

Распределенная оптимизация

Методы оптимизации

Александр Безносиков

Московский физико-технический институт

14 декабря 2023



Современные проблемы обучения

- Экспоненциальный рост размеров моделей и объемов данных.

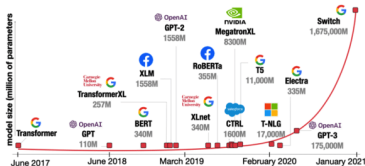
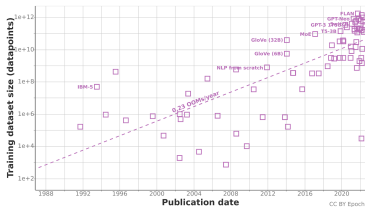


Figure: Динамика роста современных языковых моделей



Разновидности распределенного обучения

- Кластерное обучение (крупные игроки): обучаем в пределах одного большого и мощного вычислительного кластера
- Коллаборативное обучение (все игроки): объединяем вычислительные ресурсы по сети Интернет

Разновидности распределенного обучения

- Кластерное обучение (крупные игроки): обучаем в пределах одного большого и мощного вычислительного кластера
- Коллаборативное обучение (все игроки): объединяем вычислительные ресурсы по сети Интернет
- Федеративное обучение (другая парадигма): обучаемся на локальных данных пользователей, используя их вычислительные мощности

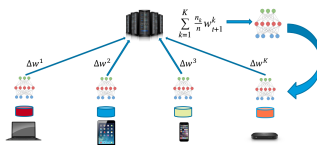


Figure: Федеративное обучение

Самая популярная распределенная постановка

- Постановка (горизонтальная, оффлайн):

$$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{M} \sum_{m=1}^M f_m(w) := \frac{1}{M} \sum_{m=1}^M \frac{1}{n_m} \sum_{i=1}^{n_m} l(g(w, x_i), y_i).$$

- w – веса модели, g – модель, l – функция потерь.
- Данные разделены между M вычислительными устройствами, на каждом устройстве m своя локальная подвыборка $\{x_i, y_i\}_{i=1}^{n_m}$ размера n_m .
- В фокусе этого доклада.

Общаемся через сервер

- Посмотрим на примере, как обычный неопределенный GD становится централизованным.

Алгоритм 1 Централизованный GD

Вход: Размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$, количество итераций K

```
1: for  $k = 0, 1, \dots, K - 1$  do
2:   Отправить  $w_k$  всем рабочим                                ▷ выполняется сервером
3:   for  $i = 1, \dots, n$  параллельно do
4:     Принять  $w_k$  от мастера                                ▷ выполняется рабочими
5:     Вычислить градиент  $\nabla f_m(w_k)$  в точке  $w_k$         ▷ выполняется рабочими
6:     Отправить  $\nabla f_m(w_k)$  мастеру                        ▷ выполняется рабочими
7:   end for
8:   Принять  $\nabla f_m(w_k)$  от всех рабочих                    ▷ выполняется сервером
9:   Вычислить  $\nabla f(w_k) = \frac{1}{M} \sum_{m=1}^M \nabla f_m(w_k)$         ▷ выполняется сервером
10:   $w_{k+1} = w_k - \gamma \nabla f(w_k)$                             ▷ выполняется сервером
11: end for
```

Выход: w^K

С чем и за что боремся?

- **Вопрос:** распределенность нужна для параллелизации, но почему не получается достигнуть полного распараллеливания?

С чем и за что боремся?

- **Вопрос:** распределенность нужна для параллелизации, но почему не получается достигнуть полного распараллеливания?
- Коммуникационные затраты являются бесполезной тратой времени.
- Проблема коммуникационного узкого места актуальна для всех постановок распределенного обучения.
- Существует много способов борьбы за эффективные коммуникации.

Несмещённая компрессия (квантизация)

Несмещённая компрессия (квантизация)

Будем называть стохастический оператор $Q(x)$ оператором несмещённой компрессии (квантизации), если для любого $x \in \mathbb{R}^d$ выполняется:

$$\mathbb{E}[Q(x)] = x, \quad \mathbb{E}[\|Q(x)\|_2^2] \leq \omega \|x\|_2^2,$$

где $\omega \geq 1$.

Случайная спарсификация (выбор случайных компонент)

$$\text{Randk}(x) = \frac{d}{k} \sum_{i \in S} [x]_i e_i,$$


- **Вопрос:** зачем нужен множитель $\frac{d}{k}$? Для несмещенности.

Несмещённая компрессия: примеры

- **Вопрос:** Чему равно ω для случайной спарсификации?

Несмещённая компрессия: примеры

- **Вопрос:** Чему равно ω для случайной спарсификации? $\frac{d}{k}$.
Каждая координата попадет в $Q(x)$ с вероятностью $\frac{k}{d}$, поэтому

$$\begin{aligned}\mathbb{E} [\|\mathcal{Q}(x)\|^2] &= \mathbb{E} \left[\sum_{i=1}^d [\mathcal{Q}(x)]_i^2 \right] \\ &= \frac{d^2}{k^2} \left[\sum_{i=1}^d \frac{k}{d} [x]_i^2 \right] \\ &= \frac{d}{k} \|x\|^2.\end{aligned}$$

Здесь $[\cdot]_i$ – i -ая координата вектора.

Трёхуровневая ℓ_2 -квантизация

$[\mathcal{Q}(x)]_i = \|x\|_2 \text{sign}(x_i) \xi_i$, $i = 1, \dots, d$, где $[\cdot]_i$ — i -я компонента вектора, и ξ_i — случайная величина, имеющая распределение Бернулли с параметром $\frac{|x_i|}{\|x\|_2}$, т. е.

$$\xi_i = \begin{cases} 1 & \text{с вероятностью } \frac{|x_i|}{\|x\|_2}, \\ 0 & \text{с вероятностью } 1 - \frac{|x_i|}{\|x\|_2}. \end{cases}$$

Таким образом, если мы хотим передать вектор $Q(x)$, то нам нужно передать вектор, состоящий из нулей и ± 1 , и вещественное число $\|x\|_2$, причём вероятность обнуления компоненты тем больше, чем компонента меньше по модулю. Можно показать, что данный оператор является несмещённой компрессией с константой $\omega = \sqrt{d}$.

- **Вопрос:** Будет ли округление несмещенным оператором?
- **Вопрос:** Какое округление кажется наиболее естественным для вычислений на компьютере?

Несмещенная компрессия: идея

- Самая простая идея, которая приходит в голову, состоит в том, чтобы использовать параллельный GD, но к градиентам, пересылаемым от рабочих на сервер, применять несмещённую компрессию.

Квантизированный GD (QGD)

Алгоритм 1 QGD

Вход: размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$, количество итераций K

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

2: Отправить w_k всем рабочим

- ▷ выполняется сервером

3: **for** $m = 1, \dots, M$ параллельно **do**

4: Принять w_k от мастера

- ▷ выполняется рабочими

5: Вычислить градиент $\nabla f_m(w_k)$ в точке w_k \triangleright выполняется рабочими

6: Независимо сгенерировать $g_{k,m} = Q(\nabla f_m(w_k))$ \triangleright выполняется

рабочими

7: Отправить $g_{k,m}$ мастеру

- ▷ выполняется рабочими

```
8:   end for
```

9: Принять $g_{k,m}$ от всех рабочих

- ▷ выполняется сервером

10: Вычислить $g_k = \frac{1}{M} \sum_{m=1}^M g_{k,m}$

- ▷ выполняется сервером

11: $w_{k+1} = w_k - \gamma g_k$

- ▷ выполняется сервером

12: end for

Выход: w^K

Несмещенная компрессия: доказательство

- Будем доказывать в случае, когда все f_m являются L -гладкими и μ -сильно выпуклыми.
- Рассмотрим одну итерацию метода:

$$\|w_{k+1} - w^*\|^2 = \|w_k - w^*\|^2 - 2\gamma \langle g_k, w_k - w^* \rangle + \|g_k\|^2.$$

Несмещенная компрессия: доказательство

- Будем доказывать в случае, когда все f_m являются L -гладкими и μ -сильно выпуклыми.
- Рассмотрим одну итерацию метода:

$$\|w_{k+1} - w^*\|^2 = \|w_k - w^*\|^2 - 2\gamma \langle g_k, w_k - w^* \rangle + \|g_k\|^2.$$

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

$$\begin{aligned}\mathbb{E} [\|w_{k+1} - w^*\|^2 \mid w_k] &= \|w_k - w^*\|^2 - 2\gamma \langle \mathbb{E}[g_k \mid w_k], w_k - w^* \rangle \\ &\quad + \gamma^2 \mathbb{E} [\|g_k\|^2 \mid w_k].\end{aligned}$$

Несмещенная компрессия: доказательство

- Работаем с $\mathbb{E}[g_k \mid w_k]$:

$$\begin{aligned}\mathbb{E}[g_k \mid w_k] &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}[g_{k,m} \mid w_k] \\ &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}[\mathbb{E}[\mathcal{Q}(\nabla f_m(w_k)) \mid \nabla f_m(w_k)] \mid w_k] \\ &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}[\nabla f_m(w_k) \mid w_k] = \frac{1}{M} \sum_{m=1}^M \nabla f_m(w_k) = \nabla f(w_k).\end{aligned}$$

Несмещенная компрессия: доказательство

- Работаем с $\mathbb{E}[g_k \mid w_k]$:

$$\begin{aligned}\mathbb{E}[g_k \mid w_k] &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}[g_{k,m} \mid w_k] \\ &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}[\mathbb{E}[\mathcal{Q}(\nabla f_m(w_k)) \mid \nabla f_m(w_k)] \mid w_k] \\ &= \frac{1}{M} \sum_{m=1}^M \mathbb{E}[\nabla f_m(w_k) \mid w_k] = \frac{1}{M} \sum_{m=1}^M \nabla f_m(w_k) = \nabla f(w_k).\end{aligned}$$

- Работаем с $\mathbb{E}[\|g_k\|^2 \mid w_k]$:

$$\mathbb{E}[\|g_k\|^2 \mid w_k] = \mathbb{E}\left[\left\|\frac{1}{M} \sum_{m=1}^M g_{k,m}\right\|^2 \mid x_k\right] = \frac{1}{M^2} \mathbb{E}\left[\left\|\sum_{m=1}^M g_{k,m}\right\|^2 \mid w^k\right]$$

