We consider distributed first-order optimization procedures (i.e. gradients). Let $f$ be a convex and geographically distant datacenters called multi-step primal-dual (MSPD) and its corresponding optimal convergence rate. Under global regularity, we provide a simple yet efficient algorithm called Distributed Randomized Smoothing (DRS) based on a local smoothing of the objective function, and show that DRS is within a $d^{1/4}$ multiplicative factor of the optimal convergence rate, where $d$ is the underlying dimension.

### Background & Setting

#### 1. Motivations

**Goals**
- Improved algorithms for distributed and decentralized optimization
- Impact of the communication network on learning
- Structure of the network vs. algorithm efficiency

**Applications**
- Machine Learning on the cloud
- Learning using geographically distant datacenters
- Empirical risk minimization: learning using

#### 2. Distributed optimization on networks

**Optimization problem**
Let $f_i$ be convex and $L_i$-Lipschitz functions. We consider minimizing the average of the local functions.

$$\min_{\theta} f(\theta) = \frac{1}{n} \sum_{i=1}^{n} f_i(\theta)$$

**Optimization procedures**
We consider distributed first-order optimization procedures (i.e. gradients).

**Network communications**
Let $G = (\mathcal{V}, \mathcal{E})$ be a connected simple graph of $n$ computing units and diameter $\Delta$, each having access to a function $f_i(\theta)$ over $\theta \in \mathbb{R}^d$. Local computations take a unit of time, while communications takes a time $\tau$. The average of the Lipschitz constants of local functions $f_i$.

### Local regularity

#### 3. Decentralized setting

- Local communication is performed through gossip (Boyd et al., 2006).
- Node $i$ knows $\sum_{j} W_{ij} x_j = (W x)_i$, where $W$ verifies:
  1. $W$ is an $n \times n$ symmetric positive semi-definite matrix.
  2. The kernel of $W$ is the set of constant vectors: $\ker(W) = \text{Span}(1)$.
- Let $\gamma(W) = \max_{i,j} |W_{ij}| / \lambda_i(W)$ be the (normalized) eigengap of $W$.

#### 4. Decentralized algorithms

**Optimal convergence rate**
For any $\gamma > 0$, there exists a gossip matrix $W$ of eigengap $\gamma$ and functions $f_i$ such that the time to reach a precision $\varepsilon > 0$ is lower bounded by

$$\Omega \left( \frac{R L}{\varepsilon} \frac{\tau}{\sqrt{\varepsilon}} \right),$$

where $R$ is the diameter of the search space $\Theta$, $L = \sqrt{\sum_{i=1}^{n} L_i^2}$ is the RMS average of the Lipschitz constants of local functions $f_i$.

**Optimal algorithm**
Novel algorithm called Multi-Step Primal-Dual (MSPD):
- Similar to the recent DCS algorithm, [Lan et al., 2017]
- Primal-dual reformulation,

$$\min_{\theta, \lambda} \max_{\omega} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) - \text{tr} \lambda^\top W \Theta W.$$

**Chambolle-Pock** algorithm for saddle-point optimization,

- Minimizing $f(\theta) = \Phi(\theta) + \sum_{i=1}^{n} \lambda_i f_i(\theta)$
- $\theta_{t+1} = \argmin_{\theta} \frac{1}{n} \sum_{i=1}^{n} f_i(\theta) - \theta_{t}^\top \lambda_{t+1} + \frac{1}{2n} \|\theta - \theta_{t}\|^2, \forall i \in \{1, \ldots, n\}$.

- Approximation of the proximal operator using subgradient descent, $\sum_{i=1}^{n} f_i(\theta)$ where $f_i$ is the RMS average of the local functions.

**Gossip accelerated using Chebyshev polynomials**

$$W \leftarrow \Phi_N(W)$$

where $\Phi_N$ is a polynomial of degree at most $K$ and $\gamma(\Phi_N(W))$ is maximal.

### Global regularity

#### 6. Centralized algorithms

**Optimal convergence rate**
For any graph of diameter $\Delta$ and any block-box procedure, there exist functions $f_i$ such that the time to reach a precision $\varepsilon > 0$ is lower bounded by

$$\Omega \left( \frac{R L}{\varepsilon} \frac{\tau}{\sqrt{\varepsilon}} \right),$$

where $L_{i}$ is the Lipschitz constant of the average (global) function $f$.

**Efficient algorithm: Distributed Randomized Smoothing (DRS)**
- Extension of Randomized Smoothing to dist. opt., [Duchi et al., 2012]
- Uses extra computation steps to smooth $f$ and improve convergence, $f(\theta) = \Phi(\theta + \gamma N(0, I))$ where $N \sim N(0, I)$,

$$O\left( \frac{R L}{\varepsilon} \frac{\tau}{\sqrt{\varepsilon}} \right).$$

### Conclusion

- Optimal decentralized convergence rate: $\Theta \left( \frac{R L}{\varepsilon} \frac{\tau}{\sqrt{\varepsilon}} \right)$
- Optimal centralized convergence rate:
  - Lower bound: $\Omega \left( \frac{R L}{\varepsilon} \frac{\tau}{\sqrt{\varepsilon}} \right)$
  - Upper bound: $O \left( \frac{R L}{\varepsilon} \frac{\tau}{\sqrt{\varepsilon}} \right)$
- Early stages are fast and rely on efficient communication.
- Late stages are slow and do not depend on the communication network.

### References


---

Kevin Scaman\textsuperscript{1} Francis Bach\textsuperscript{2} Sébastien Bubeck\textsuperscript{3} Yin Tat Lee\textsuperscript{4,5} Laurent Massoulié\textsuperscript{2,5}

\textsuperscript{1}Huawei Noah’s Ark Lab, \textsuperscript{2}INRIA, ENS, PSL, \textsuperscript{3}Microsoft Research, \textsuperscript{4}University of Washington, \textsuperscript{5}MSR-INRIA Joint Centre