Optimal Statistical Rates for Decentralised Non-Parametric Regression with Linear Speed-Up

Patrick Rebeschini & Dominic Richards

University of Oxford Department of Statistics

EPSRC Engineering and Physical Sciences Research Council

OxWaSP Oxford-Warwick Statistics Programme

UNIVERSITY OF OXFORD

MRC Medical Research Council

Desiderata for Distributed Machine Learning

Suppose n agents solving a machine learning problem. Properties of an ideal distributed algorithm:

Statistics: retain optimal statistical precision

Runtime: speed-up over single-agent due to parallel computing: ideally factor n

Communication: fixed cost per agent per step, ideally independent of n

Consensus Optimisation for Decentralised Learning

Network of agents $G = (V, E)$, each with data points sampled i.i.d $(x_{i,v}, y_{i,v})$ perform linear regression. Therefore each agent wishes to minimise with coefficient $\boldsymbol{\omega} \in \mathbb{R}^d$

$$F(\boldsymbol{\omega}) = \frac{1}{\mathrm{n}} \sum_{\mathrm{v} \in \mathrm{V}} \underbrace{\frac{1}{\mathrm{m}} \sum_{\mathrm{i} \in [\mathrm{m}]} (\langle \boldsymbol{\omega}, x_{i,v} \rangle - y_{i,v})^2}_{\text{Empirical Loss for Agent } v} = \underbrace{\frac{1}{\mathrm{n}} \sum_{\mathrm{v} \in \mathrm{V}} F_v(\boldsymbol{\omega})}_{\text{Consensus Optimisation}},$$

where F_v is function held by agent v.

Consensus Optimisation

Suppose F_v arbitrary and each agent wants to minimise $\frac{1}{\mathrm{n}} \sum_{\mathrm{v} \in \mathrm{V}} F_v$.

→ Agents alternate: local gradient descent steps on F_v and local averaging on network [4]

✓ Low communication cost per agent for sparse graphs

✓ Robust/Decentralised as no single node responsible for disseminating information

✗ Performance depends on network topology [2, 5].

Graphs with smaller spectral gap benefit less from decentralisation

In machine learning $\mathbf{F_v}$ are often not arbitrary

In our case, concentration states for $v \in V$

$$F_v(\boldsymbol{\omega}) = \frac{1}{\mathrm{m}} \sum_{\mathrm{i} \in [\mathrm{m}]} (\langle \boldsymbol{\omega}, x_{i,v} \rangle - y_{i,v})^2 \xrightarrow{\mathrm{m} \to \infty} \underbrace{\int_{X \times Y} (\langle \boldsymbol{\omega}, x \rangle - y)^2 d\rho(x, y)}_{\text{Test Risk}} =: \mathcal{E}(\boldsymbol{\omega})$$

where ρ is the distribution of the data points.

In large data scenario, all functions are converging to the same quantity → Test Risk.

Main Question:

Can concentration speed-up consensus learning for any network topology ?

Non-parametric Statistical Assumptions

Use tools from non-parametric regression [1].

Predictor minimising Test Risk over set of linear predictors $x \to \langle \omega, \cdot \rangle$ denoted f_H.

Noise: Exists $M \in (0, \infty)$, $\nu \in (1, \infty)$ so for any $\ell \in \mathbb{N}$ we have $\int_Y y^{2\ell} d\rho(x|y) \leq \nu \ell! M^\ell$.

Difficulty of estimation problem:

For $f \in L^2(H, \rho_X)$ let $\mathcal{L}_\rho(f) = \int_X \langle x \cdot \rangle f(x) d\rho_X(x)$. Exists $r > 0$ such that $\|\mathcal{L}_\rho^{-r} f_H\| < \infty$.

Spectrum of covariance operator:

Exists $\gamma \in (0, 1]$, $c_\gamma > 0$ such that $\mathrm{Tr}(\mathcal{L}_\rho(\mathcal{L}_\rho + \lambda)^{-1}) \leq c_\gamma \lambda^{-\gamma}$ for all $\lambda > 0$.

