Ускоренные методы для задач федеративного обучения в выпуклом и сильно выпуклом случае. Методы редукции дисперсии и адаптивные моментные методы для невыпуклых задач федеративного обучения. Невыпуклая оптимизация.

Современные методы теории информации, оптимизации и управления

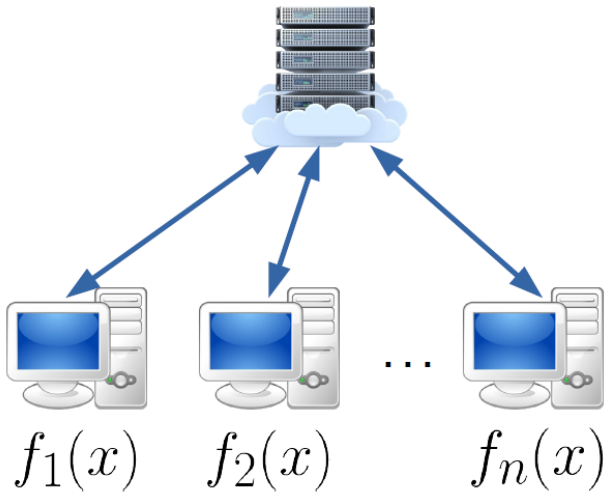
Кураторы проектов: Эдуард Горбунов, Павел Двуреченский
Студенты: Владислав Томинин, Ярослав Томинин, Максим Ткачев, Игорь Соколов, Марина Данилова

Московский Физико-Технический Институт, Институт Проблем Управления РАН

22 августа 2020

Задача федеративного обучения

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$



Типичный вид $f_i(x)$:

$$f_i(x) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_{\xi_i}(x)]$$

$$f_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} f_{ij}(x)$$

Классические подходы: Minibatch SGD и Local-SGD

Алгоритм 1 `Minibatch SGD`

Вход: stepsize $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, batchsize $\tau \geq 1$

for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$ in parallel
    Sample $\nabla f_{\xi_i^k(1)}(x_i^k), \ldots, \nabla f_{\xi_i^k(\tau)}(x_i^k)$ independently from other nodes
    $g_i^k = \frac{1}{\tau} \sum_{j=1}^{\tau} f_{\xi_i^k(j)}(x_i^k)$
  end for
  $g^k = \frac{1}{n} \sum_{i=1}^{n} g_i^k$
  $x^{k+1} = x^k - \gamma g^k$
end for $=0$

Алгоритм 2 `Local-SGD`

Вход: stepsize $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, local steps parameter $\tau \geq 1$

for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$ in parallel
    Sample $g_i^k = \nabla f_{\xi_i^k}(x_i^k)$ independently from other nodes
    if $k + 1 \mod \tau = 0$ **then**
    $x_i^{k+1} = x^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \left(x_i^k - \gamma g_i^k\right)$
    иначе
    $x_i^{k+1} = x_i^k - \gamma g_i^k$
    end if
  end for
end for $=0$

Основные случаи

Homogeneous data

$f_i(x) = \ldots = f_n(x)$

Heterogeneous data

$f_i(x)$ не обязательно одинаковы

Project 1: Algorithm

Алгоритм 3 `GAF with Inexact Minimization of` h_k

1: **Input:** x_0 (initial estimate); $\mathcal{M}$ (optimization method); μ (strong convexity constant); κ (parameter for h_k); K (number of iterations); (δ_k) (approximation errors);
2: **Initialization:** $y_0 = x_0; q = \frac{\mu}{\mu+\kappa}$; $\alpha_0 = 1$ if $\mu = 0$ or $\alpha_0 = \sqrt{q}$ if $\mu \neq 0$;
3: **for** $k = 0$, **to** K do
4: Consider a surrogate h_k satisfying $(\mathcal{H}_1), (\mathcal{H}_2)$ and obtain x_k, x_k^* using $\mathcal{M}$ satisfying $(\mathcal{H}_4)$;
5: Compute α_k in $(0, 1)$ by solving the equation $\alpha_k^2 = (1 - \alpha_k)\alpha_{k-1}^2 + q\alpha_k$.
6: Update the extrapolated sequence

$$y_k = x_k^* + \beta_k(x_k^* - x_{k-1}) + \frac{(\kappa + \mu)(1 - \alpha_k)}{\kappa}(x_k - x_k^*) \text{ with } \beta_k = \frac{\alpha_{k-1}(1 - \alpha_{k-1})}{\alpha_{k-1}^2 + \alpha_k} \tag{1}$$

7: **end for**
8: **Output:** x_k (final estimate). $=0$

Project 1: Convergence rate

1. $(\mathcal{H}_1)$ h_k is $(\kappa+\mu)-$ strongly convex, where μ is the strong convexity parameter of f;
2. $(\mathcal{H}_2)$ $\mathbb{E}\left[h_k(x)|\mathcal{F}_{k-1}\right] \leq F(x) + \frac{\kappa}{2}\|x - y_{k-1}\|^2$ for $x = \alpha_{k-1}x^* + (1-\alpha_{k-1})x_{k-1}$, which is deterministic given the past information $\mathcal{F}_{k-1}$ up to iteration $k-1$ and α_{k-1} is given in Alg.3
3. $(\mathcal{H}_3)$ for x_k^*- minimum of h_k exists x_k such as $\mathbb{E}[F(x_k)] \leq \mathbb{E}[h_k^*] + \delta_k$ holds true
4. $(\mathcal{H}_4)$ given $\varepsilon_k \geq 0$, $\mathcal{M}$ can provide a point x_k such that $\mathbb{E}\left[h_k(x_k) - h_k^*\right]$

Theorem

Consider Alg. 3. Then, for any $\gamma \in (0,1]$,

$$\mathbb{E}\left[f(x_k) - f^\star\right] \leq \begin{cases} \left(1-\frac{\sqrt{q}}{2}\right)^k \left(2\left(f(x_0) - f^\star\right) + 4\sum_{j=1}^k \left(1-\frac{\sqrt{q}}{2}\right)^{-j}\left(\delta_j + \frac{\varepsilon_j}{\sqrt{q}}\right)\right) & \text{if } \mu \neq 0 \\ \frac{2e^{1+\gamma}}{(k+1)^2}\left(\kappa\|x_0 - x^\star\|^2 + \sum_{j=1}^k (j+1)^2\delta_j + \frac{(j+1)^{3+\gamma}\varepsilon_j}{\gamma}\right) & \text{if } \mu = 0 \end{cases} \tag{2}$$

Project 1: Catalyst

Catalyst consists in solving approximately a sequence of sub-problems and updating sequences $(x_k)_{k\geq 0}$ and $(y_k)_{k\geq 0}$ by

$$x_k \approx \operatorname*{argmin}_{x\in\mathbb{R}^d}\{h_k(x) \stackrel{\triangle}{=} f(x) + \frac{\kappa}{2}\|x - y_{k-1}\|^2\} \quad \textit{and} \quad y_k = x_k + \beta_k(x_k - x_{k-1}) \tag{3}$$

For h_k from (3) with $\kappa = L - \mu$ the assumptions $(\mathcal{H}_1), (\mathcal{H}_2), (\mathcal{H}_3)$ are satisfied with $\delta_k = 0$.
As the optimization method $\mathcal{M}$ one consider the following methods: Local-SGD, FedAc, Local-SVRG, SCAFFOLD.

