sigma_k = 0, \rho = 1, C = 0, D_2 = 0.$$

Применяя Теорему 1, получаем, что при $\gamma \leq \frac{1}{2\mathcal{L}}$ для всех $N \geq 0$ выполняется неравенство:

$$\mathbb{E}[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2] \leq (1-\gamma\mu)^N\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma\sigma^2}{\mu}.$$

Для prox-SGD-SR с постоянным шагом мы наблюдаем линейную сходимость лишь к окрестности решения, в то время как prox-GD сходится линейно к точному решению, но его итерации гораздо дороже. В попытке взять «лучшее» от SGD и GD мы приходим к методам редукции дисперсии.

prox-SGD-star

Начнём мы со следующего, сугубо теоретического метода, который предполагает, что ему известны $\nabla f_1(\mathbf{x}^*), \ldots, \nabla f_m(\mathbf{x}^*)$ и $\nabla f(\mathbf{x}^*)$. На шаге k:

- Случайно равновероятно выбирается индекс j из множества $\{1, 2, \ldots, m\}$
- Вычисляется $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)$
- $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$

Отметим, что $\mathbb{E}[-\nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)] = -\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) = 0$. Однако эта часть стох. градиента позволяет редуцировать дисперсию.

prox-SGD-star

Начнём мы со следующего, сугубо теоретического метода, который предполагает, что ему известны $\nabla f_1(\mathbf{x}^*), \ldots, \nabla f_m(\mathbf{x}^*)$ и $\nabla f(\mathbf{x}^*)$. На шаге k:

- Случайно равновероятно выбирается индекс j из множества $\{1, 2, \ldots, m\}$
- Вычисляется $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)$
- $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$

Отметим, что $\mathbb{E}[-\nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)] = -\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) = 0$. Однако эта часть стох. градиента позволяет редуцировать дисперсию.

prox-SGD-star

Начнём мы со следующего, сугубо теоретического метода, который предполагает, что ему известны $\nabla f_1(\mathbf{x}^*), \ldots, \nabla f_m(\mathbf{x}^*)$ и $\nabla f(\mathbf{x}^*)$. На шаге k:

- Случайно равновероятно выбирается индекс j из множества $\{1, 2, \ldots, m\}$
- Вычисляется $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)$
- $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$

Отметим, что $\mathbb{E}[-\nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)] = -\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) = 0$. Однако эта часть стох. градиента позволяет редуцировать дисперсию.

prox-SGD-star

Начнём мы со следующего, сугубо теоретического метода, который предполагает, что ему известны $\nabla f_1(\mathbf{x}^*), \ldots, \nabla f_m(\mathbf{x}^*)$ и $\nabla f(\mathbf{x}^*)$. На шаге k:

- Случайно равновероятно выбирается индекс j из множества $\{1, 2, \ldots, m\}$
- Вычисляется $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)$
- $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$

Отметим, что $\mathbb{E}[-\nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)] = -\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) = 0$. Однако эта часть стох. градиента позволяет редуцировать дисперсию.

prox-SGD-star

Начнём мы со следующего, сугубо теоретического метода, который предполагает, что ему известны $\nabla f_1(\mathbf{x}^*), \ldots, \nabla f_m(\mathbf{x}^*)$ и $\nabla f(\mathbf{x}^*)$. На шаге k:

- Случайно равновероятно выбирается индекс j из множества $\{1, 2, \ldots, m\}$
- Вычисляется $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)$
- $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$

Отметим, что $\mathbb{E}[-\nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)] = -\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) = 0$. Однако эта часть стох. градиента позволяет редуцировать дисперсию.

prox-SGD-star

Начнём мы со следующего, сугубо теоретического метода, который предполагает, что ему известны $\nabla f_1(\mathbf{x}^*), \ldots, \nabla f_m(\mathbf{x}^*)$ и $\nabla f(\mathbf{x}^*)$. На шаге k:

- Случайно равновероятно выбирается индекс j из множества $\{1, 2, \ldots, m\}$
- Вычисляется $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)$
- $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$

Отметим, что $\mathbb{E}[-\nabla f_j(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)] = -\frac{1}{m}\sum_{i=1}^m \nabla f_i(\mathbf{x}^*) + \nabla f(\mathbf{x}^*) = 0$. Однако эта часть стох. градиента позволяет редуцировать дисперсию.

prox-SGD-star

Можно показать, что в случае, когда $f_i(\mathbf{x})$ — выпуклые и L-гладкие функции, выполняется неравенство

$$\mathbb{E}[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k] \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*),$$

то есть prox-SGD-star удовлетворяет Предположению 1 с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0,$$

откуда следует, что при $\gamma \leq \frac{1}{L}$ выполняется неравенство:

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

В частности, если $\gamma = \frac{1}{L}$, то prox-SGD-star сходится со скоростью $O\left(\kappa \ln \frac{R^2}{\varepsilon}\right)$ асимптотически точно.

Мы получили ровно то, что хотели, но, к сожалению, с одним большим недостатком: метод бесполезен на практике.

prox-SGD-star

Можно показать, что в случае, когда $f_i(\mathbf{x})$ — выпуклые и L-гладкие функции, выполняется неравенство

$$\mathbb{E}[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k] \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*),$$

то есть `prox-SGD-star` удовлетворяет Предположению 1 с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0,$$

откуда следует, что при $\gamma \leq \frac{1}{L}$ выполняется неравенство:

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

В частности, если $\gamma = \frac{1}{L}$, то `prox-SGD-star` сходится со скоростью $O\left(\kappa \ln \frac{R^2}{\varepsilon}\right)$ асимптотически точно.

Мы получили ровно то, что хотели, но, к сожалению, с одним большим недостатком: метод бесполезен на практике.

prox-SGD-star

Можно показать, что в случае, когда $f_i(\mathbf{x})$ — выпуклые и L-гладкие функции, выполняется неравенство

$$\mathbb{E}[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k] \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*),$$

то есть `prox-SGD-star` удовлетворяет Предположению 1 с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0,$$

откуда следует, что при $\gamma \leq \frac{1}{L}$ выполняется неравенство:

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

В частности, если $\gamma = \frac{1}{L}$, то `prox-SGD-star` сходится со скоростью $O\left(\kappa \ln \frac{R^2}{\varepsilon}\right)$ асимптотически точно.

Мы получили ровно то, что хотели, но, к сожалению, с одним большим недостатком: метод бесполезен на практике.

prox-SGD-star

Можно показать, что в случае, когда $f_i(\mathbf{x})$ — выпуклые и L-гладкие функции, выполняется неравенство

$$\mathbb{E}[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k] \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*),$$

то есть `prox-SGD-star` удовлетворяет Предположению 1 с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0,$$

откуда следует, что при $\gamma \leq \frac{1}{L}$ выполняется неравенство:

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

В частности, если $\gamma = \frac{1}{L}$, то `prox-SGD-star` сходится со скоростью $O\left(\kappa \ln \frac{R^2}{\varepsilon}\right)$ асимптотически точно.

Мы получили ровно то, что хотели, но, к сожалению, с одним большим недостатком: метод бесполезен на практике.

prox-SGD-star

Можно показать, что в случае, когда $f_i(\mathbf{x})$ — выпуклые и L-гладкие функции, выполняется неравенство

$$\mathbb{E}[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k] \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*),$$

то есть `prox-SGD-star` удовлетворяет Предположению 1 с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0,$$

откуда следует, что при $\gamma \leq \frac{1}{L}$ выполняется неравенство:

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

В частности, если $\gamma = \frac{1}{L}$, то `prox-SGD-star` сходится со скоростью $O\left(\kappa \ln \frac{R^2}{\varepsilon}\right)$ асимптотически точно.

Мы получили ровно то, что хотели, но, к сожалению, с одним большим недостатком: метод бесполезен на практике.

prox-SGD-star

Можно показать, что в случае, когда $f_i(\mathbf{x})$ — выпуклые и L-гладкие функции, выполняется неравенство

$$\mathbb{E}[\|\mathbf{g}^k - \nabla f(\mathbf{x}^*)\|_2^2 \mid \mathbf{x}^k] \leq 2LV_f(\mathbf{x}^k, \mathbf{x}^*),$$

то есть `prox-SGD-star` удовлетворяет Предположению 1 с параметрами

$$A = L, \quad B = C = D_1 = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0,$$

откуда следует, что при $\gamma \leq \frac{1}{L}$ выполняется неравенство:

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

В частности, если $\gamma = \frac{1}{L}$, то `prox-SGD-star` сходится со скоростью $O\left(\kappa \ln \frac{R^2}{\varepsilon}\right)$ асимптотически точно.

Мы получили ровно то, что хотели, но, к сожалению, с одним большим недостатком: метод бесполезен на практике.

prox-SVRG

Johnson, R. and Zhang, T. **Accelerating stochastic gradient descent using predictive variance reduction.** *In Advances in neural information processing systems*, pp. 315-323, *(NIPS 2013)*.

$$\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{w}) + \nabla f(\mathbf{w}),$$

где точка $\mathbf{w}$ периодически (раз в $\sim m$ итераций) обновляется.

Вместо градиентов функций f_i в точке $\mathbf{x}^*$ используются градиенты функций f_i в точке $\mathbf{w}$, которая приближается к решению по ходу работы метода.

prox-SVRG

Johnson, R. and Zhang, T. **Accelerating stochastic gradient descent using predictive variance reduction.** *In Advances in neural information processing systems*, pp. 315-323, *(NIPS 2013)*.

$$\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{w}) + \nabla f(\mathbf{w}),$$

где точка $\mathbf{w}$ периодически (раз в $\sim m$ итераций) обновляется.

Вместо градиентов функций f_i в точке $\mathbf{x}^*$ используются градиенты функций f_i в точке $\mathbf{w}$, которая приближается к решению по ходу работы метода.

prox-SVRG

Johnson, R. and Zhang, T. **Accelerating stochastic gradient descent using predictive variance reduction.** *In Advances in neural information processing systems*, pp. 315-323, *(NIPS 2013)*.

$$\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{w}) + \nabla f(\mathbf{w}),$$

где точка $\mathbf{w}$ периодически (раз в $\sim m$ итераций) обновляется.

Вместо градиентов функций f_i в точке $\mathbf{x}^*$ используются градиенты функций f_i в точке $\mathbf{w}$, которая приближается к решению по ходу работы метода.

prox-SVRG

Алгоритм 4 prox-SVRG

Вход: размер шага $\gamma > 0$, стартовая точка $\mathbf{x}^0 \in \mathbb{R}^d$, количество эпох N, длина внутреннего цикла r

Пусть $\mathbf{w} = \mathbf{x}^0$
Вычислить $\nabla f(\mathbf{w})$
for $s = 0, 1, \ldots, N-1$ **do**
  for $k = 0, 1, \ldots, r-1$ **do**
    Случайно равновероятно выбрать индекс j из множества $\{1, 2, \ldots, m\}$
    $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{w}) + \nabla f(\mathbf{w})$
    $\mathbf{x}^{k+1} = \mathrm{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$
  end for
  $\mathbf{w} = \mathbf{x}^0 = \frac{1}{r}\sum_{k=1}^{r} \mathbf{x}^k$
end for

prox-L-SVRG

Чуть более удобно анализировать видоизменённую версию, у которой длина внутреннего цикла — случайное число.

Hofmann, T., Lucchi, A., Lacoste-Julien, S. and McWilliams, B. **Variance reduced stochastic gradient descent with neighbors.** *In Advances in Neural Information Processing Systems*, pp. 2305-2313, *(NIPS 2015)*.

Kovalev, D., Horváth, S. and Richtárik, P. **Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop.** *In Algorithmic Learning Theory*, pp. 451-467, *(ALT 2020)*.

prox-L-SVRG

Алгоритм 5 prox-L-SVRG

Вход: размер шага $\gamma > 0$, стартовая точка $\mathbf{x}^0 \in \mathbb{R}^d$, количество итераций N, вероятность подсчёта полного градиента $p \in (0, 1]$

1: Пусть $\mathbf{w}^0 = \mathbf{x}^0$
2: Вычислить $\nabla f(\mathbf{w}^0)$
3: **for** $k = 0, 1, \ldots, N-1$ **do**
4: Случайно равновероятно выбрать индекс j из множества $\{1, 2, \ldots, m\}$
5: $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{w}^k) + \nabla f(\mathbf{w}^k)$
6: $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$
7: $\mathbf{w}^{k+1} = \begin{cases} \mathbf{x}^k & \text{с вероятностью } p, \\ \mathbf{w}^k & \text{с вероятностью } 1-p \end{cases}$
8: **end for**

С вероятностью p нужно вычислить $m+1$ градиент слагаемых в сумме, а с вероятностью $1-p$ нужно вычислить 2 градиента.

prox-L-SVRG

Алгоритм 5 prox-L-SVRG

Вход: размер шага $\gamma > 0$, стартовая точка $\mathbf{x}^0 \in \mathbb{R}^d$, количество итераций N, вероятность подсчёта полного градиента $p \in (0, 1]$

1: Пусть $\mathbf{w}^0 = \mathbf{x}^0$
2: Вычислить $\nabla f(\mathbf{w}^0)$
3: **for** $k = 0, 1, \ldots, N-1$ **do**
4: Случайно равновероятно выбрать индекс j из множества $\{1, 2, \ldots, m\}$
5: $\mathbf{g}^k = \nabla f_j(\mathbf{x}^k) - \nabla f_j(\mathbf{w}^k) + \nabla f(\mathbf{w}^k)$
6: $\mathbf{x}^{k+1} = \operatorname{prox}_{\gamma R}\left(\mathbf{x}^k - \gamma \mathbf{g}^k\right)$
7: $\mathbf{w}^{k+1} = \begin{cases} \mathbf{x}^k & \text{с вероятностью } p, \\ \mathbf{w}^k & \text{с вероятностью } 1-p \end{cases}$
8: **end for**

С вероятностью p нужно вычислить $m+1$ градиент слагаемых в сумме, а с вероятностью $1-p$ нужно вычислить 2 градиента.

prox-L-SVRG

Можно показать, что prox-L-SVRG удовлетворяет Предположению 1 с параметрами

$$A = 2L, \quad B = 2, \quad \rho = p, \quad C = Lp, \quad D_1 = D_2 = 0,$$

$$\sigma_k^2 = \frac{1}{m}\sum_{i=1}^{m} \|\nabla f_i(\mathbf{w}^k) - \nabla f_i(\mathbf{x}^*)\|_2^2.$$

Применяя Теорему 1 с $M = \frac{2B}{\rho} = \frac{4}{p}$, получаем, что при $\gamma = \frac{1}{A+CM} = \frac{1}{6L}$ выполняется

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\frac{\mu}{6L}, \frac{p}{2}\right\}\right)^N T^0.$$

При $p = \frac{1}{m}$ получаем, что prox-L-SVRG сходится со скоростью $O\left(\left(m + \frac{L}{\mu}\right)\ln\frac{R^2}{\varepsilon}\right)$ асимптотически точно.

prox-L-SVRG

Можно показать, что prox-L-SVRG удовлетворяет Предположению 1 с параметрами

$$A = 2L, \quad B = 2, \quad \rho = p, \quad C = Lp, \quad D_1 = D_2 = 0,$$

$$\sigma_k^2 = \frac{1}{m}\sum_{i=1}^{m} \|\nabla f_i(\mathbf{w}^k) - \nabla f_i(\mathbf{x}^*)\|_2^2.$$

Применяя Теорему 1 с $M = \frac{2B}{\rho} = \frac{4}{p}$, получаем, что при $\gamma = \frac{1}{A+CM} = \frac{1}{6L}$ выполняется

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\frac{\mu}{6L}, \frac{p}{2}\right\}\right)^N T^0.$$

При $p = \frac{1}{m}$ получаем, что prox-L-SVRG сходится со скоростью $O\left(\left(m + \frac{L}{\mu}\right)\ln\frac{R^2}{\varepsilon}\right)$ асимптотически точно.

prox-L-SVRG

Можно показать, что prox-L-SVRG удовлетворяет Предположению 1 с параметрами

$$A = 2L, \quad B = 2, \quad \rho = p, \quad C = Lp, \quad D_1 = D_2 = 0,$$

$$\sigma_k^2 = \frac{1}{m}\sum_{i=1}^{m} \|\nabla f_i(\mathbf{w}^k) - \nabla f_i(\mathbf{x}^*)\|_2^2.$$

Применяя Теорему 1 с $M = \frac{2B}{\rho} = \frac{4}{p}$, получаем, что при $\gamma = \frac{1}{A+CM} = \frac{1}{6L}$ выполняется

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\frac{\mu}{6L}, \frac{p}{2}\right\}\right)^N T^0.$$

При $p = \frac{1}{m}$ получаем, что prox-L-SVRG сходится со скоростью $O\left(\left(m + \frac{L}{\mu}\right)\ln\frac{R^2}{\varepsilon}\right)$ асимптотически точно.

prox-L-SVRG

Можно показать, что prox-L-SVRG удовлетворяет Предположению 1 с параметрами

$$A = 2L, \quad B = 2, \quad \rho = p, \quad C = Lp, \quad D_1 = D_2 = 0,$$

$$\sigma_k^2 = \frac{1}{m}\sum_{i=1}^{m} \|\nabla f_i(\mathbf{w}^k) - \nabla f_i(\mathbf{x}^*)\|_2^2.$$

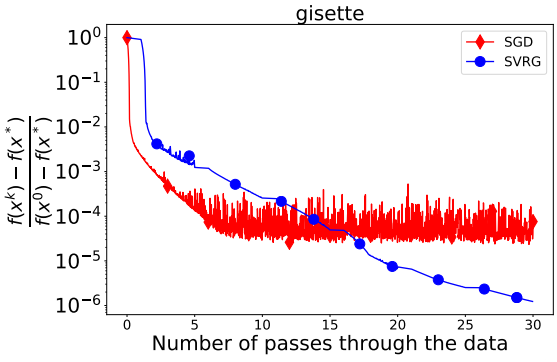
Применяя Теорему 1 с $M = \frac{2B}{\rho} = \frac{4}{p}$, получаем, что при $\gamma = \frac{1}{A+CM} = \frac{1}{6L}$ выполняется

$$\mathbb{E}[T^N] \leq \left(1 - \min\left\{\frac{\mu}{6L}, \frac{p}{2}\right\}\right)^N T^0.$$

При $p = \frac{1}{m}$ получаем, что prox-L-SVRG сходится со скоростью $O\left(\left(m + \frac{L}{\mu}\right)\ln\frac{R^2}{\varepsilon}\right)$ асимптотически точно.

SGD и SVRG на задаче логистической регрессии

$$\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-y_i\cdot(\mathbf{A}\mathbf{x})_i\right)\right)+\frac{l_2}{2}\|\mathbf{x}\|_2^2+l_1\|\mathbf{x}\|_1\to\min_{\mathbf{x}\in\mathbb{R}^n},$$



Некоторые методы, покрываемые Предположением 1

Problem	Method	Alg #	Citation	VR?	AS?	Quant?	RCD?	Section	Result
(1)+(2)	SGD	Alg 1	[26]	✗	✗	✗	✗	A.1	Cor A.1
(1)+(3)	SGD-SR	Alg 2	[6]	✗	✓	✗	✗	A.2	Cor A.2
(1)+(3)	SGD-MB	Alg 3	**NEW**	✗	✗	✗	✗	A.3	Cor A.3
(1)+(3)	SGD-star	Alg 4	**NEW**	✓	✓	✗	✗	A.4	Cor A.4
(1)+(3)	SAGA	Alg 5	[5]	✓	✗	✗	✗	A.5	Cor A.5
(1)+(3)	N-SAGA	Alg 6	**NEW**	✗	✗	✗	✗	A.6	Cor A.6
(1)	SEGA	Alg 7	[11]	✓	✗	✗	✓	A.7	Cor A.7
(1)	N-SEGA	Alg 8	**NEW**	✗	✗	✗	✓	A.8	Cor A.8
(1)+(3)	SVRG[a]	Alg 9	[15]	✓	✗	✗	✗	A.9	Cor A.9
(1)+(3)	L-SVRG	Alg 10	[13, 18]	✓	✗	✗	✗	A.10	Cor A.10
(1)+(3)	DIANA	Alg 11	[20, 14]	✗	✗	✓	✗	A.11	Cor A.11
(1)+(3)	DIANA[b]	Alg 12	[20, 14]	✓	✗	✓	✗	A.11	Cor A.12
(1)+(3)	Q-SGD-SR	Alg 13	**NEW**	✗	✓	✓	✗	A.12	Cor A.13
(1)+(3)+(4)	VR-DIANA	Alg 14	[14]	✓	✗	✓	✗	A.13	Cor A.15
(1)+(3)	JacSketch	Alg 15	[9]	✓	✓✗	✗	✗	A.14	Cor A.16

Table 1: List of specific existing (in some cases generalized) and new methods which fit our general analysis framework. VR = variance reduced method, AS = arbitrary sampling, Quant = supports gradient quantization, RCD = randomized coordinate descent type method. [a] Special case of SVRG with 1 outer loop only; [b] Special case of DIANA with 1 node and quantization of exact gradient.

Обсуждение

- Недавно этот подход был адаптирован для случая выпуклых (не сильно) функций в статье Khaled, A., Sebbouh, O., Loizou, N., Gower, R.M. and Richtárik, P., 2020. **Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization.** *arXiv preprint* arXiv:2006.11573.
- Ускоренные стохастические методы данным подходом не покрываются.
- Методы со смещённым стохастическим градиентом (например, SAG) тоже не покрываются описанным подходом.

Нижние оценки:
Woodworth, B.E. and Srebro, N. **Tight complexity bounds for optimizing composite objectives.** *In Advances in neural information processing systems*, pp. 3639-3647, *(NIPS 2016)*.
Оптимальные методы:
Allen-Zhu, Z., 2017. **Katyusha: The first direct acceleration of stochastic gradient methods.** The Journal of Machine Learning Research, 18(1), pp.8194-8244.

Обсуждение

- Недавно этот подход был адаптирован для случая выпуклых (не сильно) функций в статье Khaled, A., Sebbouh, O., Loizou, N., Gower, R.M. and Richtárik, P., 2020. **Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization.** *arXiv preprint* arXiv:2006.11573.
- Ускоренные стохастические методы данным подходом не покрываются.
- Методы со смещённым стохастическим градиентом (например, SAG) тоже не покрываются описанным подходом.

Нижние оценки:
Woodworth, B.E. and Srebro, N. **Tight complexity bounds for optimizing composite objectives.** *In Advances in neural information processing systems*, pp. 3639-3647, *(NIPS 2016)*.
Оптимальные методы:
Allen-Zhu, Z., 2017. **Katyusha: The first direct acceleration of stochastic gradient methods.** The Journal of Machine Learning Research, 18(1), pp.8194-8244.

Обсуждение

- Недавно этот подход был адаптирован для случая выпуклых (не сильно) функций в статье Khaled, A., Sebbouh, O., Loizou, N., Gower, R.M. and Richtárik, P., 2020. **Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization.** *arXiv preprint* arXiv:2006.11573.
- Ускоренные стохастические методы данным подходом не покрываются.
- Методы со смещённым стохастическим градиентом (например, SAG) тоже не покрываются описанным подходом.

Нижние оценки:
Woodworth, B.E. and Srebro, N. **Tight complexity bounds for optimizing composite objectives.** *In Advances in neural information processing systems*, pp. 3639-3647, *(NIPS 2016)*.
Оптимальные методы:
Allen-Zhu, Z., 2017. **Katyusha: The first direct acceleration of stochastic gradient methods.** The Journal of Machine Learning Research, 18(1), pp.8194-8244.

Обсуждение

- Недавно этот подход был адаптирован для случая выпуклых (не сильно) функций в статье Khaled, A., Sebbouh, O., Loizou, N., Gower, R.M. and Richtárik, P., 2020. **Unified Analysis of Stochastic Gradient Methods for Composite Convex and Smooth Optimization.** *arXiv preprint* arXiv:2006.11573.
- Ускоренные стохастические методы данным подходом не покрываются.
- Методы со смещённым стохастическим градиентом (например, SAG) тоже не покрываются описанным подходом.

Нижние оценки:
Woodworth, B.E. and Srebro, N. **Tight complexity bounds for optimizing composite objectives.** *In Advances in neural information processing systems*, pp. 3639-3647, *(NIPS 2016)*.
Оптимальные методы:
Allen-Zhu, Z., 2017. **Katyusha: The first direct acceleration of stochastic gradient methods.** The Journal of Machine Learning Research, 18(1), pp.8194-8244.

