# A Parameter-free Decentralized Algorithm for Composite Convex Optimization

Xiaokai Chen

Purdue University

Email: Chen4373@purdue.edu

Co-authers: Ilya Kuruzov, Gesualdo Scutari & Alexander Gasnikov

The 64th IEEE Conference on Decision and Control (CDC2025)

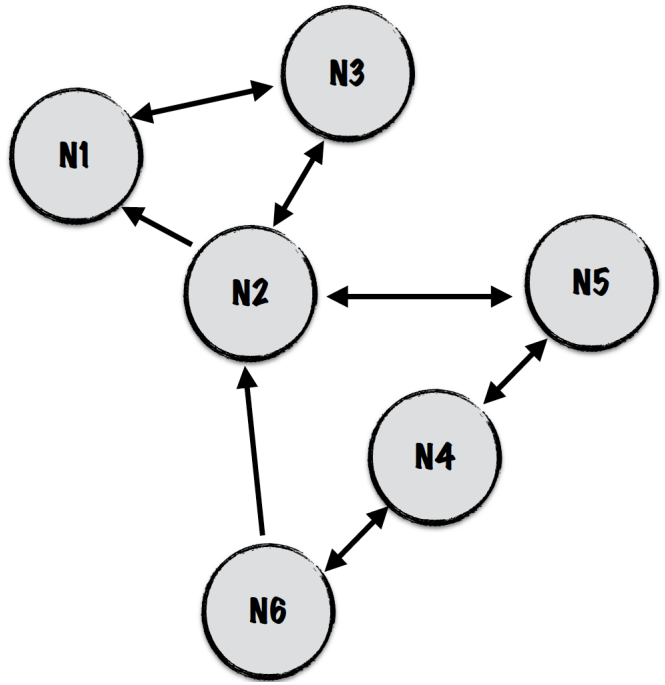
# Decentralized Optimization: Vanilla Formulation

$$\min_{x\in\mathbb{R}^d} \quad u(x) \triangleq \sum_{i=1}^{m} f_i(x) + r_i(x) \qquad \text{(P)}$$

- Each agent $i$ has access only to $f_i$, $r_i$
  - each $f_i$ is locally smooth and convex,
  - each $r_i$ is extended-value convex, and possibly nonsmooth.
- The graph network is connected
  - $\mathcal{N}_i \triangleq$ neighbors of agent $i$.

Meshed Networks (M-Nets)

Decentralized

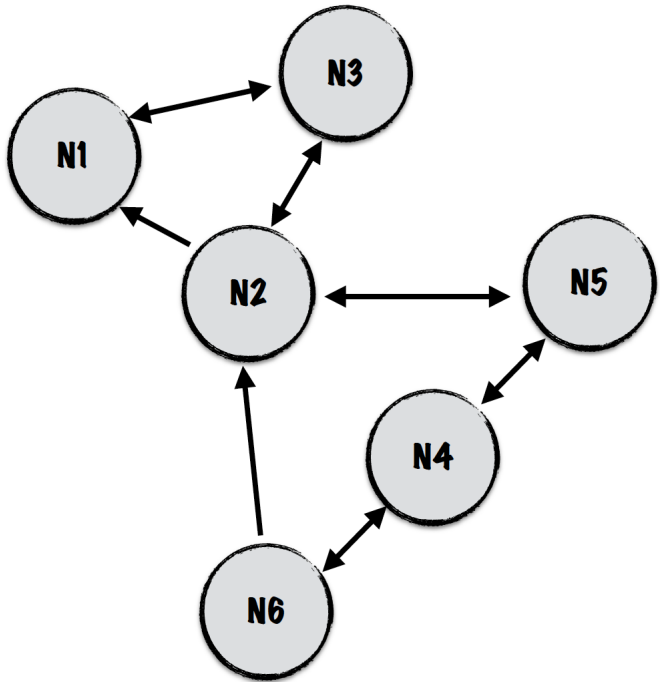


# Decentralized Optimization: Vanilla Formulation

$$\min_{x\in\mathbb{R}^d} \quad u(x) \triangleq \sum_{i=1}^{m} f_i(x) + r_i(x) \qquad \text{(P)}$$

- Each agent $i$ has access only to $f_i$, $r_i$
  - each $f_i$ is locally smooth and convex,
  - each $r_i$ is extended-value convex, and possibly nonsmooth.
- The graph network is connected
  - $\mathcal{N}_i \triangleq$ neighbors of agent $i$.

