# Towards Faster Decentralized Stochastic Optimization with Communication Compression


Rustem Islamov*[1] Yuan Gao*[2,3] Sebastian U. Stich[2]

*Equal contribution [1]University of Basel [2]CISPA Helmholtz Center for Information Security [3]University of Saarland


## Problem Formulation

We want to solve the finite-sum optimization problem

$f^\star = \arg\min_x f(x)$

Lower bounded non-convex

# clients

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

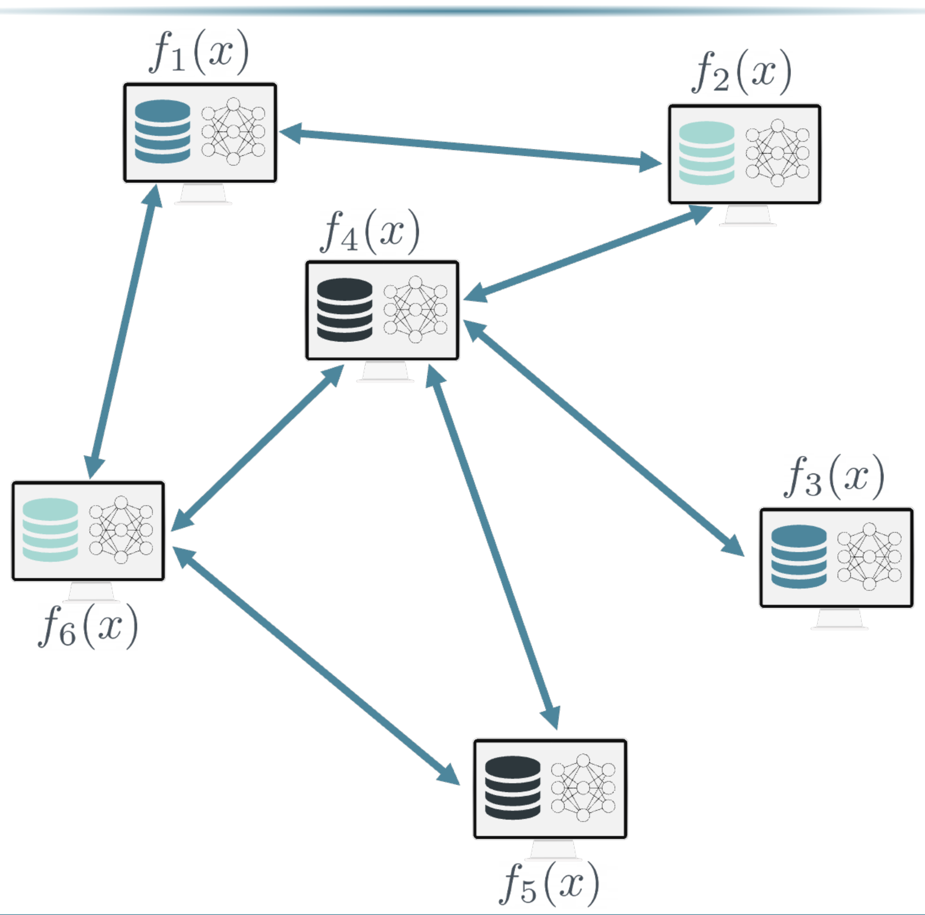
# model parameters

Empirical risk/loss

Local loss function $f_i(x) = \mathbb{E}_{\xi\sim\mathcal{D}_i}[f_\xi(x)]$

- This problem has many applications in machine learning, data science and engineering.

## Decentralized Communication Network




## Motivation

There is no algorithm that can achieve an optimal asymptotic convergence rate in the decentralized distributed optimization under assumptions **A1-A2** with contractive compression and without data heterogeneity bounds.

