Linear Convergence for Distributed Optimization Without Strong Convexity

Xinlei Yi

Joint work with Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H. Johansson

December, 2020

School of Electrical Engineering and Computer Science
KTH Royal Institute of Technology
Stockholm, Sweden

Distributed optimization

A network of agents cooperatively solve a global optimization problem, where

- each agent i has a local private objective $f_i(x)$

Distributed optimization

A network of agents cooperatively solve a global optimization problem, where

- each agent i has a local private objective $f_i(x)$
- all agents collaborate together to find the solution to minimize $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

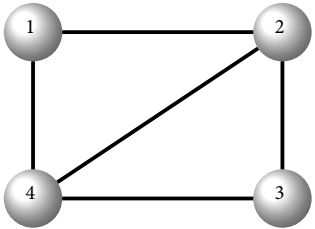
$$\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

Distributed optimization

A network of agents cooperatively solve a global optimization problem, where

- each agent i has a local private objective $f_i(x)$
- all agents collaborate together to find the solution to minimize $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
- agents exchange information through the underlying communication network $\mathcal{G}$

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

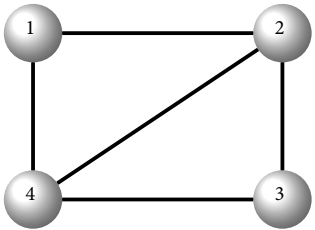


Distributed optimization

A network of agents cooperatively solve a global optimization problem, where

- each agent i has a local private objective $f_i(x)$
- all agents collaborate together to find the solution to minimize $f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$
- agents exchange information through the underlying communication network $\mathcal{G}$

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



- Distributed optimization summarizes many popular machine learning models, e.g., deep learning and federated learning [Dean et al, NeurIPS, 2012]
- Distributed algorithms outperform centralized algorithms in some applications, e.g., training neural networks [Lian et al, NeurIPS, 2017]

Motivation

Existing algorithms

Continuous- and discrete-time distributed algorithms

Motivation

Existing algorithms

Continuous- and discrete-time distributed algorithms

Existing result

A standard assumption for proving exponential/linear convergence of existing distributed algorithms is

strong convexity of the cost functions

Motivation

Existing algorithms

Continuous- and discrete-time distributed algorithms

Existing result

A standard assumption for proving exponential/linear convergence of existing distributed algorithms is

strong convexity of the cost functions

Question

Could strong convexity be relaxed?

For example, quadratic functions may be not strongly convex.

Motivation

Existing algorithms

Continuous- and discrete-time distributed algorithms

Existing result

A standard assumption for proving exponential/linear convergence of existing distributed algorithms is

strong convexity of the cost functions

Question

Could strong convexity be relaxed?
For example, quadratic functions may be not strongly convex.

Answer in our paper

Yes, it can be relaxed by the Polyak–Łojasiewicz condition.

Polyak–Łojasiewicz condition

Polyak–Łojasiewicz (P–Ł) condition

The function $f(x)$ satisfies the P–Ł condition with constant $\nu > 0$ if

$$\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*), \ \forall x \in \mathbb{R}^p.$$

Polyak–Łojasiewicz condition

Polyak–Łojasiewicz (P–Ł) condition

The function $f(x)$ satisfies the P–Ł condition with constant $\nu > 0$ if

$$\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*), \ \forall x \in \mathbb{R}^p.$$

- Examples: $f(x) = \|Ax\|^2$ and $f(x) = x^2 + 3\sin^2(x)$.

Polyak–Łojasiewicz condition

Polyak–Łojasiewicz (P–Ł) condition

The function $f(x)$ satisfies the P–Ł condition with constant $\nu > 0$ if

$$\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*), \ \forall x \in \mathbb{R}^p.$$

- Examples: $f(x) = \|Ax\|^2$ and $f(x) = x^2 + 3\sin^2(x)$.
- P–Ł does not imply convexity.
- P–Ł implies that every stationary point is a global minimizer.
- Every strongly convex function satisfies P–Ł.
- It is difficult to verify P–Ł in general.

