Distributed Methods with Absolute Compression and Error Compensation

Marina Danilova
ICS RAS, MIPT

Eduard Gorbunov
MIPT

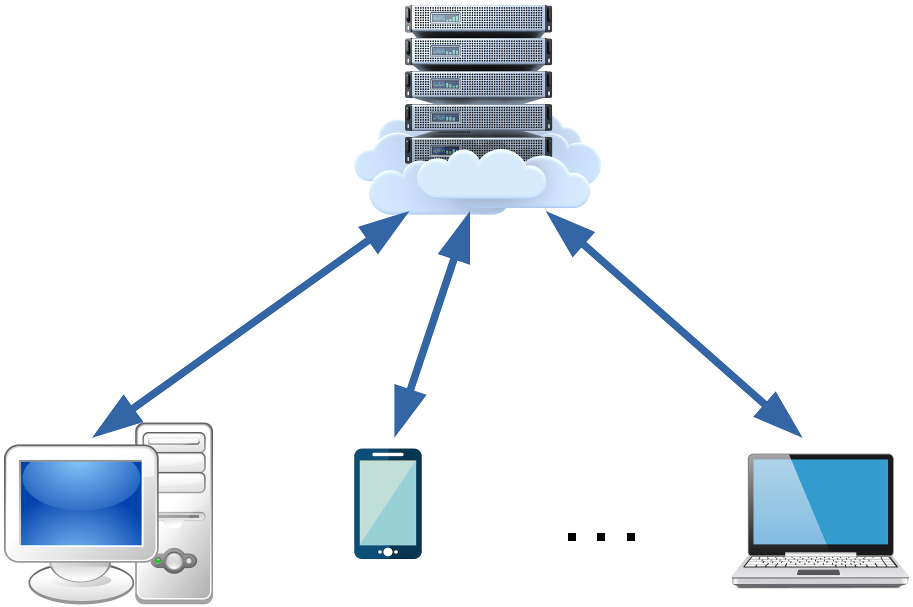
MOTOR 2022, Petrozavodsk, Russia

July 3, 2022

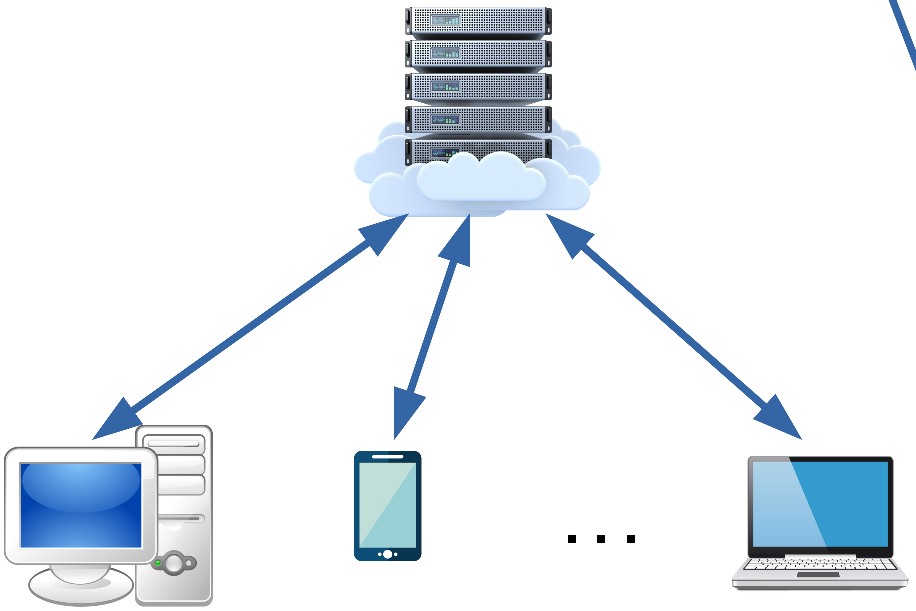
Outline

1. Distributed Optimization with Compression
2. Error-Compensated SGD (EC-SGD) and Absolute Compression
3. EC-SGD with Arbitrary Sampling and Absolute Compression
4. EC-SGD with Variance Reduction and Absolute Compression
5. Unified Analysis

1. Distributed Optimization



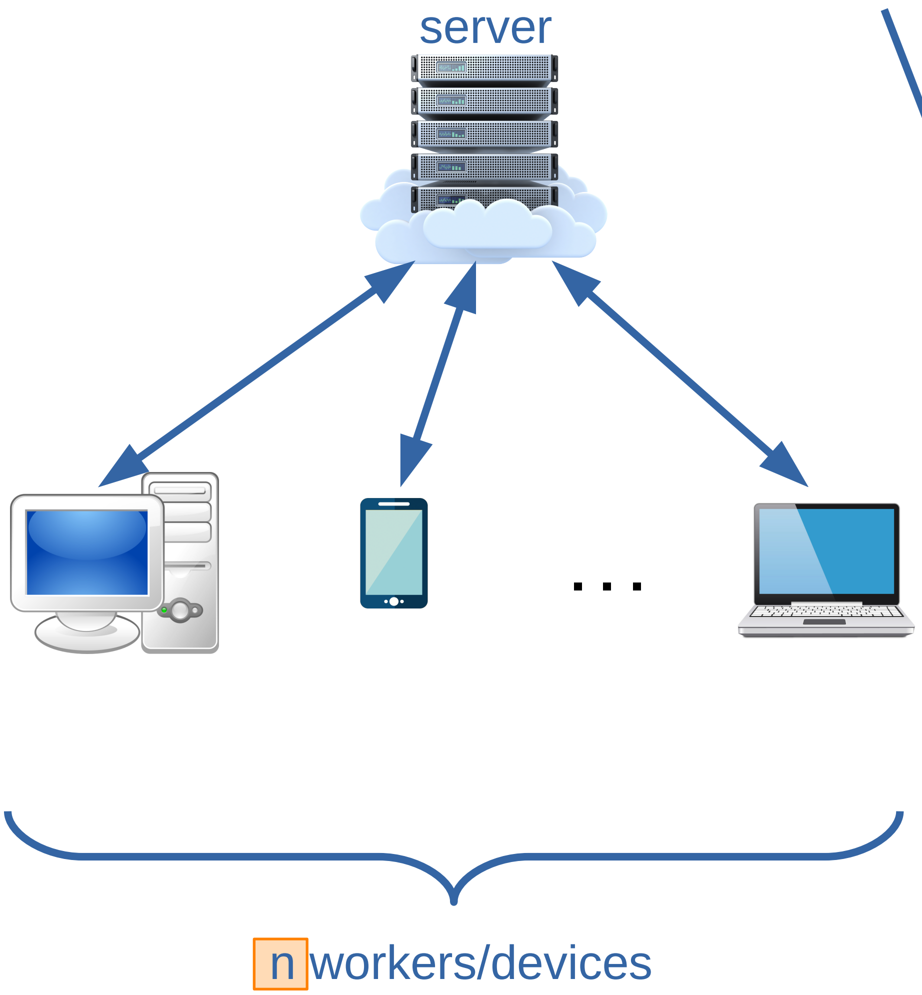
$$\min_{x\in\mathbb{R}^d}\left\{f(x):=\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right\}$$



