Распределенные методы использующие сжатые коммуникации для решения вариационных неравенств

Александр Безносиков

MIPT

2 ноября 2022

A. Beznosikov, P. Richtárik, M. Diskin, M. Ryabinin, A. Gasnikov. Distributed Methods with Compressed Communication for Solving Variational Inequalities, with Theoretical Guarantees [3]

Распределенное вариационное неравенство

Определение

$$\textit{Найти } z^* \in \mathbb{R}^d \textit{ такую, что } \langle F(z^*), z - z^* \rangle \geq 0, \; \forall z \in \mathbb{R}^d,$$

где $F : \mathbb{R}^d \to \mathbb{R}^d$ *некоторый оператор. Мы предполагаем, что* F распределен *между* M *рабочими/агентами/устройствами:*

$$F(z) \overset{def}{=} \frac{1}{M} \sum_{m=1}^{M} F_m(z),$$

где $F_m : \mathbb{R}^d \to \mathbb{R}^d$ *для всех* $m \in \{1, 2, \dots, M\}$.

Эта формулировка эквивалентна:

$$\text{Найти } z^* \in \mathbb{R}^d \text{ такую, что } F(z^*) = 0.$$

Вариацинное неравенство

- Задача минимизация:

$$\min_{z \in \mathbb{R}^d} f(z).$$

Мы берем $F(z) \overset{\text{def}}{=} \nabla f(z)$. Ищем $\nabla f(z^*)$.

- Седловая задача:

$$\min_{x \in \mathbb{R}^{d_x}} \min_{y \in \mathbb{R}^{d_y}} g(x, y).$$

Здесь $F(z) \overset{\text{def}}{=} F(x, y) = [\nabla_x g(x, y), -\nabla_y g(x, y)]$. Ищем $\nabla_x g(x^*, y^*) = 0$.

- Поиск стационарной точки оператора:

$$\text{Найти } z^* \in \mathbb{R}^d \text{ такую, что } T(z^*) = z^*,$$

где $T : \mathbb{R}^d \to \mathbb{R}^d$ оператор. Берем $F(z) = z - T(z)$.

Вариационное неравенство: пример

- Задача минимизация:

$$\min_{z \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i, z), y_i),$$

где $\{x_i, y_i\}_{i=1}^n$ – данные, f – модель в параметрами z, l – функция потерь.
Распределенный вариант:

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

- Седловая задача:

$$\min_{z \in \mathbb{R}^d} \max_{\|\delta_i\| \le e} \frac{1}{n} \sum_{i=1}^{n} l(f(x_i + \delta_i, z), y_i),$$

где δ_i – состязательный шум.

Распределенные задачи

- Кластерное обучение
- Федеративное обучение

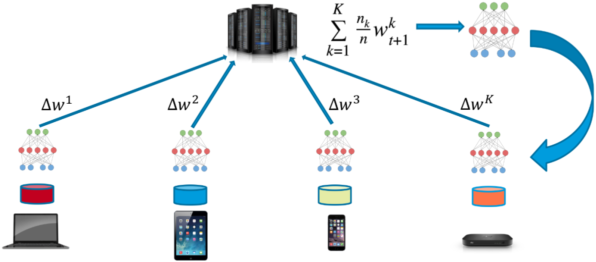


Figure: Централизованное распределенное/федеративное обучение

Распределенные задачи

- Кластерное обучение
- Федеративное обучение

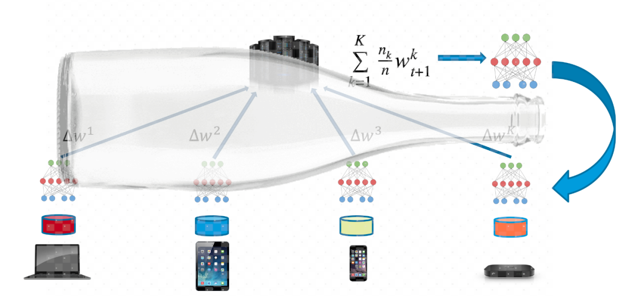


Figure: Централизованное распределенное/федеративное обучение

Квантизация и компрессия

Определение (Квантизация)