Выбор шага в SGD: адаптивный подбор

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В случае, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ — L-гладкая выпуклая функция, $L \in [\underline{L}, \overline{L}]$ и есть доступ к стохастическому градиенту $\nabla f(\mathbf{x}, \xi)$ с равномерно ограниченной дисперсией σ^2, можно получить адаптивный по размерам мини-батча r_k и шага γ_k вариант SGD, итерация внешнего цикла которого имеет вид:

- $L_{k+1} := \frac{L_k}{4}$ (для теории $L_{k+1} := \max\left\{\frac{L_k}{2}, \underline{L}\right\}$)
- Повторять
  - $L_{k+1} := 2L_{k+1}$ (для теории $L_{k+1} := \min\left\{2L_{k+1}, \overline{L}\right\}$), $r_{k+1} = \max\left\{\frac{\sigma_0}{L_{k+1}\varepsilon}, 1\right\}$ (для теории выражение совпадает с точностью до логарифмов и множителя $\frac{\overline{L}^2}{\underline{L}^2}$)
  - $x^{k+1} = x^k - \frac{1}{2L_{k+1}} \nabla^{r_{k+1}} f(\mathbf{x}^k) = x^k - \frac{1}{2L_{k+1} r_{k+1}} \sum_{i=1}^{r_{k+1}} \nabla f(\mathbf{x}^k, \xi_i^{k+1})$
- Пока не выполнится неравенство:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon}{2}.$$

Выбор шага в SGD: адаптивный подбор

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В случае, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ — L-гладкая выпуклая функция, $L \in [\underline{L}, \overline{L}]$ и есть доступ к стохастическому градиенту $\nabla f(\mathbf{x}, \xi)$ с равномерно ограниченной дисперсией σ^2, можно получить адаптивный по размерам мини-батча r_k и шага γ_k вариант SGD, итерация внешнего цикла которого имеет вид:

- $L_{k+1} := \frac{L_k}{4}$ (для теории $L_{k+1} := \max\left\{\frac{L_k}{2}, \underline{L}\right\}$)
- Повторять
  - $L_{k+1} := 2L_{k+1}$ (для теории $L_{k+1} := \min\left\{2L_{k+1}, \overline{L}\right\}$), $r_{k+1} = \max\left\{\frac{\sigma_0}{L_{k+1}\varepsilon}, 1\right\}$ (для теории выражение совпадает с точностью до логарифмов и множителя $\frac{\overline{L}^2}{\underline{L}^2}$)
  - $x^{k+1} = x^k - \frac{1}{2L_{k+1}} \nabla^{r_{k+1}} f(\mathbf{x}^k) = x^k - \frac{1}{2L_{k+1} r_{k+1}} \sum_{i=1}^{r_{k+1}} \nabla f(\mathbf{x}^k, \xi_i^{k+1})$
- Пока не выполнится неравенство:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon}{2}.$$

Выбор шага в SGD: адаптивный подбор

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В случае, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ — L-гладкая выпуклая функция, $L \in [\underline{L}, \overline{L}]$ и есть доступ к стохастическому градиенту $\nabla f(\mathbf{x}, \xi)$ с равномерно ограниченной дисперсией σ^2, можно получить адаптивный по размерам мини-батча r_k и шага γ_k вариант SGD, итерация внешнего цикла которого имеет вид:

- $L_{k+1} := \frac{L_k}{4}$ (для теории $L_{k+1} := \max\left\{\frac{L_k}{2}, \underline{L}\right\}$)
- Повторять
  - $L_{k+1} := 2L_{k+1}$ (для теории $L_{k+1} := \min\left\{2L_{k+1}, \overline{L}\right\}$), $r_{k+1} = \max\left\{\frac{\sigma \mathbf{0}}{L_{k+\mathbf{1}}\varepsilon}, 1\right\}$ (для теории выражение совпадает с точностью до логарифмов и множителя $\frac{\overline{L}^{\mathbf{2}}}{\underline{L}^{\mathbf{2}}}$)
  - $x^{k+1} = x^k - \frac{1}{2L_{k+1}} \nabla^{r_{k+1}} f(\mathbf{x}^k) = x^k - \frac{1}{2L_{k+1} r_{k+1}} \sum_{i=1}^{r_{k+1}} \nabla f(\mathbf{x}^k, \xi_i^{k+1})$
- Пока не выполнится неравенство:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon}{2}.$$

Выбор шага в SGD: адаптивный подбор

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В случае, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ — L-гладкая выпуклая функция, $L \in [\underline{L}, \overline{L}]$ и есть доступ к стохастическому градиенту $\nabla f(\mathbf{x}, \xi)$ с равномерно ограниченной дисперсией σ^2, можно получить адаптивный по размерам мини-батча r_k и шага γ_k вариант SGD, итерация внешнего цикла которого имеет вид:

- $L_{k+1} := \frac{L_k}{4}$ (для теории $L_{k+1} := \max\left\{\frac{L_k}{2}, \underline{L}\right\}$)
- Повторять
  - $L_{k+1} := 2L_{k+1}$ (для теории $L_{k+1} := \min\left\{2L_{k+1}, \overline{L}\right\}$), $r_{k+1} = \max\left\{\frac{\sigma_0}{L_{k+1}\varepsilon}, 1\right\}$ (для теории выражение совпадает с точностью до логарифмов и множителя $\frac{\overline{L}^2}{\underline{L}^2}$)
  - $x^{k+1} = x^k - \frac{1}{2L_{k+1}}\nabla^{r_{k+1}} f(\mathbf{x}^k) = x^k - \frac{1}{2L_{k+1} r_{k+1}} \sum_{i=1}^{r_{k+1}} \nabla f(\mathbf{x}^k, \xi_i^{k+1})$
- Пока не выполнится неравенство:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon}{2}.$$

Выбор шага в SGD: адаптивный подбор

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В случае, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ — L-гладкая выпуклая функция, $L \in [\underline{L}, \overline{L}]$ и есть доступ к стохастическому градиенту $\nabla f(\mathbf{x}, \xi)$ с равномерно ограниченной дисперсией σ^2, можно получить адаптивный по размерам мини-батча r_k и шага γ_k вариант SGD, итерация внешнего цикла которого имеет вид:

- $L_{k+1} := \frac{L_k}{4}$ (для теории $L_{k+1} := \max\left\{\frac{L_k}{2}, \underline{L}\right\}$)
- Повторять
  - $L_{k+1} := 2L_{k+1}$ (для теории $L_{k+1} := \min\left\{2L_{k+1}, \overline{L}\right\}$), $r_{k+1} = \max\left\{\frac{\sigma_{\mathbf{0}}}{L_{k+\mathbf{1}}\varepsilon}, 1\right\}$ (для теории выражение совпадает с точностью до логарифмов и множителя $\frac{\overline{L}^{\mathbf{2}}}{\underline{L}^{\mathbf{2}}}$)
  - $x^{k+1} = x^k - \frac{1}{2L_{k+\mathbf{1}}}\nabla^{r_{k+\mathbf{1}}} f(\mathbf{x}^k) = x^k - \frac{1}{2L_{k+\mathbf{1}}r_{k+\mathbf{1}}}\sum_{i=1}^{r_{k+\mathbf{1}}} \nabla f(\mathbf{x}^k, \xi_i^{k+1})$
- Пока не выполнится неравенство:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+\mathbf{1}}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon}{2}.$$

Выбор шага в SGD: адаптивный подбор

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В случае, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)]$ — L-гладкая выпуклая функция, $L \in [\underline{L}, \overline{L}]$ и есть доступ к стохастическому градиенту $\nabla f(\mathbf{x}, \xi)$ с равномерно ограниченной дисперсией σ^2, можно получить адаптивный по размерам мини-батча r_k и шага γ_k вариант SGD, итерация внешнего цикла которого имеет вид:

- $L_{k+1} := \frac{L_k}{4}$ (для теории $L_{k+1} := \max\left\{\frac{L_k}{2}, \underline{L}\right\}$)
- Повторять
  - $L_{k+1} := 2L_{k+1}$ (для теории $L_{k+1} := \min\left\{2L_{k+1}, \overline{L}\right\}$), $r_{k+1} = \max\left\{\frac{\sigma_{\mathbf{0}}}{L_{k+\mathbf{1}}\varepsilon}, 1\right\}$ (для теории выражение совпадает с точностью до логарифмов и множителя $\frac{\overline{L}^{\mathbf{2}}}{\underline{L}^{\mathbf{2}}}$)
  - $x^{k+1} = x^k - \frac{1}{2L_{k+\mathbf{1}}}\nabla^{r_{k+\mathbf{1}}} f(\mathbf{x}^k) = x^k - \frac{1}{2L_{k+\mathbf{1}}r_{k+\mathbf{1}}}\sum_{i=1}^{r_{k+\mathbf{1}}} \nabla f(\mathbf{x}^k, \xi_i^{k+1})$
- Пока не выполнится неравенство:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+\mathbf{1}}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon}{2}.$$

Выбор шага в SGD: адаптивный подбор

Описанный метод находит точку $\bar{\mathbf{x}}^N$, такую что $\mathbb{E}[f(\bar{\mathbf{x}}^N) - f(\mathbf{x}^*)] \leq \varepsilon$, за

$$N = O\left(\frac{\bar{L}R_0^2}{\varepsilon}\right) \text{ итераций,}$$

причём суммарное число обращений к оракулу равняется

$$\widetilde{O}\left(\frac{\sigma_0^2 R_0^2 \bar{L}^3}{\varepsilon^2 \underline{L}^3}\right).$$

Описанную процедуру можно ускорить (на базе метода подобных треугольников) и улучшить оценку на число итераций до

$$N = O\left(\sqrt{\frac{\bar{L}R_0^2}{\varepsilon}}\right).$$

Выбор шага в SGD: адаптивный подбор

Описанный метод находит точку $\bar{\mathbf{x}}^N$, такую что $\mathbb{E}[f(\bar{\mathbf{x}}^N) - f(\mathbf{x}^*)] \leq \varepsilon$, за

$$N = O\left(\frac{\bar{L}R_0^2}{\varepsilon}\right) \text{ итераций,}$$

причём суммарное число обращений к оракулу равняется

$$\widetilde{O}\left(\frac{\sigma_0^2 R_0^2 \bar{L}^3}{\varepsilon^2 \underline{L}^3}\right).$$

Описанную процедуру можно ускорить (на базе метода подобных треугольников) и улучшить оценку на число итераций до

$$N = O\left(\sqrt{\frac{\bar{L}R_0^2}{\varepsilon}}\right).$$

Выбор шага в SGD: адаптивный подбор

Описанный метод находит точку $\overline{\mathbf{x}}^N$, такую что $\mathbb{E}[f(\overline{\mathbf{x}}^N) - f(\mathbf{x}^*)] \leq \varepsilon$, за

$$N = O\left(\frac{\overline{L}R_0^2}{\varepsilon}\right) \text{ итераций,}$$

причём суммарное число обращений к оракулу равняется

$$\widetilde{O}\left(\frac{\sigma_0^2 R_0^2 \overline{L}^3}{\varepsilon^2 \underline{L}^3}\right).$$

Описанную процедуру можно ускорить (на базе метода подобных треугольников) и улучшить оценку на число итераций до

$$N = O\left(\sqrt{\frac{\overline{L}R_0^2}{\varepsilon}}\right).$$

Выбор шага в SGD: адаптивный подбор

Описанный метод находит точку $\overline{\mathbf{x}}^N$, такую что $\mathbb{E}[f(\overline{\mathbf{x}}^N) - f(\mathbf{x}^*)] \leq \varepsilon$, за

$$N = O\left(\frac{\overline{L}R_0^2}{\varepsilon}\right) \text{ итераций,}$$

причём суммарное число обращений к оракулу равняется

$$\widetilde{O}\left(\frac{\sigma_0^2 R_0^2 \overline{L}^3}{\varepsilon^2 \underline{L}^3}\right).$$

Описанную процедуру можно ускорить (на базе метода подобных треугольников) и улучшить оценку на число итераций до

$$N = O\left(\sqrt{\frac{\overline{L}R_0^2}{\varepsilon}}\right).$$

Выбор шага в SGD: стохастическое правило Поляка

Loizou, N., Vaswani, S., Laradji, I. and Lacoste-Julien, S. **Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence.** *arXiv preprint* arXiv:2002.10542.

Рассмотрим случай, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$, $f_i(\mathbf{x})$ — μ_i-сильно выпуклая и L_i-гладкая функция (для всех $i = 1, \ldots, m$), ограниченная снизу константой f_i^*, и хотя бы одна $\mu_i > 0$. Можно показать, что SGD со стохастическим правилом Поляка для выбора размера шага

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_j(\mathbf{x}^k), \quad \gamma_k = \min\left\{\frac{f_j(\mathbf{x}^k) - f_j^*}{c\|\nabla f_j(\mathbf{x}^k)\|^2}, \gamma_b\right\},$$

где j выбирается случайно и равновероятно из $\{1, \ldots, m\}$, выполняется

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \leq (1 - \overline{\mu}\alpha)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma_b \sigma^2}{\overline{\mu}\alpha},$$

где $\alpha = \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$, $\overline{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$, $\sigma^2 = f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$.

Выбор шага в SGD: стохастическое правило Поляка

Loizou, N., Vaswani, S., Laradji, I. and Lacoste-Julien, S. **Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence.** *arXiv preprint* arXiv:2002.10542.

Рассмотрим случай, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$, $f_i(\mathbf{x})$ — μ_i-сильно выпуклая и L_i-гладкая функция (для всех $i = 1, \ldots, m$), ограниченная снизу константой f_i^*, и хотя бы одна $\mu_i > 0$. Можно показать, что SGD со стохастическим правилом Поляка для выбора размера шага

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_j(\mathbf{x}^k), \quad \gamma_k = \min\left\{\frac{f_j(\mathbf{x}^k) - f_j^*}{c\|\nabla f_j(\mathbf{x}^k)\|^2}, \gamma_b\right\},$$

где j выбирается случайно и равновероятно из $\{1, \ldots, m\}$, выполняется

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \leq (1 - \overline{\mu}\alpha)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma_b \sigma^2}{\overline{\mu}\alpha},$$

где $\alpha = \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$, $\overline{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$, $\sigma^2 = f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$.

Выбор шага в SGD: стохастическое правило Поляка

Loizou, N., Vaswani, S., Laradji, I. and Lacoste-Julien, S. **Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence.** *arXiv preprint* arXiv:2002.10542.

Рассмотрим случай, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$, $f_i(\mathbf{x})$ — μ_i-сильно выпуклая и L_i-гладкая функция (для всех $i = 1, \ldots, m$), ограниченная снизу константой f_i^*, и хотя бы одна $\mu_i > 0$. Можно показать, что SGD со стохастическим правилом Поляка для выбора размера шага

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_j(\mathbf{x}^k), \quad \gamma_k = \min\left\{\frac{f_j(\mathbf{x}^k) - f_j^*}{c\|\nabla f_j(\mathbf{x}^k)\|^2}, \gamma_b\right\},$$

где j выбирается случайно и равновероятно из $\{1, \ldots, m\}$, выполняется

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \leq (1 - \overline{\mu}\alpha)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma_b \sigma^2}{\overline{\mu}\alpha},$$

где $\alpha = \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$, $\overline{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$, $\sigma^2 = f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$.

Выбор шага в SGD: стохастическое правило Поляка

Loizou, N., Vaswani, S., Laradji, I. and Lacoste-Julien, S. **Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence.** *arXiv preprint* arXiv:2002.10542.

Рассмотрим случай, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$, $f_i(\mathbf{x})$ — μ_i-сильно выпуклая и L_i-гладкая функция (для всех $i = 1, \ldots, m$), ограниченная снизу константой f_i^*, и хотя бы одна $\mu_i > 0$. Можно показать, что SGD со стохастическим правилом Поляка для выбора размера шага

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_j(\mathbf{x}^k), \quad \gamma_k = \min\left\{\frac{f_j(\mathbf{x}^k) - f_j^*}{c\|\nabla f_j(\mathbf{x}^k)\|^2}, \gamma_b\right\},$$

где j выбирается случайно и равновероятно из $\{1, \ldots, m\}$, выполняется

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \leq (1 - \overline{\mu}\alpha)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma_b \sigma^2}{\overline{\mu}\alpha},$$

где $\alpha = \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$, $\overline{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$, $\sigma^2 = f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$.

Выбор шага в SGD: стохастическое правило Поляка

Loizou, N., Vaswani, S., Laradji, I. and Lacoste-Julien, S. **Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence.** *arXiv preprint* arXiv:2002.10542.

Рассмотрим случай, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$, $f_i(\mathbf{x})$ — μ_i-сильно выпуклая и L_i-гладкая функция (для всех $i = 1, \ldots, m$), ограниченная снизу константой f_i^*, и хотя бы одна $\mu_i > 0$. Можно показать, что SGD со стохастическим правилом Поляка для выбора размера шага

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_j(\mathbf{x}^k), \quad \gamma_k = \min\left\{\frac{f_j(\mathbf{x}^k) - f_j^*}{c\|\nabla f_j(\mathbf{x}^k)\|^2}, \gamma_b\right\},$$

где j выбирается случайно и равновероятно из $\{1, \ldots, m\}$, выполняется

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \leq (1 - \overline{\mu}\alpha)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma_b \sigma^2}{\overline{\mu}\alpha},$$

где $\alpha = \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$, $\overline{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$, $\sigma^2 = f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$.

Выбор шага в SGD: стохастическое правило Поляка

Loizou, N., Vaswani, S., Laradji, I. and Lacoste-Julien, S. **Stochastic polyak step-size for SGD: An adaptive learning rate for fast convergence.** *arXiv preprint* arXiv:2002.10542.

Рассмотрим случай, когда $R(\mathbf{x}) \equiv 0$, $f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x})$, $f_i(\mathbf{x})$ — μ_i-сильно выпуклая и L_i-гладкая функция (для всех $i = 1, \ldots, m$), ограниченная снизу константой f_i^*, и хотя бы одна $\mu_i > 0$. Можно показать, что SGD со стохастическим правилом Поляка для выбора размера шага

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_j(\mathbf{x}^k), \quad \gamma_k = \min\left\{\frac{f_j(\mathbf{x}^k) - f_j^*}{c\|\nabla f_j(\mathbf{x}^k)\|^2}, \gamma_b\right\},$$

где j выбирается случайно и равновероятно из $\{1, \ldots, m\}$, выполняется

$$\mathbb{E}\|\mathbf{x}^k - \mathbf{x}^*\|_2^2 \leq (1 - \overline{\mu}\alpha)^k \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{2\gamma_b \sigma^2}{\overline{\mu}\alpha},$$

где $\alpha = \min\{\frac{1}{2cL_{\max}}, \gamma_b\}$, $\overline{\mu} = \frac{1}{m}\sum_{i=1}^m \mu_i$, $\sigma^2 = f^* - \frac{1}{m}\sum_{i=1}^m f_i^*$.

Задача невыпуклой оптимизации

Рассмотрим следующую задачу:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \tag{11}$$

- $f(\mathbf{x})$ — L-гладкая функция
- $f(\mathbf{x})$ — ограничена снизу некоторой константой f^*

Плохие новости:

- В общем случае задача (11) может не иметь решения.
- Задача (11) может быть многоэкстремальной.
- В общем случае есть только эвристики, позволяющие найти глобальный минимум, если он существует (мультистарт, метод иммитации отжига, генетические алгоритмы и др.)

Задача невыпуклой оптимизации

Рассмотрим следующую задачу:

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \tag{11}$$

- $f(\mathbf{x})$ — L-гладкая функция
- $f(\mathbf{x})$ — ограничена снизу некоторой константой f^*

Плохие новости:

- В общем случае задача (11) может не иметь решения.
- Задача (11) может быть многоэкстремальной.
- В общем случае есть только эвристики, позволяющие найти глобальный минимум, если он существует (мультистарт, метод иммитации отжига, генетические алгоритмы и др.)

Задача невыпуклой оптимизации

Рассмотрим следующую задачу:

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \tag{11}$$

- $f(\mathbf{x})$ — L-гладкая функция
- $f(\mathbf{x})$ — ограничена снизу некоторой константой f^*

Плохие новости:

- В общем случае задача (11) может не иметь решения.
- Задача (11) может быть многоэкстремальной.
- В общем случае есть только эвристики, позволяющие найти глобальный минимум, если он существует (мультистарт, метод иммитации отжига, генетические алгоритмы и др.)

Задача невыпуклой оптимизации

Рассмотрим следующую задачу:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \tag{11}$$

- $f(\mathbf{x})$ — L-гладкая функция
- $f(\mathbf{x})$ — ограничена снизу некоторой константой f^*

Плохие новости:

- В общем случае задача (11) может не иметь решения.
- Задача (11) может быть многоэкстремальной.
- В общем случае есть только эвристики, позволяющие найти глобальный минимум, если он существует (мультистарт, метод иммитации отжига, генетические алгоритмы и др.)

Задача невыпуклой оптимизации

Рассмотрим следующую задачу:

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x}) \tag{11}$$

- $f(\mathbf{x})$ — L-гладкая функция
- $f(\mathbf{x})$ — ограничена снизу некоторой константой f^*

Плохие новости:

- В общем случае задача (11) может не иметь решения.
- Задача (11) может быть многоэкстремальной.
- В общем случае есть только эвристики, позволяющие найти глобальный минимум, если он существует (мультистарт, метод иммитации отжига, генетические алгоритмы и др.)

Задача невыпуклой оптимизации

Рассмотрим следующую задачу:

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \tag{11}$$

- $f(\mathbf{x})$ — L-гладкая функция
- $f(\mathbf{x})$ — ограничена снизу некоторой константой f^*

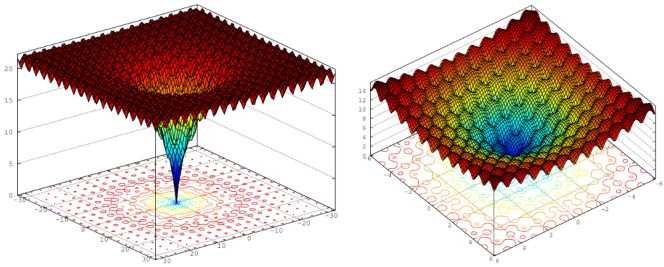
Плохие новости:

- В общем случае задача (11) может не иметь решения.
- Задача (11) может быть многоэкстремальной.
- В общем случае есть только эвристики, позволяющие найти глобальный минимум, если он существует (мультистарт, метод иммитации отжига, генетические алгоритмы и др.)

Многоэкстремальная задача

$$f(x) = 20 + e - 20e^{-\frac{1}{5}\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}} - e^{\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)}$$

Dimensions: n
Domain: $-32.768 \leq x_i \leq 32.768$
Global Optimum: $f(x) = 0.0$ at $x = (0.0, 0.0, \ldots, 0.0)$
Operator: AckleyEvaluator
Charts:



(a) [-32.768, 32.768] (b) [-6.0, 6.0]

Figure 1: Ackley function plots.

Figure: https://dev.heuristiclab.com/trac.fcgi/wiki/Documentation/Reference

Задача невыпуклой оптимизации

Хорошие новости:

- В некоторых задачах достаточно найти любой локальный минимум (например, такие задачи возникают в биологии, медицине).
- Есть содержательная (с нижними оценками и оптимальными методами) теория для поиска *локальных минимумов*, а точнее для поиска ε-стационарных точек: $\|\nabla f(\hat{\mathbf{x}})\|_2 \leq \varepsilon$ или $\mathbb{E}[\|\nabla f(\hat{\mathbf{x}})\|_2^2] \leq \varepsilon^2$.

Задача невыпуклой оптимизации

Хорошие новости:

- В некоторых задачах достаточно найти любой локальный минимум (например, такие задачи возникают в биологии, медицине).
- Есть содержательная (с нижними оценками и оптимальными методами) теория для поиска *локальных минимумов*, а точнее для поиска ε-стационарных точек: $\|\nabla f(\hat{\mathbf{x}})\|_2 \leq \varepsilon$ или $\mathbb{E}[\|\nabla f(\hat{\mathbf{x}})\|_2^2] \leq \varepsilon^2$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_{\xi}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j\left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_\xi[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_{\xi}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j\left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_{\xi}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j\left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_{\xi}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j\left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_{\xi}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j\left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Задача невыпуклой оптимизации: 3 основных случая

1. Доступен полный градиент функции $f(\mathbf{x})$.
2. Целевая функция имеет вид
$$f(\mathbf{x}) = \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)] \tag{12}$$
и есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём $\mathbb{E}_{\xi}\left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$ для всех $\mathbf{x} \in \mathbb{R}^n$.
3. Целевая функция имеет вид
$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}), \tag{13}$$
причём функции $f_i(\mathbf{x})$ являются L-гладкими для всех $i = \overline{1, m}$, подсчёт градиентов r слагаемых стоит $O(r)$ времени и для индекса ξ, который выбирается случайно равновероятно из множества $\{1, \ldots, m\}$ выполняется $\mathbb{E}_j\left[\|\nabla f_j(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2\right] \leq \sigma^2$, причём допускается ситуация, что $\sigma = +\infty$.

Общий взгляд на оптимальные методы первого порядка

Как это ни странно, но для всех трёх случаев итерация оптимального метода будет иметь вид:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k, \tag{14}$$

где $\gamma_k \leq \frac{1}{2L}$, а $\mathbf{g}^k$ — некоторый стох. градиент, который, вообще говоря, может быть смещённым.

Общий взгляд на оптимальные методы первого порядка

Как это ни странно, но для всех трёх случаев итерация оптимального метода будет иметь вид:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k, \tag{14}$$

где $\gamma_k \leq \frac{1}{2L}$, а $\mathbf{g}^k$ — некоторый стох. градиент, который, вообще говоря, может быть смещённым.

Общий взгляд на оптимальные методы первого порядка

Функция f является L-гладкой, поэтому

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) + \langle \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \langle \nabla f(x^k) - \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2,
\end{aligned}
$$

где последнее неравенство следует из неравенства Фенхеля-Юнга:

$$
\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \; \forall \eta > 0 \quad \langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2\eta}\|\mathbf{a}\|_2^2 + \frac{\eta}{2}\|\mathbf{b}\|_2^2.
$$

Общий взгляд на оптимальные методы первого порядка

Функция f является L-гладкой, поэтому

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) + \langle \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \langle \nabla f(x^k) - \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2,
\end{aligned}
$$

где последнее неравенство следует из неравенства Фенхеля-Юнга:

$$
\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \; \forall \eta > 0 \quad \langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2\eta}\|\mathbf{a}\|_2^2 + \frac{\eta}{2}\|\mathbf{b}\|_2^2.
$$

Общий взгляд на оптимальные методы первого порядка

Функция f является L-гладкой, поэтому

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\ &= f(\mathbf{x}^k) + \langle \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \langle \nabla f(x^k) - \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\ &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2, \end{aligned}$$

где последнее неравенство следует из неравенства Фенхеля-Юнга:

$$\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \; \forall \eta > 0 \quad \langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2\eta}\|\mathbf{a}\|_2^2 + \frac{\eta}{2}\|\mathbf{b}\|_2^2.$$

Общий взгляд на оптимальные методы первого порядка

Функция f является L-гладкой, поэтому

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) + \langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\ &= f(\mathbf{x}^k) + \langle \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \langle \nabla f(x^k) - \mathbf{g}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{L}{2}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\ &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right)\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2, \end{aligned}$$

где последнее неравенство следует из неравенства Фенхеля-Юнга:

$$\forall \mathbf{a}, \mathbf{b} \in \mathbb{R}^n \; \forall \eta > 0 \quad \langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{1}{2\eta}\|\mathbf{a}\|_2^2 + \frac{\eta}{2}\|\mathbf{b}\|_2^2.$$

Общий взгляд на оптимальные методы первого порядка

Поскольку $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k$ и $\gamma_k \leq \frac{1}{2L}$, мы можем упростить предыдущее неравенство:

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right) \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{L\gamma_k^2}{2}\right) \|\mathbf{g}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{\gamma_k}{4}\right) \|\mathbf{g}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \frac{\gamma_k}{2} \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2. \qquad (15)
\end{aligned}
$$

Без дополнительных предположений на $\mathbf{g}^k$ получить что-то полезное не удаётся.