Polyak–Łojasiewicz condition

Polyak–Łojasiewicz (P–Ł) condition

The function $f(x)$ satisfies the P–Ł condition with constant $\nu > 0$ if

$$\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*), \ \forall x \in \mathbb{R}^p.$$

- Examples: $f(x) = \|Ax\|^2$ and $f(x) = x^2 + 3\sin^2(x)$.
- P–Ł does not imply convexity.
- P–Ł implies that every stationary point is a global minimizer.
- Every strongly convex function satisfies P–Ł.
- It is difficult to verify P–Ł in general.

One special case

If $g : \mathbb{R}^p \to \mathbb{R}$ is a strongly convex function and $A \in \mathbb{R}^{p \times p}$ is a matrix, then $f(x) = g(Ax)$ satisfies the P–Ł condition.

Derivation of primal–dual gradient descent algorithm

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

Derivation of primal–dual gradient descent algorithm

$$\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{is equivalent to} \quad \min_{\boldsymbol{x}\in\mathbb{R}^{np}} \tilde{f}(\boldsymbol{x}) := \sum_{i=1}^{n} f_i(x_i)$$

subject to $\boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np}$,

where $\boldsymbol{x} = \mathrm{col}(x_1, \ldots, x_n)$ and $\boldsymbol{L} = L \otimes \mathbf{I}_p$ with L being the Laplacian.

Derivation of primal–dual gradient descent algorithm

$$\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{is equivalent to} \quad \min_{\boldsymbol{x}\in\mathbb{R}^{np}} \tilde{f}(\boldsymbol{x}) := \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } \boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np},$$

where $\boldsymbol{x} = \mathrm{col}(x_1, \ldots, x_n)$ and $\boldsymbol{L} = L \otimes \mathbf{I}_p$ with L being the Laplacian.

Associated augmented Lagrangian:

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x},$$

where $\boldsymbol{u} \in \mathbb{R}^{np}$ is the dual variable, $\alpha, \beta > 0$ are parameters.

Derivation of primal–dual gradient descent algorithm

$$\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{is equivalent to} \quad \min_{\boldsymbol{x}\in\mathbb{R}^{np}} \tilde{f}(\boldsymbol{x}) := \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } \boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np},$$

where $\boldsymbol{x} = \mathrm{col}(x_1, \ldots, x_n)$ and $\boldsymbol{L} = L \otimes \mathbf{I}_p$ with L being the Laplacian.

Associated augmented Lagrangian:

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x},$$

where $\boldsymbol{u} \in \mathbb{R}^{np}$ is the dual variable, $\alpha, \beta > 0$ are parameters.

Minimize $\mathcal{A}(\boldsymbol{x}, \boldsymbol{u})$ with a primal–dual algorithm:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \frac{\partial \mathcal{A}(\boldsymbol{x}_k, \boldsymbol{u}_k)}{\partial \boldsymbol{x}_k} = \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{L}^{1/2}\boldsymbol{u}_k + \nabla \tilde{f}(\boldsymbol{x}_k)),$$

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \eta \frac{\partial \mathcal{A}(\boldsymbol{x}_k, \boldsymbol{u}_k)}{\partial \boldsymbol{u}_k} = \boldsymbol{u}_k + \eta\beta \boldsymbol{L}^{1/2}\boldsymbol{x}_k,$$

where $\eta > 0$ is a fixed stepsize.

Derivation of primal–dual gradient descent algorithm

$$\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \quad \text{is equivalent to} \quad \min_{\boldsymbol{x}\in\mathbb{R}^{np}} \tilde{f}(\boldsymbol{x}) := \sum_{i=1}^{n} f_i(x_i)$$

$$\text{subject to } \boldsymbol{L}^{1/2}\boldsymbol{x} = \boldsymbol{0}_{np},$$

where $\boldsymbol{x} = \mathrm{col}(x_1, \ldots, x_n)$ and $\boldsymbol{L} = L \otimes \mathbf{I}_p$ with L being the Laplacian.

Associated augmented Lagrangian:

$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x},$$

where $\boldsymbol{u} \in \mathbb{R}^{np}$ is the dual variable, $\alpha, \beta > 0$ are parameters.