Project 1: Our contributions, Homogeneous Case

Algorithm	Homogeneous
Acc. Local-SGD	$\exp\left(-\sqrt{\frac{\mu}{L}}K\right)(f(x_0)-f^*)+\frac{D_1 L^{1/2}}{n\mu^{3/2}K}+\frac{L(\tau-1)D_1}{\mu^2K^2}$
Local-SGD	$\exp\left(-\frac{\mu}{L}K\right)(f(x_0)-f^*)+\frac{D_1}{n\mu K}+\frac{L(\tau-1)D_1}{\mu^2K^2}$
Acc. FedAc	$\exp\left(-\sqrt{\frac{\mu}{L}}K\right)(f(x_0)-f^*)+\frac{D_1 L^{1/2}}{n\mu^{3/2}K}+\frac{L^{3/2}D_1}{\mu^{5/2}KK_c^2}$
FedAc	$\exp\left(\min\left\{-\frac{\mu K}{L},-\sqrt{\frac{\mu K K_c}{L}}\right\}\right)(f(x_0)-f^*)+\frac{D_1}{\mu nK}+\frac{LD_1}{\mu^2KK_c^2}$

Table: Convergence rates in strongly convex case, $\mathbb{E}\left[f\left(x_k\right)-f^\star\right]\leq\ldots$

Our contributions, Heterogeneous Case

Algorithm	Heterogeneous
Acc. Local-SGD	$\tau(f(x_0) - f^*)\exp\left(-\sqrt{\frac{\mu}{L}}K_c\right) + \frac{D_1' L^{1/2}}{\mu^{3/2}\tau K_c} + \frac{L(\tau-1)^2 D_1}{\mu^2\tau^2 K_c^2}$
Local-SGD	$\tau L\left\|x^0 - x^*\right\|^2\exp\left(-\frac{\mu}{L}K_c\right) + \frac{D_1'}{\mu\tau K_c} + \frac{L(\tau-1)^2 D_1}{\mu^2\tau^2 K_c^2}$
Acc. Local-SVRG	$\frac{mL}{\mu}(f(x_0) - f^*)\exp\left(-\sqrt{\frac{\mu}{L}}\frac{K}{m}\right)$
Local-SVRG	$\frac{mL}{\mu}(f(x_0) - f^*)\exp\left(-\frac{\mu}{L}\frac{K}{m}\right)$
Acc. SCAFFOLD	$(f(x_0) - f^*)\exp\left(-p\sqrt{\frac{\mu}{L}}K\right) + \frac{D_1 L^{1/2}}{\mu^{3/2} nK} + \frac{L(1-p)D_1}{p^2\mu^2 K^2}$
SCAFFOLD	$\frac{LR^2}{p}\exp\left(-p\frac{\mu}{L}K\right) + \frac{D_1}{\mu nK} + \frac{L(1-p)D_1}{p^2\mu^2 K^2}$

Table: Convergence rates in strongly convex case, $\mathbb{E}\left[f(x_k) - f^\star\right] \leq \ldots$

Project 2: FedAdam

Алгоритм 4 FedAdam

Input: $x_0, v_{-1} \geq \tau^2$, $\beta_1, \beta_2 \in [0, 1)$
for $t = 0, \cdots, T-1$
  Sample a subset $\mathcal{S}$ of clients
  $x_i^t = x_t$
  for each client $i \in \mathcal{S}$ **in parallel**
    for $e = 1, \ldots, E$
      for $b \in \mathbf{B}_i$
        $x_i^t = x_i^t - \eta_l \nabla f_i(x_i^t; b)$
      end for
    end for
    $\Delta_i^t = x_i^t - x_t$
  end for
  $n = \sum_{i \in \mathcal{S}} n_i$, $\Delta_t = \sum_{i \in \mathcal{S}} \frac{n_i}{n} \Delta_i^t$
  $\Delta_t = \beta_1 \Delta_{t-1} + (1 - \beta_1)\Delta_t$
  $v_t = \beta_2 v_{t-1} + (1 - \beta_2)\Delta_t^2$
  $x_{t+1} = x_t + \eta \frac{\Delta_t}{\sqrt{v_t} + \tau}$
end for =0

Project 2: Local-Adam

Алгоритм 5 Local-Adam

Вход: stepsize $\gamma > 0$, initial vector $x^{\mathbf{0}} \in \mathbb{R}^d$, $v^{\mathbf{0}} = m^{\mathbf{0}} = 0 \in \mathbb{R}^d$, local steps parameter $\tau \geq 1$, $\beta_{\mathbf{1}} = 0.9$, $\beta_{\mathbf{2}} = 0.99$

for $k = 0, 1, \ldots$
  for $i = 1, \ldots, n$ in parallel
    Sample $g_i^k = \nabla f_{\xi_i^k}(x_i^k)$ independently from other nodes
    if $k + 1 \mod \tau = 0$ **then**
    $\mathbf{m}_{ij}^{k+\mathbf{1}} = \mathbf{m}_j^{k+\mathbf{1}} = \frac{\mathbf{1}}{n} \sum_{i=\mathbf{1}}^n \mathbf{m}_{ij}^k, \quad j = \overline{1, d}$
    $\mathbf{v}_{ij}^{k+\mathbf{1}} = \mathbf{v}_j^{k+\mathbf{1}} = \frac{\mathbf{1}}{n} \sum_{i=\mathbf{1}}^n \mathbf{v}_{ij}^k, \quad j = \overline{1, d}$
    $\mathbf{x}_{ij}^{k+\mathbf{1}} = \mathbf{x}_j^{k+\mathbf{1}} = \frac{\mathbf{1}}{n} \sum_{i=\mathbf{1}}^n \mathbf{x}_{ij}^k, \quad j = \overline{1, d}$
    иначе
    $\mathbf{m}_{ij}^k = \beta_{\mathbf{1}} \mathbf{m}_{ij}^{k-\mathbf{1}} + (1 - \beta_{\mathbf{1}}) \mathbf{g}_{ij}^k, \quad \hat{\mathbf{m}}_{ij}^k = \frac{\mathbf{m}_{ij}^k}{\mathbf{1} - (\beta_{\mathbf{1}})^k}$
    $\mathbf{v}_{ij}^k = \beta_{\mathbf{2}} \mathbf{v}_{ij}^{k-\mathbf{1}} + (1 - \beta_{\mathbf{2}}) \left(\mathbf{g}_{ij}^k\right)^{\mathbf{2}}, \quad \hat{\mathbf{v}}_{ij}^k = \frac{\mathbf{v}_{ij}^k}{\mathbf{1} - (\beta_{\mathbf{2}})^k}$
    $\mathbf{x}_{ij}^{k+\mathbf{1}} = \mathbf{x}_{ij}^k - \frac{\gamma}{\sqrt{\hat{\mathbf{v}}_{ij}^k + \delta}} \hat{\mathbf{m}}_{ij}^k, \quad j = \overline{1, d}, \quad \delta = 10^{-\mathbf{8}}$
    end if
  end for
end for $=0$

Project 2: Local-SPIDER

Алгоритм 6 Local-SPIDER

Input: **Initial iterate** $\mathbf{x}_{i,m}^{\mathbf{0}} = \mathbf{x}^{\mathbf{0}}, \mathbf{v}_{i,m}^{\mathbf{0}} = \nabla f\left(\mathbf{x}^{\mathbf{0}}\right)\ \ i \in [\mathbf{1} : N]$

for $s = \mathbf{0}$ **to** $S - \mathbf{1}$

  $\mathbf{x}_{i,\mathbf{0}}^{s+\mathbf{1}} = \mathbf{x}_{i,m}^{s},\ \ i \in [\mathbf{1} : N]$

  $\mathbf{v}_{i,\mathbf{0}}^{s+\mathbf{1}} = \mathbf{v}_{i,m}^{s},\ \ i \in [\mathbf{1} : N]$