Общий взгляд на оптимальные методы первого порядка

Поскольку $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k$ и $\gamma_k \leq \frac{1}{2L}$, мы можем упростить предыдущее неравенство:

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right) \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{L\gamma_k^2}{2}\right) \|\mathbf{g}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{\gamma_k}{4}\right) \|\mathbf{g}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \frac{\gamma_k}{2} \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2. \qquad (15)
\end{aligned}
$$

Без дополнительных предположений на $\mathbf{g}^k$ получить что-то полезное не удаётся.

Общий взгляд на оптимальные методы первого порядка

Поскольку $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k$ и $\gamma_k \leq \frac{1}{2L}$, мы можем упростить предыдущее неравенство:

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right) \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{L\gamma_k^2}{2}\right) \|\mathbf{g}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{\gamma_k}{4}\right) \|\mathbf{g}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \frac{\gamma_k}{2} \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2. \qquad (15)
\end{aligned}
$$

Без дополнительных предположений на $\mathbf{g}^k$ получить что-то полезное не удаётся.

Общий взгляд на оптимальные методы первого порядка

Поскольку $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k$ и $\gamma_k \leq \frac{1}{2L}$, мы можем упростить предыдущее неравенство:

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right) \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{L\gamma_k^2}{2}\right) \|\mathbf{g}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{\gamma_k}{4}\right) \|\mathbf{g}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \frac{\gamma_k}{2} \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2. \qquad (15)
\end{aligned}
$$

Без дополнительных предположений на $\mathbf{g}^k$ получить что-то полезное не удаётся.

Общий взгляд на оптимальные методы первого порядка

Поскольку $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k$ и $\gamma_k \leq \frac{1}{2L}$, мы можем упростить предыдущее неравенство:

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{1}{4\gamma_k} + \frac{L}{2}\right) \|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{L\gamma_k^2}{2}\right) \|\mathbf{g}^k\|_2^2 \\
&\leq f(\mathbf{x}^k) - \gamma_k \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 + \left(\frac{\gamma_k}{4} + \frac{\gamma_k}{4}\right) \|\mathbf{g}^k\|_2^2 \\
&= f(\mathbf{x}^k) - \frac{\gamma_k}{2} \|\mathbf{g}^k\|_2^2 + \gamma_k \|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2. \qquad (15)
\end{aligned}
$$

Без дополнительных предположений на $\mathbf{g}^k$ получить что-то полезное не удаётся.

Детерминированный случай: градиентный спуск

Предположим теперь, что для всех k вектор $\mathbf{g}^k$ удовлетворяет следующему неравенству:

$$\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \leq \frac{\varepsilon^2}{10}. \tag{16}$$

В частности, $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ удовлетворяет этому предположению. Кроме того, бывают ситуации, когда для подсчёта градиента функции требуется решать некоторую задачу оптимизации с некоторой точностью.

Детерминированный случай: градиентный спуск

Предположим теперь, что для всех k вектор $\mathbf{g}^k$ удовлетворяет следующему неравенству:

$$\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \leq \frac{\varepsilon^2}{10}. \tag{16}$$

В частности, $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$ удовлетворяет этому предположению. Кроме того, бывают ситуации, когда для подсчёта градиента функции требуется решать некоторую задачу оптимизации с некоторой точностью.

Детерминированный случай: градиентный спуск

Поступим следующим образом: будем останавливать метод на итерации N, если $\|\mathbf{g}^N\|_2^2 \leq \frac{2\varepsilon^2}{5}$. В таком случае для последней точки $\mathbf{x}^N$ будет выполняться:

$$\begin{aligned}\|\nabla f(\mathbf{x}^N)\|_2^2 &= \|\nabla f(\mathbf{x}^N) - \mathbf{g}^N + \mathbf{g}^N\|_2^2 \\ &\leq 2\|\nabla f(\mathbf{x}^N) - \mathbf{g}^N\|_2^2 + 2\|\mathbf{g}^N\|_2^2 \overset{(16)}{\leq} \varepsilon^2.\end{aligned}$$

Если $\gamma_k \equiv \gamma = \frac{1}{2L}$, то метод остановится через $N \leq \frac{2L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2} = O\left(\frac{L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2}\right)$ итераций.

Детерминированный случай: градиентный спуск

Поступим следующим образом: будем останавливать метод на итерации N, если $\|\mathbf{g}^N\|_2^2 \leq \frac{2\varepsilon^2}{5}$. В таком случае для последней точки $\mathbf{x}^N$ будет выполняться:

$$\begin{aligned}\|\nabla f(\mathbf{x}^N)\|_2^2 &= \|\nabla f(\mathbf{x}^N) - \mathbf{g}^N + \mathbf{g}^N\|_2^2 \\ &\leq 2\|\nabla f(\mathbf{x}^N) - \mathbf{g}^N\|_2^2 + 2\|\mathbf{g}^N\|_2^2 \overset{(16)}{\leq} \varepsilon^2.\end{aligned}$$

Если $\gamma_k \equiv \gamma = \frac{1}{2L}$, то метод остановится через $N \leq \frac{2L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2} = O\left(\frac{L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2}\right)$ итераций.

Детерминированный случай: градиентный спуск

Поступим следующим образом: будем останавливать метод на итерации N, если $\|\mathbf{g}^N\|_2^2 \leq \frac{2\varepsilon^2}{5}$. В таком случае для последней точки $\mathbf{x}^N$ будет выполняться:

$$\begin{aligned}\|\nabla f(\mathbf{x}^N)\|_2^2 &= \|\nabla f(\mathbf{x}^N) - \mathbf{g}^N + \mathbf{g}^N\|_2^2 \\ &\leq 2\|\nabla f(\mathbf{x}^N) - \mathbf{g}^N\|_2^2 + 2\|\mathbf{g}^N\|_2^2 \overset{(16)}{\leq} \varepsilon^2.\end{aligned}$$

Если $\gamma_k \equiv \gamma = \frac{1}{2L}$, то метод остановится через $N \leq \frac{20L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2} = O\left(\frac{L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2}\right)$ итераций.

Детерминированный случай: градиентный спуск

Поступим следующим образом: будем останавливать метод на итерации N, если $\|\mathbf{g}^N\|_2^2 \leq \frac{2\varepsilon^2}{5}$. В таком случае для последней точки $\mathbf{x}^N$ будет выполняться:

$$\begin{aligned}\|\nabla f(\mathbf{x}^N)\|_2^2 &= \|\nabla f(\mathbf{x}^N) - \mathbf{g}^N + \mathbf{g}^N\|_2^2 \\ &\leq 2\|\nabla f(\mathbf{x}^N) - \mathbf{g}^N\|_2^2 + 2\|\mathbf{g}^N\|_2^2 \overset{(16)}{\leq} \varepsilon^2.\end{aligned}$$

Если $\gamma_k \equiv \gamma = \frac{1}{2L}$, то метод остановится через $N \leq \frac{2L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2} = O\left(\frac{L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2}\right)$ итераций.

Детерминированный случай: градиентный спуск

Поступим следующим образом: будем останавливать метод на итерации N, если $\|\mathbf{g}^N\|_2^2 \leq \frac{2\varepsilon^2}{5}$. В таком случае для последней точки $\mathbf{x}^N$ будет выполняться:

$$\begin{aligned}\|\nabla f(\mathbf{x}^N)\|_2^2 &= \|\nabla f(\mathbf{x}^N) - \mathbf{g}^N + \mathbf{g}^N\|_2^2 \\ &\leq 2\|\nabla f(\mathbf{x}^N) - \mathbf{g}^N\|_2^2 + 2\|\mathbf{g}^N\|_2^2 \overset{(16)}{\leq} \varepsilon^2.\end{aligned}$$

Если $\gamma_k \equiv \gamma = \frac{1}{2L}$, то метод остановится через $N \leq \frac{20L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2} = O\left(\frac{L(f(\mathbf{x}^0) - f^*)}{\varepsilon^2}\right)$ итераций.

Стохастический случай: SGD, равномерно ограниченная дисперсия

Ghadimi, S. and Lan, G., 2013. **Stochastic first-and zeroth-order methods for nonconvex stochastic programming.** *SIAM Journal on Optimization*, 23(4), pp.2341-2368.

Теперь рассмотрим второй из трёх случаев. В этом случае у нас есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём

$$\mathbb{E}_\xi[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x}), \quad \mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2 \right] \leq \sigma^2$$

В этом случае (для простоты) мы воспользуемся мини-батчингом: будем вычислять $\mathbf{g}^k$ как среднее арифметическое r независимых (для фиксированного $\mathbf{x}^k$) стох. градиентов $\nabla f(\mathbf{x}^k, \xi_1^k), \ldots, \nabla f(\mathbf{x}^k, \xi_r^k)$, посчитанных в точке $\mathbf{x}^k$:

$$\mathbf{g}^k = \frac{1}{r} \sum_{i=1}^{r} \nabla f(\mathbf{x}^k, \xi_i^k).$$

Стохастический случай: SGD, равномерно ограниченная дисперсия

Ghadimi, S. and Lan, G., 2013. **Stochastic first-and zeroth-order methods for nonconvex stochastic programming.** *SIAM Journal on Optimization*, 23(4), pp.2341-2368.

Теперь рассмотрим второй из трёх случаев. В этом случае у нас есть доступ к стох. градиенту $\nabla f(\mathbf{x}, \xi)$, причём

$$\mathbb{E}_\xi[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x}), \quad \mathbb{E}_\xi \left[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|_2^2 \right] \leq \sigma^2$$

В этом случае (для простоты) мы воспользуемся мини-батчингом: будем вычислять $\mathbf{g}^k$ как среднее арифметическое r независимых (для фиксированного $\mathbf{x}^k$) стох. градиентов $\nabla f(\mathbf{x}^k, \xi_1^k), \ldots, \nabla f(\mathbf{x}^k, \xi_r^k)$, посчитанных в точке $\mathbf{x}^k$:

$$\mathbf{g}^k = \frac{1}{r} \sum_{i=1}^{r} \nabla f(\mathbf{x}^k, \xi_i^k).$$

Стохастический случай: SGD, равномерно ограниченная дисперсия

Отсюда следует, что

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] \leq \frac{\sigma^2}{r}.$$

Таким образом, беря $r = \max\left\{1, \frac{2\sigma^2}{\varepsilon^2}\right\}$, мы получаем оценку на дисперсию:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\right\|_2^2 \mid \mathbf{x}^k\right] \leq \frac{\varepsilon^2}{2}. \tag{17}$$

Стохастический случай: SGD, равномерно ограниченная дисперсия

Отсюда следует, что

$$\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2 \mid \mathbf{x}^k\right] \quad \leq \quad \frac{\sigma^2}{r}.$$

Таким образом, беря $r = \max\left\{1, \frac{2\sigma^2}{\varepsilon^2}\right\}$, мы получаем оценку на дисперсию:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\right\|_2^2 \mid \mathbf{x}^k\right] \leq \frac{\varepsilon^2}{2}. \tag{17}$$

Стохастический случай: SGD, равномерно ограниченная дисперсия

Подставляя эту оценку в исходное неравенство и беря $\gamma_k = \frac{1}{2L}$, получаем

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^N)\|_2^2\right] \quad \leq \quad \frac{4L\left(f(\mathbf{x}^0) - f^*\right)}{N} + \frac{\varepsilon^2}{2},$$

где точка $\hat{\mathbf{x}}^N$ выбирается случайно равновероятно среди $\{\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{N-1}\}$. Таким образом, если мы возьмём $N = \frac{8L\left(f(\mathbf{x}^0) - f^*\right)}{\varepsilon^2}$, то получим $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^N)\|_2^2\right] \leq \varepsilon^2$, а общее число обращений к оракулу равняется

$$Nr = \max\left\{\frac{8L\left(f(\mathbf{x}^0) - f^*\right)}{\varepsilon^2}, \frac{16L\left(f(\mathbf{x}^0) - f^*\right)\sigma^2}{\varepsilon^4}\right\}.$$

Стохастический случай: SGD, равномерно ограниченная дисперсия

Подставляя эту оценку в исходное неравенство и беря $\gamma_k = \frac{1}{2L}$, получаем

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^N)\|_2^2\right] \quad \leq \quad \frac{4L\left(f(\mathbf{x}^0) - f^*\right)}{N} + \frac{\varepsilon^2}{2},$$

где точка $\hat{\mathbf{x}}^N$ выбирается случайно равновероятно среди $\{\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{N-1}\}$. Таким образом, если мы возьмём $N = \frac{8L\left(f(\mathbf{x}^0) - f^*\right)}{\varepsilon^2}$, то получим $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^N)\|_2^2\right] \leq \varepsilon^2$, а общее число обращений к оракулу равняется

$$Nr = \max\left\{\frac{8L\left(f(\mathbf{x}^0) - f^*\right)}{\varepsilon^2}, \frac{16L\left(f(\mathbf{x}^0) - f^*\right)\sigma^2}{\varepsilon^4}\right\}.$$

Стохастический случай: SGD, равномерно ограниченная дисперсия

Подставляя эту оценку в исходное неравенство и беря $\gamma_k = \frac{1}{2L}$, получаем

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^N)\|_2^2\right] \quad \leq \quad \frac{4L\left(f(\mathbf{x}^0) - f^*\right)}{N} + \frac{\varepsilon^2}{2},$$

где точка $\hat{\mathbf{x}}^N$ выбирается случайно равновероятно среди $\{\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{N-1}\}$. Таким образом, если мы возьмём $N = \frac{8L\left(f(\mathbf{x}^0) - f^*\right)}{\varepsilon^2}$, то получим $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^N)\|_2^2\right] \leq \varepsilon^2$, а общее число обращений к оракулу равняется

$$Nr = \max\left\{\frac{8L\left(f(\mathbf{x}^0) - f^*\right)}{\varepsilon^2}, \frac{16L\left(f(\mathbf{x}^0) - f^*\right)\sigma^2}{\varepsilon^4}\right\}.$$

Минимизация суммы: SPIDER

Рассмотрим последний из трёх случаев и следующий метод, который называется SPIDER:

$$r_k = r = \max\left\{1, \frac{20\sigma^2}{\varepsilon^2}\right\}, \tag{18}$$

$$q = \min\{r, m\}, \tag{19}$$

$$\mathbf{g}^k = \begin{cases} \frac{1}{r}\sum\limits_{j=1}^{r} \nabla f_{\xi_{k,j}}(\mathbf{x}^k), \text{ если } r < m \text{ и } r \text{ делит } k, \\ \nabla f(\mathbf{x}^k), \text{ если } m \leq r \text{ и } m \text{ делит } k, \\ \nabla f_{\xi_k}(\mathbf{x}^k) - \nabla f_{\xi_k}(\mathbf{x}^{k-1}) + \mathbf{g}^{k-1}, \text{ иначе} \end{cases} \tag{20}$$

$$\gamma_k = \gamma = \frac{1}{10L\sqrt{q}}. \tag{21}$$

Раз в q итераций стох. градиент вычисляется достаточно точно, а на других итерациях происходит дополнительное вычисление градиента одного слагаемого, посчитанного в текущей и предыдущей точке.

Минимизация суммы: SPIDER

Рассмотрим последний из трёх случаев и следующий метод, который называется SPIDER:

$$r_k = r = \max\left\{1, \frac{20\sigma^2}{\varepsilon^2}\right\}, \tag{18}$$

$$q = \min\{r, m\}, \tag{19}$$

$$\mathbf{g}^k = \begin{cases} \frac{1}{r}\sum\limits_{j=1}^{r} \nabla f_{\xi_{k,j}}(\mathbf{x}^k), \text{ если } r < m \text{ и } r \text{ делит } k, \\ \nabla f(\mathbf{x}^k), \text{ если } m \leq r \text{ и } m \text{ делит } k, \\ \nabla f_{\xi_k}(\mathbf{x}^k) - \nabla f_{\xi_k}(\mathbf{x}^{k-1}) + \mathbf{g}^{k-1}, \text{ иначе} \end{cases} \tag{20}$$

$$\gamma_k = \gamma = \frac{1}{10L\sqrt{q}}. \tag{21}$$

Раз в q итераций стох. градиент вычисляется достаточно точно, а на других итерациях происходит дополнительное вычисление градиента одного слагаемого, посчитанного в текущей и предыдущей точке.

Минимизация суммы: SPIDER

Рассмотрим последний из трёх случаев и следующий метод, который называется SPIDER:

$$r_k = r = \max\left\{1, \frac{20\sigma^2}{\varepsilon^2}\right\}, \tag{18}$$

$$q = \min\{r, m\}, \tag{19}$$

$$\mathbf{g}^k = \begin{cases} \frac{1}{r}\sum\limits_{j=1}^{r} \nabla f_{\xi_{k,j}}(\mathbf{x}^k), \text{ если } r < m \text{ и } r \text{ делит } k, \\ \nabla f(\mathbf{x}^k), \text{ если } m \leq r \text{ и } m \text{ делит } k, \\ \nabla f_{\xi_k}(\mathbf{x}^k) - \nabla f_{\xi_k}(\mathbf{x}^{k-1}) + \mathbf{g}^{k-1}, \text{ иначе} \end{cases} \tag{20}$$

$$\gamma_k = \gamma = \frac{1}{10L\sqrt{q}}. \tag{21}$$

Раз в q итераций стох. градиент вычисляется достаточно точно, а на других итерациях происходит дополнительное вычисление градиента одного слагаемого, посчитанного в текущей и предыдущей точке.

Минимизация суммы: SPIDER

Рассмотрим последний из трёх случаев и следующий метод, который называется SPIDER:

$$r_k = r = \max\left\{1, \frac{20\sigma^2}{\varepsilon^2}\right\}, \tag{18}$$

$$q = \min\{r, m\}, \tag{19}$$

$$\mathbf{g}^k = \begin{cases} \frac{1}{r}\sum\limits_{j=1}^{r} \nabla f_{\xi_{k,j}}(\mathbf{x}^k), \text{ если } r < m \text{ и } r \text{ делит } k, \\ \nabla f(\mathbf{x}^k), \text{ если } m \leq r \text{ и } m \text{ делит } k, \\ \nabla f_{\xi_k}(\mathbf{x}^k) - \nabla f_{\xi_k}(\mathbf{x}^{k-1}) + \mathbf{g}^{k-1}, \text{ иначе} \end{cases} \tag{20}$$

$$\gamma_k = \gamma = \frac{1}{10L\sqrt{q}}. \tag{21}$$

Раз в q итераций стох. градиент вычисляется достаточно точно, а на других итерациях происходит дополнительное вычисление градиента одного слагаемого, посчитанного в текущей и предыдущей точке.

Минимизация суммы: SPIDER

Таким образом, получается, что за q итераций метода происходит вычисление $3q$ стох градиентов, то есть стоимость q итераций равняется $O(q)$. Идея анализа практически такая же, как и для SGD: нужно оценить $\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2\right]$, но не пытаться малость этого слагаемого, а оценить его через сумму градиентов в предыдущих точках.

Минимизация суммы: SPIDER

Таким образом, получается, что за q итераций метода происходит вычисление $3q$ стох градиентов, то есть стоимость q итераций равняется $O(q)$. Идея анализа практически такая же, как и для SGD: нужно оценить $\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2\right]$, но не пытаться малость этого слагаемого, а оценить его через сумму градиентов в предыдущих точках.

Минимизация суммы: SPIDER

Первый шаг состоит в том, чтобы рекурренто оценить $\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2\right]$ через сумму квадратов норм градиентов за текущую эпоху:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\right\|_2^2\right] \leq \sum_{l=1}^{p} 9L^2\gamma^2\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+l})\|_2^2\right] + \frac{11\varepsilon^2}{200}. \tag{22}$$

Теперь осталось лишь правильным образом соединить полученные неравенства.

Минимизация суммы: SPIDER

Первый шаг состоит в том, чтобы рекурренто оценить $\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2\right]$ через сумму квадратов норм градиентов за текущую эпоху:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\right\|_2^2\right] \leq \sum_{l=1}^{p} 9L^2\gamma^2\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+l})\|_2^2\right] + \frac{11\varepsilon^2}{200}. \tag{22}$$

Теперь осталось лишь правильным образом соединить полученные неравенства.

Минимизация суммы: SPIDER

Первый шаг состоит в том, чтобы рекурренто оценить $\mathbb{E}\left[\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\|_2^2\right]$ через сумму квадратов норм градиентов за текущую эпоху:

$$\mathbb{E}\left[\left\|\mathbf{g}^k - \nabla f(\mathbf{x}^k)\right\|_2^2\right] \quad \leq \quad \sum_{l=1}^{p} 9L^2\gamma^2\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+l})\|_2^2\right] + \frac{11\varepsilon^2}{200}. \tag{22}$$

Теперь осталось лишь правильным образом соединить полученные неравенства.

Минимизация суммы: SPIDER

Начнём с небольшого видоизменения неравенства (15):

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \frac{\gamma_k}{2}\|\mathbf{g}^k\|_2^2 + \gamma_k\|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 \\ &\leq f(\mathbf{x}^k) - \frac{\gamma}{4}\|\nabla f(\mathbf{x}^k)\|_2^2 + \frac{3\gamma}{2}\|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2, \end{aligned}$$

где мы воспользовались неравенством $\|\mathbf{a}+\mathbf{b}\|_2^2 \geq \frac{1}{2}\|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2$, которое выполняется для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ (в частности, мы использовали $\mathbf{a} = \nabla f(\mathbf{x}^k)$ и $\mathbf{b} = \mathbf{g}^k - \nabla f(\mathbf{x}^k)$).

Минимизация суммы: SPIDER

Начнём с небольшого видоизменения неравенства (15):

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \frac{\gamma_k}{2}\|\mathbf{g}^k\|_2^2 + \gamma_k\|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 \\ &\leq f(\mathbf{x}^k) - \frac{\gamma}{4}\|\nabla f(\mathbf{x}^k)\|_2^2 + \frac{3\gamma}{2}\|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2, \end{aligned}$$

где мы воспользовались неравенством $\|\mathbf{a}+\mathbf{b}\|_2^2 \geq \frac{1}{2}\|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2$, которое выполняется для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ (в частности, мы использовали $\mathbf{a} = \nabla f(\mathbf{x}^k)$ и $\mathbf{b} = \mathbf{g}^k - \nabla f(\mathbf{x}^k)$).

Минимизация суммы: SPIDER

Начнём с небольшого видоизменения неравенства (15):

$$\begin{aligned} f(\mathbf{x}^{k+1}) &\leq f(\mathbf{x}^k) - \frac{\gamma_k}{2}\|\mathbf{g}^k\|_2^2 + \gamma_k\|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2 \\ &\leq f(\mathbf{x}^k) - \frac{\gamma}{4}\|\nabla f(\mathbf{x}^k)\|_2^2 + \frac{3\gamma}{2}\|\nabla f(\mathbf{x}^k) - \mathbf{g}^k\|_2^2, \end{aligned}$$

где мы воспользовались неравенством $\|\mathbf{a}+\mathbf{b}\|_2^2 \geq \frac{1}{2}\|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2$, которое выполняется для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ (в частности, мы использовали $\mathbf{a} = \nabla f(\mathbf{x}^k)$ и $\mathbf{b} = \mathbf{g}^k - \nabla f(\mathbf{x}^k)$).

Минимизация суммы: SPIDER

Теперь возьмём от полученного неравенства полное мат. ожидание (учитываем, что $k = aq + p$):

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}^{aq+p+1})] &\leq \mathbb{E}[f(\mathbf{x}^{aq+p})] - \frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+p})\|_2^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\|\mathbf{g}^{aq+p} - \nabla f(\mathbf{x}^{aq+p})\|_2^2\right] \\ &\leq \mathbb{E}[f(\mathbf{x}^{aq+p})] - \frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+p})\|_2^2\right] + \frac{3\gamma}{2}\sum_{l=1}^{p} 9L^2\gamma^2\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+l})\|_2^2\right] + \frac{33\gamma\varepsilon^2}{400}.\end{aligned}$$

Сразу отметим, что это неравенство выполняется для всех целых неотрицательных a и $p \in \{0, 1, \ldots, q-1\}$.

Минимизация суммы: SPIDER

Теперь возьмём от полученного неравенства полное мат. ожидание (учитываем, что $k = aq + p$):

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}^{aq+p+1})] &\leq \mathbb{E}[f(\mathbf{x}^{aq+p})] - \frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+p})\|_2^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\|\mathbf{g}^{aq+p} - \nabla f(\mathbf{x}^{aq+p})\|_2^2\right] \\ &\leq \mathbb{E}[f(\mathbf{x}^{aq+p})] - \frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+p})\|_2^2\right] + \frac{3\gamma}{2}\sum_{l=1}^{p} 9L^2\gamma^2\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+l})\|_2^2\right] + \frac{33\gamma\varepsilon^2}{400}.\end{aligned}$$

Сразу отметим, что это неравенство выполняется для всех целых неотрицательных a и $p \in \{0, 1, \ldots, q-1\}$.