Стохастический оператор $Q : \mathbb{R}^d \to \mathbb{R}^d$ *называется* квантизацией *если существует константа* $q \geq 1$ *такая, что*

$$Q(z) = z, \quad \mathrm{E}\|Q(z)\|^2 \leq q\|z\|^2, \quad \forall z \in \mathbb{R}^d.$$

Ожидаемое/среднее сжатие (насколько меньше занимает в памяти сжатый вектор): $\beta^{-1} \overset{\text{def}}{=} \frac{\mathrm{E}\|Q(z)\|_{\mathbf{bits}}}{\|z\|_{\mathbf{bits}}}$. Отметим, что $\beta \geq 1$.

Примеры: случайный выбор координат.

Определение (Компрессия)

(Стохастический) оператор $C : \mathbb{R}^d \to \mathbb{R}^d$ *называется* компрессией, *если существует* $\delta \geq 1$ *такая, что*

$$\mathrm{E}\|C(z) - z\|^2 \leq (1 - 1/\delta)\|z\|^2, \quad \forall z \in \mathbb{R}^d.$$

Ожидаемое/среднее сжатие (насколько меньше занимает в памяти сжатый вектор): $\beta^{-1} \overset{\text{def}}{=} \frac{\mathrm{E}\|C(z)\|_{\mathbf{bits}}}{\|z\|_{\mathbf{bits}}}$. Отметим, что $\beta \geq 1$.

Предположения

Определение (Липшецевость)

Каждый оператор F_m является L-Липшецевым, если для любых $z_1, z_2 \in \mathbb{R}^d$ мы имеем $\|F_m(z_1) - F_m(z_2)\| \leq L\|z_1 - z_2\|$.

Для минимизации и седел, эти свойства эквиваленты гладкости.

Definition (Монотонность)

(SM) Сильная монотонность. Оператор F является μ-сильно монотонным, если для любых $z_1, z_2 \in \mathbb{R}^d$ мы имеем $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq \mu \|z_1 - z_2\|^2$.
(M) Монотонность. Оператор F является монотонным, если для любых $z_1, z_2 \in \mathbb{R}^d$ мы имеем $\langle F(z_1) - F(z_2), z_1 - z_2 \rangle \geq 0$.
(NM) Minty/немонотонность. Оператор F удовлетворяет условию Minty, если существует $z^* \in \mathbb{R}^d$ такая, что для любой $z \in \mathbb{R}^d$ мы имеем $\langle F(z), z - z^* \rangle \geq 0$.

Для минимизации это эквивалентно (сильной) выпуклости, а для седловых задач (сильной) выпуклости–(сильной) вогнутости.

Идеи

- Идея первая: использовать уже имеющиеся методы для задач минимизации.
- Например, квантизованный градиентный спуск

$$z^{k+1} = z^k - \gamma \cdot \frac{1}{M} \sum_{m=1}^{M} Q(F_m(z^k))$$

- Суть:
  1) пересылаем на сервер $Q(F_m(z^k))$ на сервер,
  2) сервер делает апдейт: $z^{k+1} = z^k - \gamma \cdot \frac{1}{M} \sum_{m=1}^{M} Q(F_m(z^k))$,
  3) сервер рассылает всем устройствам z^{k+1}.

Идеи

- Идея первая: использовать уже имеющиеся методы для задач минимизации.
- Получается оценка на число итераций в сильно монотонном случае с $F_m(z^*) = 0$:
$$\mathcal{O}\left(q \cdot \frac{L^2}{\mu^2}\right).$$
Оценка на число бит:
$$\mathcal{O}\left(\frac{q}{\beta} \cdot \frac{L^2}{\mu^2}\right).$$
- Это не очень хорошая оценка. Более того, в монотонном случае вообще не получится доказать сходимость.

