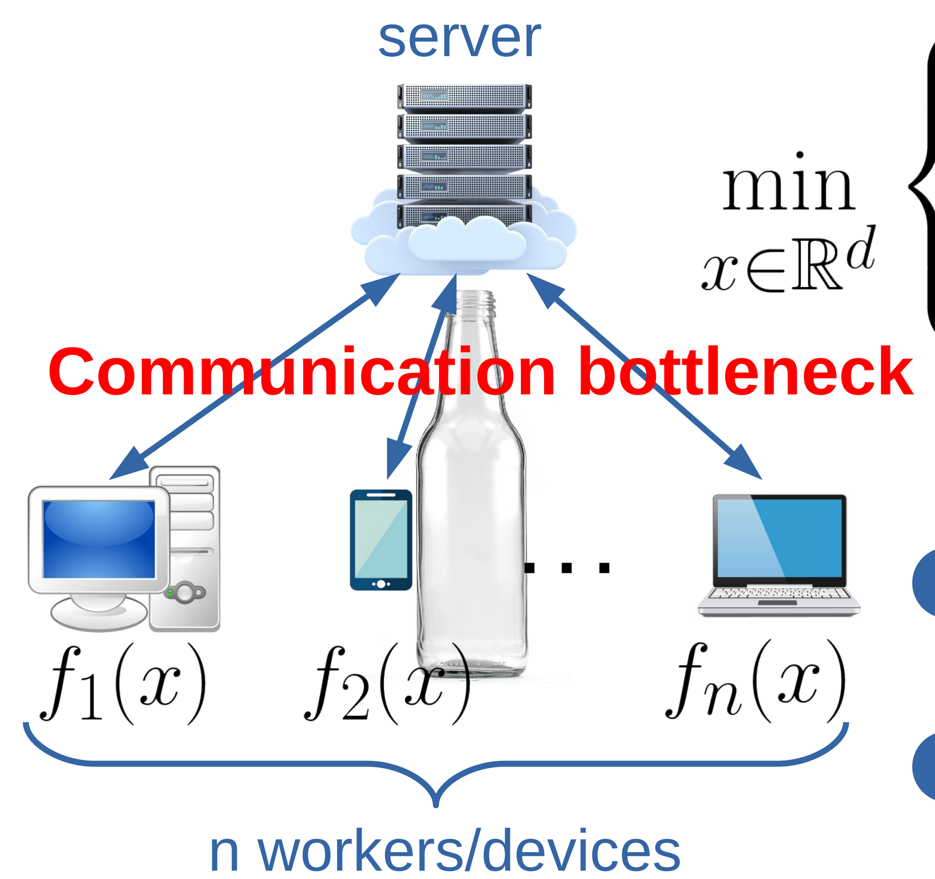
Local SGD: Unified Theory and New Efficient Methods

Eduard Gorbunov[1,2,3], Filip Hanzely[4], Peter Richtárik[3]
[1]MIPT (Russia), [2]Yandex (Russia), [3]KAUST (Saudi Arabia), [4]TTIC (United States)

1. The Problem



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

Options for $f_i(x)$:

- $f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right]$
- $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$