Минимизация суммы: SPIDER

Теперь возьмём от полученного неравенства полное мат. ожидание (учитываем, что $k = aq + p$):

$$\begin{aligned}\mathbb{E}[f(\mathbf{x}^{aq+p+1})] &\leq \mathbb{E}[f(\mathbf{x}^{aq+p})] - \frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+p})\|_2^2\right] + \frac{3\gamma}{2}\mathbb{E}\left[\|\mathbf{g}^{aq+p} - \nabla f(\mathbf{x}^{aq+p})\|_2^2\right] \\ &\leq \mathbb{E}[f(\mathbf{x}^{aq+p})] - \frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+p})\|_2^2\right] + \frac{3\gamma}{2}\sum_{l=1}^{p} 9L^2\gamma^2\mathbb{E}\left[\|\nabla f(\mathbf{x}^{aq+l})\|_2^2\right] + \frac{33\gamma\varepsilon^2}{400}.\end{aligned}$$

Сразу отметим, что это неравенство выполняется для всех целых неотрицательных a и $p \in \{0, 1, \ldots, q-1\}$.

Минимизация суммы: SPIDER

Суммируя эти неравенства и учитывая соотношения на γ_k, можно получить следующий результат:

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_2^2\right] \overset{(21)}{\leq} \frac{2000L\sqrt{q}\left(f(\mathbf{x}^0)-f^*\right)}{23(N+1)} + \frac{33\varepsilon^2}{46}.$$

Как и для SGD возьмём точку $\hat{\mathbf{x}}^{\hat{N}}$ случайно и равновероятно среди $\{\mathbf{x}^0,\ldots,\mathbf{x}^{\hat{N}}\}$ и получим:

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^{\hat{N}})\|_2^2\right] \leq \frac{2000L\sqrt{q}\left(f(\mathbf{x}^0)-f^*\right)}{23(\hat{N}+1)} + \frac{33\varepsilon^2}{46}.$$

Таким образом, после

$$\hat{N} \overset{(18),(19)}{=} \frac{4000L(f(\mathbf{x}^0)-f^*)}{13\varepsilon^2}\min\left\{\sqrt{m}, \max\left\{1, \frac{\sqrt{20}\sigma}{\varepsilon}\right\}\right\}$$

итераций метод гарантирует $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^{\hat{N}})\|_2^2\right] \leq \varepsilon^2$.

Минимизация суммы: SPIDER

Суммируя эти неравенства и учитывая соотношения на γ_k, можно получить следующий результат:

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_2^2\right] \overset{(21)}{\leq} \frac{2000L\sqrt{q}\left(f(\mathbf{x}^0)-f^*\right)}{23(N+1)} + \frac{33\varepsilon^2}{46}.$$

Как и для SGD возьмём точку $\hat{\mathbf{x}}^{\hat{N}}$ случайно и равновероятно среди $\{\mathbf{x}^0,\ldots,\mathbf{x}^{\hat{N}}\}$ и получим:

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^{\hat{N}})\|_2^2\right] \leq \frac{2000L\sqrt{q}\left(f(\mathbf{x}^0)-f^*\right)}{23(\hat{N}+1)} + \frac{33\varepsilon^2}{46}.$$

Таким образом, после

$$\hat{N} \overset{(18),(19)}{=} \frac{4000L(f(\mathbf{x}^0)-f^*)}{13\varepsilon^2}\min\left\{\sqrt{m}, \max\left\{1, \frac{\sqrt{20}\sigma}{\varepsilon}\right\}\right\}$$

итераций метод гарантирует $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^{\hat{N}})\|_2^2\right] \leq \varepsilon^2$.

Минимизация суммы: SPIDER

Суммируя эти неравенства и учитывая соотношения на γ_k, можно получить следующий результат:

$$\frac{1}{N+1}\sum_{k=0}^{N}\mathbb{E}\left[\|\nabla f(\mathbf{x}^k)\|_2^2\right] \overset{(21)}{\leq} \frac{2000L\sqrt{q}\left(f(\mathbf{x}^0)-f^*\right)}{23(N+1)} + \frac{33\varepsilon^2}{46}.$$

Как и для SGD возьмём точку $\hat{\mathbf{x}}^{\hat{N}}$ случайно и равновероятно среди $\{\mathbf{x}^0,\ldots,\mathbf{x}^{\hat{N}}\}$ и получим:

$$\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^{\hat{N}})\|_2^2\right] \leq \frac{2000L\sqrt{q}\left(f(\mathbf{x}^0)-f^*\right)}{23(\hat{N}+1)} + \frac{33\varepsilon^2}{46}.$$

Таким образом, после

$$\hat{N} \overset{(18),(19)}{=} \frac{4000L(f(\mathbf{x}^0)-f^*)}{13\varepsilon^2}\min\left\{\sqrt{m}, \max\left\{1, \frac{\sqrt{20}\sigma}{\varepsilon}\right\}\right\}$$

итераций метод гарантирует $\mathbb{E}\left[\|\nabla f(\hat{\mathbf{x}}^{\hat{N}})\|_2^2\right] \leq \varepsilon^2$.

Минимизация суммы: SPIDER

Более того, это требует

$$O\left(\frac{L(f(\mathbf{x}^0)-f^*)}{\varepsilon^2}\min\left\{\sqrt{m},\max\left\{1,\frac{\sigma}{\varepsilon}\right\}\right\}+\min\left\{m,\max\left\{1,\frac{\sigma^2}{\varepsilon^2}\right\}\right\}\right)$$

вычислений $\nabla f_\xi(\mathbf{x})$.

Сравнительная таблица с результатами

Метод	Сложность
GD	$O\left(\frac{L\left(f(\mathbf{x^0})-f^*\right)}{\varepsilon^2}\right)$
SGD	$O\left(L\left(f(\mathbf{x}^0)-f^*\right)\max\left\{\frac{1}{\varepsilon^2},\frac{\sigma^2}{\varepsilon^4}\right\}\right)$
SPIDER	$O\left(L(f(\mathbf{x}^0)-f^*)\min\left\{\frac{\sqrt{m}}{\varepsilon^2},\max\left\{\frac{1}{\varepsilon^2},\frac{\sigma}{\varepsilon^3}\right\}\right\}\right)$

Оказывается, что кажый из методов оптимален для своего случая.

Сравнительная таблица с результатами

Метод	Сложность
GD	$O\left(\frac{L\left(f(\mathbf{x^0})-f^*\right)}{\varepsilon^2}\right)$
SGD	$O\left(L\left(f(\mathbf{x}^0)-f^*\right)\max\left\{\frac{1}{\varepsilon^2},\frac{\sigma^2}{\varepsilon^4}\right\}\right)$
SPIDER	$O\left(L(f(\mathbf{x}^0)-f^*)\min\left\{\frac{\sqrt{m}}{\varepsilon^2},\max\left\{\frac{1}{\varepsilon^2},\frac{\sigma}{\varepsilon^3}\right\}\right\}\right)$

Оказывается, что кажый из методов оптимален для своего случая.

Ссылки на работы с нижними оценками сложности

- Carmon, Y., Duchi, J.C., Hinder, O. and Sidford, A., 2019. **Lower bounds for finding stationary points I.** *Mathematical Programming*, pp.1-50.
- Arjevani, Y., Carmon, Y., Duchi, J.C., Foster, D.J., Srebro, N. and Woodworth, B., 2019. **Lower bounds for non-convex stochastic optimization.** *arXiv preprint* arXiv:1912.02365.
- Fang, C., Li, C.J., Lin, Z. and Zhang, T. **Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator.** *In Advances in Neural Information Processing Systems*, pp. 689-699, *(NeurIPS 2018)*.

SGD: замечание

- На практике при обученнии нейроннных сетей не обязательно брать такие большие батчи: очень часто удачно работает политика периодически уменьшающегося размера шага даже в невыпуклом случае. Такой подход хорошо работает на практике, но до недавних пор не было формального обоснования, почему так происходит в невыпуклом случае.
- 16 апреля 2020 года на arXiv вышел препринт, в котором эта проблема была решена и показано, что на самом деле SGD сходится «лучше», чем выписанные оценки: как и в сильно выпуклом случае SGD сходится с линейной скоростью к окрестности стационарной точки, причём размер окрестности пропорционален γ, а вот скорость линейной сходимости зависит от γ чуть хитрее, чем в сильно выпуклом случае.

Препринт: Bin Shi, Weijie J. Su, Michael I. Jordan. On Learning Rates and Schrodinger Operators, arXive preprint: arXiv:2004.06977.

SGD: замечание

- На практике при обученнии нейроннных сетей не обязательно брать такие большие батчи: очень часто удачно работает политика периодически уменьшающегося размера шага даже в невыпуклом случае. Такой подход хорошо работает на практике, но до недавних пор не было формального обоснования, почему так происходит в невыпуклом случае.
- 16 апреля 2020 года на arXiv вышел препринт, в котором эта проблема была решена и показано, что на самом деле SGD сходится «лучше», чем выписанные оценки: как и в сильно выпуклом случае SGD сходится с линейной скоростью к окрестности стационарной точки, причём размер окрестности пропорционален γ, а вот скорость линейной сходимости зависит от γ чуть хитрее, чем в сильно выпуклом случае.

Препринт: Bin Shi, Weijie J. Su, Michael I. Jordan. On Learning Rates and Schrodinger Operators, arXive preprint: arXiv:2004.06977.

SGD: замечание

- На практике при обученнии нейроннных сетей не обязательно брать такие большие батчи: очень часто удачно работает политика периодически уменьшающегося размера шага даже в невыпуклом случае. Такой подход хорошо работает на практике, но до недавних пор не было формального обоснования, почему так происходит в невыпуклом случае.
- 16 апреля 2020 года на arXiv вышел препринт, в котором эта проблема была решена и показано, что на самом деле SGD сходится «лучше», чем выписанные оценки: как и в сильно выпуклом случае SGD сходится с линейной скоростью к окрестности стационарной точки, причём размер окрестности пропорционален γ, а вот скорость линейной сходимости зависит от γ чуть хитрее, чем в сильно выпуклом случае.

Препринт: Bin Shi, Weijie J. Su, Michael I. Jordan. On Learning Rates and Schrodinger Operators, arXive preprint: arXiv:2004.06977.

SGD: замечание

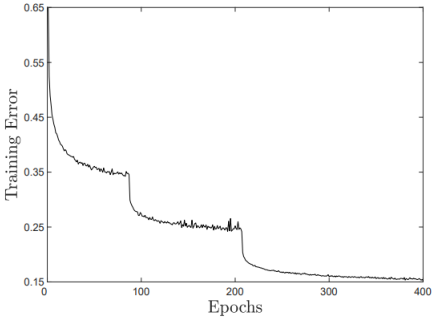


Figure 1: Training error using SGD with mini-batch size 32 to train an 8-layer convolutional neural network on CIFAR-10 [Kri09]. The first 90 epochs use a learning rate of $s = 0.006$, the next 120 epochs use $s = 0.003$, and the final 190 epochs use $s = 0.0005$. Note that the training error decreases as the learning rate s decreases and a smaller s leads to a larger number of epochs for SGD to reach a plateau. See [HZRS16] for further investigation of this phenomenon.

Результаты для задач с регуляризацией

- Ghadimi, S., Lan, G. and Zhang, H., 2016. **Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization.** *Mathematical Programming*, 155(1-2), pp.267-305.

  Для `prox-SGD` доказана та же скорость убывания $\sim \varepsilon^{-4}$ (как в случае без регуляризации) для нормы градиентного отображения, но с батчами размера $\sim \varepsilon^{-2}$.

- Davis, D. and Drusvyatskiy, D., 2019. **Stochastic model-based minimization of weakly convex functions.** *SIAM Journal on Optimization*, 29(1), pp.207-239.

  Решена проблема с большими батчами для `prox-SGD` в невыпуклом случае.

- Wang, Z., Ji, K., Zhou, Y., Liang, Y. and Tarokh, V. **SpiderBoost and momentum: Faster variance reduction algorithms.** *In Advances in Neural Information Processing Systems*, pp. 2406-2416, *(NeurIPS 2019)*.

  Предложена вариация SPIDER — `prox-SpiderBoost`.

Результаты для задач с регуляризацией

- Ghadimi, S., Lan, G. and Zhang, H., 2016. **Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization.** *Mathematical Programming*, 155(1-2), pp.267-305.

  Для prox-SGD доказана та же скорость убывания $\sim \varepsilon^{-4}$ (как в случае без регуляризации) для нормы градиентного отображения, но с батчами размера $\sim \varepsilon^{-2}$.
- Davis, D. and Drusvyatskiy, D., 2019. **Stochastic model-based minimization of weakly convex functions.** *SIAM Journal on Optimization*, 29(1), pp.207-239.

  Решена проблема с большими батчами для prox-SGD в невыпуклом случае.
- Wang, Z., Ji, K., Zhou, Y., Liang, Y. and Tarokh, V. **SpiderBoost and momentum: Faster variance reduction algorithms.** *In Advances in Neural Information Processing Systems*, pp. 2406-2416, *(NeurIPS 2019)*.

  Предложена вариация SPIDER — prox-SpiderBoost.

Результаты для задач с регуляризацией

- Ghadimi, S., Lan, G. and Zhang, H., 2016. **Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization.** *Mathematical Programming*, 155(1-2), pp.267-305.

  Для prox-SGD доказана та же скорость убывания $\sim \varepsilon^{-4}$ (как в случае без регуляризации) для нормы градиентного отображения, но с батчами размера $\sim \varepsilon^{-2}$.
- Davis, D. and Drusvyatskiy, D., 2019. **Stochastic model-based minimization of weakly convex functions.** *SIAM Journal on Optimization*, 29(1), pp.207-239.

  Решена проблема с большими батчами для prox-SGD в невыпуклом случае.
- Wang, Z., Ji, K., Zhou, Y., Liang, Y. and Tarokh, V. **SpiderBoost and momentum: Faster variance reduction algorithms.** *In Advances in Neural Information Processing Systems*, pp. 2406-2416, *(NeurIPS 2019)*.

  Предложена вариация SPIDER — prox-SpiderBoost.

Вариация SPIDER: SpiderBoost

Алгоритм 6 SpiderBoost

Вход: размер шага $\gamma > 0$, размер эпохи T, стартовая точка $\mathbf{x}^0 \in \mathbb{R}^n$, размер батча $r \geq 1$, число итераций K

for $k = 0, 1, 2, \ldots$ **do**
  if $k \mod T = 0$ **then**
    Посчитать $\mathbf{g}^k = \nabla f(\mathbf{x}^k)$
  else
    Случайно равновероятно выбрать набор I_k из $\{1, \ldots, m\}$ (допускаются повторения) размера $|I_k| = r$
    Вычислить $\mathbf{g}^k = \frac{1}{r} \sum\limits_{i \in I_k} \left(\nabla f_i(\mathbf{x}^k) - \nabla f_i(\mathbf{x}^{k-1}) \right) + \mathbf{g}^{k-1}$
  end if
  $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma \mathbf{g}^k$
end for
Выбрать ξ случайно равновероятно из $\{0, \ldots, K-1\}$
Выход: $\mathbf{x}^\xi$

Вариация SPIDER: SpiderBoost

- SpiderBoost — тот же SPIDER, но с пробатченными градиентами на всех итерациях и с более практичным размером шага γ.
- Тоже работает по нижним оценкам: если $r = L = \sqrt{m}$ и $\gamma = \frac{1}{2L}$, то метод сходится по тем же оценкам, что и SPIDER.

Вариация SPIDER: SpiderBoost

- SpiderBoost — тот же SPIDER, но с пробатченными градиентами на всех итерациях и с более практичным размером шага γ.
- Тоже работает по нижним оценкам: если $r = L = \sqrt{m}$ и $\gamma = \frac{1}{2L}$, то метод сходится по тем же оценкам, что и SPIDER.

Классические результаты для SVRG и SAGA в невыпуклом случае

Reddi, S.J., Hefny, A., Sra, S., Poczos, B. and Smola, A. **Stochastic variance reduction for nonconvex optimization.** *In International conference on machine learning*, pp. 314-323, *(ICML 2016)*.

Reddi, S.J., Sra, S., Poczos, B. and Smola, A.J. **Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization.** *In Advances in Neural Information Processing Systems*, pp. 1145-1153, *(NIPS 2016)*.

В этих статьях для SVRG и SAGA (и их проксимальных вариантов) доказываются следующая оценка на скорость сходимости:

$$O\left(\frac{L(f(\mathbf{x}^0) - f^*)m^{2/3}}{\varepsilon^2}\right).$$

SpiderBoost: пример работы метода

Далее будут приведены результаты работы SpiderBoost из оригинальной статьи: Zhe Wang et al. “SpiderBoost and Momentum: Faster Variance Reduction Algorithms”. In: Advances in Neural Information Processing Systems. 2019, pp. 2403–2413. Авторы рассмотрели 2 примера.

- Логистическая регрессия с невыпуклым регуляризатором:

$$\min_{x \in \mathbb{R}^n} \underbrace{\frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i \cdot (\mathbf{A}x)_i\right)\right)}_{\text{выпуклая функция}} + \underbrace{\alpha \sum_{i=1}^{m} \frac{x_i^2}{1 + x_i^2}}_{\text{невыпуклая функция}}. \tag{23}$$

- Робастная линейная регрессия:

$$\min_{x \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \log\left(\frac{1}{2}\left(y_i - (\mathbf{A}x)_i\right)^2 + 1\right). \tag{24}$$

SpiderBoost: пример работы метода

Далее будут приведены результаты работы SpiderBoost из оригинальной статьи: Zhe Wang et al. “SpiderBoost and Momentum: Faster Variance Reduction Algorithms”. In: Advances in Neural Information Processing Systems. 2019, pp. 2403–2413. Авторы рассмотрели 2 примера.

- Логистическая регрессия с невыпуклым регуляризатором:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \underbrace{\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-y_i\cdot(\mathbf{A}\mathbf{x})_i\right)\right)}_{\text{выпуклая функция}} + \underbrace{\alpha\sum_{i=1}^{m}\frac{x_i^2}{1+x_i^2}}_{\text{невыпуклая функция}}. \tag{23}$$

- Робастная линейная регрессия:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{1}{2}\left(y_i-(\mathbf{A}\mathbf{x})_i\right)^2+1\right). \tag{24}$$

SpiderBoost: пример работы метода

Далее будут приведены результаты работы SpiderBoost из оригинальной статьи: Zhe Wang et al. “SpiderBoost and Momentum: Faster Variance Reduction Algorithms”. In: Advances in Neural Information Processing Systems. 2019, pp. 2403–2413. Авторы рассмотрели 2 примера.

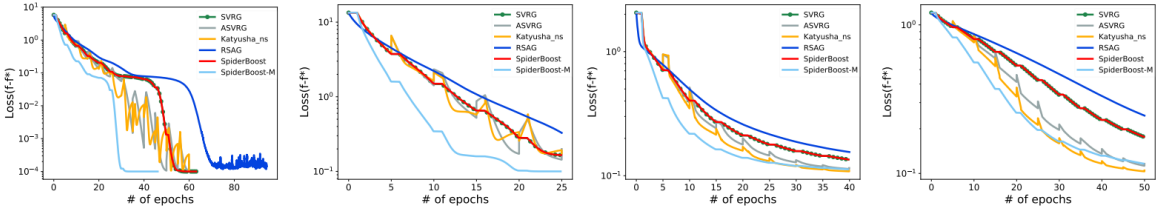
- Логистическая регрессия с невыпуклым регуляризатором:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \underbrace{\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\exp\left(-y_i\cdot(\mathbf{A}\mathbf{x})_i\right)\right)}_{\text{выпуклая функция}} + \underbrace{\alpha\sum_{i=1}^{m}\frac{x_i^2}{1+x_i^2}}_{\text{невыпуклая функция}}. \tag{23}$$

- Робастная линейная регрессия:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{1}{2}\left(y_i-(\mathbf{A}\mathbf{x})_i\right)^2+1\right). \tag{24}$$

SpiderBoost: пример работы метода



(a) Dataset: a9a (b) Dataset: w8a (c) Dataset: a9a (d) Dataset: w8a

Figure 2: (a) and (b): Logistic regression with nonconvex regularizer, (c) and (d): Robust linear regression..

Figure: `https://papers.nips.cc/paper/8511-spiderboost-and-momentum-faster-variance-reduction-algorithms.pdf`

SpiderBoost: пример работы метода

Даже на этих примерах видно, что попытка ускорить метод добавлением моментного члена улучшает сходимость метода даже в невыпуклом случае.

Метод тяжёлого шарика (Momentum-GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Изначально разрабатывался как раз для того, чтобы «проскакивать» небольшие локальные минимумы, сёдла, плато.
- Существующая теория для метода тяжёлого шарика для общих выпуклых/невыпуклых гладких функций даёт оценки скорости сходимости не лучше, чем у градиентного метода.
- Подход популярен для решения невыпуклых задач.
- https://distill.pub/2017/momentum/

Метод тяжёлого шарика (Momentum-GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Изначально разрабатывался как раз для того, чтобы «проскакивать» небольшие локальные минимумы, сёдла, плато.
- Существующая теория для метода тяжёлого шарика для общих выпуклых/невыпуклых гладких функций даёт оценки скорости сходимости не лучше, чем у градиентного метода.
- Подход популярен для решения невыпуклых задач.
- https://distill.pub/2017/momentum/

Метод тяжёлого шарика (Momentum-GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Изначально разрабатывался как раз для того, чтобы «проскакивать» небольшие локальные минимумы, сёдла, плато.
- Существующая теория для метода тяжёлого шарика для общих выпуклых/невыпуклых гладких функций даёт оценки скорости сходимости не лучше, чем у градиентного метода.
- Подход популярен для решения невыпуклых задач.
- https://distill.pub/2017/momentum/

Метод тяжёлого шарика (Momentum-GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Изначально разрабатывался как раз для того, чтобы «проскакивать» небольшие локальные минимумы, сёдла, плато.
- Существующая теория для метода тяжёлого шарика для общих выпуклых/невыпуклых гладких функций даёт оценки скорости сходимости не лучше, чем у градиентного метода.
- Подход популярен для решения невыпуклых задач.
- https://distill.pub/2017/momentum/

Метод тяжёлого шарика (Momentum-GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Изначально разрабатывался как раз для того, чтобы «проскакивать» небольшие локальные минимумы, сёдла, плато.
- Существующая теория для метода тяжёлого шарика для общих выпуклых/невыпуклых гладких функций даёт оценки скорости сходимости не лучше, чем у градиентного метода.
- Подход популярен для решения невыпуклых задач.
- https://distill.pub/2017/momentum/

Метод тяжёлого шарика (Momentum-GD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Изначально разрабатывался как раз для того, чтобы «проскакивать» небольшие локальные минимумы, сёдла, плато.
- Существующая теория для метода тяжёлого шарика для общих выпуклых/невыпуклых гладких функций даёт оценки скорости сходимости не лучше, чем у градиентного метода.
- Подход популярен для решения невыпуклых задач.
- `https://distill.pub/2017/momentum/`

Метод тяжёлого шарика (Momentum-GD)

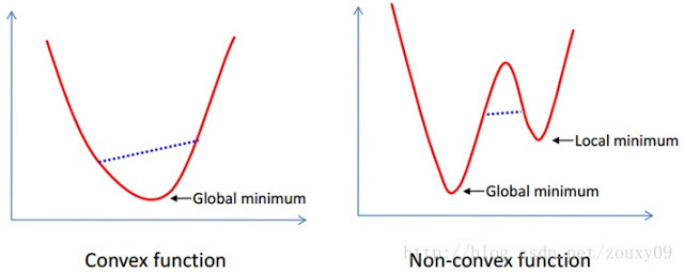


Figure: http://blog.ittraining.com.tw/2017/08/convex-function.html

Метод Нестерова (Nesterov-Momentum)

Метод тяжёлого шарика:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k) + \beta_k(\mathbf{x}^k - \mathbf{x}^{k-1}).$$

Метод Нестерова:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f(\mathbf{x}^k + \beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})) + \beta_k(\mathbf{x}^k - \mathbf{x}^{k-1}).$$

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Стохастический метод тяжёлого шарика (Momentum-SGD)

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \mathbf{g}^k + \underbrace{\beta_k(\mathbf{x}^k - \mathbf{x}^{k-1})}_{\text{momentum}}$$

- Подход очень популярен для решения невыпуклых задач.
- Теоретические гарантии – не лучше, чем у обычного SGD
- По-прежнему может «проскакивать» небольшие локальные минимумы, сёдла, плато, однако теперь важно брать размер батча не слишком маленьким, т.к. на следующих итерациях за счёт инерции это может испортить сходимость.
- Позволяет стабилизировать шаги в случае больших выбросов
- http://fa.bianp.net/teaching/2018/COMP-652/stochastic_gradient.html
- Вообще говоря, параметры γ_k и β_k нужно подбирать, но часто используется $\beta_k = 0.9, 0.95, 0.99$.

Adagrad

k-я итерация SGD:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \gamma \mathbf{g}_i^k, \quad i = \overline{1, n}.$$

k-я итерация Adagrad:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \frac{\gamma}{\sqrt{G_i^k + \delta}} \mathbf{g}_i^k, \quad i = \overline{1, n}, \text{ где}$$

$$G_i^k = \sum_{l=0}^{k} \left(\mathbf{g}_i^l\right)^2, \quad \delta = 10^{-8}.$$

Небольшое уточнение: везде далее будем считать, что то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.

- Основной плюс Adagrad — не нужно подбирать γ, т.к. метод часто хорошо работает с $\gamma = 0.01$.
- Главный недостаток — размер шага всегда уменьшается.