Meshed Networks (M-Nets)
Decentralized




Decentralized algorithms: each agent interleaves local computations with neighboring communications.

# On the Choice of the Stepsize

Convergence relays sensibly on the tuning of the stepsize

- **Theory:** Upper bounds
  - require knowledge of global optimization & network parameters, not available locally
  - are quite conservative
  - require *global* smoothness
- **Practice:**
  - manual tuning is not practical and experiment dependent
  - algorithm performance are quite sensitive to variations of the stepsize

# On the Choice of the Stepsize

Convergence relays sensibly on the tuning of the stepsize

- **Theory:** Upper bounds
  - require knowledge of global optimization & network parameters, not available locally
  - are quite conservative
  - require *global* smoothness
- **Practice:**
  - manual tuning is not practical and experiment dependent
  - algorithm performance are quite sensitive to variations of the stepsize

Open question: Can one design decentralized algorithms that *adaptively* select agents' stepsizes (with no global knowledge on network and optimization parameters)?

# Adaptive Stepsize in Decentralized Optimization

1. For smooth problems:
   - Adaptive but parameter-dependent
     - ★ Adaptive but require knowledge of network parameters: [Aldana-López et al., 2024], [Malitsky and Pock, 2018]
   - Adaptive but slow convergence
     - ★ Diminishing stepsize: [Sundhar Ram et al., 2010], [Chen, 2012], [Nedić and Olshevsky, 2014]
     - ★ Gradient normalization: [Nazari et al., 2022], [Chen et al., 2023], [Li et al., 2024]
   - Adaptive parameter-free with provable faster convergence
     - ★ line search-based method: [Kuruzov et al., 2024], [Kuruzov et al., 2025]
2. For nonsmooth problems with composite structure:
   - Implementable adaptation from centralized algorithms
     - ★ network-dependent parameters required: [Atenas et al., 2024]:
     - ★ heavy communication budget in line search procedure:[Latafat et al., 2024]

# Decentralized Setting: Why is Not so Trivial?

Decentralizing the backtracking procedure

## Warmup: Backtracking (centralized)

- Algorithm update: $x^{k+1} = \texttt{prox}_{\gamma^k, r}(x^k + \gamma^k d^k)$
- Backtracking: largest $\gamma^k > 0$ such that $f(x^k + \gamma^k d^k) \leq f(x^k) + c\gamma^k \nabla f(x^k)^\top d^k$, $c > 0$
- If $\nabla f(x^k)^\top d^k < 0 \Rightarrow u = f + r$ is a valid merit function and $d^k$ a proper search direction.

# Decentralized Setting: Why is Not so Trivial?

Decentralizing the backtracking procedure

## Warmup: Backtracking (centralized)

- Algorithm update: $x^{k+1} = \texttt{prox}_{\gamma^k, r}(x^k + \gamma^k d^k)$
- Backtracking: largest $\gamma^k > 0$ such that $f(x^k + \gamma^k d^k) \leq f(x^k) + c\gamma^k \nabla f(x^k)^\top d^k$, $c > 0$
- If $\nabla f(x^k)^\top d^k < 0 \Rightarrow u = f + r$ is a valid merit function and $d^k$ a proper search direction.

## Decentralized setting:

- How to define a "suitable" direction $d_i^k$ for each agent $i$?
- Which *local* surrogate of $f = \sum_{i=1}^m f_i$ to use in individual line-search procedures?
- Does the above yields convergence?

## Proposed approach:

Develop a **decentralized**, adaptive variant of the Davis-Yin three operator splitting coupled with a suitable **BCV metric** enabling local backtracking line-search with **provable convergence**.