Несмещенная компрессия: доказательство

- Продолжаем и применяем первое свойство (несмещенность) в определении компрессии:

$$\begin{aligned}\mathbb{E} [\|g_k\|^2 \mid w_k] &= \frac{1}{M^2} \mathbb{E} \left[\left\| \sum_{m=1}^M g_{k,m} \right\|^2 \mid w_k \right] \\ &= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\|g_{k,m}\|^2 \mid w_k] + \frac{2}{M^2} \sum_{m \neq l} \mathbb{E} [\langle g_{k,m}, g_{k,l} \rangle \mid w_k] \\ &= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\|g_{k,m}\|^2 \mid w_k] \\ &\quad + \frac{1}{M^2} \sum_{m \neq l} \mathbb{E} [\langle \mathbb{E}[g_{k,m} \mid \nabla f_m(w_k)], \mathbb{E}[g_{k,l} \mid \nabla f_l(x_k)] \rangle \mid w_k].\end{aligned}$$

Несмещенная компрессия: доказательство

- Продолжаем и применяем первое свойство (несмещенность) в определении компрессии:

$$\begin{aligned}
 \mathbb{E} [\|g_k\|^2 \mid w_k] &= \frac{1}{M^2} \mathbb{E} \left[\left\| \sum_{m=1}^M g_{k,m} \right\|^2 \mid w_k \right] \\
 &= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\|g_{k,m}\|^2 \mid w_k] + \frac{2}{M^2} \sum_{m \neq l} \mathbb{E} [\langle g_{k,m}, g_{k,l} \rangle \mid w_k] \\
 &= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\|g_{k,m}\|^2 \mid w_k] \\
 &\quad + \frac{1}{M^2} \sum_{m \neq l} \mathbb{E} [\langle \mathbb{E} [g_{k,m} \mid \nabla f_m(w_k)], \mathbb{E} [g_{k,l} \mid \nabla f_l(x_k)] \rangle \mid w_k].
 \end{aligned}$$

Несмещенная компрессия: доказательство

- Продолжаем и применяем второе свойство в определении компрессии: $\frac{1}{M}$

$$\begin{aligned} \mathbb{E} [\|g_k\|^2 \mid w_k] &= \frac{1}{M^2} \sum_{m=1}^M \mathbb{E} [\|\mathcal{Q}(\nabla f_m(w_k))\|^2 \mid w_k] \\ &\quad + \frac{1}{M^2} \sum_{m \neq l} \langle \nabla f_m(w_k), \nabla f_l(w_k) \rangle \\ &\leq \frac{\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w_k)\|^2 + \|\nabla f(w_k)\|^2 \\ &\leq \frac{2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w_k) - \nabla f_m(w^*)\|^2 \\ &\quad + \frac{2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2 + \|\nabla f(w_k) - \nabla f(w^*)\|^2. \end{aligned}$$

Несмещенная компрессия: доказательство

- Продолжаем и применяем второе свойство в определении компрессии:

$$\begin{aligned}
 \mathbb{E} [\|g_k\|^2 \mid w_k] &\leq \frac{4\omega L}{M^2} \sum_{m=1}^M (f_m(w_k) - f_m(w^*) - \langle \nabla f_m(w^*), w_k - w^* \rangle) \\
 &\quad + \frac{2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2 + 2L(f(w_k) - f(w^*)) \\
 &= \frac{4\omega L}{M} (f(w_k) - f(w^*)) \\
 &\quad + \frac{2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2 + 2L(f(w_k) - f(w^*)).
 \end{aligned}$$

Несмещенная компрессия: доказательство

- Все, что получили:

$$\begin{aligned}\mathbb{E} [\|w_{k+1} - w^*\|^2 \mid w_k] &= \|w_k - w^*\|^2 - 2\gamma \langle \mathbb{E}[g_k \mid w_k], w_k - w^* \rangle \\ &\quad + \gamma^2 \mathbb{E} [\|g_k\|^2 \mid w_k].\end{aligned}$$

$$\mathbb{E}[g_k \mid w_k] = \nabla f(w_k).$$

$$\begin{aligned} \mathbb{E} [\|g_k\|^2 \mid w_k] &\leq \frac{4\omega L}{M} (f(w_k) - f(w^*)) \\ &\quad + \frac{2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2 + 2L(f(w_k) - f(w^*)). \end{aligned}$$

Несмещенная компрессия: доказательство

- Объединяем:

$$\begin{aligned}\mathbb{E} [\|w_{k+1} - w^*\|^2 \mid w_k] &\leq \|w_k - w^*\|^2 - 2\gamma \langle \nabla f(w_k), w_k - w^* \rangle \\ &\quad + 2\gamma^2 L \left(\frac{2\omega}{M} + 1 \right) (f(w_k) - f(w^*)) \\ &\quad + \frac{2\gamma^2 \omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2.\end{aligned}$$

- Пользуемся сильной выпуклостью:

$$\begin{aligned}\mathbb{E} [\|w_{k+1} - w^*\|^2 \mid w_k] &\leq \|w_k - w^*\|^2 \\ &\quad - 2\gamma \left(\frac{\mu}{2} \|w_k - w^*\|^2 + f(w_k) - f(w^*) \right) \\ &\quad + 2\gamma^2 L \left(\frac{2\omega}{M} + 1 \right) (f(w_k) - f(w^*)) \\ &\quad + \frac{2\gamma^2 \omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2.\end{aligned}$$

Несмещенная компрессия: доказательство

- Если взять полное математическое ожидание

$$\begin{aligned}\mathbb{E} [\|w_{k+1} - w^*\|^2] &\leq (1 - \gamma\mu) \mathbb{E} [\|w_k - w^*\|^2] \\ &\quad - 2\gamma \left[1 - \gamma L \left(\frac{2\omega}{M} + 1 \right) \right] \mathbb{E} [(f(x_k) - f(x^*))] \\ &\quad + \frac{2\gamma^2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2.\end{aligned}$$

- Если $\gamma \leq L^{-1} \left(\frac{2\omega}{M} + 1 \right)^{-1}$, то

$$\begin{aligned}\mathbb{E} [\|w_{k+1} - w^*\|^2] &\leq (1 - \gamma\mu) \mathbb{E} [\|w_k - w^*\|^2] \\ &\quad + \frac{2\gamma^2\omega}{M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2.\end{aligned}$$

QGD: СХОДИМОСТЬ

Теорема (QGD)

Пусть все локальные функции f_m являются μ -сильно выпуклыми и имеют L -Липшицев градиент, тогда если $\eta \leq L^{-1} \left(\frac{2\omega}{M} + 1 \right)^{-1}$, то

$$\mathbb{E} [\|w_K - w^*\|^2] = \mathcal{O} \left((1 - \gamma\mu)^K \|w_0 - w^*\|^2 + \frac{1}{K} \cdot \frac{2\omega}{\mu M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2 \right)$$

При получении данного результата так же использовался подбор γ из работы:



Stich S. Unified Optimal Analysis of the (Stochastic) Gradient Method

QGD: СХОДИМОСТЬ

Теорема (QGD)

Пусть все локальные функции f_m являются μ -сильно выпуклыми и имеют L -Липшицев градиент, тогда если $\eta \leq L^{-1} \left(\frac{2\omega}{M} + 1 \right)^{-1}$, то

$$\mathbb{E} [\|w_K - w^*\|^2] = \mathcal{O} \left((1 - \gamma\mu)^K \|w_0 - w^*\|^2 + \frac{1}{K} \cdot \frac{2\omega}{\mu M^2} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2 \right)$$

При получении данного результата так же использовался подбор γ из работы:



Stich S. Unified Optimal Analysis of the (Stochastic) Gradient Method

- **Вопрос:** какие проблемы есть в этой оценке? (вспомните оценку сходимости GD)

QGD: СХОДИМОСТЬ

- Поведение на практике:

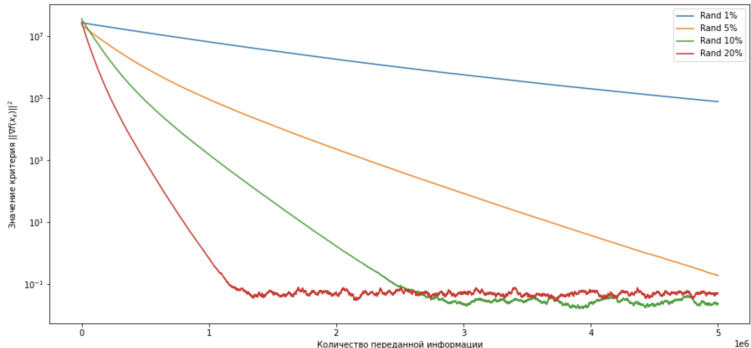


Figure: Поведение методов с несмещенным оператором сжатия и постоянным шагом

- В теории шаг подбирался хитро, при постоянном шаге теория предугадывает ровно этот же эффект – ранний «выход на плато».

Несмещенная компрессия: решаем проблему с плато

- Метод DIANA – QGD с памятью:

Алгоритм 1 DIANA (скетч)

- 1: Каждое устройство m обладает вектором "памяти" $h_0^m = 0$
- 2: Сервер хранит $h_0 = \frac{1}{M} \sum_{m=1}^M h_0^m = 0$
- 3: Досылаем на сервер сжатую версию разницы $\mathcal{Q}(\nabla f_m(w^k) - h_k^m)$
- 4: Обновляем память $h_{k+1}^m = h_k^m + \alpha \mathcal{Q}(\nabla f_m(w^k) - h_k^m)$
- 5: Сервер вычисляет $g_k = h_k + \frac{1}{M} \sum_{m=1}^M \mathcal{Q}(\nabla f_m(w^k) - h_k^m)$
- 6: Для апдейта $w_{k+1} = w_k - \gamma g_k$
- 7: Сервер обновляет $h_{k+1} = h_k + \alpha \frac{1}{M} \sum_{m=1}^M \mathcal{Q}(\nabla f_m(w^k) - h_k^m)$



Mishchenko K. et al. Distributed Learning with Compressed Gradient Differences

QGD и DIANA: сходимость

- **Вопрос:** Какие есть еще вопросы к сходимости/оценкам сходимости?

QGD и DIANA: сходимость

- **Вопрос:** Какие есть еще вопросы к сходимости/оценкам сходимости? Лучше ли вообще сходится?
- Лучшая оценка на число коммуникаций для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[1 + \frac{\omega}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Оценка на число коммуникаций для GD:

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

- С точки зрения числа коммуникаций методы с компрессией уступают базовым методам – это ожидаемо (плата за сжатие).
НО!

QGD и DIANA: сходимость

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$.

QGD и DIANA: сходимость

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[\frac{1}{\beta} + \frac{1}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Оценка на число информации для GD:

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

QGD и DIANA: сходимость

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[\frac{1}{\beta} + \frac{1}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Оценка на число информации для GD:

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

- Несмещенный компрессор доказуемо улучшает число передаваемой информации, фактор улучшения: $\left[\frac{1}{\beta} + \frac{1}{M} \right]$.