Adagrad

k-я итерация SGD:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \gamma \mathbf{g}_i^k, \quad i = \overline{1, n}.$$

k-я итерация Adagrad:

$$\begin{aligned} \mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma}{\sqrt{G_i^k + \delta}} \mathbf{g}_i^k, \quad i = \overline{1, n}, \text{ где} \\ G_i^k &= \sum_{l=0}^{k} \left(\mathbf{g}_i^l\right)^2, \quad \delta = 10^{-8}. \end{aligned}$$

Небольшое уточнение: везде далее будем считать, что то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.

- Основной плюс Adagrad — не нужно подбирать γ, т.к. метод часто хорошо работает с $\gamma = 0.01$.
- Главный недостаток — размер шага всегда уменьшается.

Adagrad

k-я итерация SGD:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \gamma \mathbf{g}_i^k, \quad i = \overline{1, n}.$$

k-я итерация Adagrad:

$$\begin{aligned} \mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma}{\sqrt{G_i^k + \delta}} \mathbf{g}_i^k, \quad i = \overline{1, n}, \text{ где} \\ G_i^k &= \sum_{l=0}^{k} \left(\mathbf{g}_i^l\right)^2, \quad \delta = 10^{-8}. \end{aligned}$$

Небольшое уточнение: везде далее будем считать, что то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.

- Основной плюс Adagrad — не нужно подбирать γ, т.к. метод часто хорошо работает с $\gamma = 0.01$.
- Главный недостаток — размер шага всегда уменьшается.

Adagrad

k-я итерация SGD:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \gamma \mathbf{g}_i^k, \quad i = \overline{1, n}.$$

k-я итерация Adagrad:

$$\begin{aligned} \mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma}{\sqrt{G_i^k + \delta}} \mathbf{g}_i^k, \quad i = \overline{1, n}, \text{ где} \\ G_i^k &= \sum_{l=0}^{k} \left(\mathbf{g}_i^l\right)^2, \quad \delta = 10^{-8}. \end{aligned}$$

Небольшое уточнение: везде далее будем считать, что то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.

- Основной плюс Adagrad — не нужно подбирать γ, т.к. метод часто хорошо работает с $\gamma = 0.01$.
- Главный недостаток — размер шага всегда уменьшается.

RMSprop

k-я итерация Adagrad:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \frac{\gamma}{\sqrt{G_i^k + \delta}}\mathbf{g}_i^k, \quad i = \overline{1,n}, \text{ где}$$

$$G_i^k = \sum_{l=0}^{k} \left(\mathbf{g}_i^l\right)^2, \quad \delta = 10^{-8}.$$

k-я итерация RMSprop:

$$\mathbf{x}_i^{k+1} = \mathbf{x}_i^k - \frac{\gamma}{\sqrt{\mathbf{v}_i^k + \delta}}\mathbf{g}_i^k, \quad i = \overline{1,n}, \text{ где}$$

$$\mathbf{v}_i^k = \beta\mathbf{v}_i^{k-1} + (1-\beta)(\mathbf{g}_i^k)^2, \quad \delta = 10^{-8}.$$

RMSprop

У RMSprop исчезает главный недостаток Adagrad — информация о прошлых градиентах со временем берётся всё с меньшим и меньшим весом. Часто в качестве β выбирают $\beta = 0.9$, а в качестве γ берут 0.001.

RMSprop

У RMSprop исчезает главный недостаток Adagrad — информация о прошлых градиентах со временем берётся всё с меньшим и меньшим весом. Часто в качестве β выбирают $\beta = 0.9$, а в качестве γ берут 0.001.

Adam

Всеми известный Adam получается из RMSprop добавлением моментума. Итак, k-я итерация Adam: $m_0 = v_0 = 0$ и

$$\begin{aligned}
\mathbf{m}_i^k &= \beta_1 \mathbf{m}_i^{k-1} + (1-\beta_1)\mathbf{g}_i^k, \quad \hat{\mathbf{m}}_i^k = \frac{\mathbf{m}_i^k}{1-\beta_1^k} \\
\mathbf{v}_i^k &= \beta_2 \mathbf{v}_i^{k-1} + (1-\beta_2)(\mathbf{g}_i^k)^2, \quad \hat{\mathbf{v}}_i^k = \frac{\mathbf{v}_i^k}{1-\beta_2^k} \\
\mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma}{\sqrt{\hat{\mathbf{v}}_i^k} + \delta}\hat{\mathbf{m}}_i^k, \quad i = \overline{1,n}, \quad \delta = 10^{-8}.
\end{aligned}$$

- Напоминание: то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.
- На практике часто выбирают $\beta = 0.9$, $\beta_2 = 0.99$ и $\gamma = 0.001$.

Adam

Всеми известный Adam получается из RMSprop добавлением моментума. Итак, k-я итерация Adam: $m_0 = v_0 = 0$ и

$$\begin{aligned}
\mathbf{m}_i^k &= \beta_1 \mathbf{m}_i^{k-1} + (1-\beta_1)\mathbf{g}_i^k, \quad \hat{\mathbf{m}}_i^k = \frac{\mathbf{m}_i^k}{1-\beta_1^k} \\
\mathbf{v}_i^k &= \beta_2 \mathbf{v}_i^{k-1} + (1-\beta_2)(\mathbf{g}_i^k)^2, \quad \hat{\mathbf{v}}_i^k = \frac{\mathbf{v}_i^k}{1-\beta_2^k} \\
\mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma}{\sqrt{\hat{\mathbf{v}}_i^k} + \delta}\hat{\mathbf{m}}_i^k, \quad i = \overline{1,n}, \quad \delta = 10^{-8}.
\end{aligned}$$

- Напоминание: то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.
- На практике часто выбирают $\beta = 0.9$, $\beta_2 = 0.99$ и $\gamma = 0.001$.

Adam

Всеми известный Adam получается из RMSprop добавлением моментума. Итак, k-я итерация Adam: $m_0 = v_0 = 0$ и

$$\begin{aligned}
\mathbf{m}_i^k &= \beta_1 \mathbf{m}_i^{k-1} + (1-\beta_1)\mathbf{g}_i^k, \quad \hat{\mathbf{m}}_i^k = \frac{\mathbf{m}_i^k}{1-\beta_1^k} \\
\mathbf{v}_i^k &= \beta_2 \mathbf{v}_i^{k-1} + (1-\beta_2)(\mathbf{g}_i^k)^2, \quad \hat{\mathbf{v}}_i^k = \frac{\mathbf{v}_i^k}{1-\beta_2^k} \\
\mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma}{\sqrt{\hat{\mathbf{v}}_i^k} + \delta}\hat{\mathbf{m}}_i^k, \quad i = \overline{1,n}, \quad \delta = 10^{-8}.
\end{aligned}$$

- Напоминание: то, что выделено красным — это показатели степени, а всё остальное — это верхние индексы.
- На практике часто выбирают $\beta = 0.9$, $\beta_2 = 0.99$ и $\gamma = 0.001$.

Adam, другой вариант записи

$$
\begin{aligned}
\mathbf{m}_i^k &= \beta_1 \mathbf{m}_i^{k-1} + (1-\beta_1)\mathbf{g}_i^k, \\
\mathbf{v}_i^k &= \beta_2 \mathbf{v}_i^{k-1} + (1-\beta_2)(\mathbf{g}_i^k)^2, \\
\mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma_k}{\sqrt{\mathbf{v}_i^k} + \delta}\mathbf{m}_i^k, \quad i = \overline{1,n}, \quad \delta = 10^{-8}, \\
\gamma_k &= \gamma(1-\beta_1)\sqrt{1-\beta_2^k}.
\end{aligned}
$$

Плохие новости: Adam может расходиться

Причём даже для выпуклых задач! Это было показано в статье:

- Sashank J. Reddi, Satyen Kale и Sanjiv Kumar, ON THE CONVERGENCE OF ADAM AND BEYOND, ICLR 2018.

Хорошие новости: авторы предложили, как исправить проблему.

Плохие новости: Adam может расходиться

Причём даже для выпуклых задач! Это было показано в статье:

- Sashank J. Reddi, Satyen Kale и Sanjiv Kumar, ON THE CONVERGENCE OF ADAM AND BEYOND, ICLR 2018.

Хорошие новости: авторы предложили, как исправить проблему.

AMSGrad

$$
\begin{aligned}
\mathbf{m}_i^k &= \beta_1 \mathbf{m}_i^{k-1} + (1-\beta_1)\mathbf{g}^k, \\
\mathbf{v}_i^k &= \beta_2 \mathbf{v}_i^{k-1} + (1-\beta_2)(\mathbf{g}_i^k)^2, \quad \hat{\mathbf{v}}_i^k = \max\{\hat{\mathbf{v}}_i^{k-1}, \mathbf{v}_i^k\} \\
\mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma_k}{\sqrt{\hat{\mathbf{v}}_i^k + \delta}} \mathbf{m}_i^k, \quad i = \overline{1,n}, \quad \delta = 10^{-8}, \\
\gamma_k &= \gamma(1-\beta_1)\sqrt{1-\beta_2^k}.
\end{aligned}
$$

На самом деле, тот контрпример, который был предложен в статье, указанной на предыдущем слайде, имеет весьма специальный вид. На практике и исходная версия алгоритма часто работает хорошо.

AMSGrad

$$
\begin{aligned}
\mathbf{m}_i^k &= \beta_1 \mathbf{m}_i^{k-1} + (1-\beta_1)\mathbf{g}^k, \\
\mathbf{v}_i^k &= \beta_2 \mathbf{v}_i^{k-1} + (1-\beta_2)(\mathbf{g}_i^k)^2, \quad \hat{\mathbf{v}}_i^k = \max\{\hat{\mathbf{v}}_i^{k-1}, \mathbf{v}_i^k\} \\
\mathbf{x}_i^{k+1} &= \mathbf{x}_i^k - \frac{\gamma_k}{\sqrt{\hat{\mathbf{v}}_i^k + \delta}} \mathbf{m}_i^k, \quad i = \overline{1,n}, \quad \delta = 10^{-8}, \\
\gamma_k &= \gamma(1-\beta_1)\sqrt{1-\beta_2^k}.
\end{aligned}
$$

На самом деле, тот контрпример, который был предложен в статье, указанной на предыдущем слайде, имеет весьма специальный вид. На практике и исходная версия алгоритма часто работает хорошо.

Какие гарантии сходимости для Adagrad, RMSprop, Adam?

В невыпуклом случае (в ситуации 2 из трёх описанных ранее) недавно появился анализ Adam и чуть раньше для Adagrad, который говорит, что эти методы сходятся не лучше, чем SGD, то есть со скоростью $O\left(\frac{L(f(\mathbf{x})-f^*)\sigma^2}{\varepsilon^4}\right)$:

Alexandre Defossez, Leon Bottou, Francis Bach, Nicolas Usunier. **On the Convergence of Adam and Adagrad**, *arXiv preprint*, arXiv:2003.02395

Другой вариант адаптивного SGD

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В невыпуклом случае тот метод, который мы обсудили в первой части доклада, меняется следующим образом.

- Размер батча задаётся формулой $r_k = r = \max\left\{\frac{8\sigma_0^2}{\varepsilon^2}, 1\right\}$.
- Правило «принятия» оценки L_{k+1} записывается в виде:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon^2}{32 L_{k+1}}.$$

Другой вариант адаптивного SGD

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В невыпуклом случае тот метод, который мы обсудили в первой части доклада, меняется следующим образом.

- Размер батча задаётся формулой $r_k = r = \max\left\{\frac{8\sigma_0^2}{\varepsilon^2}, 1\right\}$.
- Правило «принятия» оценки L_{k+1} записывается в виде:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon^2}{32 L_{k+1}}.$$

Другой вариант адаптивного SGD

Ogaltsov, A., Dvinskikh, D., Dvurechensky, P., Gasnikov, A. and Spokoiny, V. **Adaptive gradient descent for convex and non-convex stochastic optimization.** *arXiv preprint* arXiv:1911.08380.

В невыпуклом случае тот метод, который мы обсудили в первой части доклада, меняется следующим образом.

- Размер батча задаётся формулой $r_k = r = \max\left\{\frac{8\sigma_0^2}{\varepsilon^2}, 1\right\}$.
- Правило «принятия» оценки L_{k+1} записывается в виде:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \langle \nabla^{r_{k+1}} f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + L_{k+1}\|\mathbf{x}^{k+1} - \mathbf{x}^k\|_2^2 + \frac{\varepsilon^2}{32 L_{k+1}}.$$

Другой вариант адаптивного SGD

В указанной статье было доказано, что для нахождения ε-стационарной точки (в среднем) описанному алгоритму требуется

$$N = O\left(\frac{\overline{L}^2(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{\underline{L}^2\varepsilon^2}\right),$$

причём суммарное число обращений к оракулу равняется

$$O\left(\frac{\sigma_0^2\overline{L}^2(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{\underline{L}^2\varepsilon^4}\right).$$

Другой вариант адаптивного SGD

В указанной статье было доказано, что для нахождения ε-стационарной точки (в среднем) описанному алгоритму требуется

$$N = O\left(\frac{\overline{L}^2(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{\underline{L}^2\varepsilon^2}\right),$$

причём суммарное число обращений к оракулу равняется

$$O\left(\frac{\sigma_0^2\overline{L}^2(f(\mathbf{x}^0) - f(\mathbf{x}^*))}{\underline{L}^2\varepsilon^4}\right).$$

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Расширение подхода, который мы обсудили в первой части доклада, на невыпуклые задачи без регуляризации, было получено в следующих статьях.

Khaled, A. and Richtárik, P. **Better Theory for SGD in the Nonconvex World.** *arXiv preprint* arXiv:2002.03329.

Li, Z. and Richtárik, P. **A Unified Analysis of Stochastic Gradient Methods for Nonconvex Federated Optimization.** *arXiv preprint* arXiv:2006.07013.

Далее мы рассмотрим подход, описанный во второй статье, поскольку он более общий, чем подход из первой статьи.

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Предположение 2

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая SGD.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы $A_1, B_1, C_1, D_1, A_2, B_2, C_2, \rho$ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2A_1(f(\mathbf{x}^k) - f^*) + B_1\|\nabla f(\mathbf{x}^k)\|_2^2 + D_1\sigma_k^2 + C_1, \tag{25}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2A_2 V_f(\mathbf{x}^k, \mathbf{x}^*) + B_2\|\nabla f(\mathbf{x}^k)\|_2^2 + C_2. \tag{26}$$

Этот подход покрывает разные варианты SGD, а также L-SVRG и SAGA, но не покрывает оптимальные методы SPIDER.

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Предположение 2

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая SGD.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы $A_1, B_1, C_1, D_1, A_2, B_2, C_2, \rho$ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2A_1(f(\mathbf{x}^k) - f^*) + B_1\|\nabla f(\mathbf{x}^k)\|_2^2 + D_1\sigma_k^2 + C_1, \tag{25}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2A_2 V_f(\mathbf{x}^k, \mathbf{x}^*) + B_2\|\nabla f(\mathbf{x}^k)\|_2^2 + C_2. \tag{26}$$

Этот подход покрывает разные варианты SGD, а также L-SVRG и SAGA, но не покрывает оптимальные методы SPIDER.

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Предположение 2

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая SGD.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы $A_1, B_1, C_1, D_1, A_2, B_2, C_2, \rho$ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2A_1(f(\mathbf{x}^k) - f^*) + B_1\|\nabla f(\mathbf{x}^k)\|_2^2 + D_1\sigma_k^2 + C_1, \tag{25}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2A_2 V_f(\mathbf{x}^k, \mathbf{x}^*) + B_2\|\nabla f(\mathbf{x}^k)\|_2^2 + C_2. \tag{26}$$

Этот подход покрывает разные варианты SGD, а также L-SVRG и SAGA, но не покрывает оптимальные методы SPIDER.

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Предположение 2

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая SGD.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы $A_1, B_1, C_1, D_1, A_2, B_2, C_2, \rho$ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2A_1(f(\mathbf{x}^k) - f^*) + B_1\|\nabla f(\mathbf{x}^k)\|_2^2 + D_1\sigma_k^2 + C_1, \tag{25}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2A_2 V_f(\mathbf{x}^k, \mathbf{x}^*) + B_2\|\nabla f(\mathbf{x}^k)\|_2^2 + C_2. \tag{26}$$

Этот подход покрывает разные варианты SGD, а также L-SVRG и SAGA, но не покрывает оптимальные методы SPIDER.

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Предположение 2

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая SGD.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы $A_1, B_1, C_1, D_1, A_2, B_2, C_2, \rho$ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2A_1(f(\mathbf{x}^k) - f^*) + B_1\|\nabla f(\mathbf{x}^k)\|_2^2 + D_1\sigma_k^2 + C_1, \tag{25}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2A_2 V_f(\mathbf{x}^k, \mathbf{x}^*) + B_2\|\nabla f(\mathbf{x}^k)\|_2^2 + C_2. \tag{26}$$

Этот подход покрывает разные варианты SGD, а также L-SVRG и SAGA, но не покрывает оптимальные методы SPIDER.

Обобщения Предположения 1 на невыпуклые задачи без регуляризации

Предположение 2

Пусть $\{\mathbf{x}^k\}$ — последовательность точек, генерируемая SGD.

1. Для всех $k \geq 0$ стох. градиент $\mathbf{g}^k$ — это несмещённая оценка $\nabla f(\mathbf{x}^k)$: $\mathbb{E}\left[\mathbf{g}^k \mid \mathbf{x}^k\right] = \nabla f(\mathbf{x}^k)$.
2. Существуют такие неотрицательные константы $A_1, B_1, C_1, D_1, A_2, B_2, C_2, \rho$ и такая последовательность (возможно, случайных) величин $\{\sigma_k^2\}_{k\geq 0}$, что для всех $k \geq 0$ выполняется:

$$\mathbb{E}\left[\left\|\mathbf{g}^k\right\|_2^2 \mid \mathbf{x}^k\right] \leq 2A_1(f(\mathbf{x}^k) - f^*) + B_1\|\nabla f(\mathbf{x}^k)\|_2^2 + D_1\sigma_k^2 + C_1, \tag{25}$$

$$\mathbb{E}\left[\sigma_{k+1}^2 \mid \sigma_k^2\right] \leq (1-\rho)\sigma_k^2 + 2A_2 V_f(\mathbf{x}^k, \mathbf{x}^*) + B_2\|\nabla f(\mathbf{x}^k)\|_2^2 + C_2. \tag{26}$$

Этот подход покрывает разные варианты SGD, а также L-SVRG и SAGA, но не покрывает оптимальные методы SPIDER.

Структурная невыпуклость

Necoara, I., Nesterov, Y. and Glineur, F., 2019. **Linear convergence of first order methods for non-strongly convex optimization.** *Mathematical Programming*, 175(1-2), pp.69-107.

Hinder, O., Sidford, A. and Sohoni, N.S., 2019. **Near-optimal methods for minimizing star-convex functions and beyond.** *arXiv preprint* arXiv:1906.11985.

Gower, R.M., Sebbouh, O. and Loizou, N., 2020. **SGD for Structured Nonconvex Functions: Learning Rates, Minibatching and Interpolation.** *arXiv preprint* arXiv:2006.10311.

Ослабления выпуклости

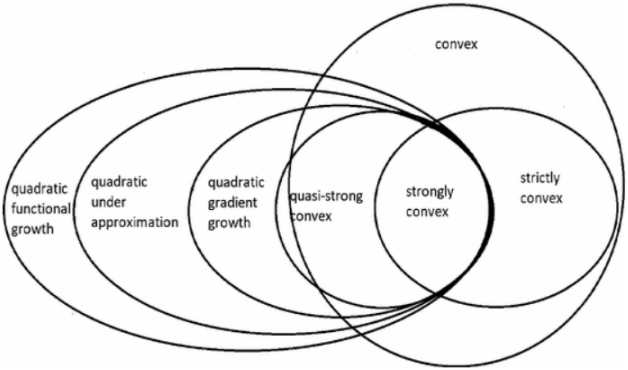


Figure: Necoara, I., Nesterov, Y. and Glineur, F., 2019. **Linear convergence of first order methods for non-strongly convex optimization.** *Mathematical Programming*, 175(1-2), pp.69-107.

Over-parameterized model

Рассмотрим задачу минимизации суммы

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{27}$$

- $\mathbf{x} \in \mathbb{R}^n$ — вектор параметров
- $f_i(\mathbf{x})$ — потери на i-м объекте из датасета
- Модель достаточно сложная (например, глубокая нейронная сеть), чтобы полностью «подогнаться» под обучающую выборку, или, как ещё часто говорят, выполняется условие интерполяции. Обычно это выполняется, когда n очень большое — отсюда и название. **На языке оптимизации это означает, что $\nabla f_i(\mathbf{x}^*) = 0$ для всех i.**
- Как это ни странно, но такие модели имеют хорошую обобщающую способность, что подтверждается эмпирически, а вот теоретическая сторона этого вопроса пока что до конца не изучена.

Over-parameterized model

Рассмотрим задачу минимизации суммы

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{27}$$

- $\mathbf{x} \in \mathbb{R}^n$ — вектор параметров
- $f_i(\mathbf{x})$ — потери на i-м объекте из датасета
- Модель достаточно сложная (например, глубокая нейронная сеть), чтобы полностью «подогнаться» под обучающую выборку, или, как ещё часто говорят, выполняется условие интерполяции. Обычно это выполняется, когда n очень большое — отсюда и название. **На языке оптимизации это означает, что $\nabla f_i(\mathbf{x}^*) = 0$ для всех i.**
- Как это ни странно, но такие модели имеют хорошую обобщающую способность, что подтверждается эмпирически, а вот теоретическая сторона этого вопроса пока что до конца не изучена.

Over-parameterized model

Рассмотрим задачу минимизации суммы

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{27}$$

- $\mathbf{x} \in \mathbb{R}^n$ — вектор параметров
- $f_i(\mathbf{x})$ — потери на i-м объекте из датасета
- Модель достаточно сложная (например, глубокая нейронная сеть), чтобы полностью «подогнаться» под обучающую выборку, или, как ещё часто говорят, выполняется условие интерполяции. Обычно это выполняется, когда n очень большое — отсюда и название. **На языке оптимизации это означает, что $\nabla f_i(\mathbf{x}^*) = 0$ для всех i.**
- Как это ни странно, но такие модели имеют хорошую обобщающую способность, что подтверждается эмпирически, а вот теоретическая сторона этого вопроса пока что до конца не изучена.

Over-parameterized model

Рассмотрим задачу минимизации суммы

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{27}$$

- $\mathbf{x} \in \mathbb{R}^n$ — вектор параметров
- $f_i(\mathbf{x})$ — потери на i-м объекте из датасета
- Модель достаточно сложная (например, глубокая нейронная сеть), чтобы полностью «подогнаться» под обучающую выборку, или, как ещё часто говорят, выполняется условие интерполяции. Обычно это выполняется, когда n очень большое — отсюда и название. **На языке оптимизации это означает, что $\nabla f_i(\mathbf{x}^*) = 0$ для всех i.**
- Как это ни странно, но такие модели имеют хорошую обобщающую способность, что подтверждается эмпирически, а вот теоретическая сторона этого вопроса пока что до конца не изучена.

Over-parameterized model

Рассмотрим задачу минимизации суммы

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{27}$$

- $\mathbf{x} \in \mathbb{R}^n$ — вектор параметров
- $f_i(\mathbf{x})$ — потери на i-м объекте из датасета
- Модель достаточно сложная (например, глубокая нейронная сеть), чтобы полностью «подогнаться» под обучающую выборку, или, как ещё часто говорят, выполняется условие интерполяции. Обычно это выполняется, когда n очень большое — отсюда и название. **На языке оптимизации это означает, что $\nabla f_i(\mathbf{x}^*) = 0$ для всех i.**
- Как это ни странно, но такие модели имеют хорошую обобщающую способность, что подтверждается эмпирически, а вот теоретическая сторона этого вопроса пока что до конца не изучена.

Over-parameterized model

Рассмотрим задачу минимизации суммы

$$f(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{x}) \to \min_{\mathbf{x}\in\mathbb{R}^n} \tag{27}$$

- $\mathbf{x} \in \mathbb{R}^n$ — вектор параметров
- $f_i(\mathbf{x})$ — потери на i-м объекте из датасета
- Модель достаточно сложная (например, глубокая нейронная сеть), чтобы полностью «подогнаться» под обучающую выборку, или, как ещё часто говорят, выполняется условие интерполяции. Обычно это выполняется, когда n очень большое — отсюда и название. **На языке оптимизации это означает, что $\nabla f_i(\mathbf{x}^*) = 0$ для всех i.**
- Как это ни странно, но такие модели имеют хорошую обобщающую способность, что подтверждается эмпирически, а вот теоретическая сторона этого вопроса пока что до конца не изучена.

Некоторые ссылки

Schmidt, M. and Roux, N.L., 2013. **Fast convergence of stochastic gradient descent under a strong growth condition.** *arXiv preprint* arXiv:1308.6370.

Ma, S., Bassily, R. and Belkin, M. **The power of interpolation: Understanding the effectiveness of SGD in modern over-parametrized learning.** *In International Conference on Machine Learning*, pp. 3325-3334 *(ICML 2018)*.

Vaswani, S., Bach, F. and Schmidt, M. **Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron.** *In The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1195-1204 *(AISTATS 2019)*.

Условие слабого роста

Кроме всего прочего, мы будем предполагать, что функции $f_i(\mathbf{x})$ выпуклые и L-гладкие.

Условие слабого роста (Weak Growth Condition = WGC)

Говорят, что функция $f(\mathbf{x})$ определённая в (27), удовлетворяет условию слабого роста с константой $\eta > 0$, если для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется неравенство:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right),$$

где j выбирается случайно и равновероятно из множества $\{1, \ldots, m\}$.