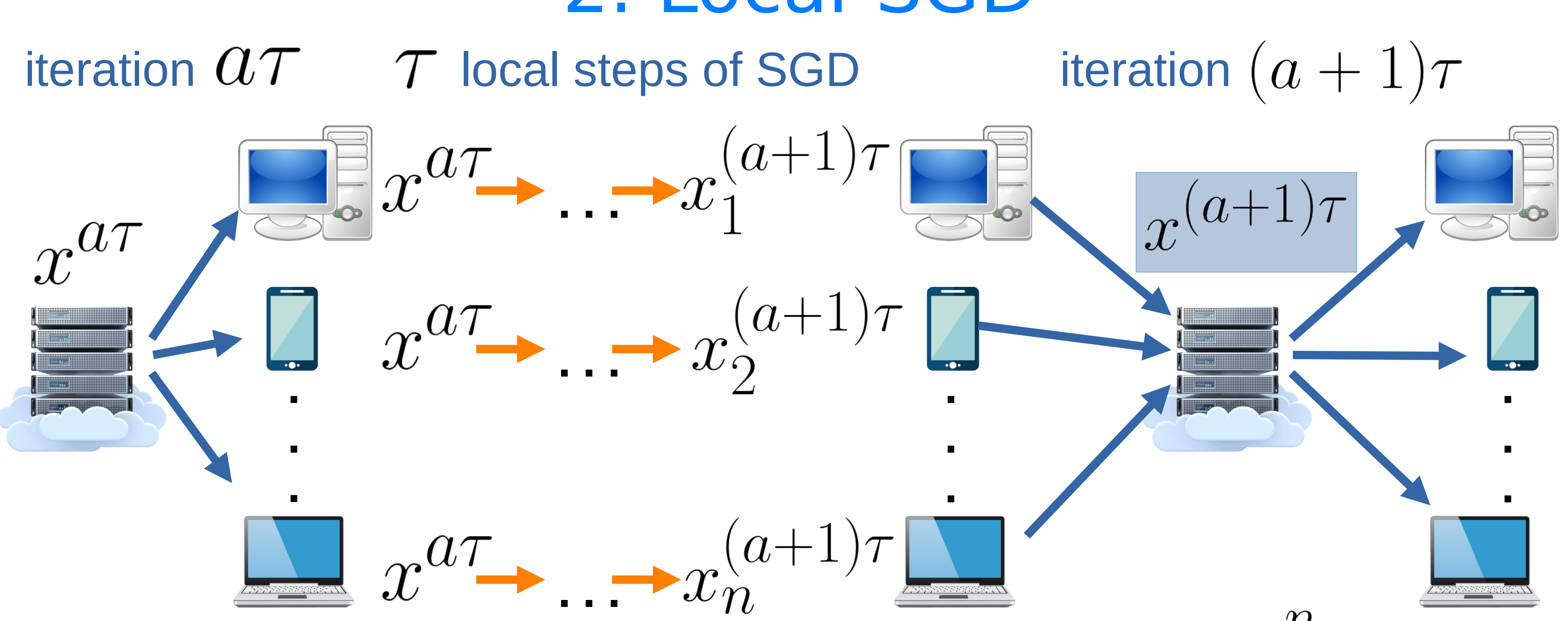
Problem: Distributed optimization / training, where n workers (devices/clients) jointly solve a problem by communicating with a central server.

Assumptions: Smoothness and quasi-strong convexity of local loss functions:

$$\|\nabla f_i(x) - \nabla f_i(y)\| \le L\|x-y\|$$

$$f_i(x^*) \ge f_i(x) + \langle\nabla f_i(x), x^* - x\rangle + \frac{\mu}{2}\|x - x^*\|^2$$

2. Local-SGD



$$x^{(a+1)\tau} = \frac{1}{n}\sum_{i=1}^n x_i^{(a+1)\tau}$$

✓ A lot of results are already known...

✗ ... but many fruitful directions are **unexplored**

- better understanding of the local shifts
- importance sampling
- variance reduction
- variable number of local steps
- general theory for multiple data similarity types

3. Our Contributions

- General theoretical framework for local first-order methods covering
  - local shifts
  - importance sampling
  - variance reduction
  - variable number of local steps
- Our framework
  - recovers tight rates for known optimizers
  - fills missing gaps for known methods
  - extends the established optimizers
- New efficient methods
  - S-Local-SVRG – the first linearly converging stochastic method with local updates in **heterogeneous case**

4. The Framework

$$x_i^{k+1} = \begin{cases} x_i^k - \gamma g_i^k, & \text{if } c_{k+1} = 0\\ \frac{1}{n}\sum_{i=1}^n\left(x_i^k - \gamma g_i^k\right), & \text{if } c_{k+1} = 1\end{cases}$$

encodes the time of communication

Examples of $\{c_k\}_{k\ge 0}$:

Constant local loop

$$c_k = \begin{cases}0, & \text{if } k \mod \tau \ne 0\\ 1, & \text{if } k \mod \tau = 0\end{cases}$$

Random local loop

$$c_k = \begin{cases}0, & \text{with prob. } 1-p\\ 1, & \text{with prob. } p\end{cases}$$

allows variable number of local steps

The assumption below covers a very broad class of methods.