Distributed Gradient Descent

Consider a simple consensus optimisation algorithm [4]:

Communication Matrix: $\mathbf{P} \in \mathbb{R}^{n \times n}$ symmetric doubly stochastic matrix support on the graph

$$\mathbf{P} = \mathbf{P}^\top, \quad \mathbf{P}\mathbf{1} = \mathbf{1} \quad \text{and for } v \neq w \quad \underbrace{\mathbf{P_{vw} \neq 0 \text{ only if } (v, w) \in E}}_{\text{Sparsity pattern matches network}}$$

Network Dependence: Let σ_2 be second largest eigenvalue in magnitude for $\mathbf{P}$.
Scaling $O((\mathbf{1} - \sigma_\mathbf{2})^{-\mathbf{1}}) = \mathbf{n^2}$ (Cycle), $\mathbf{n}$ (Grid/Random Geo.), and $\mathbf{1}$ (Complete/Expander).

Algorithm: Initalised $w_{1,v} = 0$ for $v \in V$, iterates updated for all $v \in V$

$$\boldsymbol{\omega}_{\mathbf{t+1,v}} = \underbrace{\sum_{\mathrm{w} \in \mathrm{V}} \mathbf{P_{vw}}}_{\text{Local Communication}} \underbrace{\Big(\boldsymbol{\omega}_{\mathbf{t,w}} - \eta \frac{\mathbf{1}}{\mathbf{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} (\langle \boldsymbol{\omega}_{\mathbf{t,w}}, x_{i,w} \rangle_H - y_{i,w}) x_{i,w}\Big)}_{\text{Local Gradient Descent}}$$

Implict Regularisation:
Model Complexity controlled through early stopping (Iterations t) and step size η.

Optimal Statistical Rate with Implicit Regularisation

Let aforementioned assumptions hold with $r \geq 1/2$ and $2r + \gamma \geq 2$. Fix

$$t = \underbrace{(\mathrm{nm})^{\mathbf{1}/(\mathbf{2r}+\gamma)}}_{\text{Single-Machine Iterations}} \times \begin{cases} \left(\frac{(nm)^{2r/(2r+\gamma)}}{\mathrm{m}(1-\sigma_2)^\gamma}\right)^{1/\gamma} \vee 1 & \text{if } m \geq n^{2r/\gamma} \\ \frac{(nm)^{r/(2r+\gamma)}}{\sqrt{\mathrm{m}}(1-\sigma_2)} & \text{otherwise} \end{cases}$$

$\eta = \frac{\kappa^{-2}(nm)^{1/(2r+\gamma)}}{t}$ and let $m \geq n^{\frac{2r+2+\gamma}{2r+\gamma-2}}$ then $\forall v \in V$:

$$\underbrace{\mathbf{E}[\mathcal{E}(\boldsymbol{\omega}_{\mathbf{t+1,v}})] - \inf_\omega \mathcal{E}(\omega)}_{\text{Test Error}} \lesssim \underbrace{(\mathrm{nm})^{-\mathbf{2r}/(\mathbf{2r}+\gamma)}}_{\text{Optimal Statistical Rate}}$$

Time Model

Gradient computation costs 1 unit of time.
Communication delay costs τ units of time, for some $\tau > 0$.
Distributed time per iteration = $\underbrace{m}_{\text{Local Gradient Computation}} + \underbrace{\tau + \mathrm{Deg}(P)}_{\text{Communicating/Aggreagating Neighbours Information}}$

Speed-Up defined as

$$\text{Speed-Up} = \frac{\text{Single Machine Run time}}{\text{Distributed Run time}} = \frac{\text{Single Machine Iterations}}{\text{Distributed Iterations}} \times \underbrace{\frac{nm}{m + \tau + \mathrm{Deg}(P)}}_{\text{Ratio of Time Per Iteration}}$$

Main Result: Speed-Up with Consensus Methods Utilising Concentration

Iterations required decreasing in number of samples m up to a point

$$\underbrace{\mathrm{m} \geq \frac{n^{2r/\gamma}}{(1 - \sigma_2)^{2\mathrm{r}+\gamma}} \vee n^{\frac{2r+2+\gamma}{2r+\gamma-2}}}_{\text{Sufficiently Many Samples}} \implies \text{Distributed Iterations} = \text{Single Machine Iterations}$$

$$\implies \text{Speed-Up} \overset{\tau + \mathrm{Deg}(P) = O(m)}{=} O(n)$$

therefore linear speed-up for any network topology.