parameters of the model

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

# of parameters

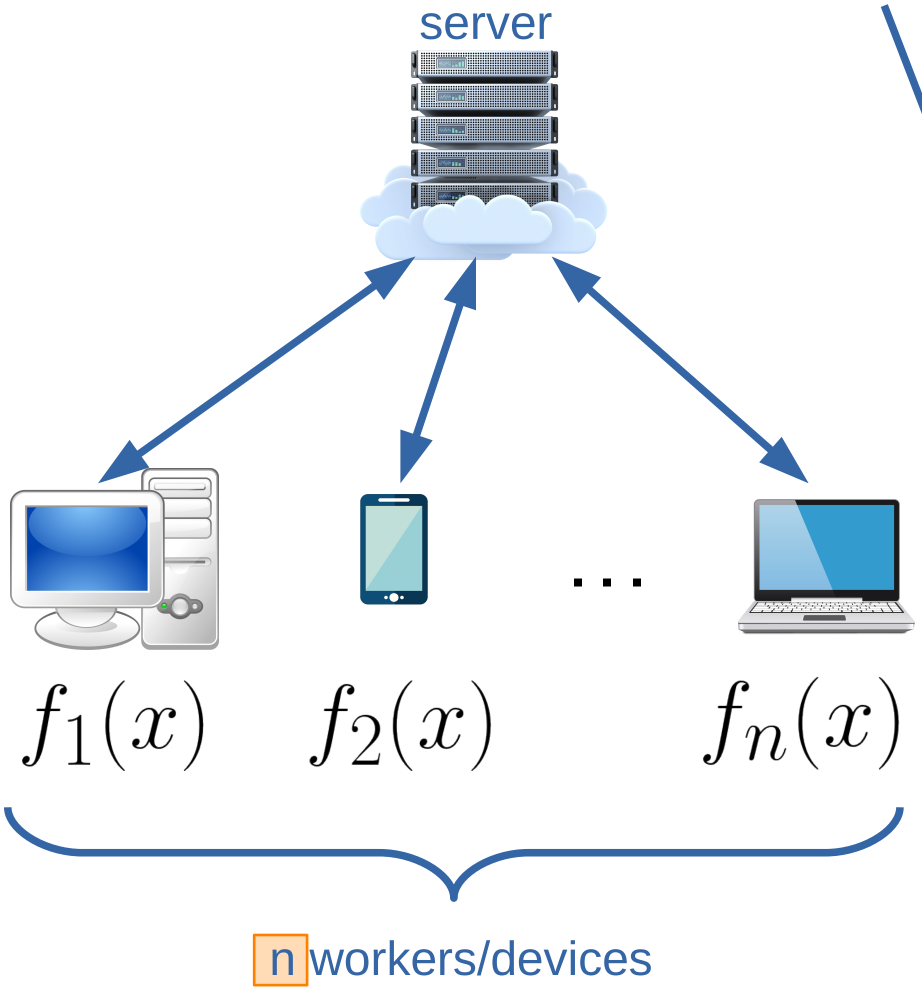


parameters of the model

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

# of parameters

# of workers/devices



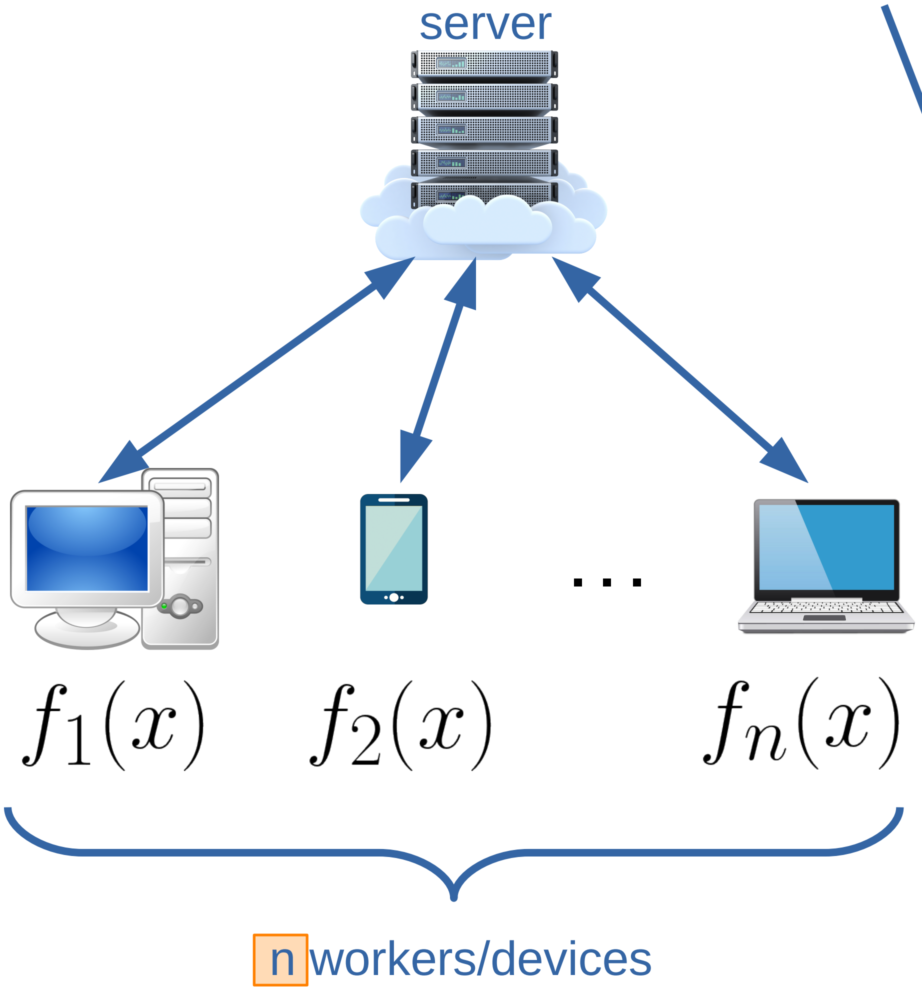
parameters of the model

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

# of parameters

# of workers/devices

loss on the data accessible by worker *i*



parameters of the model

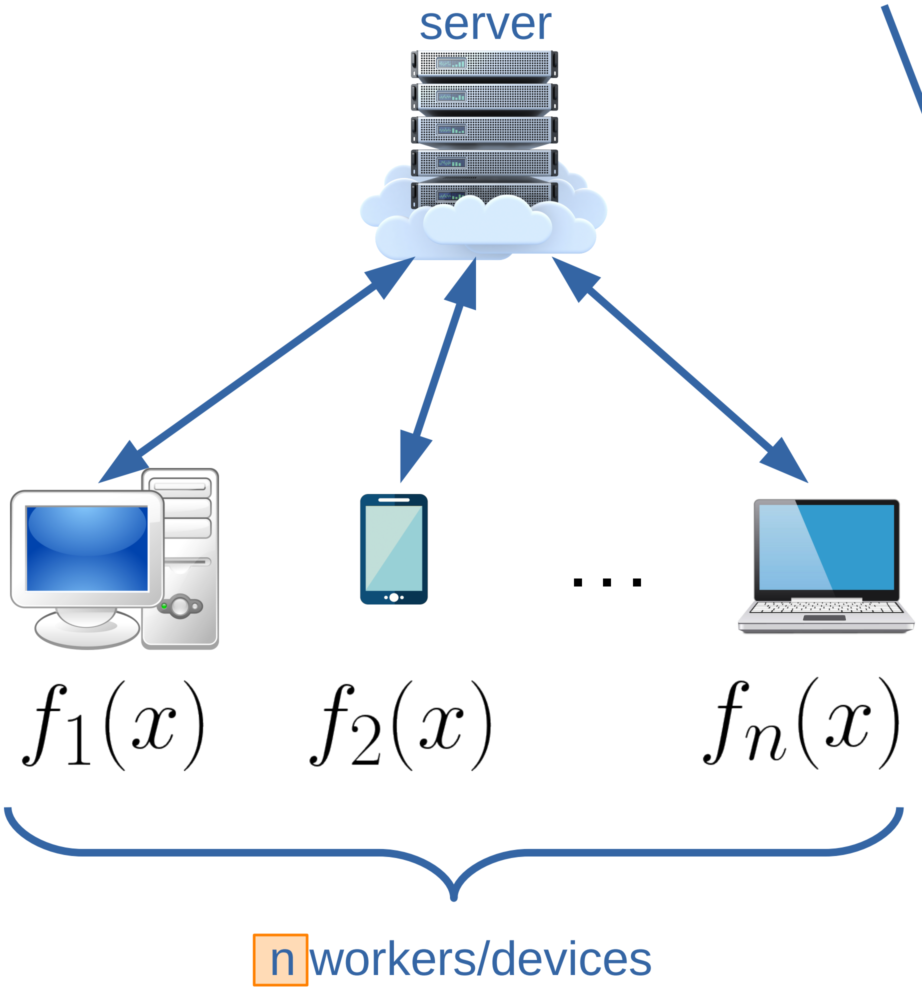
$$\min_{x\in\mathbb{R}^d}\left\{ f(x) := \frac{1}{\boxed{n}}\sum_{i=1}^{\boxed{n}} f_i(x)\right\}$$

# of parameters

# of workers/devices

loss on the data accessible by worker *i*

- $f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right]$



parameters of the model

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

# of parameters

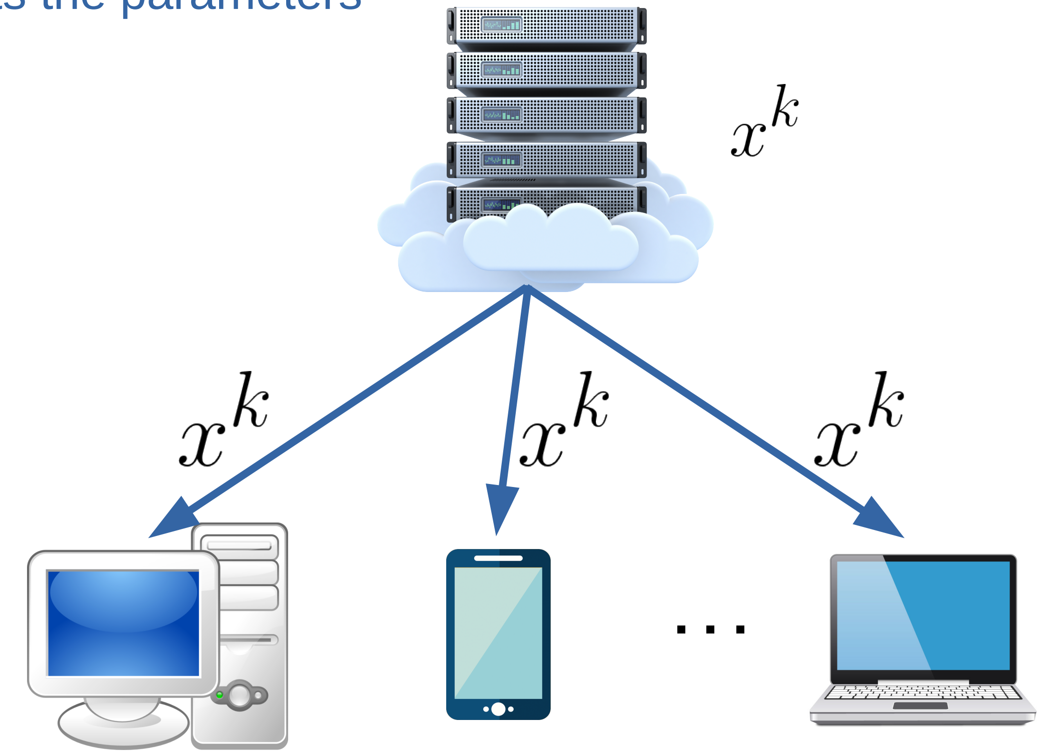
# of workers/devices

loss on the data accessible by worker *i*

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

Parallel SGD

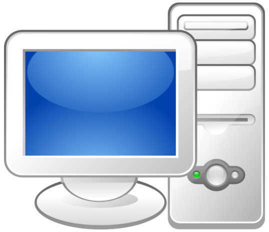
1 Server broadcasts the parameters

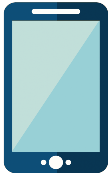


Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel

x^k





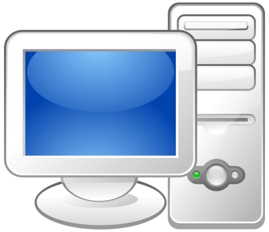
$\cdots$

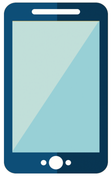
$x^k \rightarrow g_1^k \quad x^k \rightarrow g_2^k \quad x^k \rightarrow g_n^k$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel

x^k





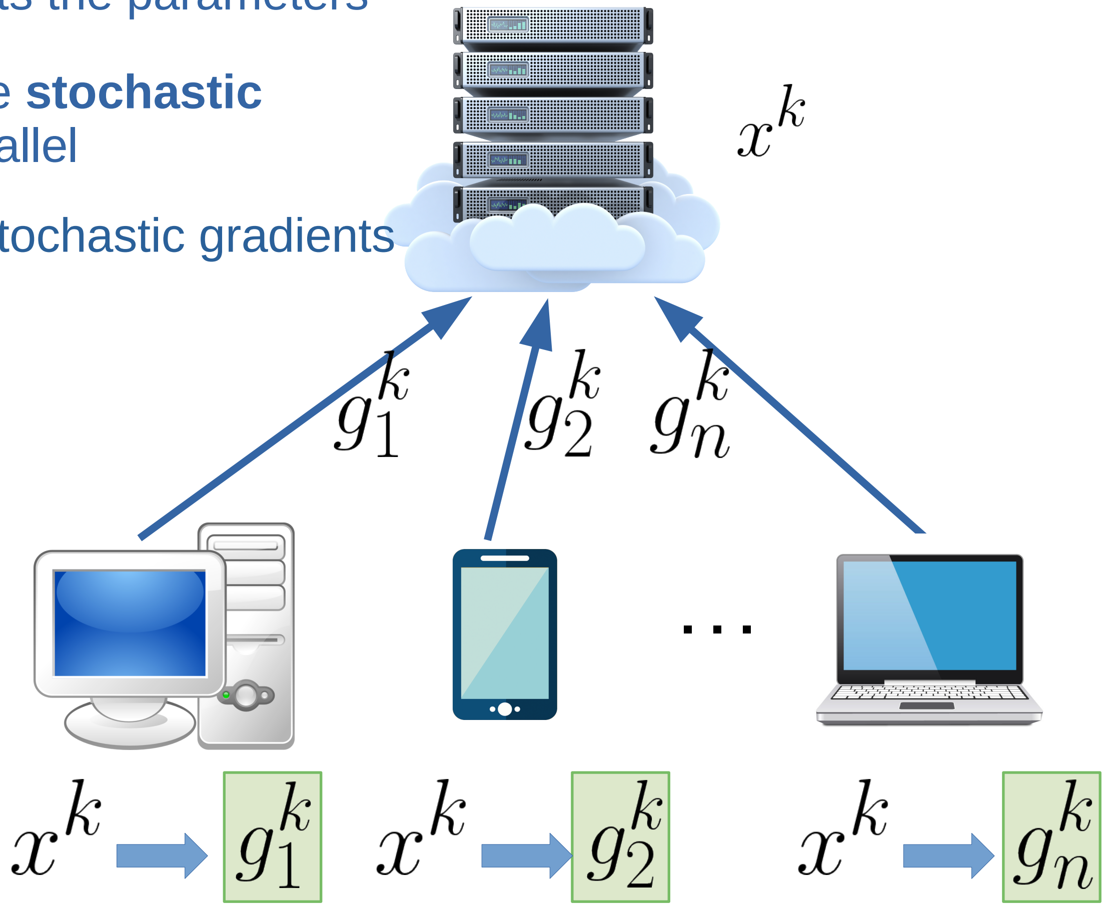
$\dots$

$x^k \rightarrow \boxed{g_1^k} \quad x^k \rightarrow \boxed{g_2^k} \quad x^k \rightarrow \boxed{g_n^k}$

$$\mathbb{E}\boxed{g_i^k} = \nabla f_i(x^k)$$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel
3. Server gathers stochastic gradients

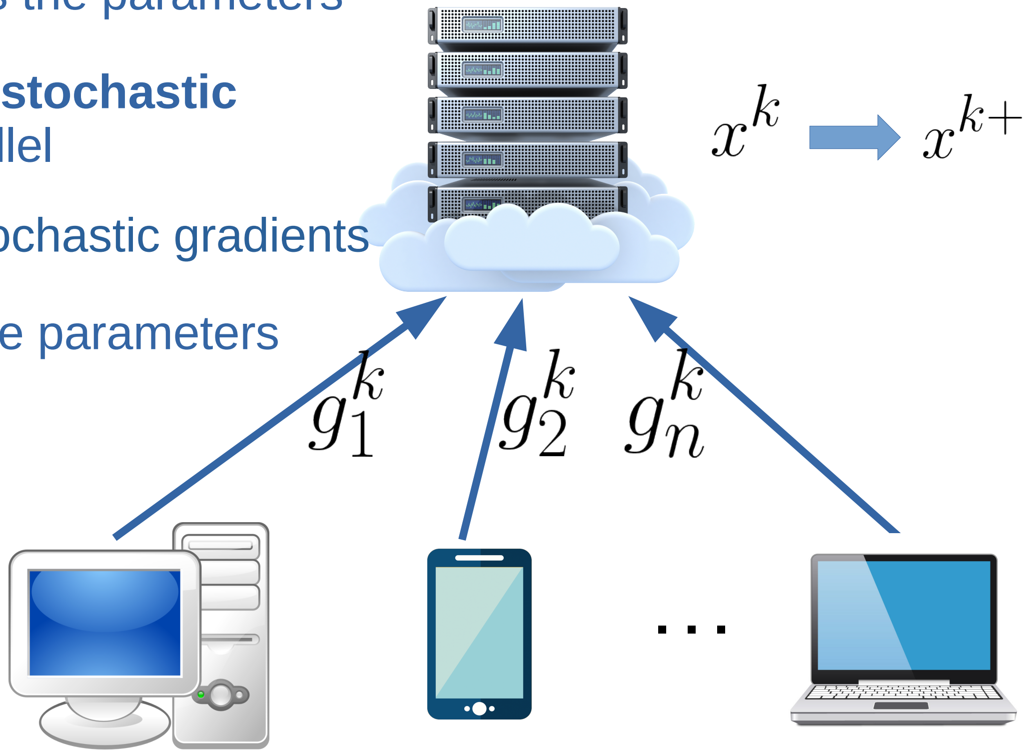


$$\mathbb{E} g_i^k = \nabla f_i(x^k)$$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel
3. Server gathers stochastic gradients
4. Server updates the parameters

$$x^k \Rightarrow x^{k+1} = x^k - \gamma \cdot \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

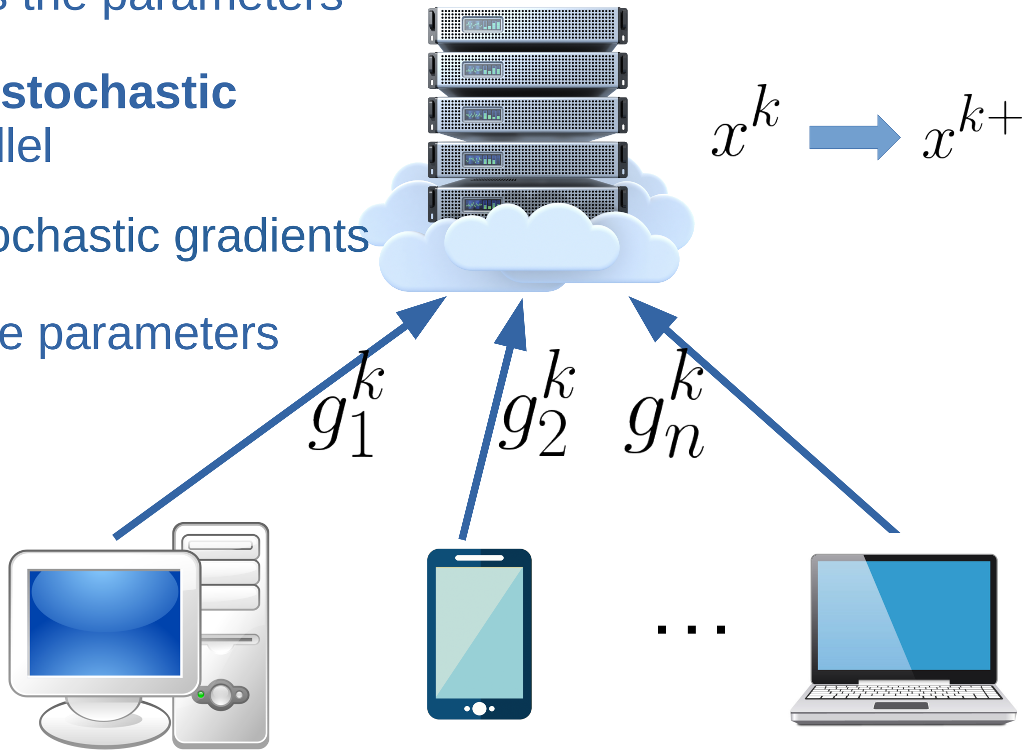


$$x^k \Rightarrow \boxed{g_1^k} \qquad x^k \Rightarrow \boxed{g_2^k} \qquad x^k \Rightarrow \boxed{g_n^k}$$

$$\mathbb{E}\boxed{g_i^k} = \nabla f_i(x^k)$$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel
3. Server gathers stochastic gradients
4. Server updates the parameters