Идеи

- Идея вторая: вставить квантизацию или компрессию в методы для вариационных неравенств.
- Например, квантизованный экстраградиентный метод

$$z^{k+1/2} = z^k - \gamma \cdot \frac{1}{M} \sum_{m=1}^{M} Q_1(F_m(z^k))$$

$$z^{k+1} = z^k - \gamma \cdot \frac{1}{M} \sum_{m=1}^{M} Q_2(F_m(z^{k+1/2}))$$

- Здесь взяты разные Q. По факту это может быть одинаковый оператор с точки зрения физики, но с разной или одинаковой случайностью.

Идеи

- Идея вторая: вставить квантизацию или компрессию в методы для вариационных неравенств.
- Получается оценка на число итераций в сильно монотонном случае с $F_m(z^*) = 0$:
$$\mathcal{O}\left(q \cdot \frac{L^2}{\mu^2}\right).$$
Оценка на число бит:
$$\mathcal{O}\left(\frac{q}{\beta} \cdot \frac{L^2}{\mu^2}\right).$$
- Эта оценка не отличается от той, что мы имели ранее и она так же не очень хорошая. Более того, в монотонном случае вообще не получится доказать сходимость.

Идеи

- Идея третья: взять за основу метод редукции дисперсии.
- Для методов минимизации это делали в работах:
  - E. Gorbunov, K. Burlachenko, Z. Li, P. Richtárik. MARINA: Faster Non-Convex Distributed Learning with Compression [4]
  - X. Qian, P. Richtárik, T. Zhang. Error Compensated Distributed SGD Can Be Accelerated [5]
- Метод редукции для ВН:
  - A. Alacaoglu, Y. Malitsky. Stochastic Variance Reduction for Variational Inequality Methods [1]

Algorithm 1 MASHA1

Parameters: Stepsize $\gamma > 0$, parameter $\tau \in (0; 1)$, number of iterations K.
Initialization: Choose $z^0 = w^0 \in \mathcal{Z}$.
Devices send $F_m(w^0)$ to server and get $F(w^0)$
for $k = 0, 1, 2, \ldots, K - 1$ **do**
  for each device m in parallel **do**
    $z^{k+1/2} = \tau z^k + (1 - \tau) w^k - \gamma F(w^k)$
    Sends $g_m^k = Q_m^{\text{dev}}(F_m(z^{k+1/2}) - F_m(w^k))$ to server
  end for
  for server **do**
    Sends to devices $g^k = Q^{\text{serv}}\left[\frac{1}{M}\sum_{m=1}^{M} g_m^k\right]$
    Sends to devices one bit b_k : 1 with probability $1 - \tau$, 0 with with probability τ
  end for
  for each device m in parallel **do**
    $z^{k+1} = z^{k+1/2} - \gamma g^k$
    If $b_k = 1$ then $w^{k+1} = z^k$, sends $F_m(w^{k+1})$ to server and gets $F(w^{k+1})$
    else $w^{k+1} = w^k$
  end for
end for

Сходимость MASHA1

Theorem

Пусть выполнено предположение о Липшецевости операторов, а также одно из предположений о монотонности. Тогда для некоторого шага γ *и* $1-\tau = {}^1\!/\!_\beta$ *справедлива следующая оценка на число бит необходимое MASHA1, чтобы достигнуть точности* ε

- *в сильно монотонном случае:* $\mathcal{O}([1+\sqrt{\frac{1}{M}+\frac{1}{\beta}}\cdot\frac{L}{\mu}]\log\frac{1}{\varepsilon})$;
- *в монотонном случае:* $\mathcal{O}(\sqrt{\frac{1}{M}+\frac{1}{\beta}}\cdot\frac{L\|z^0-z^*\|^2}{\varepsilon})$;
- *в немонотонном случае:* $\mathcal{O}([1+\frac{q}{M}]\frac{L^2\|z^0-z^*\|^2}{\varepsilon^2})$.

Сложности для методов без квантизации:

- в сильно монотонном случае: $\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon})$;
- в монотонном случае: $\mathcal{O}(\frac{L\|z^0-z^*\|^2}{\varepsilon})$;
- в немонотонном случае: $\mathcal{O}(\frac{L^2\|z^0-z^*\|^2}{\varepsilon^2})$.