Действительно, из $\nabla f_i(\mathbf{x}^*) = 0$ получаем:

$$\begin{aligned} \mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] &= \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|_2^2 = \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\ &\leq \frac{2L}{m} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \right) = 2L(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

Условие слабого роста

Кроме всего прочего, мы будем предполагать, что функции $f_i(\mathbf{x})$ выпуклые и L-гладкие.

Условие слабого роста (Weak Growth Condition = WGC)

Говорят, что функция $f(\mathbf{x})$ определённая в (27), удовлетворяет условию слабого роста с константой $\eta > 0$, если для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется неравенство:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right),$$

где j выбирается случайно и равновероятно из множества $\{1, \ldots, m\}$.

Действительно, из $\nabla f_i(\mathbf{x}^*) = 0$ получаем:

$$\begin{aligned} \mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] &= \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|_2^2 = \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\ &\leq \frac{2L}{m} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \right) = 2L(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

Условие слабого роста

Кроме всего прочего, мы будем предполагать, что функции $f_i(\mathbf{x})$ выпуклые и L-гладкие.

Условие слабого роста (Weak Growth Condition = WGC)

Говорят, что функция $f(\mathbf{x})$ определённая в (27), удовлетворяет условию слабого роста с константой $\eta > 0$, если для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется неравенство:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right),$$

где j выбирается случайно и равновероятно из множества $\{1, \ldots, m\}$.

Действительно, из $\nabla f_i(\mathbf{x}^*) = 0$ получаем:

$$\begin{aligned} \mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] &= \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|_2^2 = \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\ &\leq \frac{2L}{m} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \right) = 2L(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

Условие слабого роста

Кроме всего прочего, мы будем предполагать, что функции $f_i(\mathbf{x})$ выпуклые и L-гладкие.

Условие слабого роста (Weak Growth Condition = WGC)

Говорят, что функция $f(\mathbf{x})$ определённая в (27), удовлетворяет условию слабого роста с константой $\eta > 0$, если для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется неравенство:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right),$$

где j выбирается случайно и равновероятно из множества $\{1, \ldots, m\}$.

Действительно, из $\nabla f_i(\mathbf{x}^*) = 0$ получаем:

$$\begin{aligned} \mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] &= \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|_2^2 = \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\ &\leq \frac{2L}{m} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \right) = 2L(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

Условие слабого роста

Кроме всего прочего, мы будем предполагать, что функции $f_i(\mathbf{x})$ выпуклые и L-гладкие.

Условие слабого роста (Weak Growth Condition = WGC)

Говорят, что функция $f(\mathbf{x})$ определённая в (27), удовлетворяет условию слабого роста с константой $\eta > 0$, если для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется неравенство:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right),$$

где j выбирается случайно и равновероятно из множества $\{1, \ldots, m\}$.

Действительно, из $\nabla f_i(\mathbf{x}^*) = 0$ получаем:

$$\begin{aligned} \mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] &= \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|_2^2 = \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\ &\leq \frac{2L}{m} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \right) = 2L(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

Условие слабого роста

Кроме всего прочего, мы будем предполагать, что функции $f_i(\mathbf{x})$ выпуклые и L-гладкие.

Условие слабого роста (Weak Growth Condition = WGC)

Говорят, что функция $f(\mathbf{x})$ определённая в (27), удовлетворяет условию слабого роста с константой $\eta > 0$, если для всех $\mathbf{x} \in \mathbb{R}^n$ выполняется неравенство:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right),$$

где j выбирается случайно и равновероятно из множества $\{1, \ldots, m\}$.

Действительно, из $\nabla f_i(\mathbf{x}^*) = 0$ получаем:

$$\begin{aligned} \mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] &= \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|_2^2 = \frac{1}{m} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \\ &\leq \frac{2L}{m} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - f_i(\mathbf{x}^*) - \langle \nabla f_i(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \right) = 2L(f(\mathbf{x}) - f(\mathbf{x}^*)) \end{aligned}$$

Условие слабого роста

Таким образом, мы показали, что в рассматриваемом случае выполняется WGC с $\eta = 1$. Верно и обратное: если выполнено WGC, то выполняются и условия интерполяции:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x}^*)\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}^*) - f(\mathbf{x}^*) \right) = 0 \Longrightarrow \nabla f_i(\mathbf{x}^*) = 0 \quad \forall i = \overline{1, m}.$$

Кроме того, бывают ситуации, когда условия интерполяции не выполнены идеально точно и в таком случае работает немного другой вариант WGC:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right) + \sigma^2,$$

где $\sigma^2 \geq 0$ и предполагается достаточно маленьким.

Условие слабого роста

Таким образом, мы показали, что в рассматриваемом случае выполняется WGC с $\eta = 1$. Верно и обратное: если выполнено WGC, то выполняются и условия интерполяции:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x}^*)\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}^*) - f(\mathbf{x}^*) \right) = 0 \Longrightarrow \nabla f_i(\mathbf{x}^*) = 0 \quad \forall i = \overline{1, m}.$$

Кроме того, бывают ситуации, когда условия интерполяции не выполнены идеально точно и в таком случае работает немного другой вариант WGC:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right) + \sigma^2,$$

где $\sigma^2 \geq 0$ и предполагается достаточно маленьким.

Условие слабого роста

Таким образом, мы показали, что в рассматриваемом случае выполняется WGC с $\eta = 1$. Верно и обратное: если выполнено WGC, то выполняются и условия интерполяции:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x}^*)\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}^*) - f(\mathbf{x}^*) \right) = 0 \Longrightarrow \nabla f_i(\mathbf{x}^*) = 0 \quad \forall i = \overline{1, m}.$$

Кроме того, бывают ситуации, когда условия интерполяции не выполнены идеально точно и в таком случае работает немного другой вариант WGC:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right) + \sigma^2,$$

где $\sigma^2 \geq 0$ и предполагается достаточно маленьким.

Условие слабого роста

Таким образом, мы показали, что в рассматриваемом случае выполняется WGC с $\eta = 1$. Верно и обратное: если выполнено WGC, то выполняются и условия интерполяции:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x}^*)\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}^*) - f(\mathbf{x}^*) \right) = 0 \Longrightarrow \nabla f_i(\mathbf{x}^*) = 0 \quad \forall i = \overline{1, m}.$$

Кроме того, бывают ситуации, когда условия интерполяции не выполнены идеально точно и в таком случае работает немного другой вариант WGC:

$$\mathbb{E}_j \left[\|\nabla f_j(\mathbf{x})\|_2^2 \right] \leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*) \right) + \sigma^2,$$

где $\sigma^2 \geq 0$ и предполагается достаточно маленьким.

Условие слабого роста

Мы будем понимать под WGC более общее условие. Будем считать, что мы решаем задачу (27) при помощи SGD, у которого стох. градиент $\mathbf{g}(\mathbf{x})$ удовлетворяет следующему условию.

Условие слабого роста (Weak Growth Condition = WGC), версия 2

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq 2L\eta\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие слабого роста

Мы будем понимать под WGC более общее условие. Будем считать, что мы решаем задачу (27) при помощи SGD, у которого стох. градиент $\mathbf{g}(\mathbf{x})$ удовлетворяет следующему условию.

Условие слабого роста (Weak Growth Condition = WGC), версия 2

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq 2L\eta\left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие слабого роста

Таким образом, SGD, для которого выполняется WGC, удовлетворяет Предположению 1 из первой части лекции с параметрами

$$A = L\eta, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Применяя Теорему 1 из первой части лекции получаем, что в μ-сильно выпуклом случае с $\gamma = \frac{1}{L\eta}$ выполняется

$$\mathbb{E}\left[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2\right] \leq \left(1 - \frac{\mu}{L\eta}\right)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\sigma^2}{L\eta}.$$

Условие слабого роста

Таким образом, SGD, для которого выполняется WGC, удовлетворяет Предположению 1 из первой части лекции с параметрами

$$A = L\eta, \quad D_1 = \sigma^2, \quad B = C = D_2 = 0, \quad \rho = 1, \quad \sigma_k^2 \equiv 0.$$

Применяя Теорему 1 из первой части лекции получаем, что в μ-сильно выпуклом случае с $\gamma = \frac{1}{L\eta}$ выполняется

$$\mathbb{E}\left[\|\mathbf{x}^N - \mathbf{x}^*\|_2^2\right] \leq \left(1 - \frac{\mu}{L\eta}\right)^N \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2 + \frac{\sigma^2}{L\eta}.$$

Условие слабого роста

Если σ^2 очень мало и η не очень большое, то смысла в использования методов редукции дисперсии нет: обычный SGD будет достигать приемлемой точности быстрее. Например, это так для over-parameterized models, т.е. для многих задач в DL.

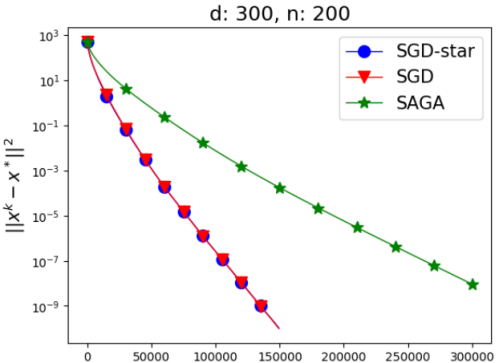
Условие слабого роста

Если σ^2 очень мало и η не очень большое, то смысла в использования методов редукции дисперсии нет: обычный SGD будет достигать приемлемой точности быстрее. Например, это так для over-parameterized models, т.е. для многих задач в DL.

Условие слабого роста: игрушечный пример

- Задача наименьших квадратов: $\mathbf{A} \in \mathbb{R}^{n \times d}$, $d \geq n$

$$\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{A}_{i:}^\top \mathbf{x} - b_i\right)^2 \to \min_{\mathbf{x} \in \mathbb{R}^d},$$



Условие сильного роста

Помимо WGC встречаются задачи, для которых выполняется и более сильное условие на стох. градиент $\mathbf{g}(\mathbf{x})$.

Условие сильного роста (Strong Growth Condition = SGC)

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2. \end{aligned}$$

Сразу же заметим, что выполнение SGC влечёт WGC:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2 = \eta \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2^2 + \sigma^2 \\ &\leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие сильного роста

Помимо WGC встречаются задачи, для которых выполняется и более сильное условие на стох. градиент $\mathbf{g}(\mathbf{x})$.

Условие сильного роста (Strong Growth Condition = SGC)

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2. \end{aligned}$$

Сразу же заметим, что выполнение SGC влечёт WGC:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2 = \eta \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2^2 + \sigma^2 \\ &\leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие сильного роста

Помимо WGC встречаются задачи, для которых выполняется и более сильное условие на стох. градиент $\mathbf{g}(\mathbf{x})$.

Условие сильного роста (Strong Growth Condition = SGC)

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2. \end{aligned}$$

Сразу же заметим, что выполнение SGC влечёт WGC:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2 = \eta \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2^2 + \sigma^2 \\ &\leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие сильного роста

Помимо WGC встречаются задачи, для которых выполняется и более сильное условие на стох. градиент $\mathbf{g}(\mathbf{x})$.

Условие сильного роста (Strong Growth Condition = SGC)

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2. \end{aligned}$$

Сразу же заметим, что выполнение SGC влечёт WGC:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2 = \eta \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2^2 + \sigma^2 \\ &\leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие сильного роста

Помимо WGC встречаются задачи, для которых выполняется и более сильное условие на стох. градиент $\mathbf{g}(\mathbf{x})$.

Условие сильного роста (Strong Growth Condition = SGC)

Существуют такие константы $\eta \geq 0$ и $\sigma^2 \geq 0$, что для всех $\mathbf{x} \in \mathbb{R}^n$ выполняются условия:

$$\begin{aligned} \mathbb{E}\left[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}\right] &= \nabla f(\mathbf{x}), \\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2. \end{aligned}$$

Сразу же заметим, что выполнение SGC влечёт WGC:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq \eta \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2 = \eta \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|_2^2 + \sigma^2 \\ &\leq 2L\eta \left(f(\mathbf{x}) - f(\mathbf{x}^*)\right) + \sigma^2. \end{aligned}$$

Условие сильного роста

Более того, если f_i — L-гладкие и f — μ-сильно выпуклая, то из WGC с константой η вытекает SGC с константой $\frac{\eta L}{\mu}$. Имеем:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq 2\mu \left(f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle\right),$$

Беря в этом неравенстве $\mathbf{y} = \mathbf{x}^*$ и пользуясь $\nabla f(\mathbf{x}^*) = 0$, получаем:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \quad \leq \quad 2L\eta(f(\mathbf{x}) - f(\mathbf{x}^*)) + \sigma^2 \leq \frac{L\eta}{\mu}\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Условие сильного роста

Более того, если f_i — L-гладкие и f — μ-сильно выпуклая, то из WGC с константой η вытекает SGC с константой $\frac{\eta L}{\mu}$. Имеем:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq 2\mu \left(f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\right),$$

Беря в этом неравенстве $\mathbf{y} = \mathbf{x}^*$ и пользуясь $\nabla f(\mathbf{x}^*) = 0$, получаем:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \quad \leq \quad 2L\eta(f(\mathbf{x}) - f(\mathbf{x}^*)) + \sigma^2 \leq \frac{L\eta}{\mu}\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Условие сильного роста

Более того, если f_i — L-гладкие и f — μ-сильно выпуклая, то из WGC с константой η вытекает SGC с константой $\frac{\eta L}{\mu}$. Имеем:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq 2\mu \left(f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\right),$$

Беря в этом неравенстве $\mathbf{y} = \mathbf{x}^*$ и пользуясь $\nabla f(\mathbf{x}^*) = 0$, получаем:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \quad \leq \quad 2L\eta(f(\mathbf{x}) - f(\mathbf{x}^*)) + \sigma^2 \leq \frac{L\eta}{\mu}\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Условие сильного роста

Более того, если f_i — L-гладкие и f — μ-сильно выпуклая, то из WGC с константой η вытекает SGC с константой $\frac{\eta L}{\mu}$. Имеем:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \geq 2\mu \left(f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\right),$$

Беря в этом неравенстве $\mathbf{y} = \mathbf{x}^*$ и пользуясь $\nabla f(\mathbf{x}^*) = 0$, получаем:

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \quad \leq \quad 2L\eta(f(\mathbf{x}) - f(\mathbf{x}^*)) + \sigma^2 \leq \frac{L\eta}{\mu}\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Условие сильного роста: 3 простых примера

Кроме всего прочего, сделаем 3 наблюдения.

1. Первое из них состоит в том, что $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ удовлетворяет условию сильного роста с параметрами $\eta = 1$ и $\sigma^2 = 0$.
2. Во-вторых, если дисперсия стох. градиента равномерно ограничена константой σ^2, то SGC выполнено с параметром $\eta = 1$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 \mid \mathbf{x}\right] \\ &\leq \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

3. В-третьих, есть примеры, когда условие сильного роста выполняется не только для минимизации суммы функций. Например, в простейшем случае покомпонентного метода ($\mathbf{g}(\mathbf{x})$ равен случайной частной производной в точке $\mathbf{x}$ или равен производной вдоль случайного направления с единичной Евклидовой сферы) SGC выполняется с $\eta = n$.

Оказывается, что условие сильного роста хорошо сочетается с ускорением.

Условие сильного роста: 3 простых примера

Кроме всего прочего, сделаем 3 наблюдения.

1. Первое из них состоит в том, что $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ удовлетворяет условию сильного роста с параметрами $\eta = 1$ и $\sigma^2 = 0$.
2. Во-вторых, если дисперсия стох. градиента равномерно ограничена константой σ^2, то SGC выполнено с параметром $\eta = 1$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 \mid \mathbf{x}\right] \\ &\leq \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

3. В-третьих, есть примеры, когда условие сильного роста выполняется не только для минимизации суммы функций. Например, в простейшем случае покомпонентного метода ($\mathbf{g}(\mathbf{x})$ равен случайной частной производной в точке $\mathbf{x}$ или равен производной вдоль случайного направления с единичной Евклидовой сферы) SGC выполняется с $\eta = n$.

Оказывается, что условие сильного роста хорошо сочетается с ускорением.

Условие сильного роста: 3 простых примера

Кроме всего прочего, сделаем 3 наблюдения.

1. Первое из них состоит в том, что $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ удовлетворяет условию сильного роста с параметрами $\eta = 1$ и $\sigma^2 = 0$.
2. Во-вторых, если дисперсия стох. градиента равномерно ограничена константой σ^2, то SGC выполнено с параметром $\eta = 1$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 \mid \mathbf{x}\right] \\ &\leq \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

3. В-третьих, есть примеры, когда условие сильного роста выполняется не только для минимизации суммы функций. Например, в простейшем случае покомпонентного метода ($\mathbf{g}(\mathbf{x})$ равен случайной частной производной в точке $\mathbf{x}$ или равен производной вдоль случайного направления с единичной Евклидовой сферы) SGC выполняется с $\eta = n$.

Оказывается, что условие сильного роста хорошо сочетается с ускорением.

Условие сильного роста: 3 простых примера

Кроме всего прочего, сделаем 3 наблюдения.

1. Первое из них состоит в том, что $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ удовлетворяет условию сильного роста с параметрами $\eta = 1$ и $\sigma^2 = 0$.
2. Во-вторых, если дисперсия стох. градиента равномерно ограничена константой σ^2, то SGC выполнено с параметром $\eta = 1$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 \mid \mathbf{x}\right] \\ &\leq \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

3. В-третьих, есть примеры, когда условие сильного роста выполняется не только для минимизации суммы функций. Например, в простейшем случае покомпонентного метода ($\mathbf{g}(\mathbf{x})$ равен случайной частной производной в точке $\mathbf{x}$ или равен производной вдоль случайного направления с единичной Евклидовой сферы) SGC выполняется с $\eta = n$.

Оказывается, что условие сильного роста хорошо сочетается с ускорением.

Условие сильного роста: 3 простых примера

Кроме всего прочего, сделаем 3 наблюдения.

1. Первое из них состоит в том, что $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ удовлетворяет условию сильного роста с параметрами $\eta = 1$ и $\sigma^2 = 0$.
2. Во-вторых, если дисперсия стох. градиента равномерно ограничена константой σ^2, то SGC выполнено с параметром $\eta = 1$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 \mid \mathbf{x}\right] \\ &\leq \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

3. В-третьих, есть примеры, когда условие сильного роста выполняется не только для минимизации суммы функций. Например, в простейшем случае покомпонентного метода ($\mathbf{g}(\mathbf{x})$ равен случайной частной производной в точке $\mathbf{x}$ или равен производной вдоль случайного направления с единичной Евклидовой сферы) SGC выполняется с $\eta = n$.

Оказывается, что условие сильного роста хорошо сочетается с ускорением.

Условие сильного роста: 3 простых примера

Кроме всего прочего, сделаем 3 наблюдения.

1. Первое из них состоит в том, что $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$ удовлетворяет условию сильного роста с параметрами $\eta = 1$ и $\sigma^2 = 0$.
2. Во-вторых, если дисперсия стох. градиента равномерно ограничена константой σ^2, то SGC выполнено с параметром $\eta = 1$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \mathbb{E}[\mathbf{g}(\mathbf{x}) \mid \mathbf{x}]\|_2^2 \mid \mathbf{x}\right] \\ &\leq \|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

3. В-третьих, есть примеры, когда условие сильного роста выполняется не только для минимизации суммы функций. Например, в простейшем случае покомпонентного метода ($\mathbf{g}(\mathbf{x})$ равен случайной частной производной в точке $\mathbf{x}$ или равен производной вдоль случайного направления с единичной Евклидовой сферы) SGC выполняется с $\eta = n$.

Оказывается, что условие сильного роста хорошо сочетается с ускорением.

Стохастический метод линейного каплинга (SLCM)

Алгоритм 7 Стохастический метод линейного каплинга (SLCM)

Вход: стартовая точка $\mathbf{x}^0 \in \mathbb{R}^d$, количество итераций N

Пусть $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{x}^0$
for $k = 0, 1, \ldots, N-1$ **do**
  Задать $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$ и $\tau_k = \frac{2}{\alpha_{k+1}L\eta^2} = \frac{2}{k+2}$
  $\mathbf{x}^{k+1} = \tau_k \mathbf{z}^k + (1-\tau_k)\mathbf{y}^k$
  Просэмплировать $\mathbf{g}(\mathbf{x}^{k+1}) = \mathbf{g}^{k+1}$, удовлетворяющий SGC
  $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} - \frac{1}{L\eta}\mathbf{g}^{k+1}$
  $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$
end for

- На каждой итерации делаем 2 градиентных шага — один шаг стандартный, а другой шаг имеет растущий рамер. После этого берём выпуклую комбинацию полученных точек.
- В детерминированном случае получаем один из вариантов метода Нестерова.

Стохастический метод линейного каплинга (SLCM)

Алгоритм 7 Стохастический метод линейного каплинга (SLCM)

Вход: стартовая точка $\mathbf{x}^0 \in \mathbb{R}^d$, количество итераций N

Пусть $\mathbf{y}^0 = \mathbf{z}^0 = \mathbf{x}^0$
for $k = 0, 1, \ldots, N-1$ **do**
  Задать $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$ и $\tau_k = \frac{2}{\alpha_{k+1}L\eta^2} = \frac{2}{k+2}$
  $\mathbf{x}^{k+1} = \tau_k \mathbf{z}^k + (1-\tau_k)\mathbf{y}^k$
  Просэмплировать $\mathbf{g}(\mathbf{x}^{k+1}) = \mathbf{g}^{k+1}$, удовлетворяющий SGC
  $\mathbf{y}^{k+1} = \mathbf{x}^{k+1} - \frac{1}{L\eta}\mathbf{g}^{k+1}$
  $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$
end for

- На каждой итерации делаем 2 градиентных шага — один шаг стандартный, а другой шаг имеет растущий рамер. После этого берём выпуклую комбинацию полученных точек.
- В детерминированном случае получаем один из вариантов метода Нестерова.

SLCM: анализ в выпуклом случае

Теорема 3

Пусть $f(\mathbf{x})$ — выпуклая L-гладкая функция и для стох. градиента $\mathbf{g}(\mathbf{x})$ выполено SGC. Тогда для любого $N > 0$ выход Алгоритма 7 удовлетворяет неравенству:

$$\mathbb{E}[f(\mathbf{y}^N) - f(\mathbf{x}^*)] \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}. \tag{28}$$

Доказательство состоит из четырёх основных шагов.

SLCM: анализ в выпуклом случае

Теорема 3

Пусть $f(\mathbf{x})$ — выпуклая L-гладкая функция и для стох. градиента $\mathbf{g}(\mathbf{x})$ выполено SGC. Тогда для любого $N > 0$ выход Алгоритма 7 удовлетворяет неравенству:

$$\mathbb{E}[f(\mathbf{y}^N) - f(\mathbf{x}^*)] \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}. \tag{28}$$

Доказательство состоит из четырёх основных шагов.

SLCM: анализ в выпуклом случае

Теорема 3

Пусть $f(\mathbf{x})$ — выпуклая L-гладкая функция и для стох. градиента $\mathbf{g}(\mathbf{x})$ выполено SGC. Тогда для любого $N > 0$ выход Алгоритма 7 удовлетворяет неравенству:

$$\mathbb{E}[f(\mathbf{y}^N) - f(\mathbf{x}^*)] \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}. \tag{28}$$

Доказательство состоит из четырёх основных шагов.