## Contractive Compression

We say that a (possibly randomized) mapping $\mathcal{C}\colon \mathbb{R}^d \to \mathbb{R}^d$ is a contractive compression operator if for some constant $0 < \alpha \le 1$ and all $\mathbf{x}\in\mathbb{R}^d$ it holds

$$\mathbb{E}\left[\|\mathcal{C}(\mathbf{x}) - \mathbf{x}\|^2\right] \le (1-\alpha)\|\mathbf{x}\|^2.$$

A classic example of contractive compression is Top-$K$ compressor.

$$\begin{pmatrix}-2\\1\\1.5\end{pmatrix} \xrightarrow{\text{Top-1}} \begin{pmatrix}-2\\0\\0\end{pmatrix},$$

which preserves top $K$ entries in magnitude. It is contractive with $\alpha = {}^K/_d$.

Table 1: Summary of convergence guarantees for decentralized methods supporting contractive compressors. We present the convergence in terms of $\mathbb{E}\|\nabla f(\mathbf{x}_{\rm out})\|^2 \le \varepsilon^2$ for specifically chosen $\mathbf{x}_{\rm out}$. Here $F^0 := \mathbb{E}[f(\mathbf{x}^0) - f^*]$, $L$ and $\ell$ are smoothness constants, $\rho$ is a spectral gap, and $\sigma^2$ is stochastic variance bound.

| Method | Asymptotic Complexity | Any Batches? | No Extra Assumptions? |
|---|---|---|---|
| Choco-SGD | $\frac{LF^0\sigma^2}{n\varepsilon^4}$ | ✓ | Bounded Gradients $\mathbb{E}[\|\nabla f_i(\mathbf{x},\xi)\|^2] \le G^2$ |
| BEER | $\frac{LF^0\sigma^2}{\alpha^2\rho^3\varepsilon^4}$ | Batch size of order $\frac{\sigma^2}{\alpha\varepsilon^2}$ | ✓ |
| CEDAS | $\frac{LF^0\sigma^2}{n\varepsilon^4}$ | ✓ | Additional Unbiased Compressor |
| DeepSqueeze | $\frac{LF^0\sigma^2}{n\varepsilon^4}$ | ✓ | Bounded Heterogeneity $n^{-1}\sum_i\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \zeta^2$ |
| DoCoM | $\frac{\ell F^0\sigma^3}{n\varepsilon^3}$ | ✓ | Sample-wise smoothness $\|\mathbf{g}^i(\mathbf{x}) - \mathbf{g}^i(\mathbf{y})\| \le \ell\|\mathbf{x}-\mathbf{y}\|$ |
| CDProxSGT | $\frac{LF^0\sigma^2}{\alpha^2\rho^2\varepsilon^4}$ | ✓ | ✓ |
| MoTEF | $\frac{LF^0\sigma^2}{n\varepsilon^4}$ | ✓ | ✓ |

## Proposed Algorithm

**Algorithm 1:** MoTEF

**Input:** $\mathbf{X}^0, \mathbf{x}_0\mathbf{1}^\top, \mathbf{G}^0, \mathbf{H}^0 = \mathbf{X}^0, \mathbf{V}^0, \gamma, \eta, \gamma, \mathcal{C}_\alpha$

**for** $t = 0, 1, \ldots T-1$ **do**

$\mathbf{X}^{t+1} = \mathbf{X}^t + \boxed{\gamma\mathbf{H}^t(\mathbf{W}-\mathbf{I})} - \boxed{\eta\mathbf{V}^t}$

$\mathbf{Q}_h^{t+1} = \mathcal{C}_\alpha(\mathbf{X}^{t+1} - \mathbf{H}^t)$

$\mathbf{H}^{t+1} = \mathbf{H}^t + \mathbf{Q}_h^{t+1}$

$\mathbf{M}^{t+1} = \boxed{(1-\lambda)\mathbf{M}^t + \lambda\nabla F(\mathbf{X}^{t+1})}$

$\mathbf{V}^{t+1} = \mathbf{V}^t + \boxed{\gamma\mathbf{G}^t(\mathbf{W}-\mathbf{I})} + \boxed{\mathbf{M}^{t+1} - \mathbf{M}^t}$