Minimize $\mathcal{A}(\boldsymbol{x}, \boldsymbol{u})$ with a primal–dual algorithm:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \frac{\partial \mathcal{A}(\boldsymbol{x}_k, \boldsymbol{u}_k)}{\partial \boldsymbol{x}_k} = \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{L}^{1/2}\boldsymbol{u}_k + \nabla \tilde{f}(\boldsymbol{x}_k)),$$

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + \eta \frac{\partial \mathcal{A}(\boldsymbol{x}_k, \boldsymbol{u}_k)}{\partial \boldsymbol{u}_k} = \boldsymbol{u}_k + \eta\beta \boldsymbol{L}^{1/2}\boldsymbol{x}_k,$$

where $\eta > 0$ is a fixed stepsize.

With $\boldsymbol{v}_k = \boldsymbol{L}^{1/2}\boldsymbol{u}_k$, this algorithm can be rewritten as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{v}_k + \nabla \tilde{f}(\boldsymbol{x}_k)),$$

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k + \eta\beta \boldsymbol{L}\boldsymbol{x}_k.$$

Distributed primal–dual gradient descent algorithm

$$\begin{aligned}\boldsymbol{x}_{k+1} =& \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{v}_k + \nabla \tilde{f}(\boldsymbol{x}_k)),\\ \boldsymbol{v}_{k+1} =& \boldsymbol{v}_k + \eta\beta \boldsymbol{L}\boldsymbol{x}_k.\end{aligned}$$

Distributed Primal–Dual Gradient Descent Algorithm

$$\begin{aligned}x_{i,k+1} &= x_{i,k} - \eta\Big(\alpha \sum\nolimits_{j=1}^{n} L_{ij}x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k})\Big),\\ v_{i,k+1} &= v_{i,k} + \eta\beta \sum\nolimits_{j=1}^{n} L_{ij}x_{j,k}.\end{aligned}$$

Distributed primal–dual gradient descent algorithm

$$\begin{aligned}\boldsymbol{x}_{k+1} =& \boldsymbol{x}_k - \eta(\alpha \boldsymbol{L}\boldsymbol{x}_k + \beta \boldsymbol{v}_k + \nabla \tilde{f}(\boldsymbol{x}_k)),\\ \boldsymbol{v}_{k+1} =& \boldsymbol{v}_k + \eta\beta \boldsymbol{L}\boldsymbol{x}_k.\end{aligned}$$

Distributed Primal–Dual Gradient Descent Algorithm

$$\begin{aligned}x_{i,k+1} &= x_{i,k} - \eta\Big(\alpha \sum\nolimits_{j=1}^{n} L_{ij}x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k})\Big),\\ v_{i,k+1} &= v_{i,k} + \eta\beta \sum\nolimits_{j=1}^{n} L_{ij}x_{j,k}.\end{aligned}$$

This algorithm is **single-loop** and communicates **only one** variable.

Linear convergence to **global optima**

Distributed Primal–Dual Gradient Descent Algorithm

$$x_{i,k+1} = x_{i,k} - \eta\Big(\alpha \sum\nolimits_{j=1}^{n} L_{ij} x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k})\Big),$$
$$v_{i,k+1} = v_{i,k} + \eta\beta \sum\nolimits_{j=1}^{n} L_{ij} x_{j,k}.$$

Theorem

If each $f_i(x)$ is smooth ($\nabla f_i(x)$ is Lipschitz continuous) and $f(x)$ satisfies P–Ł, then the primal–dual gradient descent algorithm linearly converges to a global optimum: there exists $\rho \in (0,1)$ such that

$$\underbrace{\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2}_{\text{Consensus}} + \underbrace{f(\bar{x}_k) - f^*}_{\text{Optimization}} = \mathcal{O}(\rho^k).$$

Linear convergence to **global optima**

Distributed Primal–Dual Gradient Descent Algorithm

$$x_{i,k+1} = x_{i,k} - \eta\Big(\alpha \sum\nolimits_{j=1}^{n} L_{ij} x_{j,k} + \beta v_{i,k} + \nabla f_i(x_{i,k})\Big),$$
$$v_{i,k+1} = v_{i,k} + \eta\beta \sum\nolimits_{j=1}^{n} L_{ij} x_{j,k}.$$

Theorem

If each $f_i(x)$ is smooth ($\nabla f_i(x)$ is Lipschitz continuous) and $f(x)$ satisfies P–Ł, then the primal–dual gradient descent algorithm linearly converges to a global optimum: there exists $\rho \in (0,1)$ such that

$$\underbrace{\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2}_{\text{Consensus}} + \underbrace{f(\bar{x}_k) - f^*}_{\text{Optimization}} = \mathcal{O}(\rho^k).$$

Remark

- Linear convergence is achieved without strong convexity, even without convexity.
- Parameters α, β, η do not have to depend on ν.