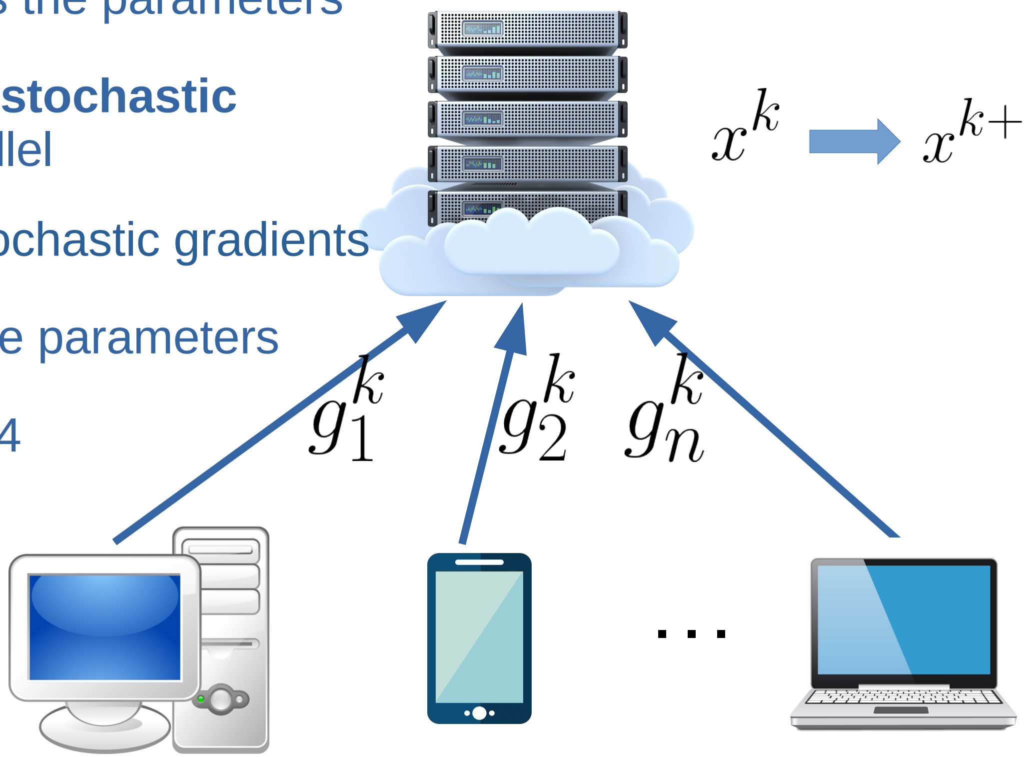
$$x^k \Rightarrow x^{k+1} = x^k - \underset{\text{stepsize}}{\gamma} \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_i^k}_{g^k}$$

$$x^k \Rightarrow \boxed{g_1^k} \quad x^k \Rightarrow \boxed{g_2^k} \quad x^k \Rightarrow \boxed{g_n^k}$$

$$\mathbb{E}\boxed{g_i^k} = \nabla f_i(x^k)$$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel
3. Server gathers stochastic gradients
4. Server updates the parameters
5. Repeat steps 1 – 4



stepsize

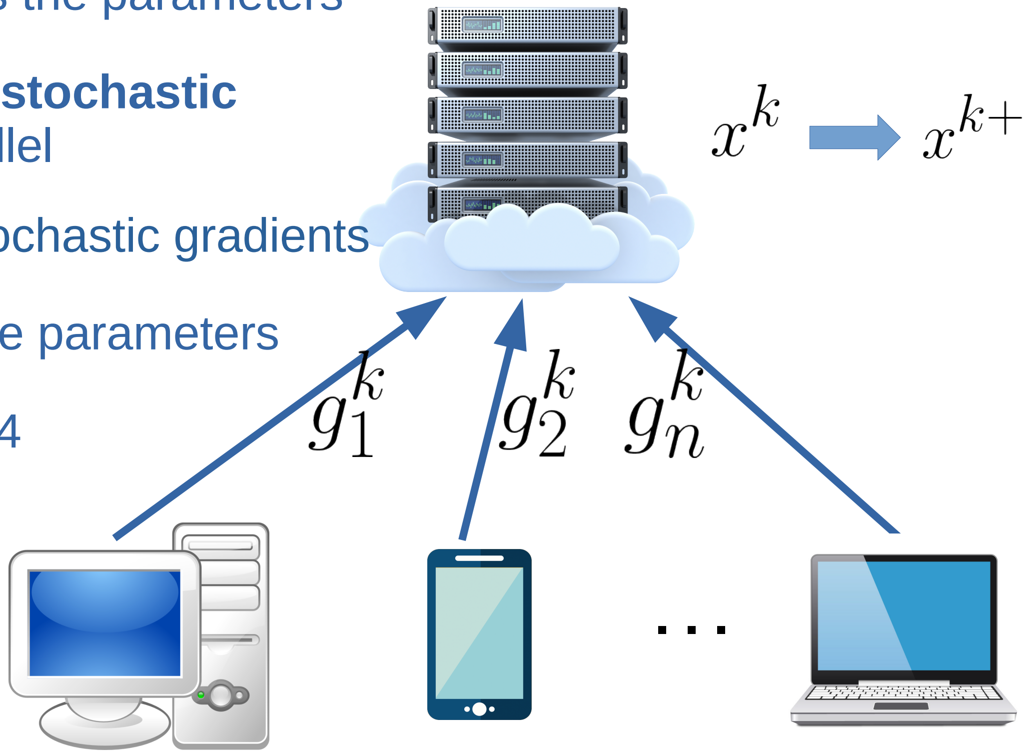
$$x^k \Rightarrow x^{k+1} = x^k - \gamma \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_i^k}_{g^k}$$

$$x^k \Rightarrow \boxed{g_1^k} \qquad x^k \Rightarrow \boxed{g_2^k} \qquad x^k \Rightarrow \boxed{g_n^k}$$

$$\mathbb{E}\boxed{g_i^k} = \nabla f_i(x^k)$$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel
3. Server gathers stochastic gradients
4. Server updates the parameters
5. Repeat steps 1 – 4



stepsize

$$x^k \Rightarrow x^{k+1} = x^k - \gamma \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_i^k}_{g^k}$$

Good news:

- Very simple algorithm
- Can be much faster than non-parallel SGD

$$x^k \Rightarrow \boxed{g_1^k} \qquad x^k \Rightarrow \boxed{g_2^k} \qquad x^k \Rightarrow \boxed{g_n^k}$$

$$\mathbb{E}\boxed{g_i^k} = \nabla f_i(x^k)$$

Parallel SGD

1. Server broadcasts the parameters
2. Devices compute **stochastic gradients** in parallel
3. Server gathers stochastic gradients
4. Server updates the parameters
5. Repeat steps 1 – 4

Good news:

- Very simple algorithm
- Can be much faster than non-parallel SGD

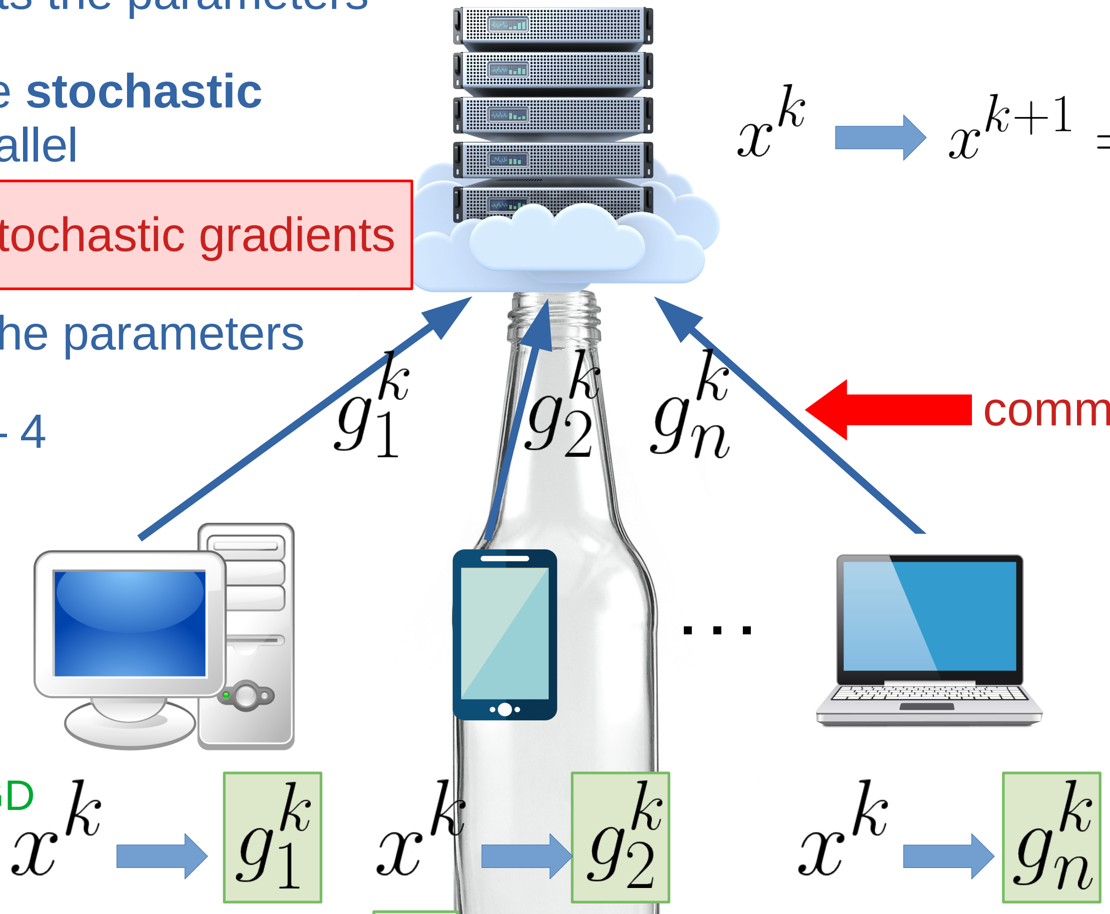
Issues:

- Overload of the server

stepsize

$$x^k \to x^{k+1} = x^k - \gamma \cdot \underbrace{\frac{1}{n}\sum_{i=1}^n g_i^k}_{g^k}$$

communication is a bottleneck



$$\mathbb{E} g_i^k = \nabla f_i(x^k)$$

Compression Operators

In this talk, we focus on **biased compression operators**

$$x \to \mathcal{C}(x) \qquad \mathbb{E}\|\mathcal{C}(x) - x\|^2 \leq (1-\delta)\|x\|^2$$

Compression Operators

In this talk, we focus on **biased compression operators**

$$x \to \mathcal{C}(x) \qquad\qquad \mathbb{E}\|\mathcal{C}(x) - x\|^2 \le (1-\delta)\|x\|^2$$

Example: TopT (for T = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 \\ -15 \\ 0 \\ 0 \\ 10 \end{pmatrix}$$

Pick T = 2 components with largest absolute value

2. Error-Compensated SGD and Absolute Compression

Error-Compensated SGD

Seide, Frank, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. **"1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns."** *In Fifteenth Annual Conference of the International Speech Communication Association*. 2014.