$\mathbf{Q}_g^{t+1} = \mathcal{C}_\alpha(\mathbf{V}^{t+1} - \mathbf{G}^t)$

$\mathbf{G}^{t+1} = \mathbf{G}^t + \mathbf{Q}_g^{t+1}$

**end**

Gossip mixing

Compressed Communication

Gradient step

Momentum

Momentum Tracking

## Assumptions & Convergence Theory

**(A1)** Let $f^\star := \mathrm{argmin}_{x\in\mathbb{R}^d} f(x) > -\infty$. Let $f$ and each $f_i$ be $L$-smooth, i.e., for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and $i\in[n]$

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L\|x-y\|,$$

**(A2)** Let stochastic gradient oracles $\mathbf{g}^i(\mathbf{x})\colon \mathbb{R}^d\to\mathbb{R}^d$ for each $f_i$ be unbiased and have bounded variance, i.e., for all $\mathbf{x}\in\mathbb{R}^d$

$$\mathbb{E}\left[\mathbf{g}^i(\mathbf{x})\right] = \nabla f_i(\mathbf{x}), \quad \mathbb{E}\left[\|\mathbf{g}^i(\mathbf{x}) - \nabla f_i(\mathbf{x})\|^2\right] \le \sigma^2.$$

**(A3)** Let the mixing matrix $\mathbf{W}\in\mathbb{R}^{n\times n}$ be symmetric ($\mathbf{W} = \mathbf{W}^\top$) and doubly stochastic ($\mathbf{W}\mathbf{1} = \mathbf{1}, \mathbf{1}^\top\mathbf{W} = \mathbf{1}^\top$) with eigenvalues $1 = |\lambda_1(\mathbf{W})| \ge |\lambda_2(\mathbf{W})| \ge \cdots \ge |\lambda_n(\mathbf{W})|$ and the spectral gap $\rho := 1 - |\lambda_2(\mathbf{W})| \in (0,1]$.

### General Non-Convex Setting

Assume that assumptions A1-A3 hold. Then there exist absolute constants $c_\eta, c_\lambda, c_\gamma$, and $\tau \le 1$ such that if we set the parameters $\gamma = c_\gamma\alpha\rho$, $\lambda = c_\lambda\alpha\rho^3\tau$, $\eta = c_\eta L^{-1}\alpha\rho^3\tau$, and choosing the initial batch size $B_{\rm init} \ge \lceil\frac{LF^0}{\sigma^2}\rceil$, then after at most

$$T = \mathcal{O}\left(\frac{\sigma^2}{n\varepsilon^4} + \frac{\sigma}{\alpha\rho^{5/2}\varepsilon^3} + \frac{1}{\alpha\rho^3\varepsilon^2}\right) LF^0 \quad (1)$$

iterations of MoTEF it holds $\mathbb{E}[\|\nabla f(\mathbf{x}_{\rm out})\|^2] \le \varepsilon^2$, where $\mathbf{x}_{\rm out}$ is chosen uniformly at random from $\{\overline{\mathbf{x}}_0, \ldots, \overline{\mathbf{x}}_{T-1}\}$, and $\mathcal{O}$ suppresses absolute constants.

### Convergence of Local Models

Assume that assumptions A1-A3 hold. Then with the same choices of parameters as above, the local models $\{\mathbf{x}_i^t\}_{i\in[n]}$ converge to the average model $\{\bar{\mathbf{x}}_t\}$. In particular, after at most

$$T = \mathcal{O}\left(\frac{\rho}{\alpha L\varepsilon^2} + \frac{\rho^8\sigma^2}{nL^3\varepsilon^4} + \frac{\rho^{7/2}L\sigma}{\alpha L^2\varepsilon^3}\right) F_0 \quad (2)$$

iterations of MoTEF, it holds that the consensus error $\Omega_T := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{x}_i^{\rm out} - \bar{\mathbf{x}}_{\rm out}\|^2] \le \varepsilon$. Moreover,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\nabla f(\mathbf{x}_i^{\rm out})\|^2] \le 2L^2\Omega_T + 2\mathbb{E}[\|\nabla f(\mathbf{x}_{\rm out})\|^2] \quad (3)$$