Сходимость MASHA2

Theorem

Пусть выполнено предположение о Липшецевости операторов, а также одно из предположений о монотонности. Тогда для некоторого шага γ *и* $1 - \tau = {}^1/_\beta$ *справедлива следующая оценка на число бит необходимое MASHA2, чтобы достигнуть точности* ε

- *в сильно монотонном случае:* $\mathcal{O}(\frac{L}{\mu} \log \frac{1}{\varepsilon})$;
- *в монотонном случае:* $\mathcal{O}(\frac{L\|z^0 - z^*\|^2}{\varepsilon})$;
- *в немонотонном случае:* $\mathcal{O}(\delta \cdot \frac{L^2\|z^0 - z^*\|^2}{\varepsilon^2})$.

Эксперименты

- Билинейная седловая задача:

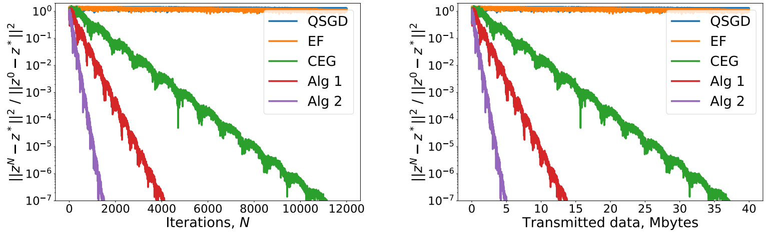
$$\min_{x \in \mathbb{R}^{d_x}} \min_{y \in \mathbb{R}^{d_y}} g(x,y) \overset{\text{def}}{=} \frac{1}{M} \sum_{m=1}^{M} g_m(x,y) \quad \text{с}$$

$$g_m(x,y) \overset{\text{def}}{=} x^\top A_m y + a_m^\top x + b_m^\top y + \frac{\lambda}{2}\|x\|^2 - \frac{\lambda}{2}\|y\|^2,$$

  где $A_m \in \mathbb{R}^{d \times d}$, $a_m, b_m \in \mathbb{R}^d$. Это задача λ-сильно выпуклая–сильно вогнутая и, более того, все функции g_m являются $\|A_m\|_2$-гладкими. Мы берем $d = 100$ и случайно генерируем положительно определенные матрицы A_m и векторы a_m, b_m, λ выбирается, как $\max_m \|A_m\|_2 / 10^5$.
- Мы сравниваем наши методы с 1) QGD [2] с Rand30% квантизацией, 2)GD с техникой компенсации ошибки [6] и Top 30% компрессором, 3) квантизованным EG с Rand30% квантизацией.

Эксперименты: билинейная седловая задача

Figure: Сравнение MASHA1 и MASHA2 с QGD, GD с компенсацией ошибки, и квантизованным EG по итерациям и мегабайтам.



Ahmet Alacaoglu and Yura Malitsky.
Stochastic variance reduction for variational inequality methods.
In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic.
QSGD: Communication-efficient SGD via gradient quantization and encoding.
In *Advances in Neural Information Processing Systems*, pages 1709–1720, 2017.

Aleksandr Beznosikov, Peter Richtárik, Michael Diskin, Max Ryabinin, and Alexander Gasnikov.
Distributed methods with compressed communication for solving variational inequalities, with theoretical guarantees.
arXiv preprint arXiv:2110.03313, 2021.

Eduard Gorbunov, Konstantin P Burlachenko, Zhize Li, and Peter Richtárik.
Marina: Faster non-convex distributed learning with compression.
In *International Conference on Machine Learning*, pages 3788–3798. PMLR, 2021.

Xun Qian, Peter Richtárik, and Tong Zhang.
Error compensated distributed sgd can be accelerated.

Advances in Neural Information Processing Systems, 34:30401–30413, 2021.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi.
Sparsified sgd with memory.
arXiv preprint arXiv:1809.07599, 2018.

Thijs Vogels, Sai Praneeth Karinireddy, and Martin Jaggi.
Powersgd: Practical low-rank gradient compression for distributed optimization.
Advances In Neural Information Processing Systems 32 (Nips 2019), 32(CONF), 2019.