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[g_i^k \mid x_1^k,\ldots,x_n^k\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i\left(x_i^k\right) \qquad V_k = \frac{1}{n}\sum_{i=1}^n\left\|x_i^k - x^k\right\|^2$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2\right] \le 2A\mathbf{E}\left[f\left(x^k\right) - f\left(x^*\right)\right] + B\mathbf{E}\left[\sigma_k^2\right] + F\mathbf{E}\left[V_k\right] + D_1$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] \le 2A'\mathbf{E}\left[f\left(x^k\right) - f\left(x^*\right)\right] + B'\mathbf{E}\left[\sigma_k^2\right] + F'\mathbf{E}\left[V_k\right] + D_1'$$

$$\mathbf{E}\left[\sigma_{k+1}^2\right] \le (1-\rho)\mathbf{E}\left[\sigma_k^2\right] + 2C\mathbf{E}\left[f\left(x^k\right) - f\left(x^*\right)\right] + G\mathbf{E}\left[V_k\right] + D_2$$

$$2L\sum_{k=0}^K w_k\mathbf{E}\left[V_k\right] \le \frac{1}{2}\sum_{k=0}^K w_k\mathbf{E}\left[f\left(x^k\right) - f\left(x^*\right)\right] + 2LH\mathbf{E}\sigma_0^2 + 2LD_3\gamma^2 W_K$$

weights: $w_k = \frac{1}{\left(1 - \min\left\{\gamma\mu, \frac{\rho}{4}\right\}\right)^{k+1}}$ $\qquad W_K = \sum_{k=0}^K w_k$

Reflects smoothness properties of the problem and noises introduced by stochastic gradients and functions dissimilarity

Describes the process of local shifts' learning and variance reduction

Bounds the workers iterates' discrepancy

5. Main Theorem: Simplified Version

$$\bar{x}^K = \frac{1}{W_K}\sum_{k=0}^K w_k x^k$$

depends only on the starting point and stepsize

linear function

$$\mathbf{E}\left[f(\bar{x}^K)\right] - f(x^*) \le \left(1 - \min\left\{\gamma\mu, \frac{\rho}{4}\right\}\right)^K\frac{\Phi^0(x^0,\gamma)}{\gamma} + \gamma\Psi^0(D_1', D_2, D_3)$$

6. New Method: Shifted Local-SVRG

Finite-sum case: $f_i(x) = \frac{1}{m}\sum_{j=1}^m f_{ij}(x)$

$$x_i^{k+1} = \begin{cases} x_i^k - \gamma g_i^k, & \text{with prob. } 1-p\\ \frac{1}{n}\sum_{i=1}^n\left(x_i^k - \gamma g_i^k\right), & \text{with prob. } p\end{cases}$$

$$g_i^k = \nabla f_{i,\boxed{j_i}}(x_i^k) - \nabla f_{i,\boxed{j_i}}\left(\boxed{y^k}\right) + \nabla f\left(\boxed{y^k}\right)$$

$\boxed{j_i} \sim \{1,\ldots,m\}$ uniformly at random

$$\boxed{y^{k+1}} = \begin{cases} x^k, & \text{with prob. } \boxed{q}\\ \boxed{y^k}, & \text{with prob. } 1-q\end{cases} \qquad \boxed{q} = \frac{1}{m}$$

Iteration complexity: $\mathcal{O}\left(\left(m + \frac{L}{p\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{(1-p)L\max L_{ij}}}{p\mu}\right)\log\frac{1}{\varepsilon}\right)$

The first linearly converging local method for heterogeneous data!

It is just an example. In fact, our approach covers a lot of different setups, methods and even the algorithms without local updates.

7. Numerical Experiments

We conducted several numerical experiments on logistic regression problem with L_2-regularization

$$f_i(x) = \frac{1}{m}\sum_{j=1}^m \log\left(1 + \exp\left(\langle a_{(i-1)m+j}, x\rangle \cdot b_{(i-1)m+j}\right)\right) + \frac{\mu}{2}\|x\|^2$$