	Cycle	Grid	R. Geom.	Complete	Expander
Speed-Up	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$
Communication	$O(1)$	$O(1)$	$O(1)$	$O(n)$	$O(1)$

Speed-Up with Single-Step Consensus Methods

Single-Step methods typically require iterations to scale with inverse spectral gap e.g. [2, 5]

$$\text{Distributed Iterations} = \text{Single Machine Iterations} \times (1 - \sigma_2)$$

$$\Downarrow$$

$$\text{Speed-Up} = \frac{nm}{m + \tau + \mathrm{Deg}(P)} (1 - \sigma_2) \overset{\tau + \mathrm{Deg}(P) = O(m)}{=} O(n(1 - \sigma_2))$$

therefore linear speed-up restricted to well connected topologies.

	Cycle	Grid	R. Geom.	Complete	Expander
Speed-Up	$O(1/n)$	$O(1)$	$O(1)$	$O(n)$	$O(n)$
Communication	$O(1)$	$O(1)$	$O(1)$	$O(n)$	$O(1)$

Speed-Up with Multi-Step Consensus Methods

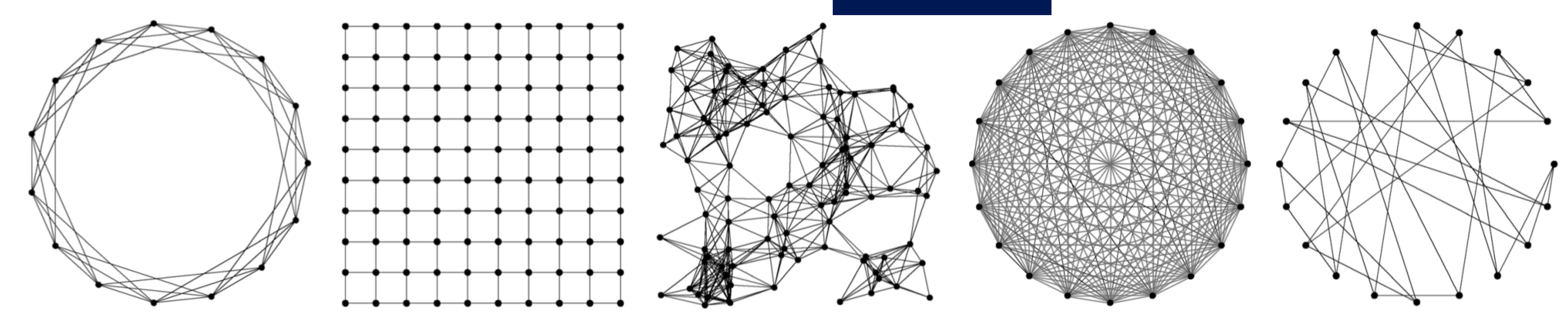
Multi-Step methods perform multiple communication steps between gradient descent steps [5],

$$\text{Distributed Iterations} = \text{Single Machine Iterations}$$

But communication rounds scales with inverse spectral gap.

	Cycle	Grid	R. Geom.	Complete	Expander
Speed-Up	$O(n)$	$O(n)$	$O(n)$	$O(n)$	$O(n)$
Communication	$O(n)$	$O(\sqrt{n})$	$O(\sqrt{n})$	$O(n)$	$O(1)$

Therefore trade off between Speed-Up and Communication Cost.