SLCM: анализ в выпуклом случае, шаг 1

Из L-гладкости функции f имеем:

$$f(\mathbf{y}^{k+1}) \leq f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^{k+1} - \mathbf{x}^{k+1} \rangle + \frac{L}{2}\|\mathbf{y}^{k+1} - \mathbf{x}^{k+1}\|_2^2$$
$$= f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{g}^{k+1} \rangle + \frac{1}{2L\eta^2}\|\mathbf{g}^{k+1}\|_2^2.$$

Пусть $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{z}^k, \mathbf{y}^k]$. Тогда $\mathbb{E}_k[\mathbf{g}^{k+1}] = \nabla f(\mathbf{x}^{k+1})$, поскольку $\mathbf{x}^{k+1}$ однозначно определяется через $\mathbf{y}^k$ и $\mathbf{z}^k$. Применяя SGC, получим:

$$\mathbb{E}_k[f(\mathbf{y}^{k+1})] \leq f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbb{E}_k[\mathbf{g}^{k+1}] \rangle + \frac{1}{2L\eta^2}\mathbb{E}_k\left[\|\mathbf{g}^{k+1}\|_2^2\right]$$
$$\leq f(\mathbf{x}^{k+1}) - \frac{1}{2L\eta}\|\nabla f(\mathbf{x}^{k+1})\|_2^2 + \frac{\sigma^2}{2L\eta^2},$$
$$\|\nabla f(\mathbf{x}^{k+1})\|_2^2 \leq 2L\eta\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta}$$

SLCM: анализ в выпуклом случае, шаг 1

Из L-гладкости функции f имеем:

$$\begin{aligned} f(\mathbf{y}^{k+1}) &\leq f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^{k+1} - \mathbf{x}^{k+1} \rangle + \frac{L}{2}\|\mathbf{y}^{k+1} - \mathbf{x}^{k+1}\|_2^2 \\ &= f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{g}^{k+1} \rangle + \frac{1}{2L\eta^2}\|\mathbf{g}^{k+1}\|_2^2. \end{aligned}$$

Пусть $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{z}^k, \mathbf{y}^k]$. Тогда $\mathbb{E}_k[\mathbf{g}^{k+1}] = \nabla f(\mathbf{x}^{k+1})$, поскольку $\mathbf{x}^{k+1}$ однозначно определяется через $\mathbf{y}^k$ и $\mathbf{z}^k$. Применяя SGC, получим:

$$\begin{aligned} \mathbb{E}_k[f(\mathbf{y}^{k+1})] &\leq f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbb{E}_k[\mathbf{g}^{k+1}] \rangle + \frac{1}{2L\eta^2}\mathbb{E}_k\left[\|\mathbf{g}^{k+1}\|_2^2\right] \\ &\leq f(\mathbf{x}^{k+1}) - \frac{1}{2L\eta}\|\nabla f(\mathbf{x}^{k+1})\|_2^2 + \frac{\sigma^2}{2L\eta^2}, \\ \|\nabla f(\mathbf{x}^{k+1})\|_2^2 &\leq 2L\eta\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta} \end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 1

Из L-гладкости функции f имеем:

$$\begin{aligned} f(\mathbf{y}^{k+1}) &\leq f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^{k+1} - \mathbf{x}^{k+1} \rangle + \frac{L}{2}\|\mathbf{y}^{k+1} - \mathbf{x}^{k+1}\|_2^2 \\ &= f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{g}^{k+1} \rangle + \frac{1}{2L\eta^2}\|\mathbf{g}^{k+1}\|_2^2. \end{aligned}$$

Пусть $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{z}^k, \mathbf{y}^k]$. Тогда $\mathbb{E}_k[\mathbf{g}^{k+1}] = \nabla f(\mathbf{x}^{k+1})$, поскольку $\mathbf{x}^{k+1}$ однозначно определяется через $\mathbf{y}^k$ и $\mathbf{z}^k$. Применяя SGC, получим:

$$\begin{aligned} \mathbb{E}_k[f(\mathbf{y}^{k+1})] &\leq f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbb{E}_k[\mathbf{g}^{k+1}] \rangle + \frac{1}{2L\eta^2}\mathbb{E}_k\left[\|\mathbf{g}^{k+1}\|_2^2\right] \\ &\leq f(\mathbf{x}^{k+1}) - \frac{1}{2L\eta}\|\nabla f(\mathbf{x}^{k+1})\|_2^2 + \frac{\sigma^2}{2L\eta^2}, \\ \|\nabla f(\mathbf{x}^{k+1})\|_2^2 &\leq 2L\eta\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta} \end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 1

Из L-гладкости функции f имеем:

$$\begin{aligned} f(\mathbf{y}^{k+1}) &\leq f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^{k+1} - \mathbf{x}^{k+1} \rangle + \frac{L}{2}\|\mathbf{y}^{k+1} - \mathbf{x}^{k+1}\|_2^2 \\ &= f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{g}^{k+1} \rangle + \frac{1}{2L\eta^2}\|\mathbf{g}^{k+1}\|_2^2. \end{aligned}$$

Пусть $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{z}^k, \mathbf{y}^k]$. Тогда $\mathbb{E}_k[\mathbf{g}^{k+1}] = \nabla f(\mathbf{x}^{k+1})$, поскольку $\mathbf{x}^{k+1}$ однозначно определяется через $\mathbf{y}^k$ и $\mathbf{z}^k$. Применяя SGC, получим:

$$\begin{aligned} \mathbb{E}_k[f(\mathbf{y}^{k+1})] &\leq f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbb{E}_k[\mathbf{g}^{k+1}] \rangle + \frac{1}{2L\eta^2}\mathbb{E}_k\left[\|\mathbf{g}^{k+1}\|_2^2\right] \\ &\leq f(\mathbf{x}^{k+1}) - \frac{1}{2L\eta}\|\nabla f(\mathbf{x}^{k+1})\|_2^2 + \frac{\sigma^2}{2L\eta^2}, \\ \|\nabla f(\mathbf{x}^{k+1})\|_2^2 &\leq 2L\eta\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta} \end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 1

Из L-гладкости функции f имеем:

$$\begin{aligned} f(\mathbf{y}^{k+1}) &\leq f(\mathbf{x}^{k+1}) + \langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^{k+1} - \mathbf{x}^{k+1} \rangle + \frac{L}{2}\|\mathbf{y}^{k+1} - \mathbf{x}^{k+1}\|_2^2 \\ &= f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{g}^{k+1} \rangle + \frac{1}{2L\eta^2}\|\mathbf{g}^{k+1}\|_2^2. \end{aligned}$$

Пусть $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathbf{z}^k, \mathbf{y}^k]$. Тогда $\mathbb{E}_k[\mathbf{g}^{k+1}] = \nabla f(\mathbf{x}^{k+1})$, поскольку $\mathbf{x}^{k+1}$ однозначно определяется через $\mathbf{y}^k$ и $\mathbf{z}^k$. Применяя SGC, получим:

$$\begin{aligned} \mathbb{E}_k[f(\mathbf{y}^{k+1})] &\leq f(\mathbf{x}^{k+1}) - \frac{1}{L\eta}\langle \nabla f(\mathbf{x}^{k+1}), \mathbb{E}_k[\mathbf{g}^{k+1}] \rangle + \frac{1}{2L\eta^2}\mathbb{E}_k\left[\|\mathbf{g}^{k+1}\|_2^2\right] \\ &\leq f(\mathbf{x}^{k+1}) - \frac{1}{2L\eta}\|\nabla f(\mathbf{x}^{k+1})\|_2^2 + \frac{\sigma^2}{2L\eta^2}, \\ \|\nabla f(\mathbf{x}^{k+1})\|_2^2 &\leq 2L\eta\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta} \end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 2

Так как $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$, то

$$\begin{aligned}\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle &= -\langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{z}^k - \mathbf{x}^*\rangle \\ &= \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2,\end{aligned}$$

так как $-\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \frac{1}{2}\|\mathbf{a}+\mathbf{b}\|_2^2$ для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Снова воспользуемся $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$:

$$\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle = \frac{\alpha_{k+1}^2}{2}\|\mathbf{g}^{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2.$$

Возьмём условное мат. ожидание $\mathbb{E}_k[\cdot]$ от левой и правой частей:

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle &\leq \frac{\alpha_{k+1}^2\eta}{2}\|\nabla f^{k+1}\|_2^2 + \frac{\alpha_{k+1}^2\sigma^2}{2} + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 \\ &\quad - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 2

Так как $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$, то

$$\begin{aligned}\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle &= -\langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{z}^k - \mathbf{x}^*\rangle \\ &= \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2,\end{aligned}$$

так как $-\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \frac{1}{2}\|\mathbf{a}+\mathbf{b}\|_2^2$ для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Снова воспользуемся $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$:

$$\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle = \frac{\alpha_{k+1}^2}{2}\|\mathbf{g}^{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2.$$

Возьмём условное мат. ожидание $\mathbb{E}_k[\cdot]$ от левой и правой частей:

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \leq\ & \frac{\alpha_{k+1}^2\eta}{2}\|\nabla f^{k+1}\|_2^2 + \frac{\alpha_{k+1}^2\sigma^2}{2} + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 \\ & - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 2

Так как $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$, то

$$\begin{aligned}\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle &= -\langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{z}^k - \mathbf{x}^*\rangle \\ &= \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2,\end{aligned}$$

так как $-\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \frac{1}{2}\|\mathbf{a}+\mathbf{b}\|_2^2$ для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Снова воспользуемся $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$:

$$\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle = \frac{\alpha_{k+1}^2}{2}\|\mathbf{g}^{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2.$$

Возьмём условное мат. ожидание $\mathbb{E}_k[\cdot]$ от левой и правой частей:

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle &\leq \frac{\alpha_{k+1}^2\eta}{2}\|\nabla f^{k+1}\|_2^2 + \frac{\alpha_{k+1}^2\sigma^2}{2} + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 \\ &\quad - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 2

Так как $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$, то

$$\begin{aligned}\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle &= -\langle \mathbf{z}^{k+1} - \mathbf{z}^k, \mathbf{z}^k - \mathbf{x}^*\rangle \\ &= \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^*\|_2^2,\end{aligned}$$

так как $-\langle \mathbf{a}, \mathbf{b}\rangle = \frac{1}{2}\|\mathbf{a}\|_2^2 + \frac{1}{2}\|\mathbf{b}\|_2^2 - \frac{1}{2}\|\mathbf{a}+\mathbf{b}\|_2^2$ для всех $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Снова воспользуемся $\mathbf{z}^{k+1} = \mathbf{z}^k - \alpha_{k+1}\mathbf{g}^{k+1}$:

$$\alpha_{k+1}\langle \mathbf{g}^{k+1}, \mathbf{z}^k - \mathbf{x}^*\rangle = \frac{\alpha_{k+1}^2}{2}\|\mathbf{g}^{k+1}\|_2^2 + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2.$$

Возьмём условное мат. ожидание $\mathbb{E}_k[\cdot]$ от левой и правой частей:

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \leq{}& \frac{\alpha_{k+1}^2\eta}{2}\|\nabla f^{k+1}\|_2^2 + \frac{\alpha_{k+1}^2\sigma^2}{2} + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 \\ &- \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 2

Используя неравенство

$$\|\nabla f(\mathbf{x}^{k+1})\|_2^2 \leq 2L\eta\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta},$$

полученное на шаге 1, мы имеем:

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle &\leq L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ &\quad + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 2

Используя неравенство

$$\|\nabla f(\mathbf{x}^{k+1})\|_2^2 \leq 2L\eta \left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \frac{\sigma^2}{\eta},$$

полученное на шаге 1, мы имеем:

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \quad &\leq \quad L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ &\quad + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 3

Из выпуклости f:

$$\begin{aligned}\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) &\leq \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^*\rangle \\ &= \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{z}^k\rangle \\ &\quad + \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \\ &\overset{\text{Ш. 2}}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^k - \mathbf{x}^{k+1}\rangle \\ &\quad + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ &\quad + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],\end{aligned}$$

где мы воспользовались

$$\mathbf{x}^{k+1} = \tau_k\mathbf{z}^k + (1-\tau_k)\mathbf{y}^k \Longleftrightarrow \tau_k(\mathbf{x}^{k+1} - \mathbf{z}^k) = (1-\tau_k)(\mathbf{y}^k - \mathbf{x}^{k+1})$$

SLCM: анализ в выпуклом случае, шаг 3

Из выпуклости f:

$$\begin{aligned}\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) &\leq \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^*\rangle \\ &= \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{z}^k\rangle \\ &\quad + \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \\ &\overset{\text{ш. 2}}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^k - \mathbf{x}^{k+1}\rangle \\ &\quad + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ &\quad + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],\end{aligned}$$

где мы воспользовались

$$\mathbf{x}^{k+1} = \tau_k\mathbf{z}^k + (1-\tau_k)\mathbf{y}^k \Longleftrightarrow \tau_k(\mathbf{x}^{k+1} - \mathbf{z}^k) = (1-\tau_k)(\mathbf{y}^k - \mathbf{x}^{k+1})$$

SLCM: анализ в выпуклом случае, шаг 3

Из выпуклости f:

$$\begin{aligned}
\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) &\leq \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^*\rangle \\
&= \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{z}^k\rangle \\
&\quad + \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \\
&\overset{\text{ш. 2}}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^k - \mathbf{x}^{k+1}\rangle \\
&\quad + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\
&\quad + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],
\end{aligned}$$

где мы воспользовались

$$\mathbf{x}^{k+1} = \tau_k\mathbf{z}^k + (1-\tau_k)\mathbf{y}^k \iff \tau_k(\mathbf{x}^{k+1} - \mathbf{z}^k) = (1-\tau_k)(\mathbf{y}^k - \mathbf{x}^{k+1})$$

SLCM: анализ в выпуклом случае, шаг 3

Из выпуклости f:

$$\begin{aligned}
\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) &\leq \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{x}^*\rangle \\
&= \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{x}^{k+1} - \mathbf{z}^k\rangle \\
&\quad + \alpha_{k+1}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{z}^k - \mathbf{x}^*\rangle \\
&\overset{\text{ш. 2}}{\leq} \frac{(1-\tau_k)\alpha_{k+1}}{\tau_k}\langle \nabla f(\mathbf{x}^{k+1}), \mathbf{y}^k - \mathbf{x}^{k+1}\rangle \\
&\quad + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\
&\quad + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],
\end{aligned}$$

где мы воспользовались

$$\mathbf{x}^{k+1} = \tau_k\mathbf{z}^k + (1-\tau_k)\mathbf{y}^k \Longleftrightarrow \tau_k(\mathbf{x}^{k+1} - \mathbf{z}^k) = (1-\tau_k)(\mathbf{y}^k - \mathbf{x}^{k+1})$$

SLCM: анализ в выпуклом случае, шаг 3

Применяя равенство $\tau_k = \frac{2}{\alpha_{k+1} L \eta^2}$ и выпуклость f, мы получаем

$$\begin{aligned}\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) \quad \leq \quad & (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1})(f(\mathbf{y}^k) - f(\mathbf{x}^{k+1})) \\ & + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ & + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],\end{aligned}$$

что можно переписать в эквивалентном виде:

$$\begin{aligned}\alpha_{k+1} f(\mathbf{x}^*) + \alpha_{k+1}^2\sigma^2 \quad \geq \quad & \alpha_{k+1}^2 L\eta^2 \mathbb{E}_k[f(\mathbf{y}^{k+1})] - (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1}) f(\mathbf{y}^k) \\ & + \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 3

Применяя равенство $\tau_k = \frac{2}{\alpha_{k+1} L \eta^2}$ и выпуклость f, мы получаем

$$\begin{aligned}\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) \quad \leq \quad & (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1})(f(\mathbf{y}^k) - f(\mathbf{x}^{k+1})) \\ & + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ & + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],\end{aligned}$$

что можно переписать в эквивалентном виде:

$$\begin{aligned}\alpha_{k+1} f(\mathbf{x}^*) + \alpha_{k+1}^2\sigma^2 \quad \geq \quad & \alpha_{k+1}^2 L\eta^2 \mathbb{E}_k[f(\mathbf{y}^{k+1})] - (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1}) f(\mathbf{y}^k) \\ & + \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 3

Применяя равенство $\tau_k = \frac{2}{\alpha_{k+1} L \eta^2}$ и выпуклость f, мы получаем

$$\begin{aligned}\alpha_{k+1}\left(f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*)\right) \quad \leq \quad & (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1})(f(\mathbf{y}^k) - f(\mathbf{x}^{k+1})) \\ & + L\eta^2\alpha_{k+1}^2\left(f(\mathbf{x}^{k+1}) - \mathbb{E}_k[f(\mathbf{y}^{k+1})]\right) + \alpha_{k+1}^2\sigma^2 \\ & + \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2 - \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right],\end{aligned}$$

что можно переписать в эквивалентном виде:

$$\begin{aligned}\alpha_{k+1} f(\mathbf{x}^*) + \alpha_{k+1}^2\sigma^2 \quad \geq \quad & \alpha_{k+1}^2 L\eta^2 \mathbb{E}_k[f(\mathbf{y}^{k+1})] - (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1}) f(\mathbf{y}^k) \\ & + \frac{1}{2}\mathbb{E}_k\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{z}^k - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 4

Возьмём полное мат. ожидание от неравенства, полученного на шаге 3, и сложим результат для $k = 0, 1, \ldots, N-1$:

$$\begin{aligned}\sum_{k=0}^{N-1} \left(f(\mathbf{x}^*)\alpha_{k+1} + \sigma^2\alpha_{k+1}^2\right) \quad \geq \quad & \sum_{k=0}^{N-1} \alpha_{k+1}^2 L\eta^2 \mathbb{E}[f(\mathbf{y}^{k+1})] \\ & - \sum_{k=0}^{N-1} (\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1})\mathbb{E}[f(\mathbf{y}^k)] \\ & + \underbrace{\sum_{k=0}^{N-1} \left(\frac{1}{2}\mathbb{E}\left[\|\mathbf{z}^{k+1} - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}^k - \mathbf{x}^*\|_2^2\right]\right)}_{\frac{1}{2}\mathbb{E}[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2] - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 4

Вспоминаем, что $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$, откуда

$$\begin{aligned}\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1} - \alpha_k^2 L\eta^2 &= \frac{k^2+4k+4}{4L\eta^2} - \frac{k+2}{2L\eta^2} - \frac{k^2+2k+1}{4L\eta^2} \\ &= -\frac{1}{4L\eta^2}.\end{aligned}$$

Следовательно,

$$\begin{aligned}\sum_{k=0}^{N-1} \left(f(\mathbf{x}^*)\alpha_{k+1} + \sigma^2\alpha_{k+1}^2\right) \geq & \underbrace{\sum_{k=0}^{N-1} \left(\alpha_{k+1}^2 L\eta^2 \mathbb{E}[f(\mathbf{x}^{k+1})] - \alpha_k^2 L\eta^2 \mathbb{E}[f(\mathbf{y}^k)]\right)}_{\alpha_N^2 L\eta^2 \mathbb{E}[f(\mathbf{x}^N)] - \frac{1}{4L\eta^2} f(\mathbf{y}^0)} \\ & + \frac{1}{4L\eta^2} \sum_{k=0}^{N-1} \mathbb{E}[f(\mathbf{y}^k)] \\ & + \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 4

Вспоминаем, что $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$, откуда

$$\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1} - \alpha_k^2 L\eta^2 = \frac{k^2+4k+4}{4L\eta^2} - \frac{k+2}{2L\eta^2} - \frac{k^2+2k+1}{4L\eta^2} = -\frac{1}{4L\eta^2}.$$

Следовательно,

$$\sum_{k=0}^{N-1} \left(f(\mathbf{x}^*)\alpha_{k+1} + \sigma^2\alpha_{k+1}^2\right) \geq \underbrace{\sum_{k=0}^{N-1} \left(\alpha_{k+1}^2 L\eta^2 \mathbb{E}[f(\mathbf{x}^{k+1})] - \alpha_k^2 L\eta^2 \mathbb{E}[f(\mathbf{y}^k)]\right)}_{\alpha_N^2 L\eta^2 \mathbb{E}[f(\mathbf{x}^N)] - \frac{1}{4L\eta^2} f(\mathbf{y}^0)}$$
$$+ \frac{1}{4L\eta^2} \sum_{k=0}^{N-1} \mathbb{E}[f(\mathbf{y}^k)]$$
$$+ \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.$$

SLCM: анализ в выпуклом случае, шаг 4

Вспоминаем, что $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$, откуда

$$\begin{aligned}\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1} - \alpha_k^2 L\eta^2 &= \frac{k^2+4k+4}{4L\eta^2} - \frac{k+2}{2L\eta^2} - \frac{k^2+2k+1}{4L\eta^2} \\ &= -\frac{1}{4L\eta^2}.\end{aligned}$$

Следовательно,

$$\begin{aligned}\sum_{k=0}^{N-1}\left(f(\mathbf{x}^*)\alpha_{k+1} + \sigma^2\alpha_{k+1}^2\right) \geq & \underbrace{\sum_{k=0}^{N-1}\left(\alpha_{k+1}^2 L\eta^2\mathbb{E}[f(\mathbf{x}^{k+1})] - \alpha_k^2 L\eta^2\mathbb{E}[f(\mathbf{y}^k)]\right)}_{\alpha_N^2 L\eta^2\mathbb{E}[f(\mathbf{x}^N)] - \frac{1}{4L\eta^2}f(\mathbf{y}^0)} \\ & + \frac{1}{4L\eta^2}\sum_{k=0}^{N-1}\mathbb{E}[f(\mathbf{y}^k)] \\ & + \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 4

Вспоминаем, что $\alpha_{k+1} = \frac{k+2}{2L\eta^2}$, откуда

$$\begin{aligned}\alpha_{k+1}^2 L\eta^2 - \alpha_{k+1} - \alpha_k^2 L\eta^2 &= \frac{k^2+4k+4}{4L\eta^2} - \frac{k+2}{2L\eta^2} - \frac{k^2+2k+1}{4L\eta^2} \\ &= -\frac{1}{4L\eta^2}.\end{aligned}$$

Следовательно,

$$\begin{aligned}\sum_{k=0}^{N-1}\left(f(\mathbf{x}^*)\alpha_{k+1} + \sigma^2\alpha_{k+1}^2\right) \geq & \underbrace{\sum_{k=0}^{N-1}\left(\alpha_{k+1}^2 L\eta^2\mathbb{E}[f(\mathbf{x}^{k+1})] - \alpha_k^2 L\eta^2\mathbb{E}[f(\mathbf{y}^k)]\right)}_{\alpha_N^2 L\eta^2\mathbb{E}[f(\mathbf{x}^N)] - \frac{1}{4L\eta^2}f(\mathbf{y}^0)} \\ & + \frac{1}{4L\eta^2}\sum_{k=0}^{N-1}\mathbb{E}[f(\mathbf{y}^k)] \\ & + \frac{1}{2}\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right] - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 4

Теперь воспользуемся явным видом α_{k+1}, чтобы оценить суммы в левой части:

$$\sum_{k=0}^{N-1} \alpha_{k+1} = \frac{1}{2L\eta^2} \sum_{k=0}^{N-1} (k+2) = \frac{N(N+3)}{4L\eta^2},$$

$$\sum_{k=0}^{N-1} \alpha_{k+1}^2 = \frac{1}{4L^2\eta^4} \sum_{k=0}^{N-1} (k+2)^2 \leq \frac{(N+1)^3}{4L^2\eta^4}$$

Получаем:

$$\frac{N(N+3)}{4L\eta^2} f(\mathbf{x}^*) + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \geq \frac{(N+1)^2}{4L\eta^2} \mathbb{E}[f(\mathbf{y}^N)] + \frac{1}{4L\eta^2} \sum_{k=1}^{N-1} \mathbb{E}\underbrace{[f(\mathbf{y}^k)]}_{\geq f(\mathbf{x}^*)}$$

$$+ \frac{1}{2} \underbrace{\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right]}_{\geq 0} - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.$$

SLCM: анализ в выпуклом случае, шаг 4

Теперь воспользуемся явным видом α_{k+1}, чтобы оценить суммы в левой части:

$$\sum_{k=0}^{N-1} \alpha_{k+1} = \frac{1}{2L\eta^2} \sum_{k=0}^{N-1} (k+2) = \frac{N(N+3)}{4L\eta^2},$$

$$\sum_{k=0}^{N-1} \alpha_{k+1}^2 = \frac{1}{4L^2\eta^4} \sum_{k=0}^{N-1} (k+2)^2 \leq \frac{(N+1)^3}{4L^2\eta^4}$$

Получаем:

$$\frac{N(N+3)}{4L\eta^2} f(\mathbf{x}^*) + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \geq \frac{(N+1)^2}{4L\eta^2} \mathbb{E}[f(\mathbf{y}^N)] + \frac{1}{4L\eta^2} \sum_{k=1}^{N-1} \mathbb{E}\underbrace{[f(\mathbf{y}^k)]}_{\geq f(\mathbf{x}^*)}$$

$$+ \frac{1}{2} \underbrace{\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right]}_{\geq 0} - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.$$

SLCM: анализ в выпуклом случае, шаг 4

Теперь воспользуемся явным видом α_{k+1}, чтобы оценить суммы в левой части:

$$\sum_{k=0}^{N-1} \alpha_{k+1} = \frac{1}{2L\eta^2} \sum_{k=0}^{N-1} (k+2) = \frac{N(N+3)}{4L\eta^2},$$

$$\sum_{k=0}^{N-1} \alpha_{k+1}^2 = \frac{1}{4L^2\eta^4} \sum_{k=0}^{N-1} (k+2)^2 \leq \frac{(N+1)^3}{4L^2\eta^4}$$

Получаем:

$$\frac{N(N+3)}{4L\eta^2} f(\mathbf{x}^*) + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \geq \frac{(N+1)^2}{4L\eta^2} \mathbb{E}[f(\mathbf{y}^N)] + \frac{1}{4L\eta^2} \sum_{k=1}^{N-1} \mathbb{E}\underbrace{[f(\mathbf{y}^k)]}_{\geq f(\mathbf{x}^*)}$$

$$+ \frac{1}{2} \underbrace{\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right]}_{\geq 0} - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.$$

SLCM: анализ в выпуклом случае, шаг 4

Теперь воспользуемся явным видом α_{k+1}, чтобы оценить суммы в левой части:

$$\sum_{k=0}^{N-1} \alpha_{k+1} = \frac{1}{2L\eta^2} \sum_{k=0}^{N-1} (k+2) = \frac{N(N+3)}{4L\eta^2},$$

$$\sum_{k=0}^{N-1} \alpha_{k+1}^2 = \frac{1}{4L^2\eta^4} \sum_{k=0}^{N-1} (k+2)^2 \leq \frac{(N+1)^3}{4L^2\eta^4}$$

Получаем:

$$\frac{N(N+3)}{4L\eta^2} f(\mathbf{x}^*) + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \geq \frac{(N+1)^2}{4L\eta^2} \mathbb{E}[f(\mathbf{y}^N)] + \frac{1}{4L\eta^2} \sum_{k=1}^{N-1} \mathbb{E}\underbrace{[f(\mathbf{y}^k)]}_{\geq f(\mathbf{x}^*)}$$

$$+ \frac{1}{2} \underbrace{\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right]}_{\geq 0} - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.$$

SLCM: анализ в выпуклом случае, шаг 4

Теперь воспользуемся явным видом α_{k+1}, чтобы оценить суммы в левой части:

$$\begin{aligned}\sum_{k=0}^{N-1} \alpha_{k+1} &= \frac{1}{2L\eta^2}\sum_{k=0}^{N-1}(k+2) = \frac{N(N+3)}{4L\eta^2},\\ \sum_{k=0}^{N-1} \alpha_{k+1}^2 &= \frac{1}{4L^2\eta^4}\sum_{k=0}^{N-1}(k+2)^2 \leq \frac{(N+1)^3}{4L^2\eta^4}\end{aligned}$$

Получаем:

$$\begin{aligned}\frac{N(N+3)}{4L\eta^2} f(\mathbf{x}^*) + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \quad \geq \quad & \frac{(N+1)^2}{4L\eta^2}\mathbb{E}[f(\mathbf{y}^N)] + \frac{1}{4L\eta^2}\sum_{k=1}^{N-1}\mathbb{E}\underbrace{[f(\mathbf{y}^k)]}_{\geq f(\mathbf{x}^*)}\\ & + \frac{1}{2}\underbrace{\mathbb{E}\left[\|\mathbf{z}^N - \mathbf{x}^*\|_2^2\right]}_{\geq 0} - \frac{1}{2}\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2.\end{aligned}$$

SLCM: анализ в выпуклом случае, шаг 4

Далее, переписываем полученное неравенство:

$$\begin{aligned}\frac{(N+1)^2}{4L\eta^2}\mathbb{E}[f(\mathbf{y}^N)] &\leq \left(\frac{N(N+3)}{4L\eta^2} - \frac{N-1}{4L\eta^2}\right) f(\mathbf{x}^*) + \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2} \\ &\quad + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \\ &= \frac{(N+1)^2}{4L\eta^2} f(\mathbf{x}^*) + \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2} + \frac{\sigma^2(N+1)^3}{4L^2\eta^4}.\end{aligned}$$

Перегруппируем слагаемые и получим требуемый результат:

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}$$

SLCM: анализ в выпуклом случае, шаг 4

Далее, переписываем полученное неравенство:

$$\begin{aligned}\frac{(N+1)^2}{4L\eta^2}\mathbb{E}[f(\mathbf{y}^N)] &\leq \left(\frac{N(N+3)}{4L\eta^2} - \frac{N-1}{4L\eta^2}\right) f(\mathbf{x}^*) + \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2} \\ &\quad + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \\ &= \frac{(N+1)^2}{4L\eta^2} f(\mathbf{x}^*) + \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2} + \frac{\sigma^2(N+1)^3}{4L^2\eta^4}.\end{aligned}$$

Перегруппируем слагаемые и получим требуемый результат:

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}$$

SLCM: анализ в выпуклом случае, шаг 4

Далее, переписываем полученное неравенство:

$$\begin{aligned}\frac{(N+1)^2}{4L\eta^2}\mathbb{E}[f(\mathbf{y}^N)] &\leq \left(\frac{N(N+3)}{4L\eta^2} - \frac{N-1}{4L\eta^2}\right) f(\mathbf{x}^*) + \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2} \\ &\quad + \frac{\sigma^2(N+1)^3}{4L^2\eta^4} \\ &= \frac{(N+1)^2}{4L\eta^2} f(\mathbf{x}^*) + \frac{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{2} + \frac{\sigma^2(N+1)^3}{4L^2\eta^4}.\end{aligned}$$

Перегруппируем слагаемые и получим требуемый результат:

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}$$

SLCM: сходимость при $\sigma = 0$

Если $\sigma^2 = 0$, то получаем

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2}.$$

Иными словами, для достижения $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$ требуется $N = O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций.