## Experiments

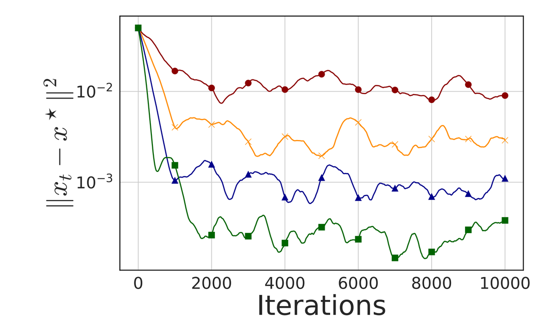



Figure 1: **Linear speedup** of MoTEF in number of clients $n$.

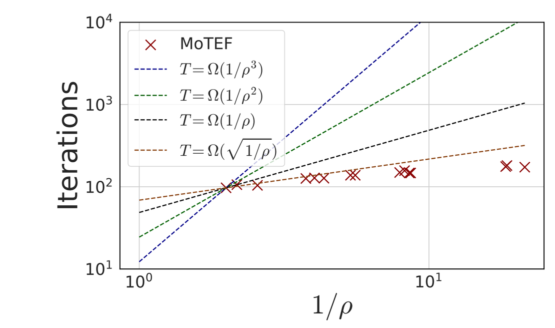



Figure 2: Empirical $\mathcal{O}(\sqrt{{}^1/_\rho})$ scaling of MoTEF to reach an error of $10^{-3}$; compared to $\mathcal{O}({}^1/_{\rho^3})$ scaling.