QGD и DIANA: сходимость

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[\frac{1}{\beta} + \frac{1}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Оценка на число информации для GD:

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

- Несмещенный компрессор доказуемо улучшает число передаваемой информации, фактор улучшения: $\left[\frac{1}{\beta} + \frac{1}{M} \right]$.
- Смещенный компрессор не улучшает число передаваемой информации в общем случае.

Сервера может и не быть

- Часто на практике "централизованные коммуникации через сервер" реализованы без "сервера".
- Архитектура с AllGather/AllReduce процедурой: задан некоторый граф связей/коммуникаций, обмен сообщениями происходит согласно этому графу, в том числе можно организовать усреднение.



Chan, E. et al. Collective communication: theory, practice, and experience

Централизованные коммуникации без сервера

Operation	Before				After			
Broadcast	Node 0 x	Node 1	Node 2	Node 3	Node 0 x	Node 1 x	Node 2 x	Node 3 x
Reduce(-to-one)	Node 0 $x^{(0)}$	Node 1 $x^{(1)}$	Node 2 $x^{(2)}$	Node 3 $x^{(3)}$	Node 0 $\sum_j x^{(j)}$	Node 1	Node 2	Node 3
Scatter	Node 0 x_0 x_1 x_2 x_3	Node 1	Node 2	Node 3	Node 0 x_0	Node 1 x_1	Node 2 x_2	Node 3 x_3
Gather	Node 0 x_0	Node 1 x_1	Node 2 x_2	Node 3 x_3	Node 0 x_0 x_1 x_2 x_3	Node 1	Node 2	Node 3
Allgather	Node 0 x_0	Node 1 x_1	Node 2 x_2	Node 3 x_3	Node 0 x_0 x_1 x_2 x_3	Node 1 x_0 x_1 x_2 x_3	Node 2 x_0 x_1 x_2 x_3	Node 3 x_0 x_1 x_2 x_3
Reduce-scatter	Node 0 $x_0^{(0)}$ $x_1^{(0)}$ $x_2^{(0)}$ $x_3^{(0)}$	Node 1 $x_0^{(1)}$ $x_1^{(1)}$ $x_2^{(1)}$ $x_3^{(1)}$	Node 2 $x_0^{(2)}$ $x_1^{(2)}$ $x_2^{(2)}$ $x_3^{(2)}$	Node 3 $x_0^{(3)}$ $x_1^{(3)}$ $x_2^{(3)}$ $x_3^{(3)}$	Node 0 $\sum_j x_0^{(j)}$	Node 1 $\sum_j x_1^{(j)}$	Node 2 $\sum_j x_2^{(j)}$	Node 3 $\sum_j x_3^{(j)}$
Allreduce	Node 0 $x^{(0)}$	Node 1 $x^{(1)}$	Node 2 $x^{(2)}$	Node 3 $x^{(3)}$	Node 0 $\sum_j x^{(j)}$	Node 1 $\sum_j x^{(j)}$	Node 2 $\sum_j x^{(j)}$	Node 3 $\sum_j x^{(j)}$

Ring AllReduce

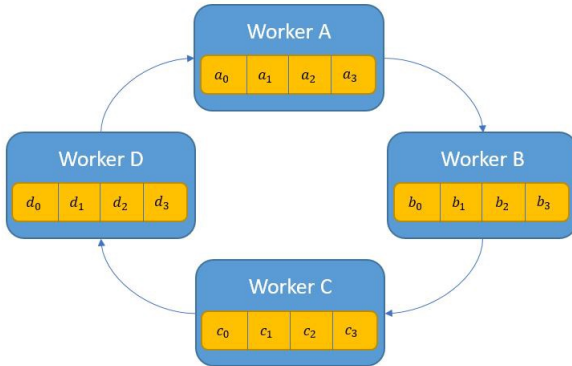


Figure: Картинка отсюда

Ring AllReduce: первый шаг суммирования

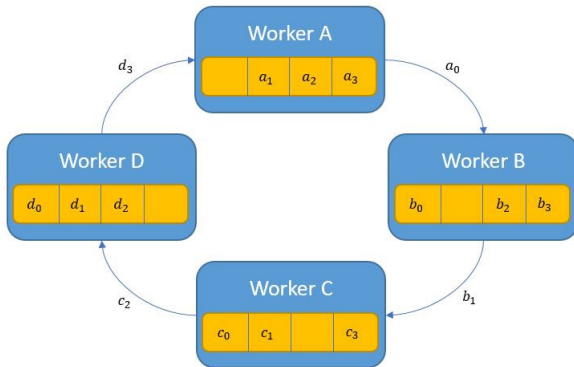


Figure: Картинка отсюда

Ring AllReduce: второй шаг суммирования

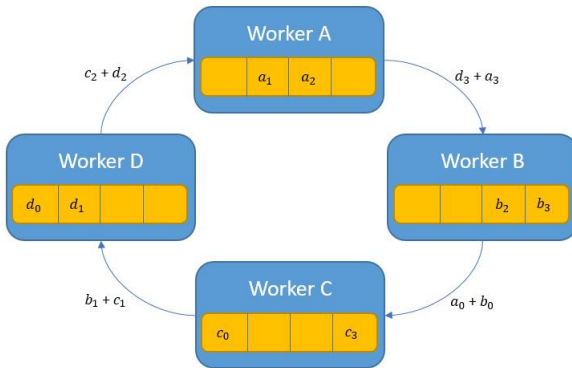


Figure: Картинка отсюда

Ring AllReduce: первый шаг распространения

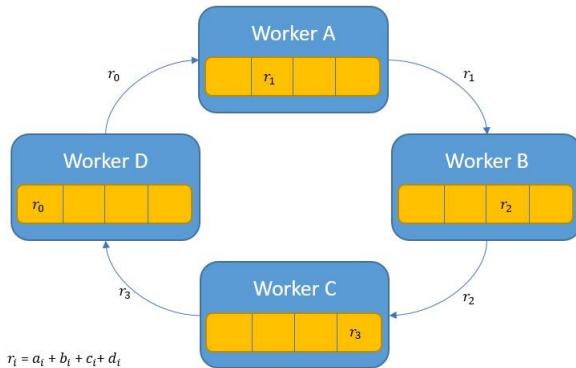


Figure: Картинка отсюда

Ring AllReduce: итог

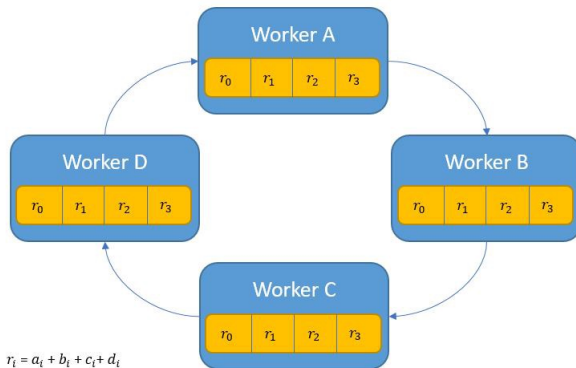


Figure: Картинка отсюда

Квантизированный GD (QGD) с AllReduce

Алгоритм 1 QGD

Вход: размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: **for** $m = 1, \dots, M$ параллельно **do**
- 3: Вычислить градиент $\nabla f_m(w_k)$ в точке w_k
- 4: Независимо сгенерировать $g_{k,m} = \mathcal{Q}(\nabla f_m(w_k))$
- 5: Запустить AllReduce $\{g_{k,m}\}$ и получить $g_k = \frac{1}{M} \sum_{m=1}^M g_{k,m}$
- 6: $w_{k+1} = w_k - \gamma g_k$
- 7: **end for**
- 8: **end for**

Выход: w^K

Квантизированный GD (QGD) с AllReduce

Алгоритм 1 QGD

Вход: размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: **for** $m = 1, \dots, M$ параллельно **do**
- 3: Вычислить градиент $\nabla f_m(w_k)$ в точке w_k
- 4: Независимо сгенерировать $g_{k,m} = \mathcal{Q}(\nabla f_m(w_k))$
- 5: Запустить AllReduce $\{g_{k,m}\}$ и получить $g_k = \frac{1}{M} \sum_{m=1}^M g_{k,m}$
- 6: $w_{k+1} = w_k - \gamma g_k$
- 7: **end for**
- 8: **end for**

Выход: w^K

- **Вопрос:** Какие проблемы могут появиться у (например) Randk?

Квантизированный GD (QGD) с AllReduce

Алгоритм 1 QGD

Вход: размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
- 2: **for** $m = 1, \dots, M$ параллельно **do**
- 3: Вычислить градиент $\nabla f_m(w_k)$ в точке w_k
- 4: Независимо сгенерировать $g_{k,m} = \mathcal{Q}(\nabla f_m(w_k))$
- 5: Запустить AllReduce $\{g_{k,m}\}$ и получить $g_k = \frac{1}{M} \sum_{m=1}^M g_{k,m}$
- 6: $w_{k+1} = w_k - \gamma g_k$
- 7: **end for**
- 8: **end for**

Выход: w^K

- **Вопрос:** Какие проблемы могут появиться у (например) Randk? Одинаковые ненулевые координаты у разных устройств могут вызывать коллизии.

PermK: делаем зависимую рандомизацию

Перестановочный компрессор (зависимый RandK)

Предположим, что $d \geq n$ и $d = qn$, где $q \geq 1$ – целое число. Пусть $\pi = (\pi_1, \dots, \pi_d)$ – случайная перестановка $\{1, \dots, d\}$. Тогда для каждого $i \in \{1, 2, \dots, n\}$ имеем следующий оператор сжатия

$$\mathcal{Q}_i(u) = n \cdot \sum_{j=q(i-1)+1}^{qi} u_{\pi_j} e_{\pi_j}.$$



Szlendak, R. et al. Permutation Compressors for Provably Faster Distributed Nonconvex Optimization

PermK: делаем зависимую рандомизацию

Перестановочный компрессор (зависимый RandK)

Предположим, что $d \geq n$ и $d = qn$, где $q \geq 1$ – целое число. Пусть $\pi = (\pi_1, \dots, \pi_d)$ – случайная перестановка $\{1, \dots, d\}$. Тогда для каждого $i \in \{1, 2, \dots, n\}$ имеем следующий оператор сжатия

$$\mathcal{Q}_i(u) = n \cdot \sum_{j=q(i-1)+1}^{qi} u_{\pi_j} e_{\pi_j}.$$



Szlendak, R. et al. Permutation Compressors for Provably Faster Distributed Nonconvex Optimization

- Дружественна к централизованным коммуникациям без сервера.
- В гомогенном случае имеют физику дешевой пересылки полного градиента.

Смещенная компрессия

- Случайный выбор – это хорошо, но и тут есть потенциал для улучшения.