Linear convergence to **global optima**

P–Ł condition: $\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*),\ \forall x \in \mathbb{R}^p$.

Theorem

If each $f_i(x)$ is smooth and $f(x)$ satisfies P–Ł, there exists $\rho \in (0,1)$ such that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* = \mathcal{O}(\rho^k).$$

Proof sketch:

Consider the nonnegative potential function:

$$V_k = \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}} + \boldsymbol{x}_k^\top \boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right) + n(f(\bar{x}_k) - f^*),$$

where $\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} = \sum_{i=1}^n \|x_{i,k} - \bar{x}_k\|^2$ and $\boldsymbol{g}_k^0 = \mathrm{col}(\nabla f_1(\bar{x}_k), \ldots, \nabla f_n(\bar{x}_k))$.

Linear convergence to **global optima**

P–Ł condition: $\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*),\ \forall x \in \mathbb{R}^p$.

Theorem

If each $f_i(x)$ is smooth and $f(x)$ satisfies P–Ł, there exists $\rho \in (0,1)$ such that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* = \mathcal{O}(\rho^k).$$

Proof sketch:

Consider the nonnegative potential function:

$$V_k = \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}} + \boldsymbol{x}_k^\top \boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right) + n(f(\bar{x}_k) - f^*),$$

where $\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} = \sum_{i=1}^n \|x_{i,k} - \bar{x}_k\|^2$ and $\boldsymbol{g}_k^0 = \mathrm{col}(\nabla f_1(\bar{x}_k), \ldots, \nabla f_n(\bar{x}_k))$.

This function is nonincreasing:

$$V_{k+1} \leq V_k - \left[c_1\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2_{\boldsymbol{K}} + c_3 n\|\nabla f(\bar{x}_k)\|^2\right].$$

Linear convergence to **global optima**

P–Ł condition: $\|\nabla f(x)\|^2 \geq \nu(f(x) - f^*),\ \forall x \in \mathbb{R}^p$.

Theorem

If each $f_i(x)$ is smooth and $f(x)$ satisfies P–Ł, there exists $\rho \in (0,1)$ such that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* = \mathcal{O}(\rho^k).$$

Proof sketch:

Consider the nonnegative potential function:

$$V_k = \frac{1}{2}\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + \frac{1}{2}\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2_{\boldsymbol{Q}+\frac{\alpha}{\beta}\boldsymbol{K}} + \boldsymbol{x}_k^\top \boldsymbol{K}\left(\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right) + n(f(\bar{x}_k) - f^*),$$

where $\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} = \sum_{i=1}^n \|x_{i,k} - \bar{x}_k\|^2$ and $\boldsymbol{g}_k^0 = \mathrm{col}(\nabla f_1(\bar{x}_k), \ldots, \nabla f_n(\bar{x}_k))$.

This function is nonincreasing:

$$V_{k+1} \leq V_k - \left[c_1\|\boldsymbol{x}_k\|^2_{\boldsymbol{K}} + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|^2_{\boldsymbol{K}} + c_3 n\|\nabla f(\bar{x}_k)\|^2\right].$$

P–Ł implies

$$\|\nabla f(\bar{x}_k)\|^2 \geq \nu(f(\bar{x}_k) - f^*).$$

Linear convergence to **global optima**

Theorem

If each $f_i(x)$ is smooth and $f(x)$ satisfies P–Ł, there exists $\rho \in (0,1)$ such that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* = \mathcal{O}(\rho^k).$$

Proof sketch:

Combine the nonincreasing property of V_k

$$V_{k+1} \le V_k - \left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\right]$$

and P–Ł condition $\|\nabla f(\bar{x}_k)\|^2 \ge \nu(f(\bar{x}_k) - f^*)$:

$$\begin{aligned} V_{k+1} &\le V_k - \left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3\nu n(f(\bar{x}_k) - f^*)\right] \\ &\le (1-c_4)V_k \le \cdots \le (1-c_4)^{k+1}V_0. \end{aligned}$$