  $\mathbf{x}_{i,\mathbf{1}}^{s+\mathbf{1}} = \mathbf{x}_{i,\mathbf{0}}^{s+\mathbf{1}} - \gamma \mathbf{v}_{i,\mathbf{0}}^{s+\mathbf{1}}$

  for $t = \mathbf{1}$ **to** $m - \mathbf{1}$

    $\mathbf{v}_{i,t}^{s+\mathbf{1}} = \frac{\mathbf{1}}{B} \sum_{\xi_{i,t}^{s+\mathbf{1}} \in \mathcal{B}_{i,t}^{s+\mathbf{1}}} \left[\nabla f_i(\mathbf{x}_{i,t}^{s+\mathbf{1}}; \xi_{i,t}^{s+\mathbf{1}}) - \nabla f_i(\mathbf{x}_{i,t-\mathbf{1}}^{s+\mathbf{1}}; \xi_{i,t}^{s+\mathbf{1}})\right] + \mathbf{v}_{i,t-\mathbf{1}}^{s+\mathbf{1}},\ \ i \in [\mathbf{1} : N] \qquad (|\mathcal{B}_{i,t}^{s+\mathbf{1}}| = B)$

    if t **mod** $l = \mathbf{0}$ **then**

      $\mathbf{x}_{i,t}^{s+\mathbf{1}} = \bar{\mathbf{x}}_t^{s+\mathbf{1}} \triangleq \frac{\mathbf{1}}{N} \sum_{j=\mathbf{1}}^{N} \mathbf{x}_{j,t}^{s+\mathbf{1}},\ \ i \in [\mathbf{1} : N]$

      $\mathbf{v}_{i,t}^{s+\mathbf{1}} = \bar{\mathbf{v}}_t^{s+\mathbf{1}} \triangleq \frac{\mathbf{1}}{N} \sum_{j=\mathbf{1}}^{N} \mathbf{v}_{j,t}^{s+\mathbf{1}},\ \ i \in [\mathbf{1} : N]$

    end if

    $\mathbf{x}_{i,t+\mathbf{1}}^{s+\mathbf{1}} = \mathbf{x}_{i,t}^{s+\mathbf{1}} - \gamma \mathbf{v}_{i,t}^{s+\mathbf{1}},\ \ i \in [\mathbf{1} : N]$

  end for

  if $s < S - \mathbf{1}$ **then**

    $\bar{\mathbf{x}}_m^{s+\mathbf{1}} = \frac{\mathbf{1}}{N} \sum_{j=\mathbf{1}}^{N} \mathbf{x}_{j,m}^{s+\mathbf{1}}$

    $\mathbf{x}_{i,m}^{s+\mathbf{1}} = \bar{\mathbf{x}}_m^{s+\mathbf{1}},\ \ i \in [\mathbf{1} : N]$

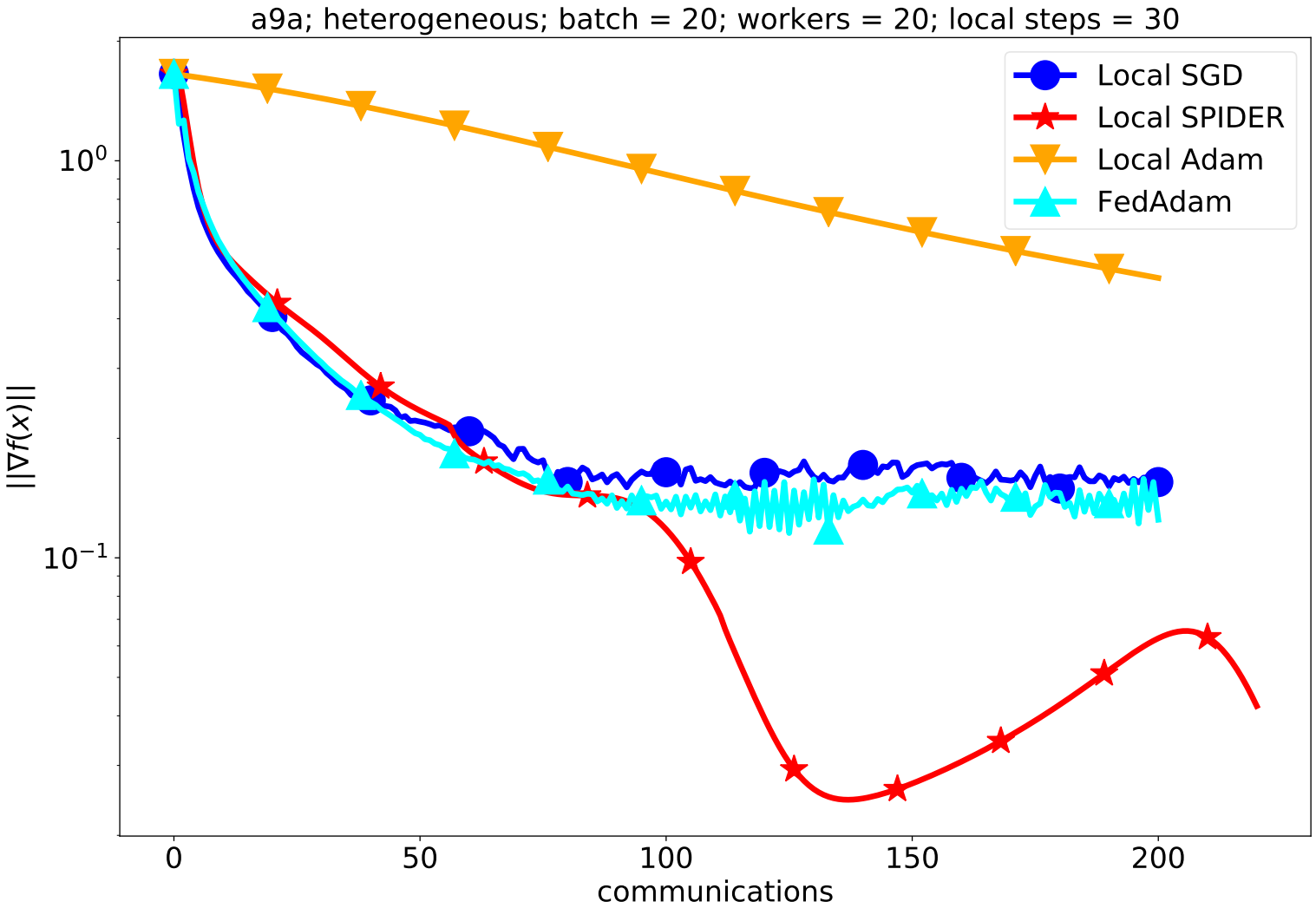
    $\bar{\mathbf{v}}_m^{s+\mathbf{1}} = \frac{\mathbf{1}}{N} \sum_{j=\mathbf{1}}^{N} \nabla f_j\left(\mathbf{x}_{j,m}^{s+\mathbf{1}}\right) = \nabla f\left(\bar{\mathbf{x}}_m^{s+\mathbf{1}}\right)$

    $\mathbf{v}_{i,m}^{s+\mathbf{1}} = \bar{\mathbf{v}}_m^{s+\mathbf{1}},\ \ i \in [\mathbf{1} : N]$