Stich, Sebastian U., Jean-Baptiste Cordonnier, and Martin Jaggi. **"Sparsified SGD with memory."** *In Advances in Neural Information Processing Systems*, pp. 4447-4458. 2018.

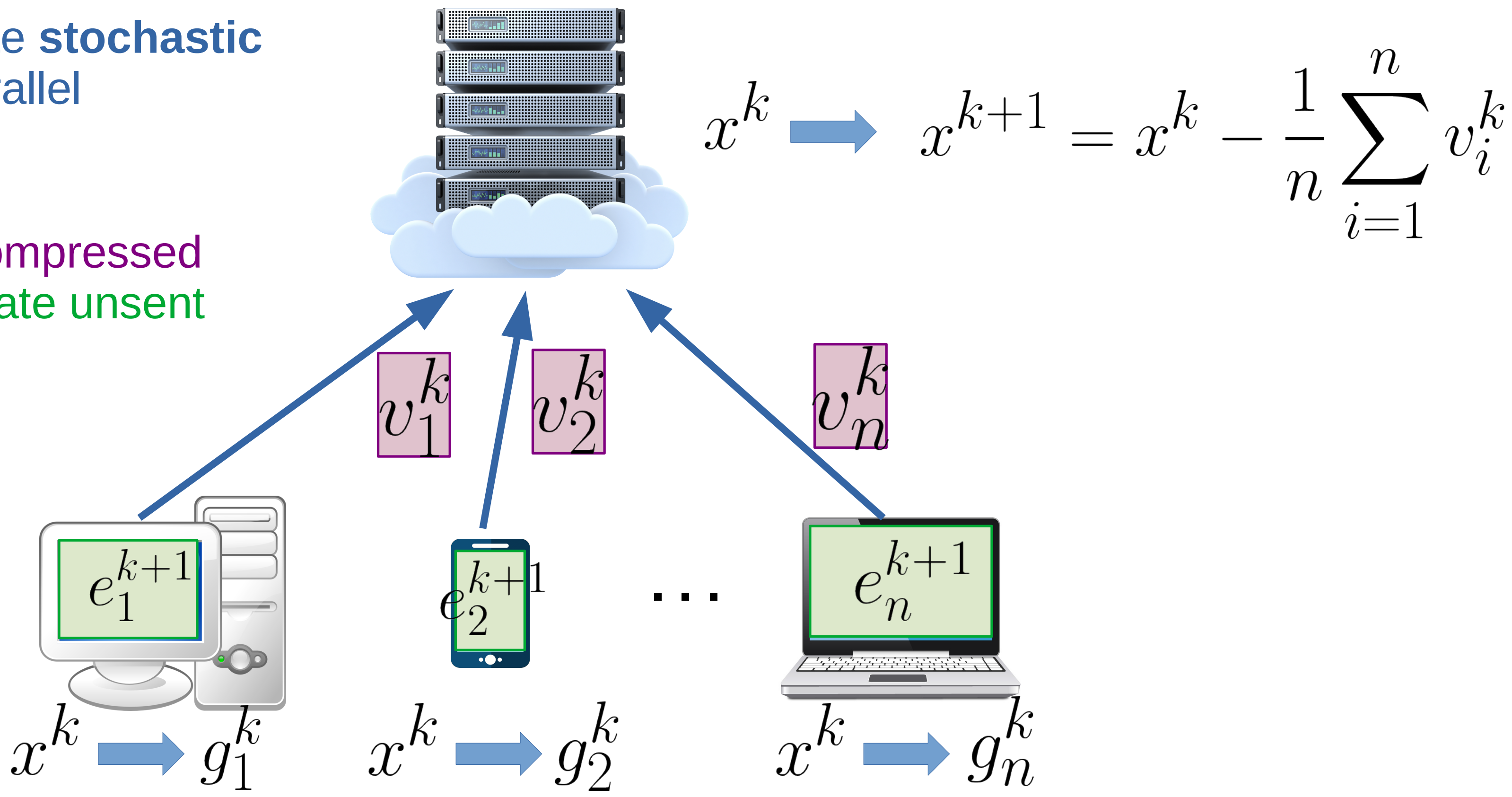
Karimireddy, Sai Praneeth, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. **"Error Feedback Fixes SignSGD and other Gradient Compression Schemes."** *In International Conference on Machine Learning*, pp. 3252-3261. 2019.

Stich, Sebastian U., and Sai Praneeth Karimireddy. **"The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication."** arXiv preprint arXiv:1909.05350 (2019).

Beznosikov, Aleksandr, Samuel Horváth, Peter Richtárik, and Mher Safaryan. **"On Biased Compression for Distributed Learning."** arXiv preprint arXiv:2002.12410 (2020).

Step k+1

1. Server broadcasts new parameters
2. Workers compute **stochastic gradients** in parallel
3. Compression
4. Devices send compressed vectors and update unsent information
5. Server gathers the information and updates the parameters
6. Repeat steps 1 – 5



$$v_i^k = \gamma \mathcal{C}\left(\frac{e_i^k}{\gamma} + g_i^k\right)$$

$$e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k$$

Revisiting EC-SGD

Sahu, Atal, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. **"Rethinking gradient sparsification as total error minimization."** *Advances in Neural Information Processing Systems 34* (2021): 8133-8146.

- In the analysis of EC-SGD, the following quantity («total error») appears (*n = 1*):

$$\sum_{k=0}^{K-1} \mathbb{E}\left\| \frac{e^k}{\gamma} + g^k - \mathcal{C}\left(\frac{e^k}{\gamma} + g^k\right)\right\|^2$$

Revisiting EC-SGD

Sahu, Atal, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. **"Rethinking gradient sparsification as total error minimization."** *Advances in Neural Information Processing Systems 34* (2021): 8133-8146.

- In the analysis of EC-SGD, the following quantity («total error») appears (*n = 1*):

$$\sum_{k=0}^{K-1} \mathbb{E} \left\| \frac{e^k}{\gamma} + g^k - \mathcal{C}\left(\frac{e^k}{\gamma} + g^k \right) \right\|^2$$

- TopT compression minimizes error on each iteration for given budget of components

Revisiting EC-SGD

Sahu, Atal, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. **"Rethinking gradient sparsification as total error minimization."** *Advances in Neural Information Processing Systems 34* (2021): 8133-8146.

- In the analysis of EC-SGD, the following quantity («total error») appears (*n* = *1*):

$$\sum_{k=0}^{K-1} \mathbb{E} \left\| \frac{e^k}{\gamma} + g^k - \mathcal{C}\left(\frac{e^k}{\gamma} + g^k\right) \right\|^2$$

- TopT compression minimizes error on each iteration for given budget of components

- Minimization of total error is intractable

Revisiting EC-SGD

Sahu, Atal, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. **"Rethinking gradient sparsification as total error minimization."** *Advances in Neural Information Processing Systems 34* (2021): 8133-8146.

- In the analysis of EC-SGD, the following quantity («total error») appears (*n = 1*):

$$\sum_{k=0}^{K-1} \mathbb{E} \left\| \frac{e^k}{\gamma} + g^k - \mathcal{C}\left(\frac{e^k}{\gamma} + g^k\right) \right\|^2$$

- TopT compression minimizes error on each iteration for given budget of components

- Minimization of total error is intractable

- Nevertheless, *in the class of absolute compressors*, for a fixed $\{a_k\}_{k\geq 0}$ one can minimize

$$\sum_{k=0}^{K-1} \mathbb{E} \left\| a_k - \mathcal{C}(a_k) \right\|^2$$

Absolute Compression

Biased compressors

$$x \to \mathcal{C}(x)$$

Contractive compressors

$$\mathbb{E}\|\mathcal{C}(x) - x\|^2 \leq (1-\delta)\|x\|^2$$

Example: TopT (for T = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 \\ -15 \\ 0 \\ 0 \\ 10 \end{pmatrix}$$

Pick T = 2 components with largest abs. value

Absolute Compression

Biased compressors

$$x \to \mathcal{C}(x)$$

Contractive compressors

$$\mathbb{E}\|\mathcal{C}(x) - x\|^2 \le (1-\delta)\|x\|^2$$

Absolute compressors

$$\mathbb{E}\|\mathcal{C}(x) - x\|^2 \le \Delta^2$$

Example: TopT (for T = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 \\ -15 \\ 0 \\ 0 \\ 10 \end{pmatrix}$$

Pick T = 2 components with largest abs. value

Absolute Compression

Biased compressors

$$x \to \mathcal{C}(x)$$

Contractive compressors

$$\mathbb{E}\|\mathcal{C}(x) - x\|^2 \leq (1-\delta)\|x\|^2$$

Example: TopT (for T = 2)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 \\ -15 \\ 0 \\ 0 \\ 10 \end{pmatrix}$$

Pick T = 2 components with largest abs. value

Absolute compressors

$$\mathbb{E}\|\mathcal{C}(x) - x\|^2 \leq \Delta^2$$

Example: Hard Threshold sparsifier (HT)

$$\begin{pmatrix} 1 \\ -15 \\ 0.2 \\ -7 \\ 10 \end{pmatrix} \Rightarrow \begin{pmatrix} 0 \\ -15 \\ 0 \\ -7 \\ 10 \end{pmatrix}$$

Pick components with abs. value at least $\lambda = 7$

EC-SGD with Absolute Compression

Sahu, Atal, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. **"Rethinking gradient sparsification as total error minimization."** *Advances in Neural Information Processing Systems 34* (2021): 8133-8146.