Смещённая компрессия

Будем называть (стохастический) оператор $C(x)$ оператором смещённой компрессии, если для любого $x \in \mathbb{R}^d$ выполняется:

$$\mathbb{E}[\|C(x) - x\|_2^2] \leq \left(1 - \frac{1}{\delta}\right) \|x\|_2^2,$$

где $\delta \geq 0$.

Смещенная компрессия: примеры

- Различные примеры компрессоров (спарсификаторы, округления и тд):



Beznosikov A. et al. On Biased Compression for Distributed Learning

- Практичный смещенный компрессор на основе итеративного SVD разложения:



Vogels T. et al. PowerSGD: Practical Low-Rank Gradient Compression for Distributed Optimization

Смещенная компрессия: идея и доказательство в случае 1 ноды

- Использовать тот же подход, что и в несмещенном случае (QGD).

1 ноды

- Использовать тот же подход, что и в несмещенном случае (QGD).
- Докажем в случае одной ноды:

$$w_{k+1} = w_k - \gamma C(\nabla f(w_k)).$$

Пусть f имеет L -Липшицев градиент и является μ -сильно выпуклой.

Смещенная компрессия: идея и доказательство в случае 1 ноды

- Использовать тот же подход, что и в несмещенном случае (QGD).
- Докажем в случае одной ноды:

$$w_{k+1} = w_k - \gamma C(\nabla f(w_k)).$$

Пусть f имеет L -Липшицев градиент и является μ -сильно выпуклой.

- Начнем с того, что воспользуемся Липшицевостью градиента:

$$\begin{aligned} f(w_{k+1}) &= f(w_k - \gamma C(\nabla f(w_k))) \\ &\leq f(w_k) + \langle \nabla f(w_k), -\gamma C(\nabla f(w_k)) \rangle + \frac{L}{2} \| -\gamma C(\nabla f(w_k)) \|^2 \\ &= f(w_k) - \gamma \langle C(\nabla f(w_k)), \nabla f(w_k) \rangle + \frac{\gamma^2 L}{2} \| C(\nabla f(w_k)) \|^2. \end{aligned}$$

Смещенная компрессия: доказательство в случае 1 ноды

- Определение компрессора:

$$\begin{aligned} & \|\nabla f(w_k)\|^2 - 2\mathbb{E}_C [\langle C(\nabla f(w_k)), \nabla f(w_k) \rangle] + \mathbb{E}_C [\|C(\nabla f(w_k))\|^2] \\ &= \mathbb{E}_C [\|C(\nabla f(w_k)) - \nabla f(w_k)\|^2] \leq \left(1 - \frac{1}{\delta}\right) \|\nabla f(w_k)\|^2. \end{aligned}$$

Смещенная компрессия: доказательство в случае 1 ноды

- Определение компрессора:

$$\begin{aligned} & \|\nabla f(w_k)\|^2 - 2\mathbb{E}_C [\langle C(\nabla f(w_k)), \nabla f(w_k) \rangle] + \mathbb{E}_C [\|C(\nabla f(w_k))\|^2] \\ &= \mathbb{E}_C [\|C(\nabla f(w_k)) - \nabla f(w_k)\|^2] \leq \left(1 - \frac{1}{\delta}\right) \|\nabla f(w_k)\|^2. \end{aligned}$$

- Откуда:

$$-\gamma \mathbb{E}_C [\langle C(\nabla f(w_k)), \nabla f(w_k) \rangle] + \frac{\gamma}{2} \mathbb{E}_C [\|C(\nabla f(w_k))\|^2] \leq -\frac{\gamma}{2\delta} \|\nabla f(w_k)\|^2.$$

Смещенная компрессия: доказательство в случае 1 ноды

- С двух предыдущих слайдов:

$$f(w_{k+1}) - \leq f(w_k) - \gamma \langle C(\nabla f(w_k)), \nabla f(w_k) \rangle + \frac{\gamma^2 L}{2} \|C(\nabla f(w_k))\|^2.$$

$$-\gamma \mathbb{E}_C [\langle C(\nabla f(w_k)), \nabla f(w_k) \rangle] + \frac{\gamma}{2} \mathbb{E}_C [\|C(\nabla f(w_k))\|^2] \leq -\frac{\gamma}{2\delta} \|\nabla f(w_k)\|^2.$$

Смещенная компрессия: доказательство в случае 1 ноды

- С двух предыдущих слайдов:

$$f(w_{k+1}) - \leq f(w_k) - \gamma \langle C(\nabla f(w_k)), \nabla f(w_k) \rangle + \frac{\gamma^2 L}{2} \|C(\nabla f(w_k))\|^2.$$

$$-\gamma \mathbb{E}_C [\langle C(\nabla f(w_k)), \nabla f(w_k) \rangle] + \frac{\gamma}{2} \mathbb{E}_C [\|C(\nabla f(w_k))\|^2] \leq -\frac{\gamma}{2\delta} \|\nabla f(w_k)\|^2.$$

- Сложим, вычтем из обеих частей $f(w^*)$ и возьмем полное мат. ожидание:

$$\begin{aligned} \mathbb{E} [f(w_{k+1}) - f(w^*)] &\leq \mathbb{E} [f(w_k) - f(w^*)] - \frac{\gamma}{2} (1 - \gamma L) \mathbb{E} [\|C(\nabla f(w_k))\|^2] \\ &\quad - \frac{\gamma}{2\delta} \mathbb{E} [\|\nabla f(w_k)\|^2]. \end{aligned}$$

Смещенная компрессия: доказательство в случае 1 ноды

- С двух предыдущих слайдов:

$$f(w_{k+1}) - \leq f(w_k) - \gamma \langle C(\nabla f(w_k)), \nabla f(w_k) \rangle + \frac{\gamma^2 L}{2} \|C(\nabla f(w_k))\|^2.$$

$$-\gamma \mathbb{E}_C [\langle C(\nabla f(w_k)), \nabla f(w_k) \rangle] + \frac{\gamma}{2} \mathbb{E}_C [\|C(\nabla f(w_k))\|^2] \leq -\frac{\gamma}{2\delta} \|\nabla f(w_k)\|^2.$$

- Сложим, вычтем из обеих частей $f(w^*)$ и возьмем полное мат. ожидание:

$$\begin{aligned} \mathbb{E} [f(\mathbf{w}_{k+1}) - f(\mathbf{w}^*)] &\leq \mathbb{E} [f(\mathbf{w}_k) - f(\mathbf{w}^*)] - \frac{\gamma}{2} (1 - \gamma L) \mathbb{E} [\|C(\nabla f(\mathbf{w}_k))\|^2] \\ &\quad - \frac{\gamma}{2\delta} \mathbb{E} [\|\nabla f(\mathbf{w}_k)\|^2]. \end{aligned}$$

- Возьмем $\gamma \leq \frac{1}{l}$:

$$\mathbb{E}[f(w_{k+1}) - f(w^*)] \leq \mathbb{E}[f(w_k) - f(w^*)] - \frac{\gamma}{2\delta} \mathbb{E}[\|\nabla f(w_k)\|^2].$$

Смещенная компрессия: доказательство в случае 1 ноды

- С предыдущего слайда:

$$\mathbb{E}[f(w_{k+1}) - f(w^*)] \leq \mathbb{E}[f(w_k) - f(w^*)] - \frac{\gamma}{2\delta} \mathbb{E}[\|\nabla f(w_k)\|^2].$$

Смещенная компрессия: доказательство в случае 1 ноды

- С предыдущего слайда:

$$\mathbb{E}[f(w_{k+1}) - f(w^*)] \leq \mathbb{E}[f(w_k) - f(w^*)] - \frac{\gamma}{2\delta} \mathbb{E}[\|\nabla f(w_k)\|^2].$$

- Сильная выпуклость (или даже более слабое условие PL):

$$2\mu(f(w_k) - f(w^*)) \leq \|\nabla f(w_k)\|^2.$$

Смещенная компрессия: доказательство в случае 1 ноды

- С предыдущего слайда:

$$\mathbb{E}[f(w_{k+1}) - f(w^*)] \leq \mathbb{E}[f(w_k) - f(w^*)] - \frac{\gamma}{2\delta} \mathbb{E}[\|\nabla f(w_k)\|^2].$$

- Сильная выпуклость (или даже более слабое условие PL):

$$2\mu(f(w_k) - f(w^*)) \leq \|\nabla f(w_k)\|^2.$$

- Соединим два предыдущих:

$$\mathbb{E}[f(w_{k+1}) - f(w^*)] \leq \left(1 - \frac{\gamma\mu}{\delta}\right) \mathbb{E}[f(w_k) - f(w^*)].$$

Смещенная компрессия: теорема в случае 1 ноды

Теорема (сходимость QGD со смещенной компрессией в случае 1 ноды)

Пусть f μ -сильно выпуклая (или PL) и имеет L -Липшицев градиент, тогда QGD для одной ноды с шагом $\gamma \leq 1/L$ и со смещенным компрессором с параметром δ сходится и выполнено:

$$f(w_K) - f(w^*) \leq \left(1 - \frac{\gamma\mu}{\delta}\right)^K (f(w_0) - f(w^*)).$$



Beznosikov A. et al. On Biased Compression for Distributed Learning

Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|^2$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|^2$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- Вопрос: где у нее оптимум?

Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|^2$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- Вопрос: где у нее оптимум? $(0, 0, 0)$.

Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|^2$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $w_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(w_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(w_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(w_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать *Top1* компрессию?

Смещенная компрессия: не так все просто

- Рассмотрим следующую распределенную задачу с $M = 3$, $d = 3$ и локальными функциями:

$$f_1(w) = \langle a, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_2(w) = \langle b, w \rangle^2 + \frac{1}{4} \|w\|^2, \quad f_3(w) = \langle c, w \rangle^2 + \frac{1}{4} \|w\|^2$$

где $a = (-3, 2, 2)$, $b = (2, -3, 2)$ и $c = (2, 2, -3)$.

- **Вопрос:** где у нее оптимум? $(0, 0, 0)$.
- Пусть стартовая точка $w_0 = (t, t, t)$ для какого-то $t > 0$. Тогда локальные градиенты:

$$\nabla f_1(w_0) = \frac{t}{2}(-11, 9, 9), \quad \nabla f_2(w_0) = \frac{t}{2}(9, -11, 9), \quad \nabla f_3(w_0) = \frac{t}{2}(9, 9, -11).$$

- **Вопрос:** как будет выглядеть шаг QGD (градиентного спуска с сжатиями), если мы будем использовать *Top1* компрессию?

$$w_1 = (t, t, t) + \eta \cdot \frac{11}{6}(t, t, t) = \left(1 + \frac{11\eta}{6}\right) w_0.$$

- Мы удаляемся от решения геометрически для любого $\eta > 0$.

Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(w_0) - C(0 + \gamma \nabla f_m(w_0)).$$

Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(w_0) - C(0 + \gamma \nabla f_m(w_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(w_1))$$

Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(w_0) - C(0 + \gamma \nabla f_m(w_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(w_1))$$

- На произвольной итерации это записывается так:

Посылка: $C(e_{k,m} + \gamma \nabla f_m(w_k))$,

$$e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(w_k) - C(e_{k,m} + \gamma \nabla f_m(w_k))$$

Смещенная компрессия: компенсация ошибки

- Попробуем запоминать то, что не передали в процессе общения:

$$e_{1,m} = 0 + \gamma \nabla f_m(w_0) - C(0 + \gamma \nabla f_m(w_0)).$$

- И добавлять это в будущие посылки:

$$C(e_{1,m} + \gamma \nabla f_m(w_1))$$

- На произвольной итерации это записывается так:

$$\text{Посылка: } C(e_{k,m} + \gamma \nabla f_m(w_k)),$$

$$e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(w_k) - C(e_{k,m} + \gamma \nabla f_m(w_k))$$

- Это техника называется компенсация ошибка (error feedback).



Stich S. et al. Sparsified SGD with memory

QGD с error feedback

Алгоритм 1 QGD с error feedback

Вход: Размер шага $\gamma > 0$, стартовая точка $w_0 \in \mathbb{R}^d$, стартовые ошибки $e_{0,m} = 0$ для всех m от 1 до M , количество итераций K

```
1: for  $k = 0, 1, \dots, K - 1$  do
2:     Отправить  $x_k$  всем рабочим                                ▷ выполняется сервером
3:     for  $m = 1, \dots, M$  параллельно do
4:         Принять  $w_k$  от мастера                                ▷ выполняется рабочими
5:         Вычислить градиент  $\nabla f(w_k)$  в точке  $w_k$            ▷ выполняется рабочими
6:         Сгенерировать  $g_{k,m} = C(e_{k,m} + \gamma \nabla f(w_k))$     ▷ выполняется рабочими
7:         Вычислить  $e_{k+1,m} = e_{k,m} + \gamma \nabla f_m(w_k) - g_{k,m}$  ▷ выполняется
            рабочими
8:         Отправить  $g_{k,m}$  мастеру                                ▷ выполняется рабочими
9:     end for
10:    Принять  $g_k$  от всех рабочих                                ▷ выполняется сервером
11:    Вычислить  $g_k = \frac{1}{M} \sum_{m=1}^M g_{k,m}$                     ▷ выполняется сервером
12:     $w_{k+1} = w_k - g_k$                                        ▷ выполняется сервером
13: end for
```

QGD с error feedback: сходимость

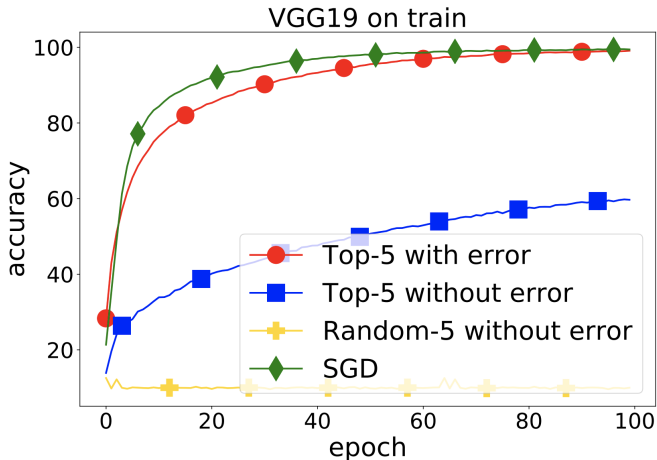


Figure: Точность в ходе обучения VGG19 на CIFAR10 с использованием разных

QGD с error feedback: сходимость

Теорема GD с error feedback

Пусть все локальные функции f_m являются μ -сильно выпуклыми и имеют L -Липшицев градиент, тогда если $\eta \leq \frac{1}{28\delta L}$, то

$$\mathbb{E}[f(\tilde{w}_K) - f(x^*)] \leq \mathcal{O}\left(\delta L \|w_0 - w^*\|^2 \exp\left(-\frac{\gamma\mu K}{2}\right) + \frac{\delta}{\mu K} \cdot \frac{1}{M} \sum_{m=1}^M \|\nabla f_m(w^*)\|^2\right).$$



Stich S. and Karimireddy S. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication



Beznosikov A. et al. On Biased Compression for Distributed Learning

- Та же самая проблема, что и у QGD – второй член в оценке

Смещенная компрессия: решение вопроса с плато

- Идея похожа на DIANA: память + сжатие разности

Алгоритм 1 EF21 (скетч)

- 1: Каждое устройство m обладает вектором "памяти" $g_0^m = 0$
- 2: Сервер хранит $h_0 = \frac{1}{M} \sum_{m=1}^M h_0^m = 0$
- 3: Досылаем на сервер сжатую версию разницы $C(\nabla f_m(w^k) - h_k^m)$
- 4: Обновляем память $h_{k+1}^m = h_k^m + C(\nabla f_m(w^k) - h_k^m)$
- 5: Сервер вычисляет $h_{k+1} = h_k + \frac{1}{M} \sum_{m=1}^M C(\nabla f_m(w^k) - h_k^m)$
- 6: Для апдейта $w_{k+1} = w_k - \gamma h_{k+1}$



Richtarik P. et al. EF21: A New, Simpler, Theoretically Better, and Practically Faster Error Feedback

Несмещенная против смещенной

- Лучшая оценка на число коммуникаций для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[1 + \frac{\omega}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Лучшая оценка на число коммуникаций для неускоренного метода со смещенной компрессией (EF-21):

$$\mathcal{O} \left([1 + \delta] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Уже обсуждалось, что эти оценки хуже, чем для базового GD.

Несмещенная против смещенной

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.

Несмещенная против смещенной

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[\frac{1}{\beta} + \frac{1}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Как уже обсуждалось, несмещенный компрессор доказуемо улучшает число передаваемой информации.

Несмещенная против смещенной

- Компрессоры сжимают информацию в β раз и типично, что $\beta \geq \omega$ и $\beta \geq \delta$.
- Лучшая оценка на число информации для неускоренного метода с несмещенной компрессией (DIANA):

$$\mathcal{O} \left(\left[\frac{1}{\beta} + \frac{1}{M} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Как уже обсуждалось, несмещенный компрессор доказуемо улучшает число передаваемой информации.
- Смещенный компрессор имеет оценку:

$$\mathcal{O} \left(\left[\frac{1}{\beta} + \frac{\delta}{\beta} \right] \frac{L}{\mu} \log \frac{1}{\varepsilon} \right).$$

- Смещенный компрессор не улучшает число передаваемой информации в общем случае. **И это открытый вопрос: как увидеть теоретическое превосходство смещенных**

Идея – больше локальных вычислений

- В базовом подходе коммуникации происходят каждую итерацию.
- Если считать (стохастические) градиенты значительно дешевле, почему бы не считать несколько раз между коммуникациями.

Локальный градиентный спуск

Идея метода:

- Делать локальные шаги:

$$x_m^{k+1} = x_m^k - \gamma \nabla f_m(x_m^k, \xi_m^k).$$

- Каждую t -ую итерацию пересылать текущий x_m^k на сервер. Сервер усредняет $x^k = \frac{1}{M} \sum_{m=1}^M x_m^k$, и пересылает x^k устройствам. Устройства обновляют: $x_m^k = x^k$.
- Централизованный SGD это Локальный SGD с $K = 1$.

Сходимость

- Типичная сходимость такого типа методов:

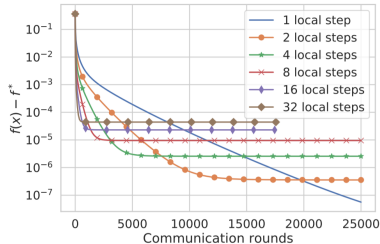


Figure: Сходимость Локального метода на практике для логистической регрессии.

- Быстрее с точки зрения коммуникаций, хуже качество придельной точности.



Khaled A. et al. Tighter Theory for Local SGD on Identical and Heterogeneous Data

Физика проблемы

- Близко к решению:

Решаем проблему

- **Вопрос:** проблема локального метода – сходимость к окрестности. Как ее можно решить?

Решаем проблему

- **Вопрос:** проблема локального метода – сходимость к окрестности. Как ее можно решить?
- Регуляризация локальной задачи:

$$\tilde{f}_m(x) := f_m(x) + \frac{\lambda}{2} \|x - v\|^2,$$

где v – некоторая референсная точка.



Karimireddy S. P. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning

А вообще чего хотим достичь?

- Нижние оценки:

$$K = \Omega \left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \right).$$

L и μ – константы гладкости и сильной выпуклости f .

- Вопрос:** какой метод даст такие оценки?

А вообще чего хотим достичь?

- Нижние оценки:

$$K = \Omega \left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \right).$$

L и μ – константы гладкости и сильной выпуклости f .

- Вопрос:** какой метод даст такие оценки? Распределенная версия метода Нестерова с 1 локальным шагом между коммуникациями.
- Отметим, что локальные методы стали для стохастических постановок.

А вообще чего хотим достичь?

- Нижние оценки:

$$K = \Omega \left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \right).$$

L и μ – константы гладкости и сильной выпуклости f .

- Вопрос:** какой метод даст такие оценки? Распределенная версия метода Нестерова с 1 локальным шагом между коммуникациями.
- Отметим, что локальные методы стали для стохастических постановок.
- Но и тут в общем случае нет улучшений.



Woodworth B. The Min-Max Complexity of Distributed Stochastic Convex Optimization with Intermittent Communication

- Но есть постановки, где локальные методы выстреливают.

Data similarity

- Распределенная задача обучения:

$$f(w) = \frac{1}{M} \sum_{m=1}^M f_m(w) = \frac{1}{M} \sum_{m=1}^M \left[\frac{1}{N} \sum_{i=1}^N \ell(w, z_i) \right],$$

где z_i – элемент выборки (x_i, y_i) , ℓ – функция потерь (в нее зашита l и g).

Data similarity

- Распределенная задача обучения:

$$f(w) = \frac{1}{M} \sum_{m=1}^M f_m(w) = \frac{1}{M} \sum_{m=1}^M \left[\frac{1}{N} \sum_{i=1}^N \ell(w, z_i) \right],$$

где z_i – элемент выборки (x_i, y_i) , ℓ – функция потерь (в нее зашита l и g).

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).

Data similarity

- Распределенная задача обучения:

$$f(w) = \frac{1}{M} \sum_{m=1}^M f_m(w) = \frac{1}{M} \sum_{m=1}^M \left[\frac{1}{N} \sum_{i=1}^N \ell(w, z_i) \right],$$

где z_i – элемент выборки (x_i, y_i) , ℓ – функция потерь (в нее зашита l и g).