- В детерминированном случае полученная оценка соответствует нижней оценке на число итераций в выпуклом случае.
- Для полноградиентного метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(m\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
- Если выполняется SGC с $\sigma^2 = 0$, то для стохастического метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

SLCM: сходимость при $\sigma = 0$

Если $\sigma^2 = 0$, то получаем

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2}.$$

Иными словами, для достижения $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$ требуется $N = O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций.

- В детерминированном случае полученная оценка соответствует нижней оценке на число итераций в выпуклом случае.
- Для полноградиентного метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(m\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
- Если выполняется SGC с $\sigma^2 = 0$, то для стохастического метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

SLCM: сходимость при $\sigma = 0$

Если $\sigma^2 = 0$, то получаем

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2}.$$

Иными словами, для достижения $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$ требуется $N = O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций.

- В детерминированном случае полученная оценка соответствует нижней оценке на число итераций в выпуклом случае.
- Для полноградиентного метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(m\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
- Если выполняется SGC с $\sigma^2 = 0$, то для стохастического метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

SLCM: сходимость при $\sigma = 0$

Если $\sigma^2 = 0$, то получаем

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2}.$$

Иными словами, для достижения $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$ требуется $N = O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций.

- В детерминированном случае полученная оценка соответствует нижней оценке на число итераций в выпуклом случае.
- Для полноградиентного метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(m\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
- Если выполняется SGC с $\sigma^2 = 0$, то для стохастического метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

SLCM: сходимость при $\sigma = 0$

Если $\sigma^2 = 0$, то получаем

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2}.$$

Иными словами, для достижения $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$ требуется $N = O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций.

- В детерминированном случае полученная оценка соответствует нижней оценке на число итераций в выпуклом случае.
- Для полноградиентного метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(m\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
- Если выполняется SGC с $\sigma^2 = 0$, то для стохастического метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

SLCM: сходимость при $\sigma = 0$

Если $\sigma^2 = 0$, то получаем

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2}.$$

Иными словами, для достижения $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$ требуется $N = O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$ итераций.

- В детерминированном случае полученная оценка соответствует нижней оценке на число итераций в выпуклом случае.
- Для полноградиентного метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(m\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
- Если выполняется SGC с $\sigma^2 = 0$, то для стохастического метода линейного каплинга, применённого к задаче минимизации суммы, число подсчётов градиентов слагаемых равняется $O\left(\eta\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

SLCM: сходимость при $\sigma > 0$

Если $\sigma^2 > 0$, то получаем, что с ростом числа итераций «шум» накапливается:

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}.$$

Чтобы сойтись с нужной точностью, нужно сделать параметр σ^2 достаточно маленким. Это можно сделать при помощи мини-батчинга.

SLCM: сходимость при $\sigma > 0$

Если $\sigma^2 > 0$, то получаем, что с ростом числа итераций «шум» накапливается:

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L\eta^2}.$$

Чтобы сойтись с нужной точностью, нужно сделать параметр σ^2 достаточно маленким. Это можно сделать при помощи мини-батчинга.

SGC и мини-батчинг

Заметим, что если $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC, то

$$\|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \eta\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2,$$

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq (\eta - 1)\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Если $\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_r(\mathbf{x})$ — независимые стох. градиенты, удовлетворяющие SGC с параметрами η и σ^2, то для $\mathbf{g}(\mathbf{x}) = \frac{1}{r}\sum\limits_{i=1}^{r} \mathbf{g}_i(\mathbf{x})$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\mathbb{E}\left[\left\|\sum_{i=1}^{r}(\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|_2^2 \mid \mathbf{x}\right] \\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right].\end{aligned}$$

SGC и мини-батчинг

Заметим, что если $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC, то

$$\|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \eta\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2,$$

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq (\eta - 1)\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Если $\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_r(\mathbf{x})$ — независимые стох. градиенты, удовлетворяющие SGC с параметрами η и σ^2, то для $\mathbf{g}(\mathbf{x}) = \frac{1}{r}\sum\limits_{i=1}^{r} \mathbf{g}_i(\mathbf{x})$:

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \\
&= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\mathbb{E}\left[\left\|\sum_{i=1}^{r}(\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|_2^2 \mid \mathbf{x}\right] \\
&= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right].
\end{aligned}$$

SGC и мини-батчинг

Заметим, что если $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC, то

$$\|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \eta\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2,$$

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq (\eta - 1)\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.$$

Если $\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_r(\mathbf{x})$ — независимые стох. градиенты, удовлетворяющие SGC с параметрами η и σ^2, то для $\mathbf{g}(\mathbf{x}) = \frac{1}{r}\sum\limits_{i=1}^{r} \mathbf{g}_i(\mathbf{x})$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\mathbb{E}\left[\left\|\sum_{i=1}^{r}(\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|_2^2 \mid \mathbf{x}\right] \\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right].\end{aligned}$$

SGC и мини-батчинг

Заметим, что если $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC, то

$$\begin{aligned}\|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \eta\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2,\\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq (\eta - 1)\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

Если $\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_r(\mathbf{x})$ — независимые стох. градиенты, удовлетворяющие SGC с параметрами η и σ^2, то для $\mathbf{g}(\mathbf{x}) = \frac{1}{r}\sum\limits_{i=1}^{r}\mathbf{g}_i(\mathbf{x})$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right]\\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\mathbb{E}\left[\left\|\sum_{i=1}^{r}(\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|_2^2 \mid \mathbf{x}\right]\\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right].\end{aligned}$$

SGC и мини-батчинг

Заметим, что если $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC, то

$$\begin{aligned}\|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \eta\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2,\\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq (\eta - 1)\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

Если $\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_r(\mathbf{x})$ — независимые стох. градиенты, удовлетворяющие SGC с параметрами η и σ^2, то для $\mathbf{g}(\mathbf{x}) = \frac{1}{r}\sum\limits_{i=1}^{r}\mathbf{g}_i(\mathbf{x})$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right]\\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\mathbb{E}\left[\left\|\sum_{i=1}^{r}(\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|_2^2 \mid \mathbf{x}\right]\\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right].\end{aligned}$$

SGC и мини-батчинг

Заметим, что если $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC, то

$$\begin{aligned}\|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] \leq \eta\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2,\\ \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &\leq (\eta - 1)\|\nabla f(\mathbf{x})\|_2^2 + \sigma^2.\end{aligned}$$

Если $\mathbf{g}_1(\mathbf{x}), \ldots, \mathbf{g}_r(\mathbf{x})$ — независимые стох. градиенты, удовлетворяющие SGC с параметрами η и σ^2, то для $\mathbf{g}(\mathbf{x}) = \frac{1}{r}\sum\limits_{i=1}^{r}\mathbf{g}_i(\mathbf{x})$:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] &= \|\nabla f(\mathbf{x})\|_2^2 + \mathbb{E}\left[\|\mathbf{g}(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right]\\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\mathbb{E}\left[\left\|\sum_{i=1}^{r}(\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x}))\right\|_2^2 \mid \mathbf{x}\right]\\ &= \|\nabla f(\mathbf{x})\|_2^2 + \frac{1}{r^2}\sum_{i=1}^{r}\mathbb{E}\left[\|\mathbf{g}_i(\mathbf{x}) - \nabla f(\mathbf{x})\|_2^2 \mid \mathbf{x}\right].\end{aligned}$$

SLCM и мини-батчинг

Таким образом,

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \left(1 + \frac{\eta - 1}{r}\right)\|\nabla f(\mathbf{x})\|_2^2 + \frac{\sigma^2}{r},$$

то есть $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC с параметрами $\eta_r = 1 + \frac{\eta-1}{r}$ и $\sigma_r^2 = \frac{\sigma^2}{r}$. Если теперь использовать мини-батчинг в SLCM, то получим оценку

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2}.$$

Для заданного уровня точности ε выберем r как

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2(N+1)}{4L\varepsilon}\right\}$$

SLCM и мини-батчинг

Таким образом,

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \left(1 + \frac{\eta - 1}{r}\right)\|\nabla f(\mathbf{x})\|_2^2 + \frac{\sigma^2}{r},$$

то есть $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC с параметрами $\eta_r = 1 + \frac{\eta-1}{r}$ и $\sigma_r^2 = \frac{\sigma^2}{r}$. Если теперь использовать мини-батчинг в SLCM, то получим оценку

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2}.$$

Для заданного уровня точности ε выберем r как

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2(N+1)}{4L\varepsilon}\right\}$$

SLCM и мини-батчинг

Таким образом,

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \left(1 + \frac{\eta - 1}{r}\right)\|\nabla f(\mathbf{x})\|_2^2 + \frac{\sigma^2}{r},$$

то есть $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC с параметрами $\eta_r = 1 + \frac{\eta-1}{r}$ и $\sigma_r^2 = \frac{\sigma^2}{r}$. Если теперь использовать мини-батчинг в SLCM, то получим оценку

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2}.$$

Для заданного уровня точности ε выберем r как

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2(N+1)}{4L\varepsilon}\right\}$$

SLCM и мини-батчинг

Таким образом,

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{x})\|_2^2 \mid \mathbf{x}\right] = \left(1 + \frac{\eta - 1}{r}\right)\|\nabla f(\mathbf{x})\|_2^2 + \frac{\sigma^2}{r},$$

то есть $\mathbf{g}(\mathbf{x})$ удовлетворяет SGC с параметрами $\eta_r = 1 + \frac{\eta-1}{r}$ и $\sigma_r^2 = \frac{\sigma^2}{r}$. Если теперь использовать мини-батчинг в SLCM, то получим оценку

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2}.$$

Для заданного уровня точности ε выберем r как

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2(N+1)}{4L\varepsilon}\right\}$$

SLCM и мини-батчинг

Тогда $\eta_r \leq 2$ и

$$\begin{aligned}\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2} \\ &\leq \frac{8L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\varepsilon}{2}.\end{aligned}$$

Если $N \geq 4\sqrt{\frac{L\|\mathbf{x}^0-\mathbf{x}^*\|_2^2}{\varepsilon}} - 1$, то

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

SLCM и мини-батчинг

Тогда $\eta_r \leq 2$ и

$$\begin{aligned}\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2} \\ &\leq \frac{8L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\varepsilon}{2}.\end{aligned}$$

Если $N \geq 4\sqrt{\frac{L\|\mathbf{x}^0-\mathbf{x}^*\|_2^2}{\varepsilon}} - 1$, то

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

SLCM и мини-батчинг

Тогда $\eta_r \leq 2$ и

$$\begin{aligned}\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2} \\ &\leq \frac{8L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\varepsilon}{2}.\end{aligned}$$

Если $N \geq 4\sqrt{\frac{L\|\mathbf{x^0}-\mathbf{x}^*\|_\mathbf{2}^\mathbf{2}}{\varepsilon}} - 1$, то

$$\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

SLCM и мини-батчинг

Итак, SLCM с батчами размера

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2 \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{\sqrt{L}\varepsilon^{3/2}}\right\}$$

за $N = \left\lceil 4\sqrt{\frac{L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{\varepsilon}} - 1 \right\rceil$ итераций гарантирует $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$. При этом общее число обращений к оракулу равняется

$$Nr = O\left(\max\left\{\eta\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}\right).$$

SLCM и мини-батчинг

Итак, SLCM с батчами размера

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2 \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{\sqrt{L}\varepsilon^{3/2}}\right\}$$

за $N = \left\lceil 4\sqrt{\frac{L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{\varepsilon}} - 1 \right\rceil$ итераций гарантирует $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$. При этом общее число обращений к оракулу равняется

$$Nr = O\left(\max\left\{\eta\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}\right).$$

SLCM и мини-батчинг

Итак, SLCM с батчами размера

$$r = \max\left\{1, \eta - 1, \frac{\sigma^2 \|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{\sqrt{L}\varepsilon^{3/2}}\right\}$$

за $N = \left\lceil 4\sqrt{\frac{L\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{\varepsilon}} - 1 \right\rceil$ итераций гарантирует $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \varepsilon$. При этом общее число обращений к оракулу равняется

$$Nr = O\left(\max\left\{\eta\sqrt{\frac{LR^2}{\varepsilon}}, \frac{\sigma^2 R^2}{\varepsilon^2}\right\}\right).$$

SLCM: сильно выпуклый случай

Как из того, что мы получили для выпуклого случая, получить оценку на скорость сходимости в μ-сильно выпуклом случае? Самый простой способ — применить технику рестартов. Мы имеем:

$$\begin{aligned}\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L r \eta_r^2} \\ f(\mathbf{y}^N) - f(\mathbf{x}^*) &\geq \frac{\mu}{2}\mathbb{E}\left[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2\right].\end{aligned}$$

Мы хотим добиться точности $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\mu R^2}{4}$.Чтобы этого достичь необходимо

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2}\right\}\right) \text{ обращений к оракулу.}$$

SLCM: сильно выпуклый случай

Как из того, что мы получили для выпуклого случая, получить оценку на скорость сходимости в μ-сильно выпуклом случае? Самый простой способ — применить технику рестартов. Мы имеем:

$$\begin{aligned} \mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{L r \eta_r^2} \\ f(\mathbf{y}^N) - f(\mathbf{x}^*) &\geq \frac{\mu}{2}\mathbb{E}\left[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2\right]. \end{aligned}$$

Мы хотим добиться точности $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\mu R^2}{4}$.Чтобы этого достичь необходимо

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2}\right\}\right) \text{ обращений к оракулу.}$$

SLCM: сильно выпуклый случай

Как из того, что мы получили для выпуклого случая, получить оценку на скорость сходимости в μ-сильно выпуклом случае? Самый простой способ — применить технику рестартов. Мы имеем:

$$\begin{aligned} \mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2} \\ f(\mathbf{y}^N) - f(\mathbf{x}^*) &\geq \frac{\mu}{2}\mathbb{E}\left[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2\right]. \end{aligned}$$

Мы хотим добиться точности $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\mu R^2}{4}$.Чтобы этого достичь необходимо

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2}\right\}\right) \text{ обращений к оракулу.}$$

SLCM: сильно выпуклый случай

Как из того, что мы получили для выпуклого случая, получить оценку на скорость сходимости в μ-сильно выпуклом случае? Самый простой способ — применить технику рестартов. Мы имеем:

$$\begin{aligned} \mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2} \\ f(\mathbf{y}^N) - f(\mathbf{x}^*) &\geq \frac{\mu}{2}\mathbb{E}\left[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2\right]. \end{aligned}$$

Мы хотим добиться точности $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\mu R^2}{4}$. Чтобы этого достичь необходимо

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2}\right\}\right) \text{ обращений к оракулу.}$$

SLCM: сильно выпуклый случай

Как из того, что мы получили для выпуклого случая, получить оценку на скорость сходимости в μ-сильно выпуклом случае? Самый простой способ — применить технику рестартов. Мы имеем:

$$\begin{aligned} \mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) &\leq \frac{2L\eta_r^2\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}{(N+1)^2} + \frac{\sigma^2(N+1)}{Lr\eta_r^2} \\ f(\mathbf{y}^N) - f(\mathbf{x}^*) &\geq \frac{\mu}{2}\mathbb{E}\left[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2\right]. \end{aligned}$$

Мы хотим добиться точности $\mathbb{E}[f(\mathbf{y}^N)] - f(\mathbf{x}^*) \leq \frac{\mu R^2}{4}$.Чтобы этого достичь необходимо

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2}\right\}\right) \text{ обращений к оракулу.}$$

SLCM: сильно выпуклый случай

Тогда получим, что

$$\mathbb{E}[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2] \leq \frac{R^2}{2}.$$

Теперь запустим тот же метод, но в качестве стартовой точки возьмём $\mathbf{y}^N$. Аналогичными рассуждениями, мы получаем, что через

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2}\right\}\right) \text{ обращений к оракулу}$$

будет выполнено

$$\mathbb{E}[\|\mathbf{y}_1^N - \mathbf{x}^*\|_2^2] \leq \frac{R^2}{4},$$

где $\mathbf{y}_1^N$ — выход алгоритма после 1-го рестарта.

SLCM: сильно выпуклый случай

Тогда получим, что

$$\mathbb{E}[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2] \leq \frac{R^2}{2}.$$

Теперь запустим тот же метод, но в качестве стартовой точки возьмём $\mathbf{y}^N$. Аналогичными рассуждениями, мы получаем, что через

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2}\right\}\right) \text{ обращений к оракулу}$$

будет выполнено

$$\mathbb{E}[\|\mathbf{y}_1^N - \mathbf{x}^*\|_2^2] \leq \frac{R^2}{4},$$

где $\mathbf{y}_1^N$ — выход алгоритма после 1-го рестарта.

SLCM: сильно выпуклый случай

Тогда получим, что

$$\mathbb{E}[\|\mathbf{y}^N - \mathbf{x}^*\|_2^2] \leq \frac{R^2}{2}.$$

Теперь запустим тот же метод, но в качестве стартовой точки возьмём $\mathbf{y}^N$. Аналогичными рассуждениями, мы получаем, что через

$$O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2}\right\}\right) \text{ обращений к оракулу}$$

будет выполнено

$$\mathbb{E}[\|\mathbf{y}_1^N - \mathbf{x}^*\|_2^2] \leq \frac{R^2}{4},$$

где $\mathbf{y}_1^N$ — выход алгоритма после 1-го рестарта.

SLCM: сильно выпуклый случай

После $t = \log_2\left(\frac{R^2}{2\varepsilon}\right)$ рестартов мы получим такую точку $\mathbf{y}_t^N$, что $\mathbb{E}[\|\mathbf{y}_t^N - \mathbf{x}^*\|_2^2] \leq \varepsilon$. При это общее число обращений к оракулу равняется

$$\begin{aligned}\sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2^l}\right\}\right) &= \sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 2^l}{\mu^2 R^2}\right\}\right) \\ &= O\left(\max\left\{\eta t\sqrt{\frac{L}{\mu}}, \frac{\sigma^2(2^t-1)}{\mu^2 R^2}\right\}\right) \\ &= O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}\ln\frac{R^2}{\varepsilon}, \frac{\sigma^2}{\mu^2\varepsilon}\right\}\right)\end{aligned}$$

SLCM: сильно выпуклый случай

После $t = \log_2\left(\frac{R^2}{2\varepsilon}\right)$ рестартов мы получим такую точку $\mathbf{y}_t^N$, что $\mathbb{E}[\|\mathbf{y}_t^N - \mathbf{x}^*\|_2^2] \leq \varepsilon$. При это общее число обращений к оракулу равняется

$$\sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2^l}\right\}\right) = \sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 2^l}{\mu^2 R^2}\right\}\right)$$

$$= O\left(\max\left\{\eta t\sqrt{\frac{L}{\mu}}, \frac{\sigma^2(2^t-1)}{\mu^2 R^2}\right\}\right)$$

$$= O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}\ln\frac{R^2}{\varepsilon}, \frac{\sigma^2}{\mu^2\varepsilon}\right\}\right)$$

SLCM: сильно выпуклый случай

После $t = \log_2\left(\frac{R^2}{2\varepsilon}\right)$ рестартов мы получим такую точку $\mathbf{y}_t^N$, что $\mathbb{E}[\|\mathbf{y}_t^N - \mathbf{x}^*\|_2^2] \leq \varepsilon$. При это общее число обращений к оракулу равняется

$$\begin{aligned}\sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2^l}\right\}\right) &= \sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 2^l}{\mu^2 R^2}\right\}\right) \\ &= O\left(\max\left\{\eta t\sqrt{\frac{L}{\mu}}, \frac{\sigma^2(2^t-1)}{\mu^2 R^2}\right\}\right) \\ &= O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}\ln\frac{R^2}{\varepsilon}, \frac{\sigma^2}{\mu^2\varepsilon}\right\}\right)\end{aligned}$$

SLCM: сильно выпуклый случай

После $t = \log_2\left(\frac{R^2}{2\varepsilon}\right)$ рестартов мы получим такую точку $\mathbf{y}_t^N$, что $\mathbb{E}[\|\mathbf{y}_t^N - \mathbf{x}^*\|_2^2] \leq \varepsilon$. При это общее число обращений к оракулу равняется

$$\begin{aligned}
\sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2}{\mu^2 R^2/2^l}\right\}\right) &= \sum_{l=0}^{t-1} O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}, \frac{\sigma^2 2^l}{\mu^2 R^2}\right\}\right) \\
&= O\left(\max\left\{\eta t\sqrt{\frac{L}{\mu}}, \frac{\sigma^2(2^t-1)}{\mu^2 R^2}\right\}\right) \\
&= O\left(\max\left\{\eta\sqrt{\frac{L}{\mu}}\ln\frac{R^2}{\varepsilon}, \frac{\sigma^2}{\mu^2\varepsilon}\right\}\right)
\end{aligned}$$

При выполнении условий интерполяции можно делать одномерный поиск и применять правило Армихо

Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G. and Lacoste-Julien, S. **Painless stochastic gradient: Interpolation, line-search, and convergence rates.** *In Advances in Neural Information Processing Systems*, pp. 3732-3745 *(NeurIPS 2019)*

Некоторые ссылки на статьи, возникшие в ходе обсуждения на семинаре

- Когда обсуждался стохастический метод тяжёлого шарика (в конце доклада), были приведены ссылки на следующие работы (в обоих работах разбирается выпуклый случай):

  Sebbouh, O., Gower, R.M. and Defazio, A., 2020. **On the convergence of the Stochastic Heavy Ball Method.** *arXiv preprint* arXiv:2006.07867.

  Taylor, A. and Bach, F. **Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions.** *In Conference on Learning Theory*, pp. 2934-2992 *(COLT 2019)*.

  Кроме того, хочется отметить работу, изучающую свойства метода тяжёлого шарика для квадратичных функций:

  Scieur, D. and Pedregosa, F., 2020. **Universal Average-Case Optimality of Polyak Momentum.** *arXiv preprint* arXiv:2002.04664.

Некоторые ссылки на статьи, возникшие в ходе обсуждения на семинаре

- Когда обсуждался стохастический метод тяжёлого шарика (в конце доклада), были приведены ссылки на следующие работы (в обоих работах разбирается выпуклый случай):

  Sebbouh, O., Gower, R.M. and Defazio, A., 2020. **On the convergence of the Stochastic Heavy Ball Method.** *arXiv preprint* arXiv:2006.07867.

  Taylor, A. and Bach, F. **Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions.** *In Conference on Learning Theory*, pp. 2934-2992 *(COLT 2019)*.
  Кроме того, хочется отметить работу, изучающую свойства метода тяжёлого шарика для квадратичных функций:

  Scieur, D. and Pedregosa, F., 2020. **Universal Average-Case Optimality of Polyak Momentum.** *arXiv preprint* arXiv:2002.04664.

Некоторые ссылки на статьи, возникшие в ходе обсуждения на семинаре

- Кроме того, в конце семинара был вопрос про нормализацию градиентов и обсуждался близкий подход — градиентный клиппинг. Упоминались следующие недавние работы про градиентный клиппинг:

  Zhang, J., He, T., Sra, S. and Jadbabaie, A. **Why Gradient Clipping Accelerates Training: A Theoretical Justification for Adaptivity.** *In International Conference on Learning Representations (ICLR 2020).*

  Zhang, J., Karimireddy, S.P., Veit, A., Kim, S., Reddi, S.J., Kumar, S. and Sra, S., 2019. **Why ADAM beats SGD for attention models.** *arXiv preprint* arXiv:1912.03194.