- Better performance in practice
- Better theoretical guarantees in some regimes under *(M, σ^2)*-bounded noise

$$\mathbb{E}_k \left\| g_i^k - \nabla f_i \left(x^k\right) \right\|^2 \leq M \left\| \nabla f_i \left(x^k\right) \right\|^2 + \sigma^2$$

EC-SGD with Absolute Compression

Sahu, Atal, Aritra Dutta, Ahmed M Abdelmoniem, Trambak Banerjee, Marco Canini, and Panos Kalnis. **"Rethinking gradient sparsification as total error minimization."** *Advances in Neural Information Processing Systems 34* (2021): 8133-8146.

- Better performance in practice
- Better theoretical guarantees in some regimes under *(M, σ²)*-bounded noise

$$\mathbb{E}_k \left\|g_i^k - \nabla f_i\left(x^k\right)\right\|^2 \leq M \left\|\nabla f_i\left(x^k\right)\right\|^2 + \sigma^2$$

- Only standard gradient estimators are analyzed, i.e., arbitrary sampling and variance reduction are not considered

Our work addresses this limitation

3. EC-SGD with Arbitrary Sampling and Absolute Compression

Arbitrary Sampling

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

Arbitrary Sampling

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

- Stochastic reformulation:

$$f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x) = \frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}(x), \quad f_{\xi_i}(x) = \frac{1}{m}\sum_{j=1}^{m} \xi_{ij} f_{ij}(x)$$

Arbitrary Sampling

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

- Stochastic reformulation:

$$f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right], \quad f_\xi(x) = \frac{1}{n}\sum_{i=1}^{n} f_{\xi_i}(x), \quad f_{\xi_i}(x) = \frac{1}{m}\sum_{j=1}^{m} \xi_{ij} f_{ij}(x)$$

- Sampling vector:

$$\xi = \left(\xi_1^\top, \ldots, \xi_n^\top\right)^\top \qquad \xi_i = \left(\xi_{i1}, \ldots, \xi_{im}\right)^\top \qquad \mathbb{E}\left[\xi_{ij}\right] = 1$$

Expected Smoothness and Examples

We assume that for all *i = 1, … , n*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\left\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}\left(x^*\right)\right\|^2\right] \leq 2\mathcal{L}\left(f_i(x) - f_i(x) - \left\langle \nabla f_i\left(x^*\right), x - x^*\right\rangle\right)$$

Expected Smoothness and Examples

We assume that for all $i = 1, \ldots, n$

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\left\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\right\|^2\right] \leq 2\mathcal{L}\left(f_i(x) - f_i(x) - \langle \nabla f_i(x^*), x - x^* \rangle\right)$$

Examples*

- Uniform sampling (US):
- Importance sampling (IS):

*all f_{ij} are assummed to be convex and L_{ij}-smooth

Expected Smoothness and Examples

We assume that for all $i = 1, \dots, n$

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\left\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\right\|^2\right] \leq 2\mathcal{L}\left(f_i(x) - f_i(x) - \left\langle \nabla f_i(x^*), x - x^* \right\rangle\right)$$

Examples*

- Uniform sampling (US): $\mathbb{P}\left\{\xi_i = m e_j\right\} = \frac{1}{m}$ $\qquad \mathcal{L} = \mathcal{L}_{\text{US}} = \max\limits_{i\in[n], j\in[m]} L_{ij}$

  j-th element of the standard basis

- Importance sampling (IS):

*all f_{ij} are assummed to be convex and L_{ij}-smooth

Expected Smoothness and Examples

We assume that for all *i = 1, … , n*

$$\mathbb{E}_{\xi_i \sim \mathcal{D}_i}\left[\left\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}\left(x^*\right)\right\|^2\right] \leq 2\mathcal{L}\left(f_i(x) - f_i(x) - \left\langle \nabla f_i\left(x^*\right), x - x^*\right\rangle\right)$$

Examples*

- Uniform sampling (US): $\mathbb{P}\left\{\xi_i = m e_j\right\} = \frac{1}{m}$ $\qquad \mathcal{L} = \mathcal{L}_{\mathrm{US}} = \max\limits_{i \in [n], j \in [m]} L_{ij}$

  j-th element of the standard basis

- Importance sampling (IS): $\mathbb{P}\left\{\xi_i = \frac{m\overline{L}_i}{L_{ij}} e_j\right\} = \frac{L_{ij}}{m\overline{L}_i}$ $\qquad \overline{L}_i = \frac{1}{m}\sum\limits_{i=j}^{m} L_{ij}$

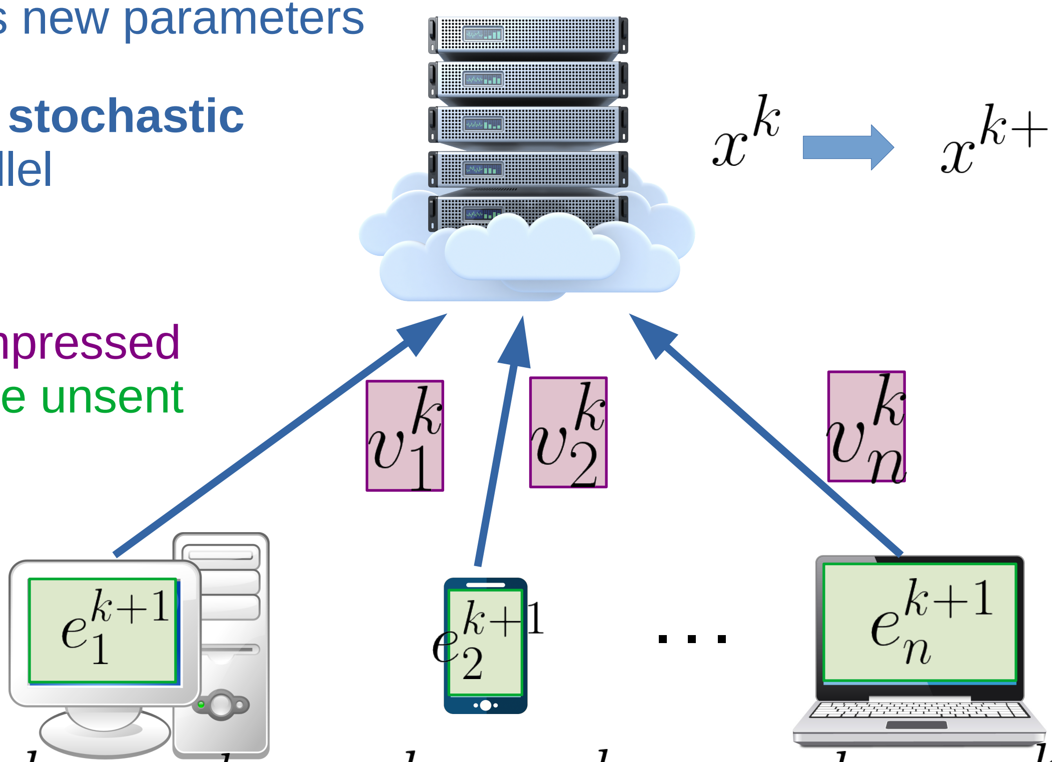
$$\mathcal{L} = \mathcal{L}_{\mathrm{IS}} = \max_{i \in [n]} \overline{L}_i$$

*all f_{ij} are assummed to be convex and L_{ij}-smooth

EC-SGD with Arbitrary Sampling

1. Server broadcasts new parameters
2. Workers compute **stochastic gradients** in parallel
3. Compression
4. Devices send compressed vectors and update unsent information
5. Server gathers the information and updates the parameters
6. Repeat steps 1 – 5

$$x^k \rightarrow x^{k+1} = x^k - \frac{1}{n}\sum_{i=1}^{n} v_i^k$$



$x^k \rightarrow g_1^k$ $\quad x^k \rightarrow g_2^k$ $\quad x^k \rightarrow g_n^k$

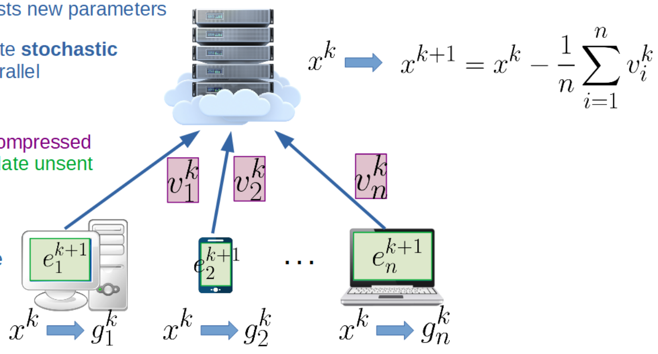
$$\boxed{v_i^k = \gamma \mathcal{C}\left(\frac{e_i^k}{\gamma} + g_i^k\right)} \qquad \boxed{e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k}$$