# Algorithm Design: Reformulation with BCV Metric

- Reformulation with BCV Metric
  - Let $x_i \in \mathbb{R}^d$ be the local copy of $x$ at agent $i$ and introduce a slack variable $\widetilde{x}_i \in \mathbb{R}^d$.
  - Denote $\mathbf{X} := [x_1, \cdots, x_m]^\top$, $\widetilde{\mathbf{X}} := [\widetilde{x}_1, \cdots, \widetilde{x}_m]^\top \in \mathbb{R}^{m\times d}$, and $\hat{\mathbf{X}} := [\mathbf{X}^\top, \widetilde{\mathbf{X}}^\top]^\top \in \mathbb{R}^{2m\times d}$.
  - Let $F(\mathbf{X}) := \sum_{i=1}^m f_i(x_i)$, $R(\mathbf{X}) := \sum_{i=1}^m r_i(x_i)$, and $\delta_{\{\mathbf{0}\}}(\bullet)$ be the indicator function of $\{\mathbf{0}\}$.
  - The original problem is equivalent to the following optimization problem

$$\min_{\mathbf{X},\widetilde{\mathbf{X}}\in\mathbb{R}^{m\times d}} \quad \underbrace{F(\mathbf{X})}_{:=\widetilde{F}(\hat{\mathbf{X}})} + \underbrace{R(\mathbf{X}) + \delta_{\{\mathbf{0}\}}(\widetilde{\mathbf{X}})}_{:=\widetilde{R}(\hat{\mathbf{X}})} + \underbrace{\delta_{\{\mathbf{0}\}}(Ł\mathbf{X} + M\widetilde{\mathbf{X}})}_{:=\widetilde{G}(\hat{\mathbf{X}})}, \tag{P'}$$

  where $Ł \in \mathbb{S}^m$ satisfied $\mathtt{null}(Ł) = \mathtt{span}(\mathbf{1}_m)$ and $M \in \mathbb{R}^m_{++}$ is the BCV metric that we have degree of freedom to choose.
  - To unlock a decentralized implementation, we can set

$$M = \sqrt{I - Ł^2}, \quad Ł = \sqrt{I - W},$$

  where $W$ is a positive definite gossip matrix.

# DATOS: Decentralized Adaptive Three Operator Splitting

(S1) Communication Step:

$$\mathbf{X}^{k+1/2} = W\mathbf{X}^k, \quad \mathbf{D}^{k+1/2} = W(\nabla F(\mathbf{X}^k) + \mathbf{S}^k + \mathbf{D}^k)$$

(S2) Decentralized line-search: Each agent update $\overline{\gamma}_i^k$ by line-search such that

$$f_i(a_i^{k+1}) \leq f(x_i^k) + \langle \nabla f_i(x_i^k), a_i^{k+1} - x_i^k \rangle + \frac{\delta}{2\overline{\gamma}_i^k} \|a_i^{k+1} - x_i^k\|^2, \quad a_i^{k+1} := x_i^{k+1/2} - \overline{\gamma}_i^k d_i^{k+1/2} \quad (1)$$

(S3) Global min-consensus:

$$\gamma^k = \min_{i\in[m]} \overline{\gamma}_i^k$$

(S4) Update of the primal and dual variables

$$\begin{aligned}
\mathbf{X}^{k+1} &= \mathtt{prox}_{\gamma^k, R}\left(\mathbf{X}^{k+1/2} - \gamma^k \mathbf{D}^{k+1/2} + \gamma^k \mathbf{S}^k\right) \\
\mathbf{S}^{k+1} &= \mathbf{S}^k + \frac{1}{\gamma^k}\left(\mathbf{X}^{k+1/2} - \mathbf{X}^{k+1} - \gamma^k \mathbf{D}^{k+1/2}\right) \\
\mathbf{D}^{k+1} &= \mathbf{D}^{k+1/2} + \frac{1}{\gamma^k}\left(\mathbf{X}^k - \mathbf{X}^{k+1/2} - \gamma^k \nabla F(\mathbf{X}^k) - \gamma^k \mathbf{S}^k\right)
\end{aligned}$$

## Convergence Analysis

- Measure of convergence: a partial primal-dual gap function $\mathcal{G}^k \geq 0$
  - $\mathcal{G}^k$ is a valid measure of the distance of the primal and dual variables from the saddle point of the primal-dual formulation of the reformed problem (P').