Detailed Error Decomposition

The Test Error is decomposed as follows when $m \geq n^{2r/\gamma}$

$$\textbf{Test Error} \lesssim \underbrace{(\eta t)^{-2r}}_{\text{Bias}} + \underbrace{\frac{(\eta t/(nm)^{1/(2r+\gamma)})^2}{(nm)^{2r/(2r+\gamma)}}}_{\text{Sample Variance}} + \underbrace{\frac{\eta^\gamma}{\mathrm{m}(\mathbf{1} - \sigma_\mathbf{2})^\gamma}}_{\text{Population Network Error}} + \underbrace{\frac{(\eta t)^{\gamma+2}}{m^2}}_{\text{Residual Network Error}}$$

Bias and **Sample Variance** align with Gradient Descent with nm samples [3]

$\implies$ Fix $\eta t = (nm)^{1/(2r+\gamma)}$.

Population and **Residual Network Error**: arise due from averaging steps with the matrix $\mathbf{P}$.

Population Network Error

Follows standard network term [2]

Decreasing with step size η

Depends on spectral gap $(\mathbf{1} - \sigma_\mathbf{2})$

Due to **concentration, decreasing with m**

Small by picking $\eta = \left(\frac{\mathrm{m}(\mathbf{1}-\sigma_\mathbf{2})^\gamma}{(nm)^{2r/(2r+\gamma)}}\right)^{1/\gamma} \vee 1.$

Residual Network Error

Higher order term

From empirical covariance multiplying the iterates at each iteration

Utilise **concentration** and **contraction** to control

$\implies$ require $m \geq n^{\frac{2r+2+\gamma}{2r+\gamma-2}}$.

Proof Sketch: Population Network Error

Let $\mathcal{T}_\rho$ be conjugate of $\mathcal{L}_\rho$ and for $k \geq 1$, let $\mathrm{N_k}$ be r.v. with zero mean concentrating to zero such that $\|(\mathcal{T}_\rho + \lambda \mathbf{I})^{-\mathbf{1}/\mathbf{2}} \mathbf{N_k}\| \lesssim \mathbf{1}/\sqrt{\lambda^\gamma \mathbf{m}}$ w.h.p. Then, with mixing time $t^\star \simeq (\mathbf{1} - \sigma_\mathbf{2})^{-1}$

$$\mathbf{E}[(\textbf{Pop. Net. Error})] \leq \mathbf{E}\Big[\Big(\sum_{k=1}^t \sigma_\mathbf{2}^{\mathbf{t-k+1}} \eta \|\mathcal{T}_\rho^{1/2}(I - \eta \mathcal{T}_\rho)^{t-k} \mathbf{N_k}\|\Big)^2\Big]$$

$$\lesssim \frac{\mathbf{1}}{\mathbf{m}\lambda^\gamma} \Big(\underbrace{\sum_{k=t-\mathbf{t}^\star}^{t}}_{\mathbf{t}^\star \text{ Poorly Mixed terms}} \sigma_\mathbf{2}^{\mathbf{t-k+1}} \eta \underbrace{\|\mathcal{T}_\rho^{1/2}(I - \eta \mathcal{T}_\rho)^{t-k}(\mathcal{T}_\rho + \lambda I)^{1/2}\|}_{\text{Contraction} \implies O(\sqrt{\lambda/(\eta(t-k))})} \Big)^2 \lesssim \frac{\log(t^\star) + \lambda \eta \mathbf{t}^\star}{\mathbf{m}\lambda^\gamma}$$

Optimised by picking $\lambda = (\eta \mathbf{t}^\star)^{-1}$.

Future Work

- **Non-parametric setting**: extending analysis to non-attainable case $r \leq 1/2$ and tighter analysis on **Residual Network Error**.
- **General Loss Function**: Squared loss yields bias/variance decomposition, concentration likely hold for more general losses.
- **Statistics/Communication trade off with sparse/randomised gossip**: linear speed-up independent of topology → agents randomly gossip at each iteration.
- **Stochastic Gradient Descent and mini-batches**: random subset of data at each iteration [3].

References

[1] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

[2] John C. Duchi, Alekh Agarwal, and Martin J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592–606, 2012.

[3] Junhong Lin and Lorenzo Rosasco. Optimal rates for multi-pass stochastic gradient methods. *Journal of Machine Learning Research*, 18(97):1–47, 2017.

[4] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[5] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th ICML*, pages 3027–3036. JMLR. org, 2017.