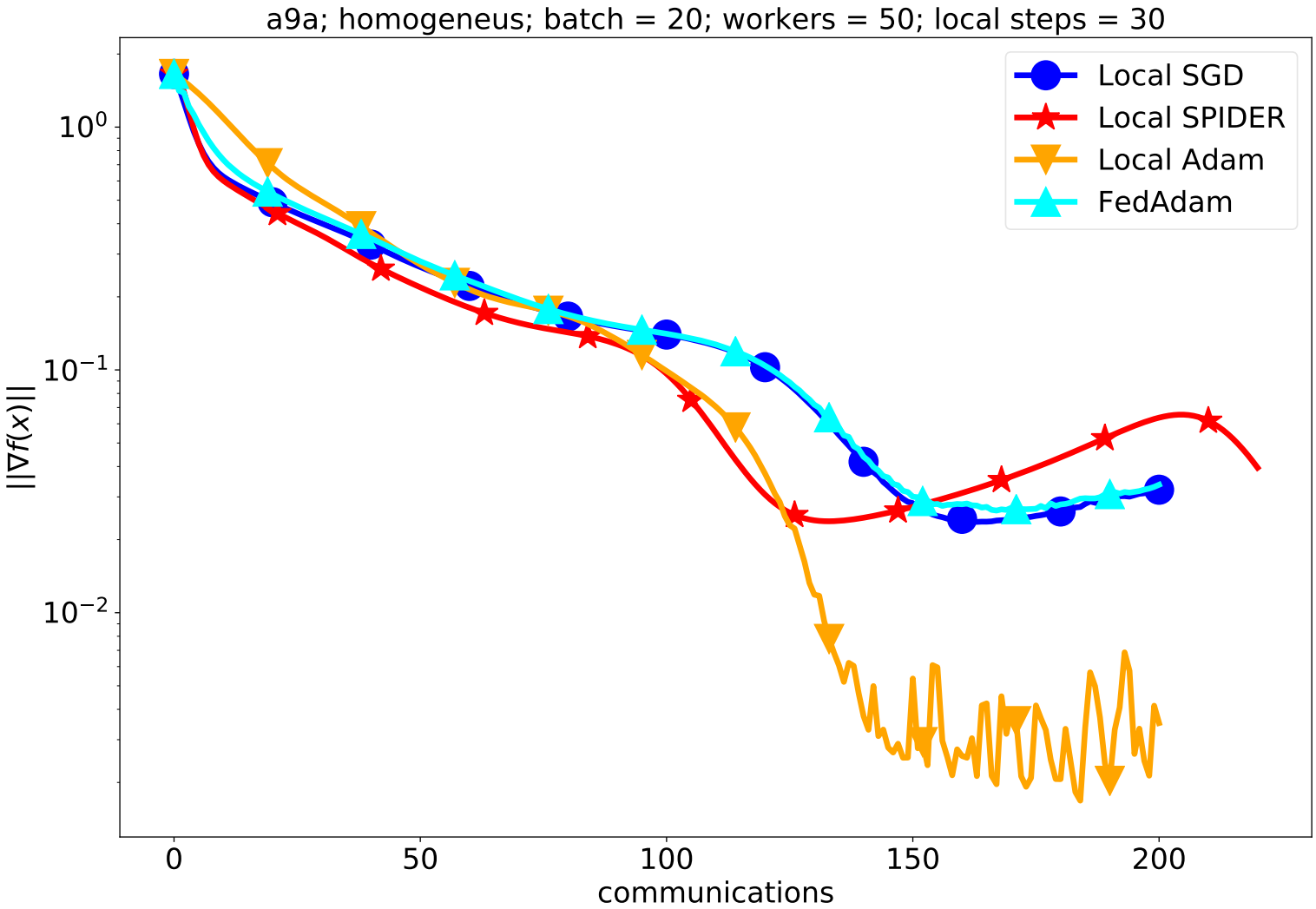
  end if

end for $=\mathbf{0}$

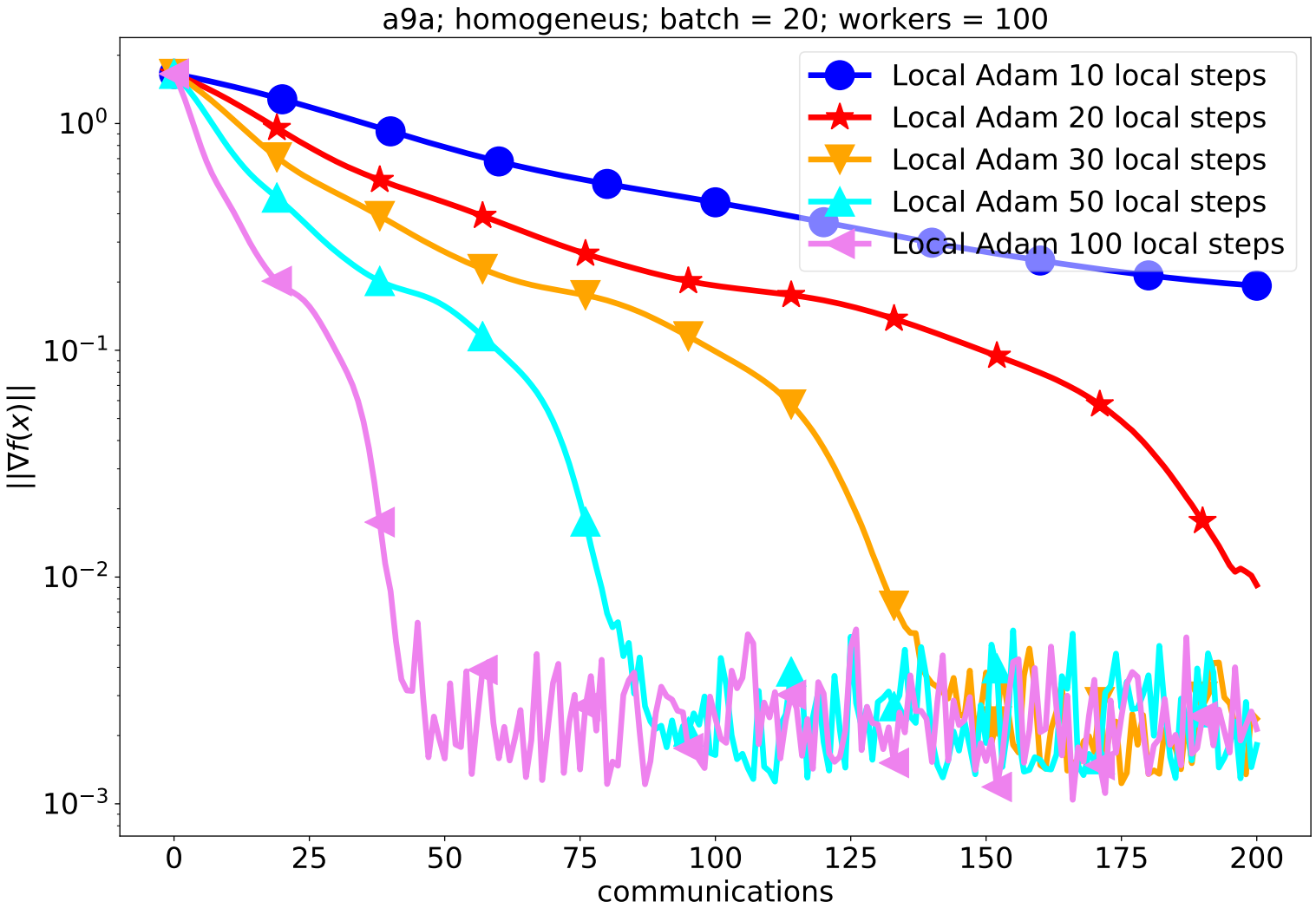
Project 2: Results



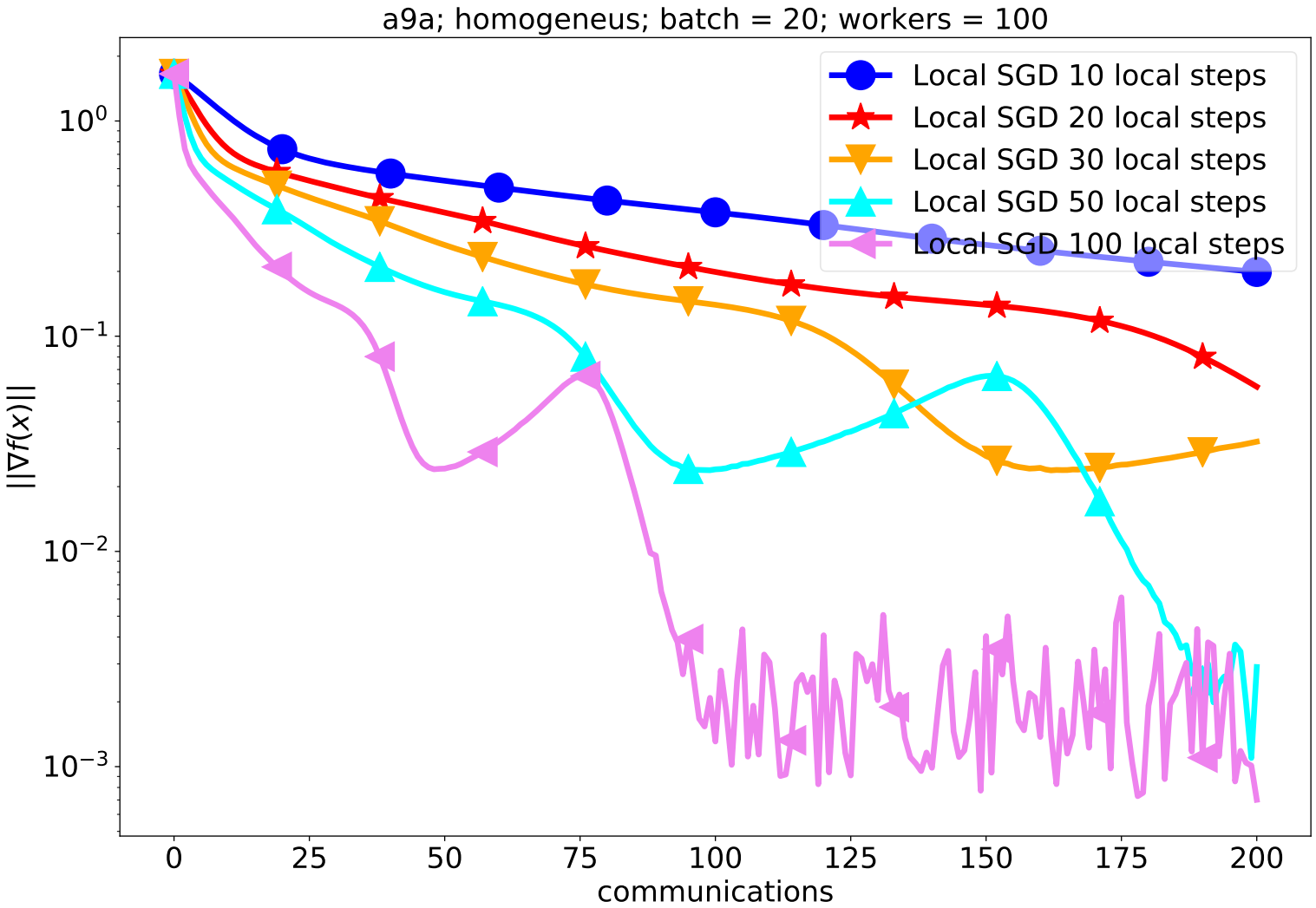
Project 2: Results



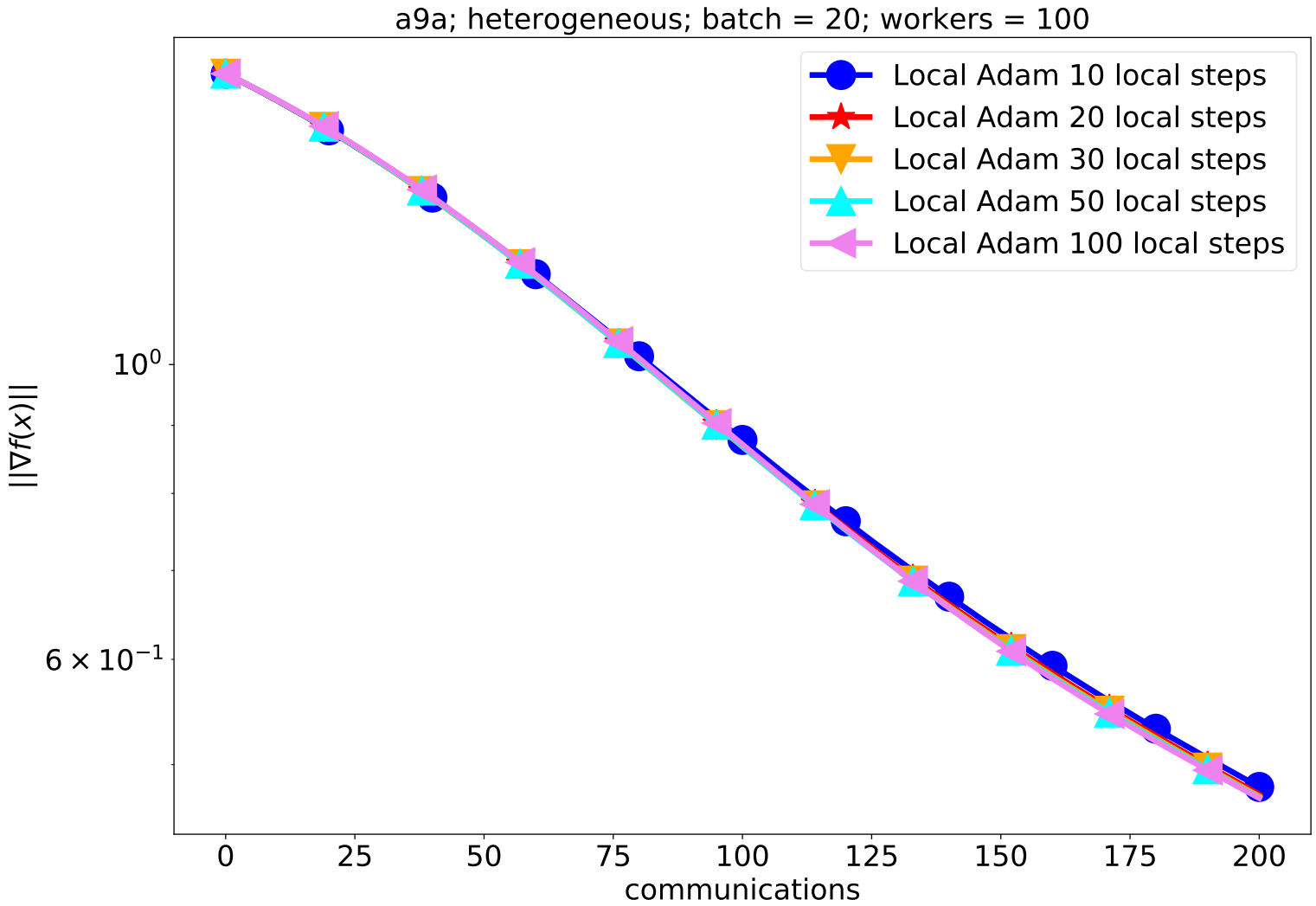
Project 2: Results



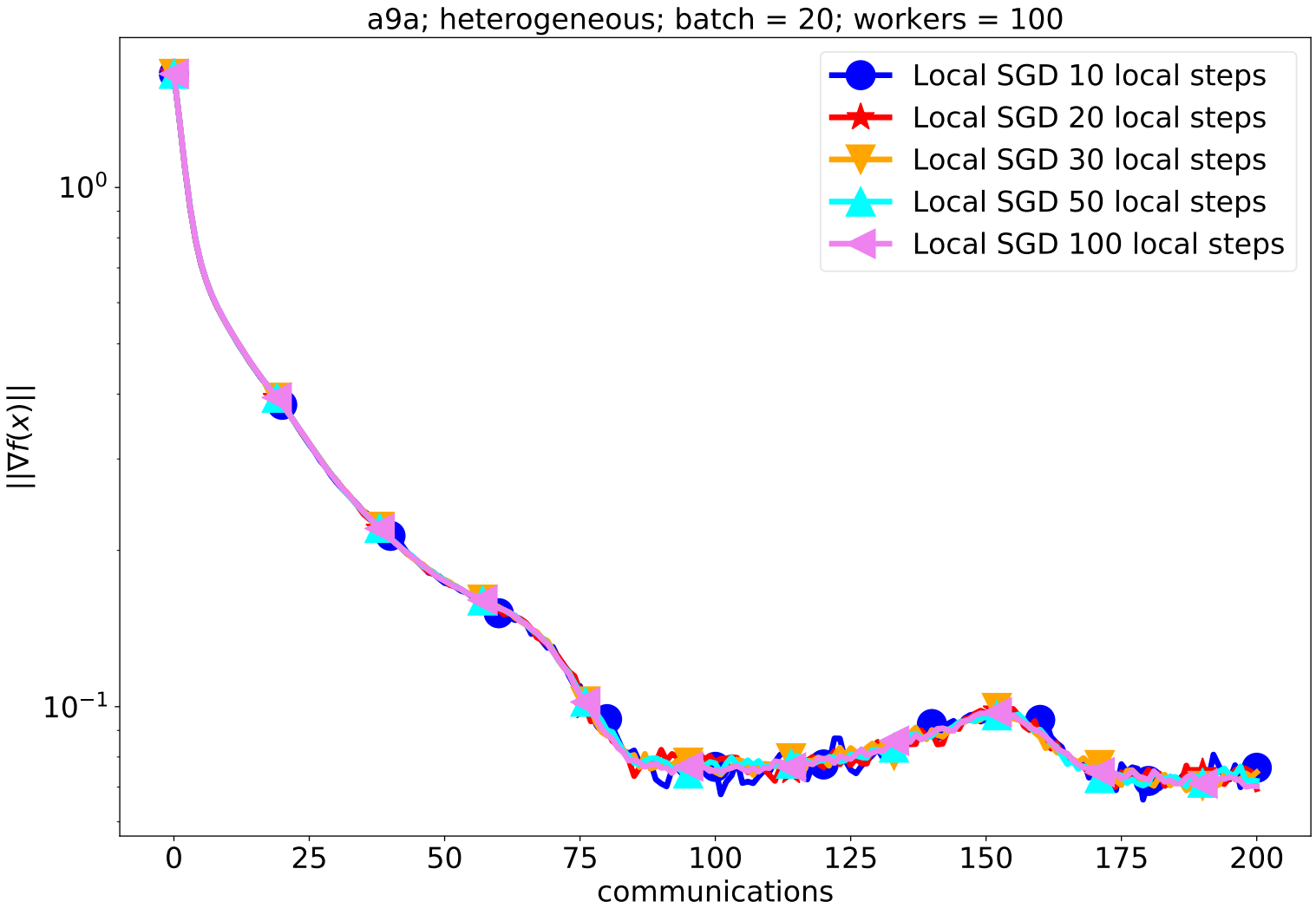
Project 2: Results



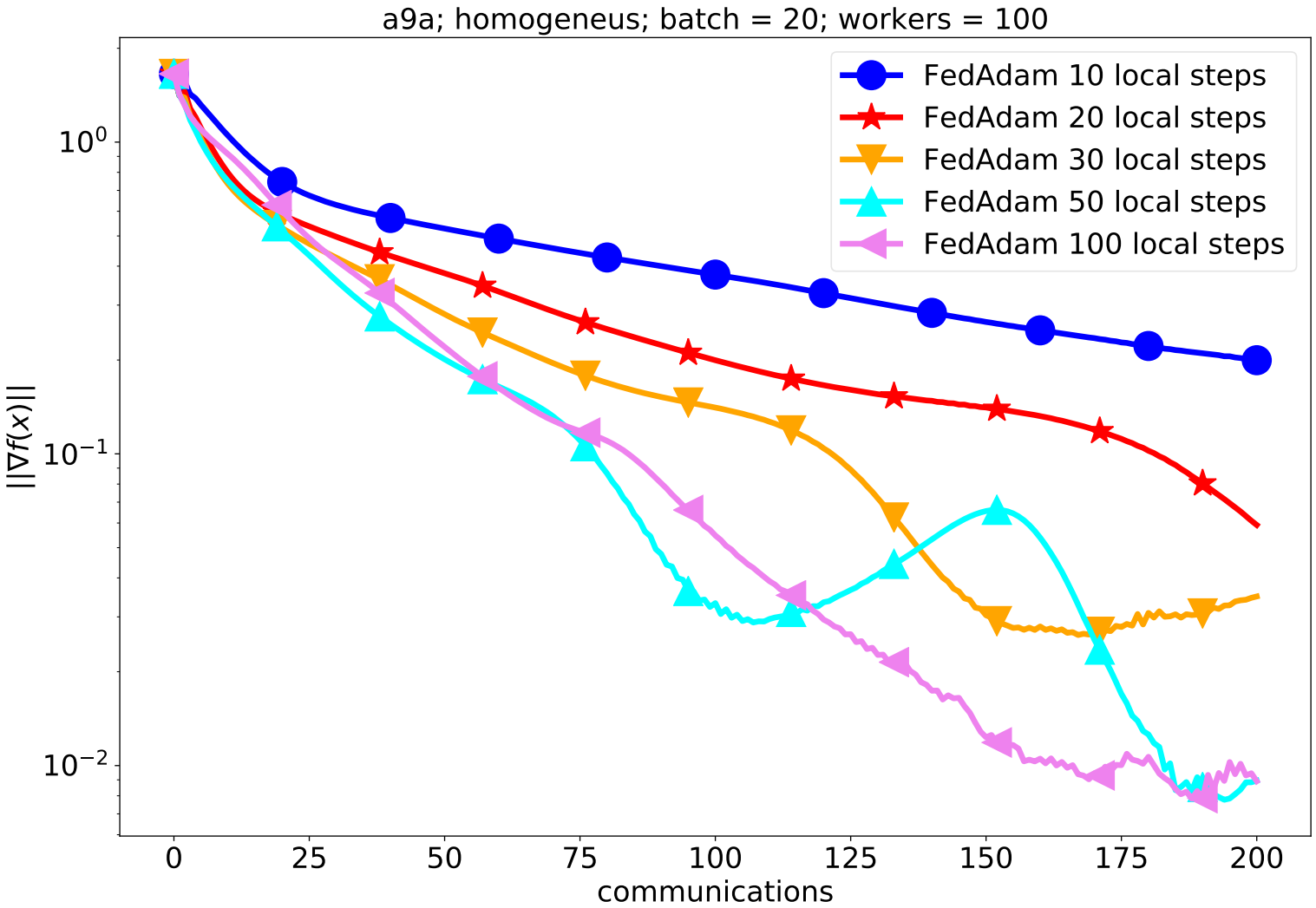
Project 2: Results



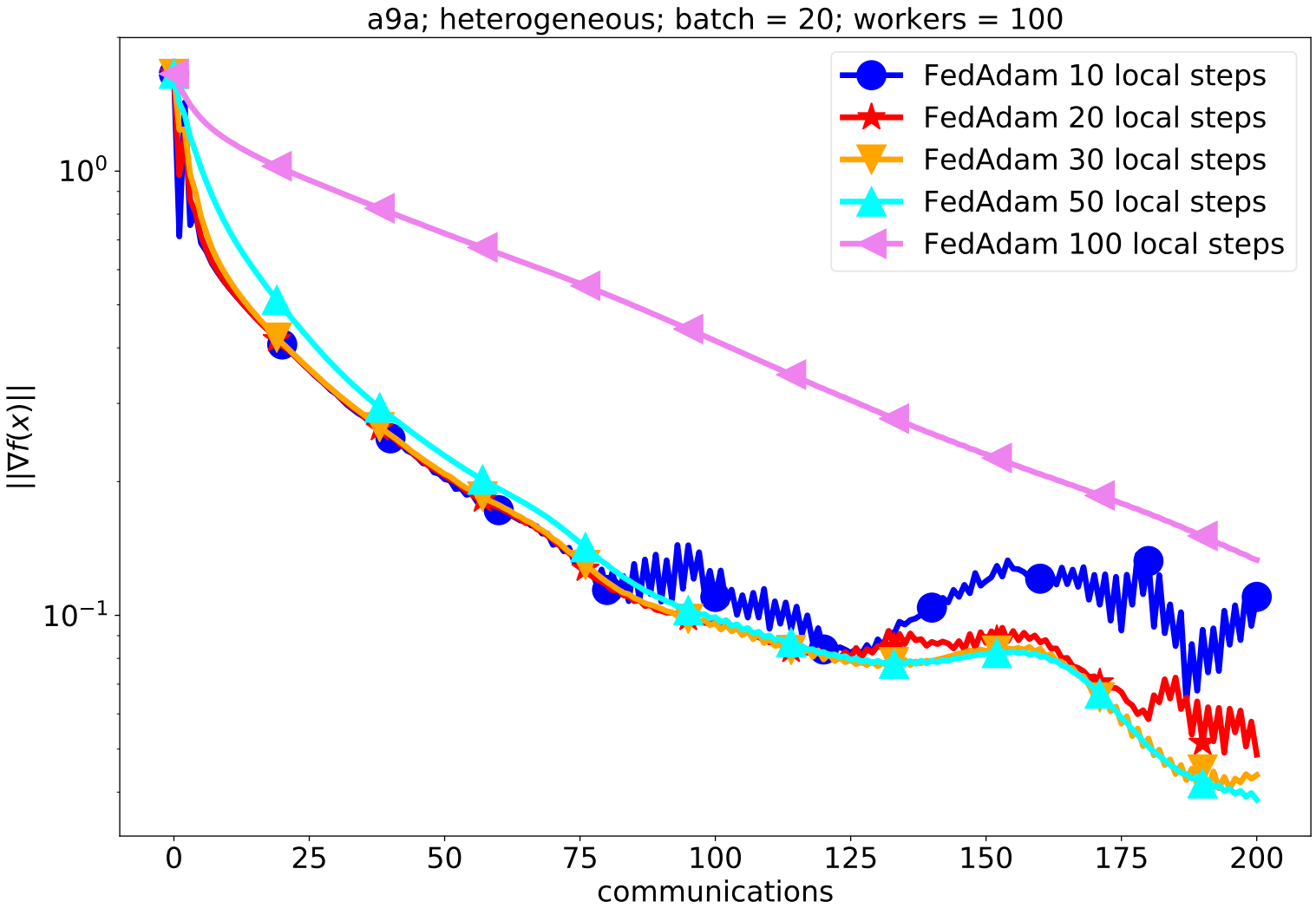
Project 2: Results



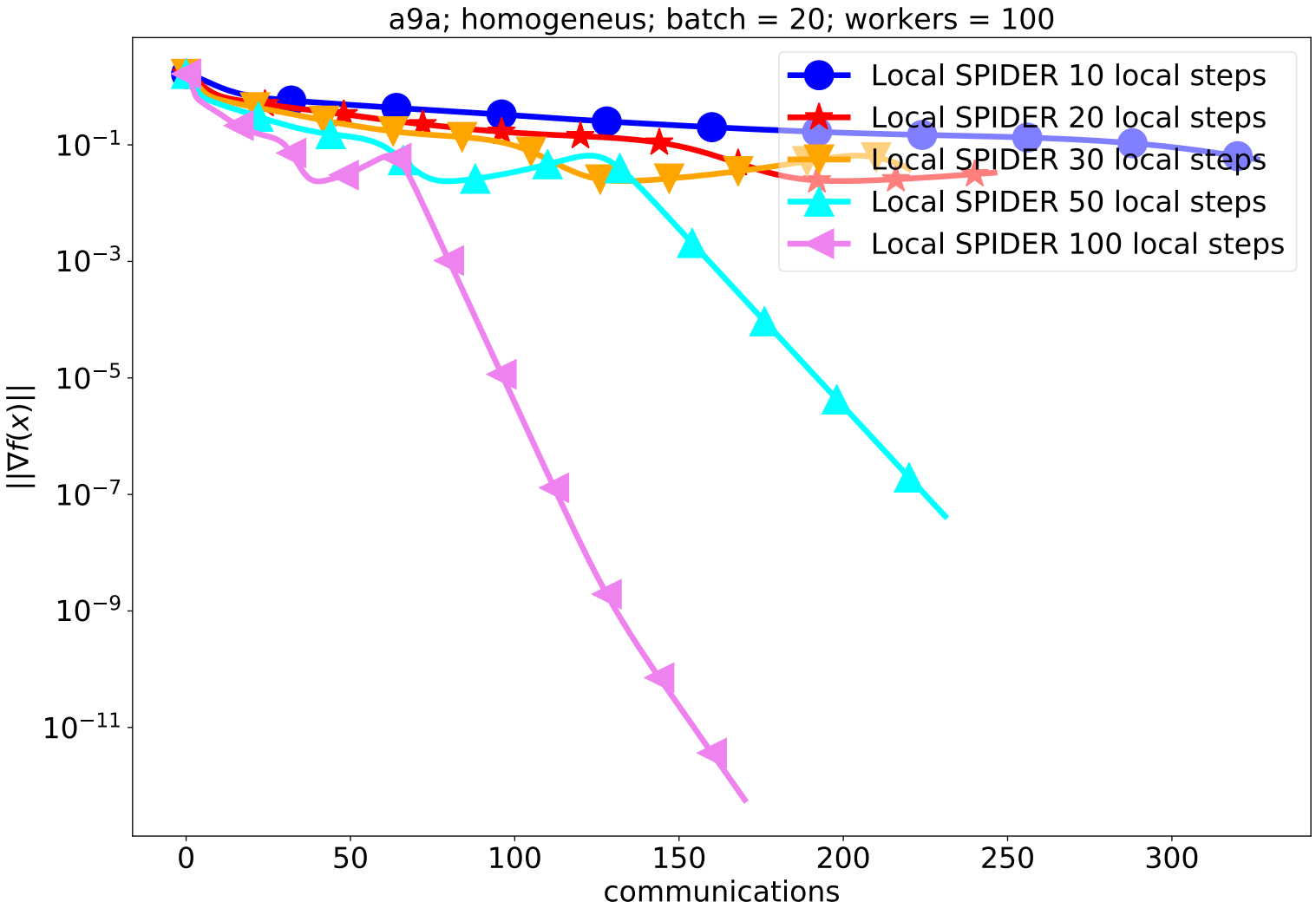
Project 2: Results



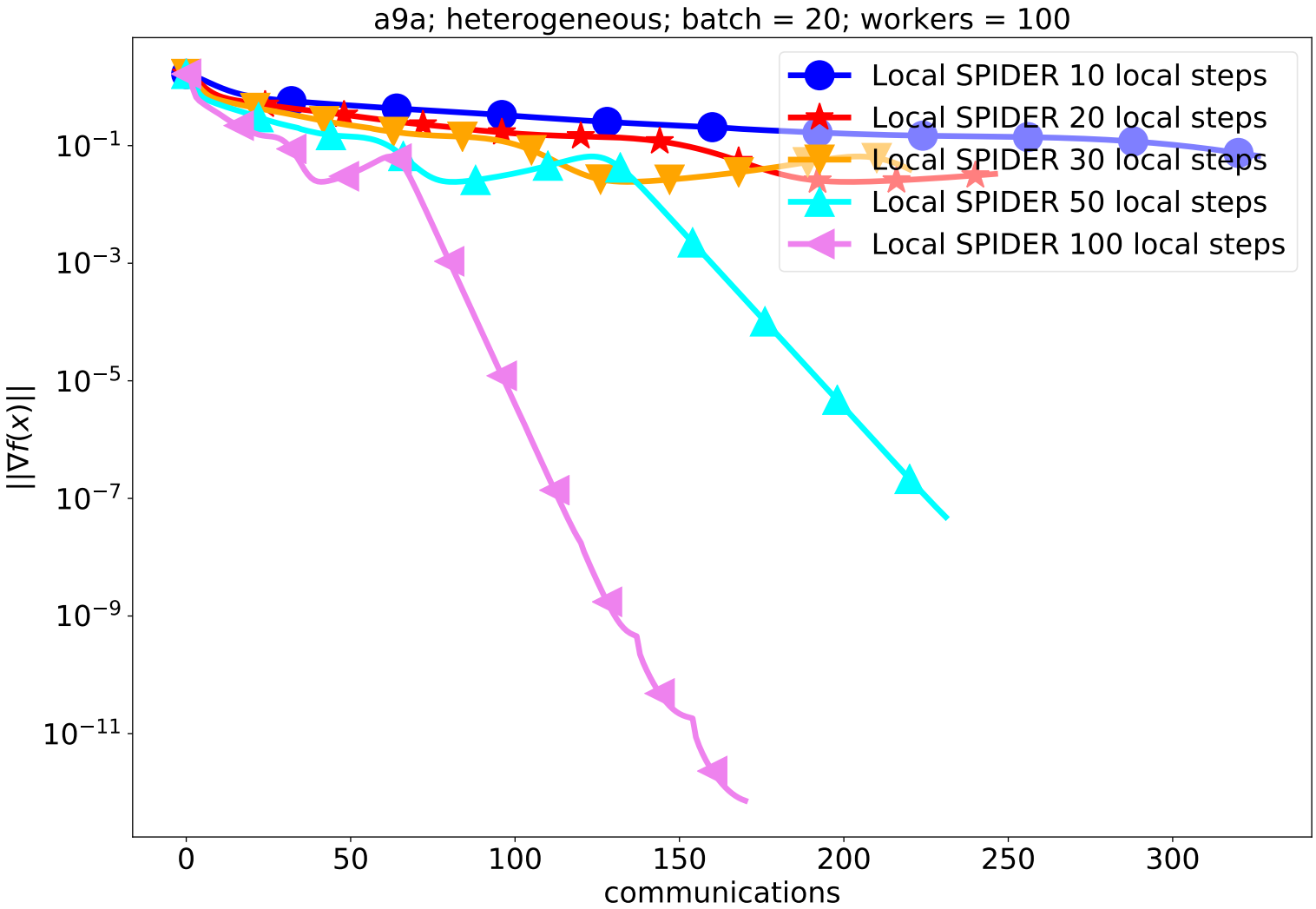