### Theorem (informal)

*Let $\{(\mathbf{X}^k, \mathbf{S}^k, \mathbf{D}^k)\}$ be the sequences generated by DATOS with the sequence of stepsizes $\{\gamma^k\}$ selected by line search procedure (1) and global min-consensus under the aforementioned assumptions, then*

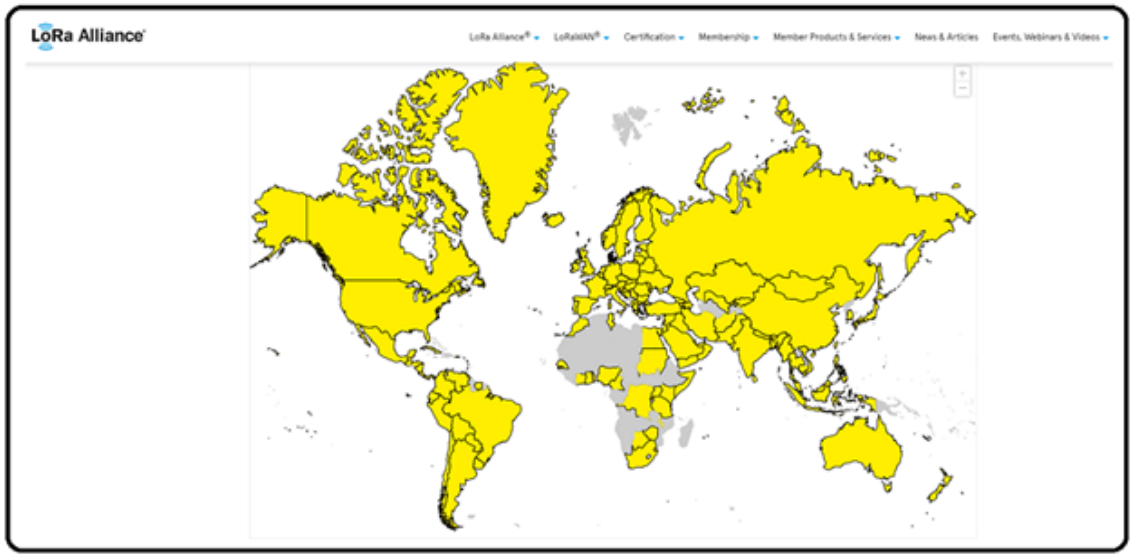
1. *The sequence $\{(\mathbf{X}^k, \mathbf{S}^k, \mathbf{D}^k)\}$ is bounded; and*
2. $$\mathcal{G}^k \leq \frac{1}{k} \cdot \frac{8R_0^2}{\min\left(\gamma^{-1}, \delta/(2L^k)\right)} = \mathcal{O}(1/k),$$

*where $L^k$ is the Lipschitz constant of $\nabla F$ over the convex hull of $\cup_{t=0}^{k-1}[\mathbf{X}^t, \mathbf{X}^{t+1/2} - \gamma^t \mathbf{D}^{t+1/2}]$ and $R_0$ is a constant associated with the initial points.*

# About the Global Min-Consensus

$$\gamma^k = \min_{i \in [m]} \overline{\gamma}_i^k$$

LoRA (Long Range, Low Power) Technology




- WiFi: transmission of vectors (high-rates, low range)
- LoRA: transmission of quantized values of $\overline{\gamma}_i^k$ reaching all agents in the network (flooding)

## Local_DATOS: From Global- to Local-min Consensus

Global min-consensus can be replaced with local min-consensus

$$\gamma^k = \min_{i\in[m]} \overline{\gamma}_i^k \longrightarrow \gamma_i^k = \min_{j\in\mathcal{N}_i} \overline{\gamma}_j^k, \quad \Gamma^k := \texttt{diag}([\gamma_1^k, \gamma_2^k, \quad, \gamma_m^k]^\top)$$