Linear convergence to **global optima**

Theorem

If each $f_i(x)$ is smooth and $f(x)$ satisfies P–Ł, there exists $\rho \in (0,1)$ such that

$$\frac{1}{n}\sum\nolimits_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* = \mathcal{O}(\rho^k).$$

Proof sketch:

Combine the nonincreasing property of V_k

$$V_{k+1} \le V_k - \left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\right]$$

and P–Ł condition $\|\nabla f(\bar{x}_k)\|^2 \ge \nu(f(\bar{x}_k) - f^*)$:

$$\begin{aligned} V_{k+1} &\le V_k - \left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3\nu n(f(\bar{x}_k) - f^*)\right] \\ &\le (1-c_4)V_k \le \cdots \le (1-c_4)^{k+1}V_0. \end{aligned}$$

Using $\frac{1}{n}\sum_{i=1}^{n} \|x_{i,k} - \bar{x}_k\|^2 + f(\bar{x}_k) - f^* \le c_5 V_k$ gives the result.

Extension: sublinear convergence to stationary points

Without P–Ł condition, the nonincreasing property of V_k still holds, i.e.,

$$V_{k+1} \leq V_k - \left[c_1 \|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2 \left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n \|\nabla f(\bar{x}_k)\|^2\right].$$

Extension: sublinear convergence to stationary points

Without P–Ł condition, the nonincreasing property of V_k still holds, i.e.,

$$V_{k+1} \leq V_k - \left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\right].$$

Sum over $k = 0, \ldots, T-1$ and rearrange terms:

$$\sum_{k=0}^{T-1}\left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\right] \leq V_0 - V_T \leq V_0.$$

Extension: sublinear convergence to stationary points

Without P–Ł condition, the nonincreasing property of V_k still holds, i.e.,

$$V_{k+1} \leq V_k - \left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\right].$$

Sum over $k = 0, \ldots, T-1$ and rearrange terms:

$$\sum_{k=0}^{T-1}\left[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\left\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\right\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\right] \leq V_0 - V_T \leq V_0.$$

Division by nT gives $\frac{1}{T}\sum_{k=0}^{T-1}\left[\frac{1}{n}\sum_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2\right] = \mathcal{O}(\frac{1}{T})$.

Extension: sublinear convergence to stationary points

Without P–Ł condition, the nonincreasing property of V_k still holds, i.e.,

$$V_{k+1} \leq V_k - \Big[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\Big\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\Big].$$

Sum over $k = 0, \ldots, T-1$ and rearrange terms:

$$\sum\nolimits_{k=0}^{T-1}\Big[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\Big\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\Big] \leq V_0 - V_T \leq V_0.$$

Division by nT gives $\frac{1}{T}\sum\nolimits_{k=0}^{T-1}\Big[\frac{1}{n}\sum\nolimits_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2\Big] = \mathcal{O}(\frac{1}{T})$.

Corollary (without P–Ł condition)

If each $f_i(x)$ is smooth, the primal–dual gradient descent algorithm converges to a stationary point sublinearly:

$$\frac{1}{T}\sum\nolimits_{k=0}^{T-1}\Big[\underbrace{\frac{1}{n}\sum\nolimits_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2}_{\text{Consensus}} + \underbrace{\|\nabla f(\bar{x}_k)\|^2}_{\text{Optimization}}\Big] = \mathcal{O}(\frac{1}{T}).$$

Extension: sublinear convergence to stationary points

Without P–Ł condition, the nonincreasing property of V_k still holds, i.e.,

$$V_{k+1} \leq V_k - \Big[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\Big\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\Big].$$

Sum over $k = 0, \ldots, T-1$ and rearrange terms:

$$\sum\nolimits_{k=0}^{T-1}\Big[c_1\|\boldsymbol{x}_k\|_{\boldsymbol{K}}^2 + c_2\Big\|\boldsymbol{v}_k + \frac{1}{\beta}\boldsymbol{g}_k^0\Big\|_{\boldsymbol{K}}^2 + c_3 n\|\nabla f(\bar{x}_k)\|^2\Big] \leq V_0 - V_T \leq V_0.$$

Division by nT gives $\frac{1}{T}\sum\nolimits_{k=0}^{T-1}\Big[\frac{1}{n}\sum\nolimits_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2 + \|\nabla f(\bar{x}_k)\|^2\Big] = \mathcal{O}(\frac{1}{T})$.