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Это дает похожесть локальных функций потерь.

Data similarity

- Распределенная задача обучения:

$$f(w) = \frac{1}{M} \sum_{m=1}^M f_m(w) = \frac{1}{M} \sum_{m=1}^M \left[\frac{1}{N} \sum_{i=1}^N \ell(w, z_i) \right],$$

где z_i – элемент выборки (x_i, y_i) , ℓ – функция потерь (в нее зашита l и g).

- Предположим, что мы можем разбить обучающую выборку равномерно между устройствами (например, если используются кластерные или коллаборативные вычисления на открытых данных).
- Это дает похожесть локальных функций потерь.
- Утверждается, что для любого w

$$\|\nabla^2 f_m(w) - \nabla^2 f(w)\| \leq \delta.$$

Матричное неравенство Хёфдинга

Теорема (Матричное неравенство Хёфдинга)

Рассмотрим конечную последовательность случайных квадратных матриц $\{X_i\}_{i=1}^N$. Пусть в этой последовательности матрицы независимы, эрмитовы и имеют размерность d . Предположим так же, что $\mathbb{E}[X_i] = 0$, и $X_i^2 \preceq A^2$ почти наверное, где A – неслучайная эрмитова матрица. Тогда с вероятностью $1 - p$ выполнено, что

$$\left\| \sum_{i=1}^N X_i \right\| \leq \sqrt{8N \|A^2\| \cdot \ln(d/p)}.$$



Tropp J. An introduction to matrix concentration inequalities



Tropp J. User-friendly tail bounds for sums of random matrices

Параметр схожести

- Локальная функция потерь:

$$f_m(w) = \frac{1}{N} \sum_{i=1}^N \ell(w, z_i).$$

- ℓ – L -гладкая (L -Липшицев градиент), выпуклая, дважды дифференцируемая функция (например, квадратичная или логрегрессия). Тогда имеем $\nabla^2 \ell(w, z_i) \preceq LI$ для любого w и z_i (здесь I – единичная матрица.).
- Распределим все данные равномерно по всем нодам.
 $X_i = \frac{1}{N} [\nabla \ell(w, z_i) - \nabla f(w)]$. Легко проверить, что все условия матричного неравенства Хёфдинга для нее выполнены, в частности, $A^2 = \frac{4L^2}{N^2} I$.

Параметр схожести: итог

- В итоге имеем

$$\|\nabla^2 f_m(w) - \nabla^2 f(w)\| \leq \delta \sim \frac{L}{\sqrt{N}}.$$

Параметр схожести: итог

- В итоге имеем

$$\|\nabla^2 f_m(w) - \nabla^2 f(w)\| \leq \delta \sim \frac{L}{\sqrt{N}}.$$

- Для квадратичных задач можно получить оценку вида:

$$\|\nabla^2 f_m(w) - \nabla^2 f(w)\| \leq \delta \sim \frac{L}{N}.$$



Hendrikx H. et al. Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization

Параметр схожести: итог

- В итоге имеем

$$\|\nabla^2 f_m(w) - \nabla^2 f(w)\| \leq \delta \sim \frac{L}{\sqrt{N}}.$$

- Для квадратичных задач можно получить оценку вида:

$$\|\nabla^2 f_m(w) - \nabla^2 f(w)\| \leq \delta \sim \frac{L}{N}.$$



Hendrikx H. et al. Statistically Preconditioned Accelerated Gradient Method for Distributed Optimization

- В любом случае следует вывод: чем больше размер локальной выборки, тем меньше параметр схожести (похожи между собой гессианы).

Метод в общем виде

- Рассмотрим зеркальный спуск:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где $V(x, y)$ – дивергенция Брегмана, порожденная функцией строго-выпуклой функцией $\varphi(x)$:

$$V(x, y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y); x - y \rangle.$$

Метод в общем виде

- Рассмотрим зеркальный спуск:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где $V(x, y)$ – дивергенция Брегмана, порожденная функцией строго-выпуклой функцией $\varphi(x)$:

$$V(x, y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y); x - y \rangle.$$

- Вопрос:** Какой метод получится, если $\varphi(x) = \frac{1}{2}\|x\|^2$?

Метод в общем виде

- Рассмотрим зеркальный спуск:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где $V(x, y)$ – дивергенция Брегмана, порожденная функцией строго-выпуклой функцией $\varphi(x)$:

$$V(x, y) = \varphi(x) - \varphi(y) - \langle \nabla \varphi(y); x - y \rangle.$$

- Вопрос:** Какой метод получится, если $\varphi(x) = \frac{1}{2}\|x\|^2$?
Градиентный спуск.

Сходимость в общем виде

Определение (относительная гладкость и сильная выпуклость)

Пусть $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ является выпуклой и дважды дифференцируемой. Будем говорить, что функция f является L_φ -гладкой и μ_φ -сильно выпуклой относительно φ , если для любого $x \in \mathbb{R}^d$ выполнено

$$\mu_\varphi \nabla^2 \varphi(x) \preceq \nabla^2 f(x) \preceq L_\varphi \nabla^2 \varphi(x),$$

или эквивалентно для любых $x, y \in \mathbb{R}^d$

$$\mu_\varphi V(x, y) \leq f(x) - f(y) - \langle \nabla f(y); x - y \rangle \leq L_\varphi V(x, y).$$



Lu H. et al. Relatively-Smooth Convex Optimization by First-Order Methods, and Applications

Сходимость в общем виде: доказательство

- Первое условие оптимальности для шага зеркального спуска:

$$\gamma \nabla f(w_k) + \nabla \varphi(w_{k+1}) - \nabla \varphi(w_k) = 0.$$

Сходимость в общем виде: доказательство

- Первое условие оптимальности для шага зеркального спуска:

$$\gamma \nabla f(w_k) + \nabla \varphi(w_{k+1}) - \nabla \varphi(w_k) = 0.$$

- Из него (здесь w^* – оптимум):

$$\langle \gamma \nabla f(w_k) + \nabla \varphi(w_{k+1}) - \nabla \varphi(w_k), w_{k+1} - w^* \rangle = 0.$$

$$\begin{aligned} \langle \gamma \nabla f(w_k), w^{k+1} - w^* \rangle &= \langle \nabla \varphi(w_k) - \nabla \varphi(w_{k+1}), w^{k+1} - w^* \rangle \\ &= V(w^*, w_k) - V(w^*, w_{k+1}) - V(w_{k+1}, w_k). \end{aligned}$$

(последнее утверждение называется теоремой Пифагора для дивергенций Брегмана и проверяется по определению)

Сходимость в общем виде: доказательство

- Первое условие оптимальности для шага зеркального спуска:

$$\gamma \nabla f(w_k) + \nabla \varphi(w_{k+1}) - \nabla \varphi(w_k) = 0.$$

- Из него (здесь w^* – оптимум):

$$\langle \gamma \nabla f(w_k) + \nabla \varphi(w_{k+1}) - \nabla \varphi(w_k), w_{k+1} - w^* \rangle = 0.$$

$$\begin{aligned} \langle \gamma \nabla f(w_k), w^{k+1} - w^* \rangle &= \langle \nabla \varphi(w_k) - \nabla \varphi(w_{k+1}), w^{k+1} - w^* \rangle \\ &= V(w^*, w_k) - V(w^*, w_{k+1}) - V(w_{k+1}, w_k). \end{aligned}$$

(последнее утверждение называется теоремой Пифагора для дивергенций Брегмана и проверяется по определению)

- Небольшие перестановки дадут:

$$\begin{aligned} \langle \gamma \nabla f(w_k), w_{k+1} - w_k \rangle + V(w_{k+1}, w_k) \\ = V(w^*, w_k) - V(w^*, w_{k+1}) - \langle \gamma \nabla f(w_k), w_k - w^* \rangle. \end{aligned}$$

Сходимость в общем виде: доказательство

- Подставим $\gamma = \frac{1}{L_\varphi}$:

$$\begin{aligned} \langle \nabla f(w_k), w^{k+1} - w^k \rangle + L_\varphi V(w_{k+1}, w_k) \\ = L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) \\ - \langle \nabla f(w_k), w_k - w^* \rangle. \end{aligned}$$

Сходимость в общем виде: доказательство

- Подставим $\gamma = \frac{1}{L_\varphi}$:

$$\begin{aligned} \langle \nabla f(w_k), w^{k+1} - w^k \rangle + L_\varphi V(w_{k+1}, w_k) \\ = L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) \\ - \langle \nabla f(w_k), w_k - w^* \rangle. \end{aligned}$$

- Воспользуемся определением гладкости относительно φ с $x = w_{k+1}$, $y = w_k$:

$$f(w_{k+1}) - f(w_k) \leq \langle \nabla f(w_k); w_{k+1} - w_k \rangle + L_\varphi V(w_{k+1}, w_k).$$

Сходимость в общем виде: доказательство

- Подставим $\gamma = \frac{1}{L_\varphi}$:

$$\begin{aligned} \langle \nabla f(w_k), w^{k+1} - w^k \rangle + L_\varphi V(w_{k+1}, w_k) \\ = L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) \\ - \langle \nabla f(w_k), w_k - w^* \rangle. \end{aligned}$$

- Воспользуемся определением гладкости относительно φ с $x = w_{k+1}$, $y = w_k$:

$$f(w_{k+1}) - f(w_k) \leq \langle \nabla f(w_k); w_{k+1} - w_k \rangle + L_\varphi V(w_{k+1}, w_k).$$

- Соединим два предыдущих:

$$f(w_{k+1}) - f(w_k) \leq L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) - \langle \nabla f(w_k), w_k - w^* \rangle$$

Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$\begin{aligned} f(w_{k+1}) - f(w_k) &\leq L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) \\ &\quad - \langle \nabla f(w_k), w_k - w^* \rangle. \end{aligned}$$

Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(w_{k+1}) - f(w_k) \leq L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) \\ - \langle \nabla f(w_k), w_k - w^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(w^*, w_k) \leq f(w^*) - f(w_k) - \langle \nabla f(w_k); w^* - w_k \rangle$$

Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(w_{k+1}) - f(w_k) \leq L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) - \langle \nabla f(w_k), w_k - w^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(w^*, w_k) \leq f(w^*) - f(w_k) - \langle \nabla f(w_k), w^* - w_k \rangle$$

- Сложим два предыдущих и немного поперемещаем:

$$f(w_{k+1}) - f(w^*) \leq (L_\varphi - \mu_\varphi) V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}).$$

Сходимость в общем виде: доказательство

- С предыдущего слайда:

$$f(w_{k+1}) - f(w_k) \leq L_\varphi V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}) - \langle \nabla f(w_k), w_k - w^* \rangle.$$

- Относительная сильная выпуклость:

$$\mu_\varphi V(w^*, w_k) \leq f(w^*) - f(w_k) - \langle \nabla f(w_k), w^* - w_k \rangle$$

- Сложим два предыдущих и немного поперемещаем:

$$f(w_{k+1}) - f(w^*) \leq (L_\varphi - \mu_\varphi) V(w^*, w_k) - L_\varphi V(w^*, w_{k+1}).$$

- В силу того, что w^* – оптимум:

$$V(w^*, w_{k+1}) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right) V(w^*, w_k).$$

Сходимость в общем виде: теорема

Теорема (сходимость зеркального спуска)

Пусть φ и f удовлетворяют определению выше, тогда зеркальный спуск с шагом $\gamma = \frac{1}{L_\varphi}$ сходится и выполнено:

$$V(w^*, w_K) \leq \left(1 - \frac{\mu_\varphi}{L_\varphi}\right)^K V(w^*, w_0).$$



Lu H. et al. Relatively-Smooth Convex Optimization by First-Order Methods, and Applications

Метод для задачи data similarity

- Зеркальный спуск:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где V — дивергенция Брегмана $V(x, y)$, порожденной функцией $\varphi(x)$ (тут нужно потребовать, чтобы f_1 была выпуклой):

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|^2.$$

Функция f_1 хранится на сервере.