### Corollary (informal)

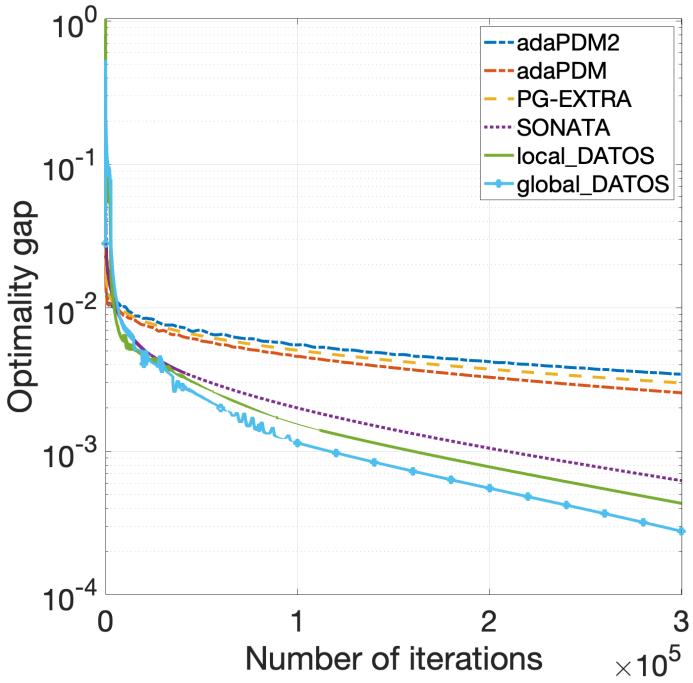
*Let $\{(\mathbf{X}^k, \mathbf{S}^k, \mathbf{D}^k)\}$ be the sequences generated by DATOS with the sequence of stepsizes $\{\Gamma^k\}$ selected by line search procedure (1) and local min-consensus under the aforementioned assumptions, then there exists a finite number $K > 0$ such that*

$$\mathcal{G}^k \leq \frac{1}{k} \cdot \frac{8R_K^2}{\min\left(\gamma^{-1}, \delta/(2L^k)\right)} = \mathcal{O}(1/k), \quad \forall k \geq K.$$

# Numerical Results

## Logistic Regression with $\ell_1-$Regularization

Erdős-Rényi ($p = 0.1$)




Erdős-Rényi ($p = 0.5$)

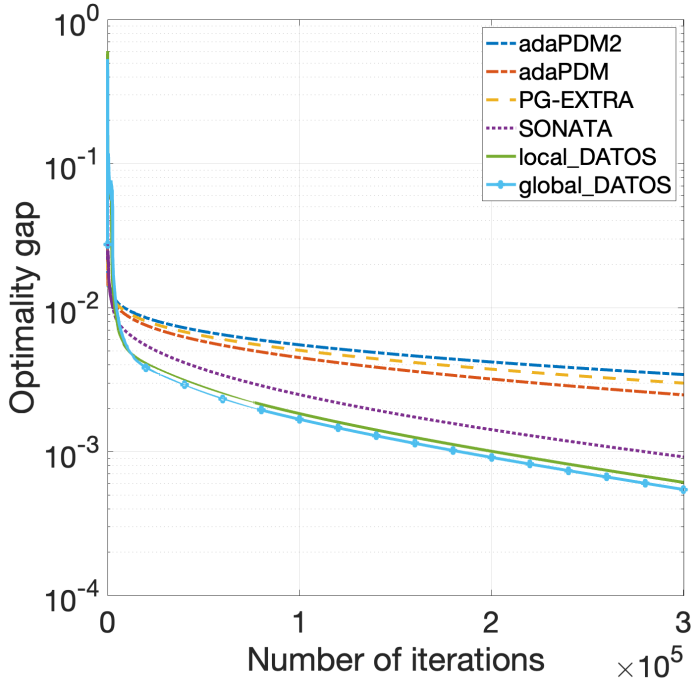



Figure 1: Logistic regression with $\ell_1$-regularization: $\frac{1}{m}\sum_{i=1}^m u(x_i) - u(x^*)$ v.s. # iterations.

