PAGE в условиях гомогенности данных

Подготовил: Максимов Роман
Участники проекта: Максимов Роман, Алимаскина Екатерина
Руководитель: Былинкин Дмитрий, МФТИ
22.04.2025

Вспомним постановку

$$\min_x h(x) = \min_x f(x) + g(x)$$

$$\Updownarrow$$

$$\min_{x,y} r(x,y) = \min_{x,y} p\begin{pmatrix}\vec{x}\\0\end{pmatrix} + q\begin{pmatrix}0\\\vec{y}\end{pmatrix}$$

$\Longleftrightarrow$

Разделение сложностей по f и g

$\Updownarrow$

Разделение сложностей по x и y

$$\|\nabla^2 f_1(x) - \nabla^2 f(x)\| \leq \delta_f$$

$$\|\nabla^2 g_1(x) - \nabla^2 g(x)\| \leq \delta_g$$

Определение похожести данных

Вдохновение

Algorithm 2 $\texttt{VRCS}^{\texttt{1ep}}(p, q, \theta, x_0)$

Input: $x_0 \in \mathbb{R}^d$
Parameters: $p, q \in (0,1),\ \theta > 0$
$T \sim \text{Geom}\,(q)$
for $t = 0, \dots, T-1$ **do**
  $i_t \sim \text{Be}(p)$
  $\xi_t = \begin{cases} \frac{1}{p}\nabla(f - f_1)(x_t) \text{ if } i_t = 1, \\ \frac{1}{1-p}\nabla(g - g_1)(x_t) \text{ if } i_t = 0 \end{cases}$
  $\zeta_t = \begin{cases} \frac{1}{p}\nabla(f - f_1)(x_0) \text{ if } i_t = 1, \\ \frac{1}{1-p}\nabla(g - g_1)(x_0) \text{ if } i_t = 0 \end{cases}$
  $e_t = \xi_t - \zeta_t + \nabla h(x_0) - \nabla h_1(x_0)$
  $x_{t+1} \approx \arg\min_{x \in \mathbb{R}^d} \left[A_\theta^t(x)\right]$, where

$$A_\theta^t(x) = \langle e_t, x\rangle + \frac{1}{2\theta}\|x - x_0\|^2 + h_1(x)$$

end for
Output: x_T

Accelerated Methods with Complexity Separation under Data Similarity for Federated Learning Problems

$$\nabla p \to \zeta = \begin{cases} \frac{1}{p}\nabla_x(f - f_1),\ p \\ \frac{1}{1-p}\nabla_y(f - f_1),\ 1 - p \end{cases}$$

Then the complexities in terms of communication rounds are

$$\mathcal{O}\left(\frac{\delta_f}{\mu}\log\frac{1}{\varepsilon}\right) \textit{ for the nodes from } M_f,$$

and

$$\mathcal{O}\left(\left(\frac{\delta_g}{\delta_f}\right)\frac{\delta_g}{\mu}\log\frac{1}{\varepsilon}\right) \textit{ for the nodes from } M_g.$$

Переход к PAGE

Algorithm 1 ProbAbilistic Gradient Estimator (PAGE)

Input: initial point x^0, stepsize η, minibatch size b, $b' < b$, probability $\{p_t\} \in (0,1]$

$g^0 = \frac{1}{b}\sum_{i \in I} \nabla f_i(x^0)$ // I denotes random minibatch samples with $|I| = b$

for $t = 0, 1, 2, \ldots$ **do**

$x^{t+1} = x^t - \eta g^t$

$$g^{t+1} = \begin{cases} \frac{1}{b}\sum\limits_{i \in I} \nabla f_i(x^{t+1}) & \text{with probability } p_t \\ g^t + \frac{1}{b'}\sum\limits_{i \in I'} (\nabla f_i(x^{t+1}) - \nabla f_i(x^t)) & \text{with probability } 1 - p_t \end{cases}$$

end for

Output: $\widehat{x}_T$ chosen uniformly from $\{x^t\}_{t \in [T]}$

PAGE: A Simple and Optimal Probabilistic Gradient Estimator for Nonconvex Optimization

Наше предложение:

Input: $x_0 \in R^d, \theta > 0, p_1, p_2 \in (0,1), T$
for $t = 0, 1, \ldots, T-1$ **do**:
Sample g_t as:

$$g_t = \begin{cases} g_{t-1} + \nabla(f - f_1)(x_t) & \text{with prob. } p_1 \\ \quad -\nabla(f - f_1)(x_{t-1}), & \\ g_{t-1} + \nabla(g - g_1)(x_t) & \text{with prob. } p_2 \\ \quad -\nabla(g - g_1)(x_{t-1}), & \\ \nabla(h - h_1)(x_t), & \text{with prob. } 1 - p_1 - p_2 \end{cases}$$

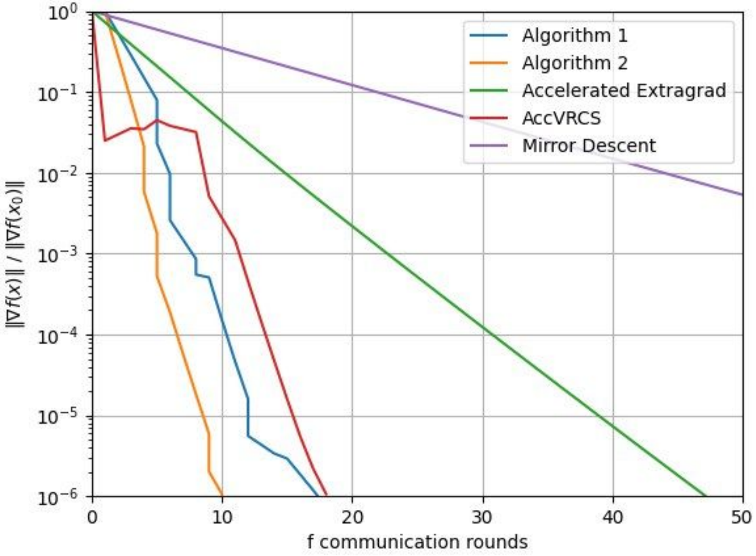
Update $x_{t+1} = \arg\min_{x \in R^d} A_t^\theta(x)$, where

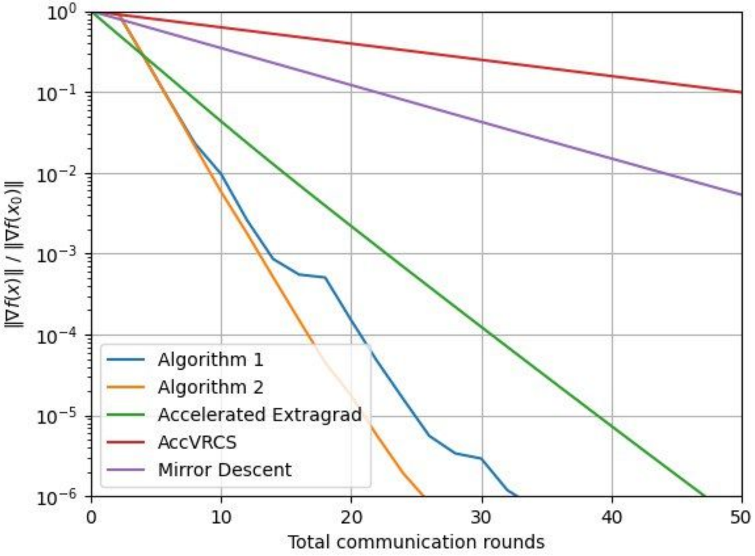
$$A_t^\theta(x) = \langle g_t, x \rangle + \frac{1}{2\theta}\|x - x_t\|^2 + h_1(x)$$

end for
Output: x_T

Начали оформлять доказательство в tex

Эксперименты: начало





Эксперименты: будущее

- Запустить ResNET на датасете CiFAR
- LoRA-адаптеры (идея)

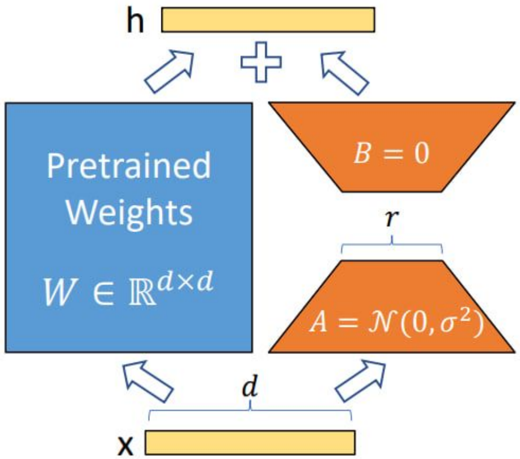


Figure 1: Our reparametrization. We only train A and B.

Спасибо за внимание!