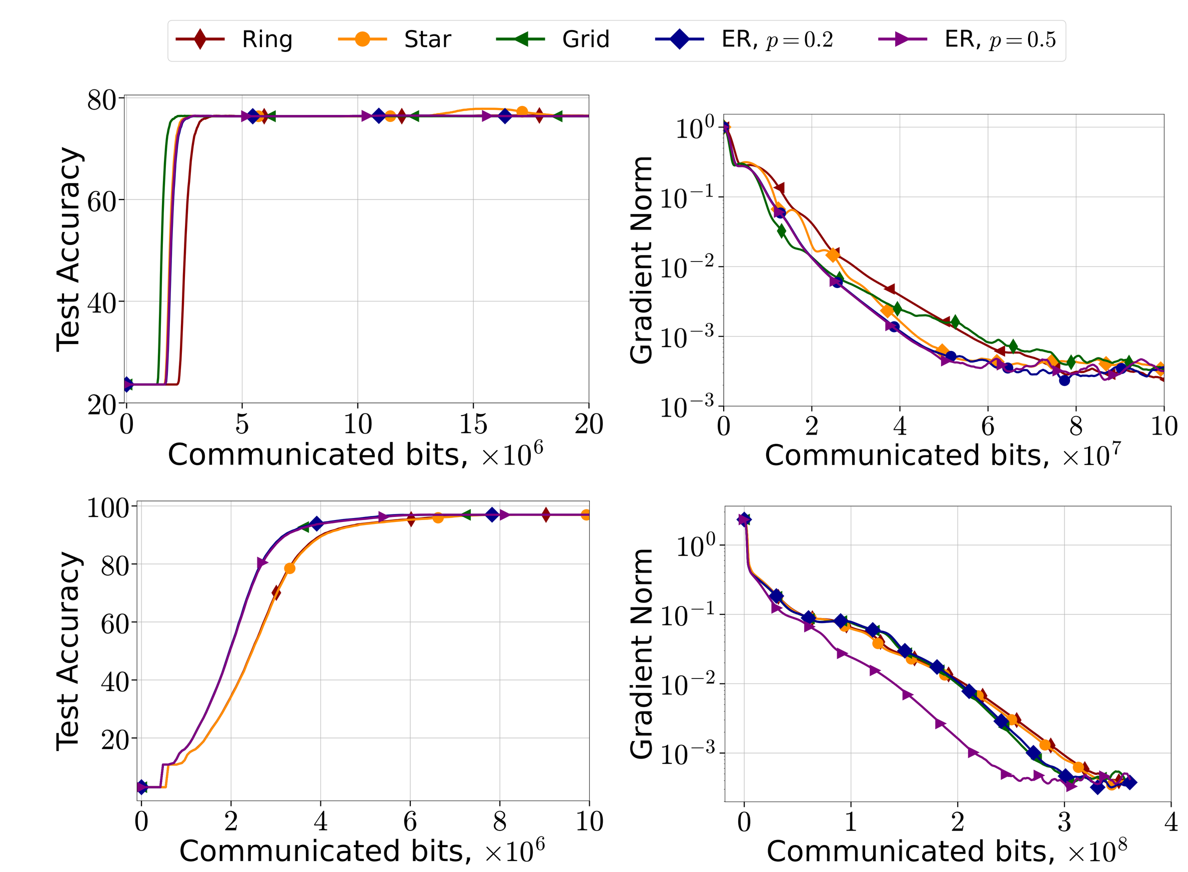



Figure 3: Performance of MoTEF changing of network topology tested on logistic regression with non-convex regularization.

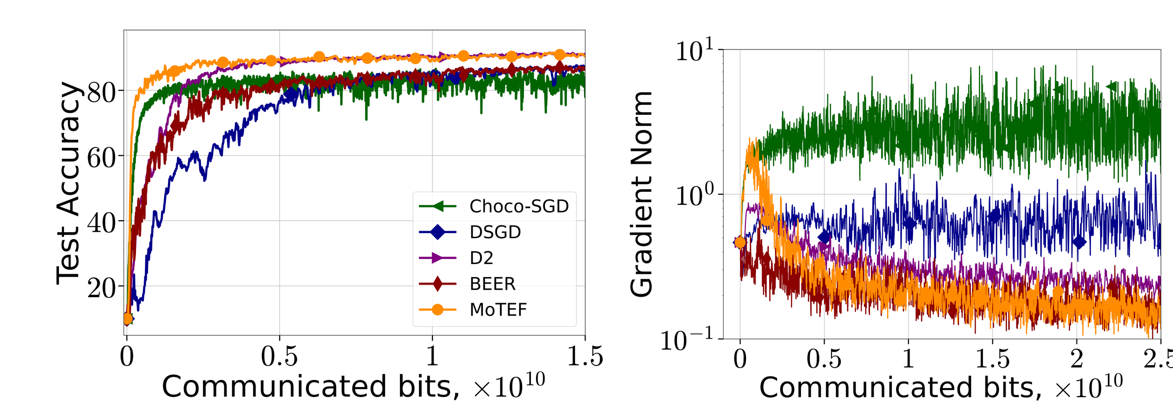



Figure 4: Comparison of MoTEF, BEER, Choco-SGD, DSGD, D2 in terms of communication complexity on training MLP with 1 hidden layer.

## References


[1] Zhao et al., BEER: Fast $O(1/T)$ rate for decentralized nonconvex optimization with communication compression, NeurIPS 2022.
[2] Richtárik et al., Ef21: A new, simpler, theoretically better, and practically faster error feedback, NeurIPS 2021.
[3] Yau & Wai, Docom: Compressed decentralized optimization with nearoptimal sample complexity, arXiv preprint arXiv:2202.00255, 2022.
[4] Koloskova et al., An improved analysis of gradient tracking for decentralized machine learning, NeurIPS 2021.