Corollary (without P–Ł condition)

If each $f_i(x)$ is smooth, the primal–dual gradient descent algorithm converges to a stationary point sublinearly:

$$\frac{1}{T}\sum\nolimits_{k=0}^{T-1}\Big[\underbrace{\frac{1}{n}\sum\nolimits_{i=1}^{n}\|x_{i,k} - \bar{x}_k\|^2}_{\text{Consensus}} + \underbrace{\|\nabla f(\bar{x}_k)\|^2}_{\text{Optimization}}\Big] = \mathcal{O}(\frac{1}{T}).$$

Remark: guaranteed convergence rate also for nonconvex cost functions.

Nonconvex distributed regularized logistic regression problem

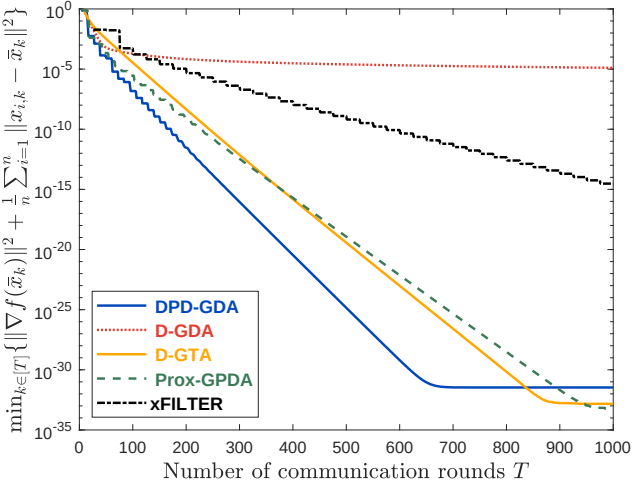
Each component function:

$$f_i(x) = \frac{n}{m}\sum_{l=1}^{m_i}(y_{il}\log(1+\exp(-x^\top z_{il})) + (1-y_{il})\log(1+\exp(x^\top z_{il}))) + \sum_{s=1}^{p}\frac{\lambda\mu[x]_s^2}{1+\mu[x]_s^2}.$$

Compared algorithms:

- Distributed primal–dual gradient descent algorithm (**DPD-GDA**) [Our paper]
- Distributed gradient descent algorithm (**D-GDA**) [Zeng & Yin, TSP, 2018]
- Distributed gradient tracking algorithm (**D-GTA**) [Qu & Li, TCNS, 2018]
- Distributed proximal primal–dual algorithm (**Prox-PDA**) [Hong et al, ICML, 2017]
- Distributed xFILTER algorithm (**xFILTER**) [Sun & Hong, TSP 2019]

Nonconvex distributed regularized logistic regression problem



- Our **DPD-GDA** gives the best performance in general.

Conclusions

- **Problem**: $\min_{x\in\mathbb{R}^p} f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)$, where each f_i could be nonconvex
- **Assumptions**: each f_i is smooth, f satisfies P–Ł, and $\mathcal{G}$ is connected

$$\text{P–Ł condition}: \ \|\nabla f(x)\|^2 \geq \nu(f(x) - f^*), \ \forall x \in \mathbb{R}^p$$

- **Method**: primal–dual gradient descent algorithm (ν is not used)
- **Result**:
  - Linear convergence with relaxed assumptions
  - Sublinear convergence for nonconvex cost functions

Conclusions

- **Problem**: $\min_{x \in \mathbb{R}^p} f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$, where each f_i could be nonconvex
- **Assumptions**: each f_i is smooth, f satisfies P–Ł, and $\mathcal{G}$ is connected

$$\text{P–Ł condition}: \ \|\nabla f(x)\|^2 \geq \nu(f(x) - f^*), \ \forall x \in \mathbb{R}^p$$

- **Method**: primal–dual gradient descent algorithm (ν is not used)
- **Result**:
  - Linear convergence with relaxed assumptions
  - Sublinear convergence for nonconvex cost functions
- **Extensions**:
  - Stochastic gradient descent algorithm
  - Zeroth-order algorithm

Thank you for your attention!