# Concluding Remarks

Contributions:

- Decentralized adaptive method for nonsmooth problems with composite structure
- Adaptive stepsize via local backtracking and (local) min-consensus
- Convergence guarantees compare favorably with nonadaptive methods

Extensions & Generalization (not presented here):

- More comprehensive analysis for DATOS: iterate convergence, linear convergence.
- Line-search free methods

On-going:

- Towards a unified parameter-free decentralized algorithmic framework
- Stochastic oracles

# Can DATOS Has Linear Convergence: Numerical Finding

Sparse Linear Regression with Elastic Net Regularization

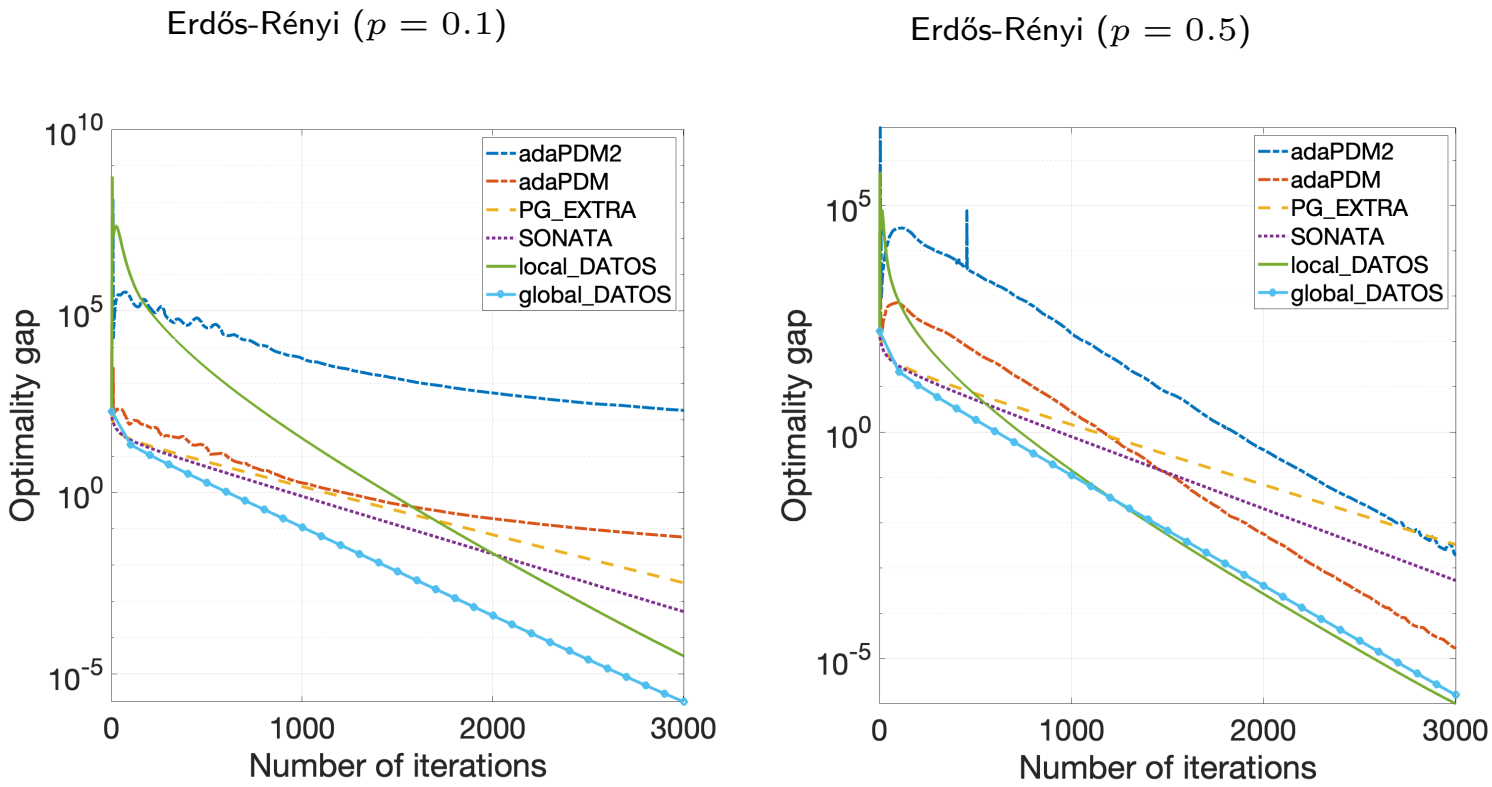



Figure 2: Sparse linear regression with elastic net regularization $\|\mathbf{X}^k - \mathbf{X}^*\|^2$ v.s. # iterations.

Impossibility: For composite problems with *general non-smooth* $r_i$, there exist counter-examples indicating TOS cannot achieve linear convergence [Davis and Yin, 2017].

Why is there a gap?

## Definition (Partly Smooth Function)

Let $r : \mathbb{R}^d \to \mathbb{R}$ be a convex, proper, lower semi-continuous function and $x \in \mathbb{R}^d$ such that $\partial r(x) \neq \emptyset$. $r$ is then said to be partly smooth at $x$ relative to a set $\mathcal{M}$ containing $x$, if

(i) **Smoothness:** $\mathcal{M}$ is a $\mathcal{C}^2$-smooth manifold around $x$, the restriction $r|_{\mathcal{M}}$ is $\mathcal{C}^2$ around $x$;