Project 2: Results



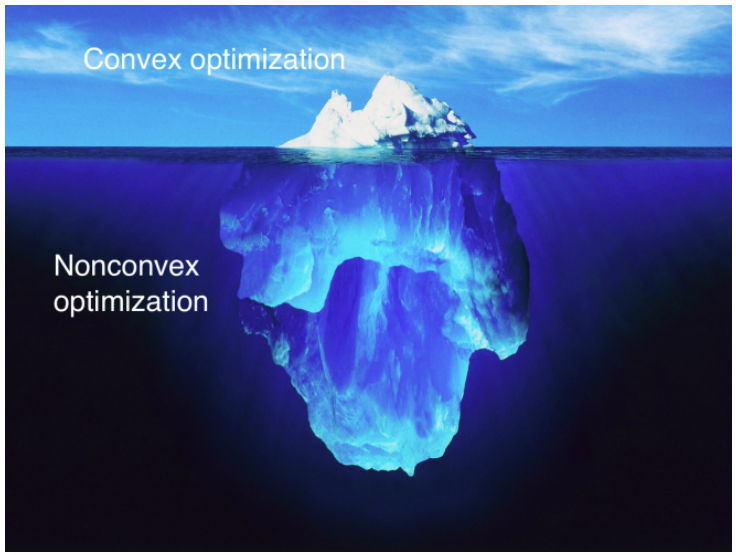
Project 2: Results



Project 2: Results



Project 3: An Overview of Nonconvex Optimization



Nonconvex optimization problems are ubiquitous in different areas. Since they have exponential in the dimension complexity, one can only guarantee convergence to a stationary point or that the method escapes saddle-point stationary points. Nevertheless, significant progress has been made recently on developing provably accurate and efficient algorithms. In this overview, we survey recent advances in global nonconvex optimization theory.