Метод для задачи data similarity

- Зеркальный спуск:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где V — дивергенция Брегмана $V(x, y)$, порожденной функцией $\varphi(x)$ (тут нужно потребовать, чтобы f_1 была выпуклой):

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|^2.$$

Функция f_1 хранится на сервере.

- **Вопрос:** Какое число коммуникаций происходит за K итераций такого зеркального спуска?

Метод для задачи data similarity

- Зеркальный спуск:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где V — дивергенция Брегмана $V(x, y)$, порожденной функцией $\varphi(x)$ (тут нужно потребовать, чтобы f_1 была выпуклой):

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|^2.$$

Функция f_1 хранится на сервере.

- **Вопрос:** Какое число коммуникаций происходит за K итераций такого зеркального спуска? K коммуникаций (количество подсчетов градиента ∇f), вычисления $\arg \min$ требуют только вычислений на сервере.

Метод для задачи data similarity

Алгоритм 1 Зеркальный спуск для задачи data similarity

Вход: Размер шага $\gamma > 0$, стартовая точка $w^0 \in \mathbb{R}^d$, количество итераций K

- 1: **for** $k = 0, 1, \dots, K - 1$ **do**
 - 2: Отправить x_k всем рабочим ▷ выполняется сервером
 - 3: **for** $m = 1, \dots, M$ параллельно **do**
 - 4: Принять w_k от мастера ▷ выполняется рабочими
 - 5: Вычислить градиент $\nabla f_m(w_k)$ в точке w_k ▷ выполняется рабочими
 - 6: Отправить $\nabla f_m(w_k)$ мастеру ▷ выполняется рабочими
 - 7: **end for**
 - 8: Принять $\nabla f_m(w_k)$ от всех рабочих ▷ выполняется сервером
 - 9: Вычислить $\nabla f(w_k) = \frac{1}{M} \sum_{m=1}^M \nabla f_m(w_k)$ ▷ выполняется сервером
 - 10: $w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\gamma \langle \nabla f(w_k), x \rangle + V(w, w_k))$ ▷ выполняется сервером
 - 11: **end for**
- Выход:** w_K

Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_{\varphi} \nabla^2 \varphi(w) \preceq \nabla^2 f(w) \preceq L_{\varphi} \nabla^2 \varphi(w),$$

Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_{\varphi} \nabla^2 \varphi(w) \preceq \nabla^2 f(w) \preceq L_{\varphi} \nabla^2 \varphi(w),$$

- В нашем случае:

$$\mu_{\varphi} (\delta I + \nabla^2 f_1(w)) \preceq \nabla^2 f(w) \preceq L_{\varphi} (\delta I + \nabla^2 f_1(w))$$

Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_{\varphi} \nabla^2 \varphi(w) \preceq \nabla^2 f(w) \preceq L_{\varphi} \nabla^2 \varphi(w),$$

- В нашем случае:

$$\mu_{\varphi} (\delta I + \nabla^2 f_1(w)) \preceq \nabla^2 f(w) \preceq L_{\varphi} (\delta I + \nabla^2 f_1(w))$$

- Найдем L_{φ} :

$$\begin{aligned} \|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta &\Rightarrow \nabla^2 f(w) - \nabla^2 f_1(w) \preceq \delta I \\ \Rightarrow \nabla^2 f(w) &\preceq \delta I + \nabla^2 f_1(w) \Rightarrow L_{\varphi} = 1. \end{aligned}$$

Сходимость для задачи data similarity: доказательство

- Напомним, что сходимость определяется через константы из соотношения:

$$\mu_\varphi \nabla^2 \varphi(w) \preceq \nabla^2 f(w) \preceq L_\varphi \nabla^2 \varphi(w),$$

- В нашем случае:

$$\mu_\varphi (\delta I + \nabla^2 f_1(w)) \preceq \nabla^2 f(w) \preceq L_\varphi (\delta I + \nabla^2 f_1(w))$$

- Найдем L_φ :

$$\begin{aligned} \|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta &\Rightarrow \nabla^2 f(w) - \nabla^2 f_1(w) \preceq \delta I \\ \Rightarrow \nabla^2 f(w) &\preceq \delta I + \nabla^2 f_1(w) \Rightarrow L_\varphi = 1. \end{aligned}$$

Сходимость для задачи data similarity: доказательство

- Найдем μ_φ . Из сильно выпуклости функции f :

$$\mu l \preceq \nabla^2 f(w) \Rightarrow \delta l \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Из $\|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta$ имеем:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \delta l.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Откуда:

$$\nabla^2 f_1(w) + \delta l \preceq \frac{2\delta + \mu}{\mu} \nabla^2 f(w) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

Сходимость для задачи data similarity: доказательство

- Найдем μ_φ . Из сильно выпуклости функции f :

$$\mu I \preceq \nabla^2 f(w) \Rightarrow \delta I \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta I.$$

Сходимость для задачи data similarity: доказательство

- Найдем μ_φ . Из сильно выпуклости функции f :

$$\mu I \preceq \nabla^2 f(w) \Rightarrow \delta I \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta I.$$

- Из $\|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta$ имеем:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \delta I.$$

Сходимость для задачи data similarity: доказательство

- Найдем μ_φ . Из сильно выпуклости функции f :

$$\mu l \preceq \nabla^2 f(w) \Rightarrow \delta l \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Из $\|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta$ имеем:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \delta l.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

Сходимость для задачи data similarity: доказательство

- Найдем μ_φ . Из сильно выпуклости функции f :

$$\mu l \preceq \nabla^2 f(w) \Rightarrow \delta l \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Из $\|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta$ имеем:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \delta l.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Откуда:

$$\nabla^2 f_1(w) + \delta l \preceq \frac{2\delta + \mu}{\mu} \nabla^2 f(w) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

Сходимость для задачи data similarity: доказательство

- Найдем μ_φ . Из сильно выпуклости функции f :

$$\mu l \preceq \nabla^2 f(w) \Rightarrow \delta l \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Из $\|\nabla^2 f_1(w) - \nabla^2 f(w)\| \leq \delta$ имеем:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \delta l.$$

- Объединяем два предыдущих пункта:

$$\nabla^2 f_1(w) - \nabla^2 f(w) \preceq \frac{2\delta}{\mu} \nabla^2 f(w) - \delta l.$$

- Откуда:

$$\nabla^2 f_1(w) + \delta l \preceq \frac{2\delta + \mu}{\mu} \nabla^2 f(w) \Rightarrow \mu_\varphi = \frac{\mu}{2\delta + \mu}.$$

Сходимость для задачи data similarity: теорема

Теорема (сходимость для задачи data similarity)

Пусть f сильно выпуклая, f_i выпуклые, а ℓ - гладкие, а $\varphi(w) = f_1(w) + \delta\|w\|^2$, тогда зеркальный спуск с шагом $\gamma = 1$ сходится и выполнено:

$$V(w^*, w_K) \leq \left(1 - \frac{\mu}{\mu + 2\delta}\right)^K V(w^*, w_0).$$

Сходимость для задачи data similarity: теорема

Теорема (сходимость для задачи data similarity)

Пусть f сильно выпуклая, f_i выпуклые, а ℓ - гладкие, а $\varphi(w) = f_1(w) + \delta \|w\|^2$, тогда зеркальный спуск с шагом $\gamma = 1$ сходится и выполнено:

$$V(w^*, w_K) \leq \left(1 - \frac{\mu}{\mu + 2\delta}\right)^K V(w^*, w_0).$$

- Это означает, что если нам необходимо достигнуть точности ε ($V(w^*, w_K) \sim \varepsilon$), то нам необходимо

$$K = \left(\left[1 + \frac{\delta}{\mu}\right] \log \frac{V(w^*, w_0)}{\varepsilon} \right) \text{ коммуникаций.}$$

Лучше?

- Оценка на число коммуникаций в условиях data similarity:

$$K = \mathcal{O} \left(\left[1 + \frac{\delta}{\mu} \right] \log \frac{1}{\varepsilon} \right).$$

- Оценка на число коммуникаций для обычного распределенного градиентного спуска:

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

Лучше?

- Оценка на число коммуникаций в условиях data similarity:

$$K = \mathcal{O} \left(\left[1 + \frac{\delta}{\mu} \right] \log \frac{1}{\varepsilon} \right).$$

- Оценка на число коммуникаций для обычного распределенного градиентного спуска:

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

- Напомним, что $\delta \sim \frac{L}{\sqrt{N}}$, т.е. может быть значительное улучшение.

Другой взгляд на зеркальный спуск

- Зеркальный спуск с $\gamma = 1$:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} (\langle \nabla f(w_k), w \rangle + V(w, w_k)),$$

где V — дивергенция Брегмана, порожденная функцией $\varphi(x)$:

$$\varphi(x) = f_1(x) + \frac{\delta}{2} \|x\|^2.$$

- Подставим $\varphi(x)$:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} \left(f_1(w) + \langle \nabla f(w_k) - \nabla f_1(w_k), w \rangle + \frac{\delta}{2} \|w - w_k\|^2 \right).$$

- Или чуть по-другому:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} \left(\frac{1}{\delta} f_1(w) + \frac{1}{2} \left\| w - \left(w_k - \frac{1}{\delta} (\nabla f(w_k) - \nabla f_1(w_k)) \right) \right\|^2 \right),$$

Итого про зеркальный спуск

- Всплыла идея регуляризации локальной подзадачи.
- Всплыла идея слайдинга \approx проксимального метода с неточностью.
- Проксимальный метод для композитной целевой функции $g_1(w) + g_2(w)$:

$$w_{k+1} = \arg \min_{w \in \mathbb{R}^d} \left(\gamma g_2(w) + \frac{1}{2} \|w - (w_k - \gamma g_1(w_k))\|^2 \right).$$

- В нашем случае, $g_1 = f - f_1$, $g_2 = f_1$.