(ii) **Sharpness:** The tangent space $\mathcal{T}_{\mathcal{M}}(x)$ coincides with the set $T_x^r := \mathtt{par}(\partial r(x))^{\perp}$;

(iii) **Continuity:** The set-valued operator $\partial r$ is continuous at $x$ relative to $\mathcal{M}$.

We denote the class of partly smooth function at $x$ relative to $\mathcal{M}$ as $PSF_x(\mathcal{M})$.

| Function | Expression | Partial smooth manifold |
|---|---|---|
| $\ell_1-$norm | $\|x\|_1$ | $\mathcal{M} = \{z \in \mathbb{R}^d : \mathtt{supp}(z) \subseteq \mathtt{supp}(x)\}$ |
| $\ell_{1,2}-$norm | $\sum_{j=1}^n \|x_{b_j}\|_2$ | $\mathcal{M} = \{z \in \mathbb{R}^d : \mathtt{supp}(z) \subseteq \mathtt{supp}(x)\}$ |
| $\ell_\infty-$norm | $\|x\|_\infty$ | $\mathcal{M} = \{z \in \mathbb{R}^d : z_{\mathcal{I}_x} \in \mathtt{sign}(x_{\mathcal{I}_x})\mathbb{R}^{\lvert\mathcal{I}_x\rvert}\}$ |
| Nuclear norm | $\sum_{i=1}^{\mathtt{rank}(x)} \sigma_i(x)$ | $\mathcal{M} = \{z \in \mathbb{R}^{m\times d} : \mathtt{rank}(z) = \mathtt{rank}(x)\}$ |
| $\ell_0$ pseudo-norm | $\|x\|_0$ | $\mathcal{M} = \{z \in \mathbb{R}^d : \mathtt{supp}(z) \subseteq \mathtt{supp}(x)\}$ |
| Rank function | $\mathtt{rank}(x)$ | $\mathcal{M} = \{z \in \mathbb{R}^{m\times d} : \mathtt{rank}(z) = \mathtt{rank}(x)\}$ |
| ... | ... | ... |

Table 1: Examples of partly smooth functions, where $\{b_j\}_{j=1}^n$ is a set of disjoint subset of indices $\{1, 2 \cdots, d\}$ and $\mathcal{I}_x := \{i : |x_i| = \|x\|_\infty\}$