EC-SGD with Arbitrary Sampling

The same method with the following estimator:

$$g_i^k = \nabla f_{\xi_i^k}(x^k)$$

1. Server broadcasts new parameters
2. Workers compute **stochastic gradients** in parallel
3. Compression
4. Devices send compressed vectors and update unsent information
5. Server gathers the information and updates the parameters
6. Repeat steps 1 – 5



$$\boxed{v_i^k = \gamma \mathcal{C}\left(\frac{e_i^k}{\gamma} + g_i^k\right)} \qquad \boxed{e_i^{k+1} = e_i^k + \gamma g_i^k - v_i^k}$$

Additional Assumptions

- Lipschitz gradients $$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$$

- Strong convexity $$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

Convergence

$$\mathbb{E} f\left(\overline{x}^K\right) - f\left(x^*\right) = \mathcal{O}\left(\left(L + \frac{\mathcal{L}}{n}\right) R_0^2 \exp\left(-\frac{\mu}{L + \mathcal{L}/n} K\right) + \frac{\boxed{\sigma_*^2}}{\mu n K} + \frac{L\Delta^2}{\mu^2 K^2}\right)$$

$$\sigma_*^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_{\xi_i}\left(x^*\right) - \nabla f_i\left(x^*\right) \right\|^2$$

Convergence

$$\mathbb{E}f\left(\overline{x}^K\right)-f\left(x^*\right)=\mathcal{O}\left(\left(L+\frac{\mathcal{L}}{n}\right)R_0^2\exp\left(-\frac{\mu}{L+\mathcal{L}/n}K\right)+\frac{\boxed{\sigma_*^2}}{\mu n K}+\frac{L\Delta^2}{\mu^2K^2}\right)$$

$$\sigma_*^2=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla f_{\xi_i}\left(x^*\right)-\nabla f_i\left(x^*\right)\right\|^2$$

Implications

- Sampling may improve the convergence on the early stages
- Better performance for HT sparsifier in comparison to TopT, when heterogeneity is large: this is verified by Sahu et al. (2021)

Experiment 1

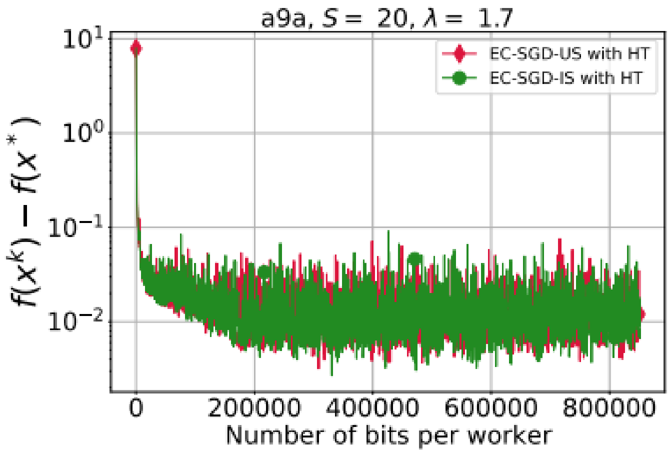
Logistic regression with l_2-regularization

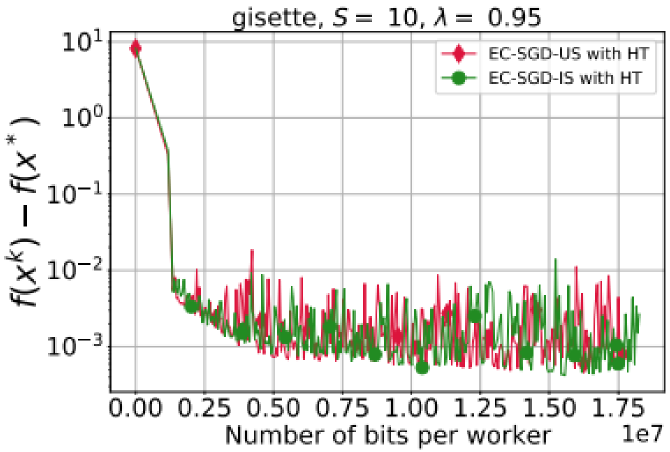
$$\min_{x\in\mathbb{R}^d}\left\{f(x)=\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\underbrace{\ln\left(1+\exp\left(-y_i\left\langle a_{ij},x\right\rangle\right)\right)+\frac{l_2}{2}\|x\|^2}_{f_{ij}(x)}\right\}$$

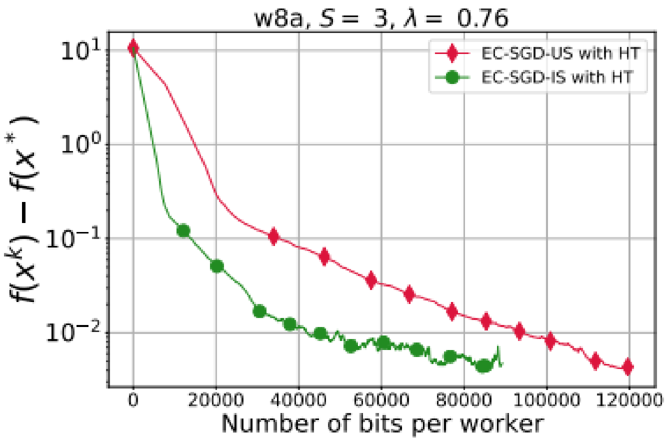
Datasets

- a9a: n = 20, m = 1600, d = 123 $\quad \mathcal{L}_{\mathrm{IS}}\approx 3.47,\quad \mathcal{L}_{\mathrm{US}}\approx 3.5$
- gisette: n = 20, m = 300, d = 5000 $\quad \mathcal{L}_{\mathrm{IS}}\approx 1164.89,\quad \mathcal{L}_{\mathrm{US}}\approx 1201.51$
- w8a: n = 20, m = 2485, d = 300 $\quad \mathcal{L}_{\mathrm{IS}}\approx 3.05,\quad \mathcal{L}_{\mathrm{US}}\approx 28.5$

Experiment 1







As expected, EC-SGD-IS converges faster than EC-SGD-US on w8a dataset.
On a9a and gisette the methods perform similarly.

4. EC-SGD with Variance Reduction and Absolute Compression

EC-SGD with Arbitrary Sampling: Reminder

Convergence of EC-SGD with Arbitrary Sampling:

$$\mathbb{E} f(\overline{x}^K) - f(x^*) = \mathcal{O}\left(\left(L + \frac{\mathcal{L}}{n}\right) R_0^2 \exp\left(-\frac{\mu}{L + \mathcal{L}/n} K\right) + \frac{\boxed{\sigma_*^2}}{\mu n K} + \frac{L\Delta^2}{\mu^2 K^2}\right)$$

$$\sigma_*^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left\| \nabla f_{\xi_i}(x^*) - \nabla f_i(x^*) \right\|^2$$

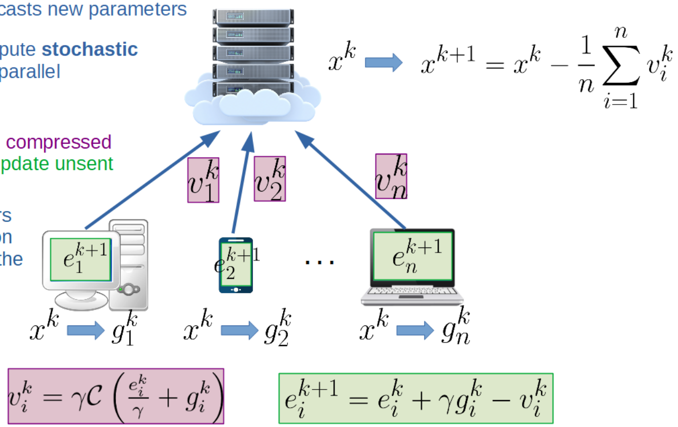
One can speed up the method via removing the variance term

EC-SGD with Variance Reducation

We consider the same algorithmic scheme, but now with estimator of Loopless Stochastic Variance Reduced Gradient (LSVRG):

$$g_i^k = \nabla f_{\xi_i^k}\left(x^k\right) - \nabla f_{\xi_i^k}\left(w^k\right) + \nabla f_i\left(w^k\right)$$

1. Server broadcasts new parameters
2. Workers compute **stochastic gradients** in parallel
3. Compression
4. Devices send compressed vectors and update unsent information
5. Server gathers the information and updates the parameters
6. Repeat steps 1 – 5



EC-SGD with Variance Reducation

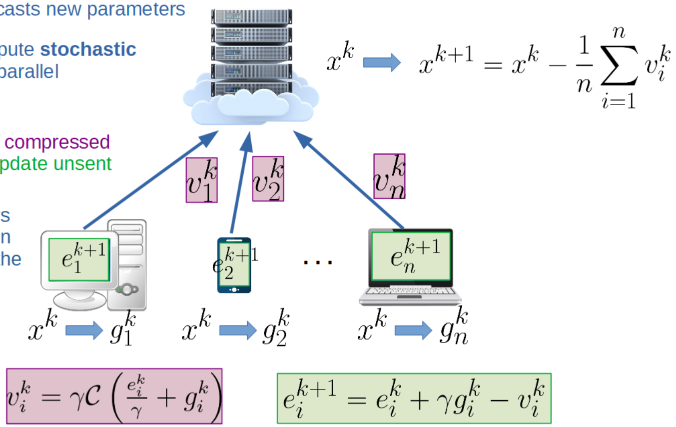
We consider the same algorithmic scheme, but now with estimator of Loopless Stochastic Variance Reduced Gradient (LSVRG):

$$g_i^k = \nabla f_{\xi_i^k}\left(x^k\right) - \nabla f_{\xi_i^k}\left(\boxed{w^k}\right) + \nabla f_i\left(\boxed{w^k}\right)$$

$$\boxed{w^{k+1}} = \begin{cases} x^k, & \text{with probability } p, \\ w^k, & \text{with probability } 1-p, \end{cases} \qquad w^0 = x^0$$