Лучше?

- Мы получили:

$$K = \mathcal{O} \left(\left[1 + \frac{\delta}{\mu} \right] \log \frac{1}{\varepsilon} \right).$$

- Но есть ведь и ускоренный градиентный метод, который дает оценки:

$$K = \mathcal{O} \left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \right).$$

- Непонятно, что лучше. Более того, можно ли ускорить метод для задачи с data similarity?
- Для задачи data similarity так же имеются нижние оценки:

$$K = \Omega \left(\sqrt{1 + \frac{\delta}{\mu}} \log \frac{1}{\varepsilon} \right),$$

т.е. предполагается возможное ускорение.

Arjevani Y. and Shamir O. Communication complexity of dis-



Оптимальный алгоритм

- У данной проблемы довольно большая история:

Reference	Communication complexity	Local gradient complexity	Order	Limitations
DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2} \log \frac{1}{\varepsilon}\right)$	— ⁽²⁾	1st	quadratic
DISCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} (\log \frac{1}{\varepsilon} + C^2 \Delta F_0) \log \frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} (\log \frac{1}{\varepsilon} + C^2 \Delta F_0) \log \frac{L}{\mu}\right)$	2nd	C - self-concordant ⁽¹⁾
AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon} \log \frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \log \frac{L}{\delta}\right)$ ⁽⁴⁾	1st	quadratic
DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu} \log \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \frac{\delta^{3/2}}{\mu^{3/2}} \log \frac{1}{\varepsilon}\right)$ ⁽⁵⁾	1st/2nd	quadratic ⁽⁶⁾
DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \frac{\delta}{\mu} \log \frac{1}{\varepsilon}\right)$ ⁽⁵⁾	1st/2nd	quadratic ⁽⁶⁾
SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu} \log \frac{1}{\varepsilon}\right)$	— ⁽²⁾	1st	decentralized
SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$ ⁽¹⁾	— ⁽²⁾	1st	M - Lipschitz hessian
DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu} \log \frac{1}{\varepsilon} + \sqrt{\frac{M \delta R_0}{\mu}}\right)$	— ⁽²⁾	2nd	M - Lipschitz hessian
ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon} + \sqrt[3]{\frac{M \delta R_0}{\mu}}\right)$	— ⁽²⁾	2nd	M - Lipschitz hessian
Acc SONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon} \log \frac{\delta}{\mu}\right)$	— ⁽²⁾	1st	decentralized
This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$	1st	

В частности, подход зеркального спуска с необычной дивергенцией называется DANE.

Оптимальный алгоритм

- У данной проблемы довольно большая история:

Reference	Communication complexity	Local gradient complexity	Order	Limitations
DANE [42]	$\mathcal{O}\left(\frac{\delta^2}{\mu^2} \log \frac{1}{\varepsilon}\right)$	— ⁽²⁾	1st	quadratic
DISCO [51]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} (\log \frac{1}{\varepsilon} + C^2 \Delta F_0) \log \frac{L}{\mu}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} (\log \frac{1}{\varepsilon} + C^2 \Delta F_0) \log \frac{L}{\mu}\right)$	2nd	C - self-concordant ⁽¹⁾
AIDE [40]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon} \log \frac{L}{\delta}\right)$	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \log \frac{L}{\delta}\right)$ ⁽⁴⁾	1st	quadratic
DANE-LS [50]	$\mathcal{O}\left(\frac{\delta}{\mu} \log \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \frac{\delta^{3/2}}{\mu^{3/2}} \log \frac{1}{\varepsilon}\right)$ ⁽⁵⁾	1st/2nd	quadratic ⁽⁶⁾
DANE-HB [50]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \frac{\delta}{\mu} \log \frac{1}{\varepsilon}\right)$ ⁽⁵⁾	1st/2nd	quadratic ⁽⁶⁾
SONATA [45]	$\mathcal{O}\left(\frac{\delta}{\mu} \log \frac{1}{\varepsilon}\right)$	— ⁽²⁾	1st	decentralized
SPAG [21]	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$ ⁽¹⁾	— ⁽²⁾	1st	M - Lipschitz hessian
DiRegINA [12]	$\mathcal{O}\left(\frac{\delta}{\mu} \log \frac{1}{\varepsilon} + \sqrt{\frac{M \delta R_0}{\mu}}\right)$	— ⁽²⁾	2nd	M - Lipschitz hessian
ACN [1]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon} + \sqrt[3]{\frac{M \delta R_0}{\mu}}\right)$	— ⁽²⁾	2nd	M - Lipschitz hessian
Acc SONATA [46]	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon} \log \frac{\delta}{\mu}\right)$	— ⁽²⁾	1st	decentralized
This paper	$\mathcal{O}\left(\sqrt{\frac{\delta}{\mu}} \log \frac{1}{\varepsilon}\right)$	$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon}\right)$	1st	

В частности, подход зеркального спуска с необычной дивергенцией называется DANE.

- Оптимальный алгоритм был предложен в 2022 году:
Kovalev D. et al. Optimal Gradient Sliding and its Application to Distributed Optimization Under Similarity



Оптимальный алгоритм

Для задачи:

$$f(w) = g_1(w) + g_2(w),$$

где $g_1 = f - f_1$ и $g_2 = f_1$.

Алгоритм 2 Accelerated Extragradient

- 1: **Input:** $w^0 = w_f^0 \in \mathbb{R}^d$
- 2: **Parameters:** $\tau \in (0, 1]$, $\eta, \theta, \alpha > 0$, $K \in \{1, 2, \dots\}$
- 3: **for** $k = 0, 1, 2, \dots, K - 1$ **do**
- 4: $w_g^k = \tau w^k + (1 - \tau) w_f^k$
- 5: $w_f^{k+1} \approx \operatorname{argmin}_{x \in \mathbb{R}^d} [\langle \nabla g_1(w_g^k), w - w_g^k \rangle + \frac{1}{2\theta} \|w - w_g^k\|^2 + g_2(w)]$
- 6: $w^{k+1} = w^k + \eta \alpha (w_f^{k+1} - w^k) - \eta \nabla g(w_f^{k+1})$
- 7: **end for**
- 8: **Output:** w^K

Три идеи

- 1 идея – Ускорение Нестерова
- 2 идея – Слайдинг
- 3 идея – Экстраградиент
- Первые две идеи понятны, ключевой является третья идея.

Оптимально. Но может еще?

- **Вопрос:** забудем на 1 слайд про распределенку, и вспомним всегда ли метод Нестерова оптимален?

Оптимально. Но может еще?

- **Вопрос:** забудем на 1 слайд про распределенку, и вспомним всегда ли метод Нестерова оптимален? Нет, если учитывать специфику, что целевая функция может иметь виды суммы

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x).$$

- **Вопрос:** какой тогда метод является оптимальным? какие у него верхние оценки сходимости?

Оптимально. Но может еще?

- **Вопрос:** забудем на 1 слайд про распределенку, и вспомним всегда ли метод Нестерова оптимален? Нет, если учитывать специфику, что целевая функция может иметь виды суммы

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x).$$

- **Вопрос:** какой тогда метод является оптимальным? какие у него верхние оценки сходимости?
- Метод называется Katyusha, он имеет следующую верхнюю оценку сходимости (оракульная сложность по вызову f_i):

$$\mathcal{O} \left(\left[n + \sqrt{n \frac{L}{\mu}} \right] \log \frac{1}{\varepsilon} \right).$$



Allen-Zhu Z. Katyusha: the first direct acceleration of stochastic gradient methods

Вопрос: А какая верхняя оценка на оракульную сложность для метода Нестерова?

Оптимально. Но может еще?

- **Вопрос:** забудем на 1 слайд про распределенку, и вспомним всегда ли метод Нестерова оптимален? Нет, если учитывать специфику, что целевая функция может иметь виды суммы

$$f(x) = \frac{1}{n} \sum_{i=1}^n f_i(x).$$

- **Вопрос:** какой тогда метод является оптимальным? какие у него верхние оценки сходимости?
- Метод называется Katyusha, он имеет следующую верхнюю оценку сходимости (оракульная сложность по вызову f_i):

$$\mathcal{O} \left(\left[n + \sqrt{n \frac{L}{\mu}} \right] \log \frac{1}{\varepsilon} \right).$$



Allen-Zhu Z. Katyusha: the first direct acceleration of stochastic gradient methods

Вопрос: А какая верхняя оценка на оракульную сложность для метода Нестерова? $\mathcal{O} \left(n \sqrt{\frac{L}{\mu}} \log \frac{1}{\varepsilon} \right)$

Редукция дисперсии для similarity

- Идея метода редукции дисперсии:

$$\begin{array}{c} \nabla f(x) \\ \downarrow \\ \nabla f_i(x) - \nabla f_i(w) + \nabla f(w), \end{array}$$

где i - генерируется случайно на каждой итерации из $[n]$, w – референсная точка, которая обновляется редко (случайно или детерминистически).

- Идея метода редукции дисперсии для data similarity:

$$\begin{array}{c} \nabla f(x) - \nabla f_1(x) \\ \downarrow \\ \nabla f_i(x) - \nabla f_i(w) + \nabla f(w) - f_1(x), \end{array}$$

где i - генерируется случайно на каждой итерации из $[M]$.

Редукция дисперсии для similarity



Beznosikov A. & Gasnikov A. Compression and data similarity: Combination of two techniques for communication-efficient solving of distributed variational inequalities



Beznosikov A. & Gasnikov A. Similarity, Compression and Local Steps: Three Pillars of Efficient Communications for Distributed Variational Inequalities



Khaled A. & Jin C. Faster federated optimization under second-order similarity

- Старая оценка:

$$\mathcal{O} \left(M \sqrt{1 + \frac{\delta}{\mu} \log \frac{1}{\varepsilon}} \right).$$

- Что можно "выбить":

$$\mathcal{O} \left(\left[M + \frac{\delta^2}{\mu^2} \right] \log \frac{1}{\varepsilon} \right) \text{ or } \mathcal{O} \left(\left[M + \sqrt{M} \frac{\delta}{\mu} \right] \log \frac{1}{\varepsilon} \right) \text{ or } \mathcal{O} \left(\left[M + M^{3/4} \sqrt{\frac{\delta}{\mu}} \right] \log \frac{1}{\varepsilon} \right)$$