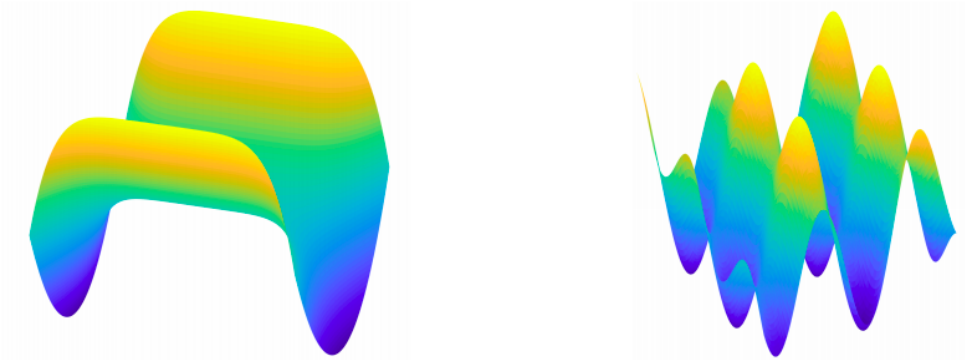
Project 3: An Overview of Nonconvex Optimization

Convergence to a local minimum

Following Yu. Nesterov and B. Polyak (2006), let's call a point x an (ε, δ) local minimum if

$$\|\nabla f(x)\|_2 \leq \varepsilon, \quad \lambda_{\min}\left(\nabla^2 f(x)\right) \geq -\delta,$$

that is, if x is an ε-stationary point and if $\nabla^2 f(x) + \delta I$ is positive semidefinite.



(a) Flat saddle points

(b) Spurious local minimizers

Figure: : Two worst case scenarios. Local methods can become trapped near local minimizers (left) or stagnate near flat saddle points (right).

Project 3: An Overview of Nonconvex Optimization

Nonconvex Problems

1. Problems with Hidden Convexity or Analytic Solutions
2. Problems with Provable Global Results
    1. Matrix Completion/Sensing
    2. Tensor Recovery/Decomposition
    3. Hidden Variable Models
    4. Phase Retrieval

Algorithms

1. Zero-order methods
    1. Deterministic methods
    2. Stochastic methods
2. First-order methods
    1. Deterministic methods
    2. Stochastic methods
    3. Structured nonconvex problems
3. Higher-order methods

Полезные ссылки

- Результаты работы над проектом 1 будут включены в одну из готовящихся статей. Со временем ссылка на статью появится вот тут: `https://eduardgorbunov.github.io/`.
- Код по проекту 2 доступен по ссылке: `https://github.com/mstkachev/Sirius_project_optimization`.
- Результаты работы над проектом 3 будут включены в обзор по невыпуклой оптимизации, который планируется завершить к середине-конце ноября 2020 года. Ссылка на обзор будет доступна вот тут: `https://marinadanya.github.io/`.