Updated with small probability *p ~ 1/m*

1. Server broadcasts new parameters
2. Workers compute **stochastic gradients** in parallel
3. Compression
4. Devices send compressed vectors and update unsent information
5. Server gathers the information and updates the parameters
6. Repeat steps 1 – 5



Convergence

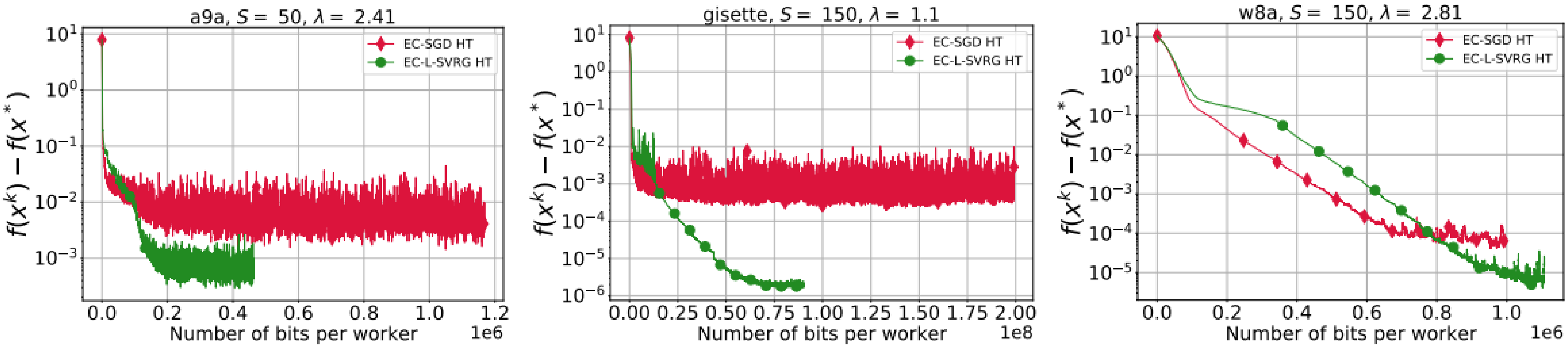
$$\mathbb{E}f(\overline{x}^K) - f(x^*) = \mathcal{O}\left(\left(L + \frac{\mathcal{L}}{n}\right)\widetilde{T}_0 \exp\left(-\min\left\{\frac{\mu}{L + \mathcal{L}/n}, \frac{1}{m}\right\}K\right) + \frac{L\Delta^2}{\mu^2 K^2}\right)$$

Implications

- Faster convergence than EC-SGD on later stages
- As for EC-SGD, the theory predicts better performance for HT sparsifier in comparison to TopT, when heterogeneity is large

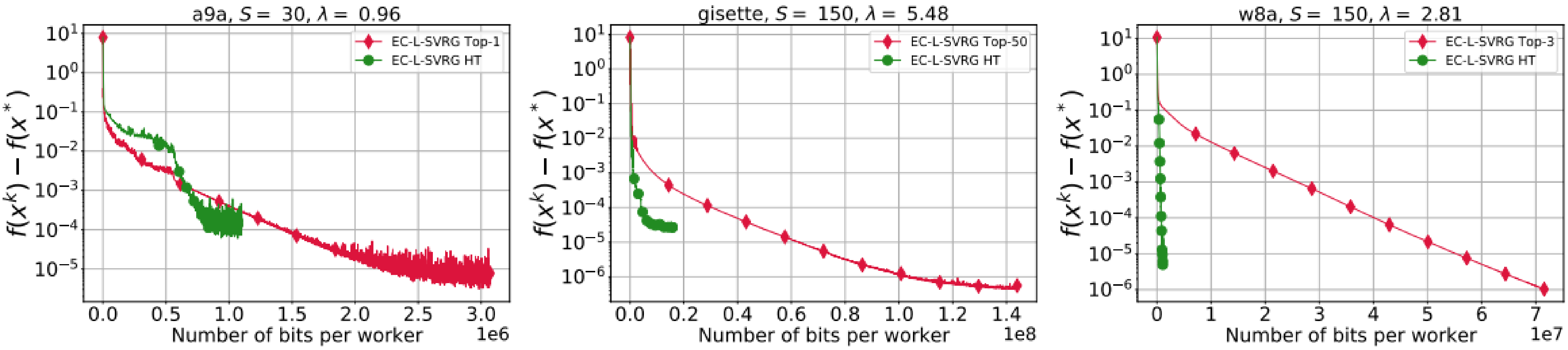
Experiment 2

We test EC-SGD and EC-LSVRG with HT sparsifier. We use the same stepsize for both methods.



As expected, EC-LSVRG achieves better accuracy than EC-SGD.

Experiment 3



EC-LSVRG with HT sparsifier achieves reasonable accuracy of the solution faster than EC-LSVRG with TopT sparsifier

5. Unified Analysis

Key Assumption

$$\mathbb{E}_k[g^k] = \nabla f(x^k), \qquad g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

$$\mathbb{E}_k\left[\left\|g^k\right\|^2\right] \leq 2A\left(f\left(x^k\right) - f\left(x^*\right)\right) + B\sigma_k^2 + D_1$$

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-\rho)\sigma_k^2 + 2C\left(f\left(x^k\right) - f\left(x^*\right)\right) + D_2$$

Key Assumption

$$\mathbb{E}_k[g^k] = \nabla f(x^k), \qquad g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

$$\boxed{\mathbb{E}_k\left[\left\|g^k\right\|^2\right] \leq 2A\left(f\left(x^k\right) - f\left(x^*\right)\right) + B\sigma_k^2 + D_1}$$

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-\rho)\sigma_k^2 + 2C\left(f\left(x^k\right) - f\left(x^*\right)\right) + D_2$$

- Reflects smoothness properties of the problem and noises introduced by compressions and stochastic gradients

Key Assumption

$$\mathbb{E}_k[g^k] = \nabla f(x^k), \qquad g^k = \frac{1}{n}\sum_{i=1}^{n} g_i^k$$

$$\mathbb{E}_k\left[\left\|g^k\right\|^2\right] \leq 2A\left(f\left(x^k\right) - f\left(x^*\right)\right) + B\sigma_k^2 + D_1$$

$$\mathbb{E}_k\left[\sigma_{k+1}^2\right] \leq (1-\rho)\sigma_k^2 + 2C\left(f\left(x^k\right) - f\left(x^*\right)\right) + D_2$$

- Reflects smoothness properties of the problem and noises introduced by compressions and stochastic gradients
- Describes the process of variance reduction of the variance coming from stochastic gradients

General Theorem

Let f be μ-strongly convex and L-smooth. Let the assumption from the previous slide hold. Then, there exists a choice of stepsize such that EC-SGD with absolute compression satisfies

$$\mathbb{E}f\left(\overline{x}^K\right) - f\left(x^*\right) \leq \frac{(1-\eta)^{K+1} 2\mathbb{E}\left[T_0\right]}{\gamma} + 2\gamma\left(D_1 + FD_2 + 3L\gamma\Delta^2\right)$$

General Theorem

Let f be μ-strongly convex and L-smooth. Let the assumption from the previous slide hold. Then, there exists a choice of stepsize such that EC-SGD with absolute compression satisfies

$$\mathbb{E} f\left(\overline{x}^K\right) - f\left(x^*\right) \leq \frac{(1-\eta)^{K+1} 2\mathbb{E}\left[T_0\right]}{\gamma} + 2\gamma\left(D_1 + F D_2 + 3L\gamma\Delta^2\right)$$

- Covers EC-SGD with Arbitrary Sampling and EC-LSVRG
- Covers the original analysis by Sahu et al. (2021)
- Can be applied for different gradient estimators satisfying the key assumption

Conclusion

Conclusion

- The first analysis of EC-SGD with arbitrary sampling and absolute compression
- The first analysis of EC-LSVRG with arbitrary sampling and absolute compression
- The general theoretical framework for analyzing EC-SGD-type methods with absolute compression is proposed
- Numerical experiments support the theoretical findings
- In the paper, we also consider (non-strongly) convex case ($\mu = 0$)

See more details in the paper: https://arxiv.org/abs/2203.02383

Extra Slides

Distributed Optimization

- Some problems cannot be solved on a single a machine in a reasonable time (deep learning models with billions of parameters and gigabytes of data)

- There exist such problems where the data that defines the optimization problem is private and distributed among several machines (federated learning)

These problems are typically solved in a distributed way

Convergence of EC-SGD

Assumptions

- Lipschitz gradients $$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|$$
- Strong convexity $$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\mu}{2}\|y - x\|^2$$

Convergence rate

$$\mathbb{E} f(\overline{x}^K) - f(x^*) = \mathcal{O}\left(\frac{L}{\boxed{\delta}\mu} R_0^2 \exp\left(-\frac{\boxed{\delta}\mu}{L}K\right) + \frac{\boxed{\sigma^2}}{\mu n K} + \frac{L(\boxed{\sigma^2} + \boxed{\zeta_*^2}/\boxed{\delta})}{\boxed{\delta}\mu^2 K^2}\right)$$

$$\boxed{\mathbb{E}\|\mathcal{C}(x) - x\|^2 \leq (1-\delta)\|x\|^2}$$

$$\boxed{\mathbb{E}\left[\|g_i^k - \nabla f_i\left(x^k\right)\|^2 \mid x^k\right] \leq \sigma^2}$$

$$\boxed{\zeta_*^2 = \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*)\|^2}$$