# Can DATOS Has Linear Convergence: Theoretical Support

## Theorem (informal)

*Under the aforementioned assumptions and further assume* $f = \sum_{i=1}^{m} f_i(x)$ *is* $\mu-$*strongly convex with* $\mu > 0$*, and each* $r_i$ *is partly smooth with respect to a manifold* $\mathcal{M}_{x^*}$*. Then with some nondegeneracy condition, there exists* $\zeta' \in (0, 1)$ *such that for any* $\zeta \in (\zeta', 1)$*, there exists* $K_\zeta > 0$ *such that for any* $k \geq K_\zeta$*,*

$$\|\hat{\mathbf{X}} - \hat{\mathbf{X}}^*\|^2 \leq \mathcal{O}(\zeta^{k-K_\zeta}). \tag{2}$$

## Theorem (informal)

*Inherit the setting of above Theorem and further assume each* $r_i$ *is partly smooth with respect to an affine manifold* $\mathcal{M}_{x^*}$*. Then with some nondegeneracy condition, there exists a finite number* $K$ *such that DATOS algorithm converges linearly for any* $k \geq K$*.*

# Appendix: Apply an Adaptive variant of TOS on the reformed problem

- Adaptive three operator splitting
  - Apply an adaptive variant of Davis-Yin three operator splitting [Davis and Yin, 2017] on the reformulation (P') reads:

$$\begin{aligned}\hat{\mathbf{A}}^{k+1} &= \mathtt{prox}_{\gamma^k,\widetilde{G}}\left(\hat{X}^k - \gamma^k\hat{\mathbf{S}}^k - \gamma^k\nabla\widetilde{F}(\hat{\mathbf{X}}^k)\right);\\ \hat{\mathbf{X}}^{k+1} &= \mathtt{prox}_{\gamma^k,\widetilde{R}}\left(\hat{\mathbf{A}}^{k+1} + \gamma^k\hat{\mathbf{S}}^k\right);\\ \hat{\mathbf{S}}^{k+1} &= \hat{\mathbf{S}}^k + \frac{1}{\gamma^k}\left(\hat{\mathbf{A}}^{k+1} - \hat{\mathbf{X}}^{k+1}\right),\end{aligned} \tag{3}$$

    where $\hat{\mathbf{A}}^k := [(\mathbf{A}^k)^\top, (\widetilde{\mathbf{A}}^k)^\top]^\top \in \mathbb{R}^{2m\times d}$ and $\hat{\mathbf{S}}^k := [(\mathbf{S}^k)^\top, (\widetilde{\mathbf{S}}^k)^\top]^\top \in \mathbb{R}^{2m\times d}$ are two intermediate variables derived by the TOS
  - For any given $\delta \in (0, 1]$, the stepsize $\gamma^k > 0$ can be selected by the largest value satisfying

$$\widetilde{F}(\hat{\mathbf{A}}^{k+1}) \le \widetilde{F}(\hat{\mathbf{X}}^k) + \langle\nabla\widetilde{F}(\hat{\mathbf{X}}^k), \hat{\mathbf{A}}^{k+1} - \hat{\mathbf{X}}^k\rangle + \frac{\delta}{2\gamma^k}\|\hat{\mathbf{A}}^{k+1} - \hat{\mathbf{X}}^k\|^2. \tag{4}$$

# Appendix: Measure of Convergence

Consider the Lagarangian function corresponding to the reformulation (P')

$$\mathcal{L}(\hat{\mathbf{X}}, \hat{\mathbf{S}}) := \widetilde{F}(\hat{\mathbf{X}}) + \widetilde{G}(\hat{\mathbf{X}}) + \langle \hat{\mathbf{S}}, \hat{\mathbf{X}} \rangle - \widetilde{R}^*(\hat{\mathbf{S}}).$$

- Partial primal-dual gap function: For any give saddle point $(\hat{\mathbf{X}}^*, \hat{\mathbf{S}}^*)$, let $\mathcal{B}_{\mathbf{X}} \times \mathcal{B}_{\mathbf{S}}$ be a bounded set containing $(\hat{\mathbf{X}}^*, \hat{\mathbf{S}}^*)$, and we introduce the the partial primal-dual gap function as

$$\mathcal{G}_{\mathcal{B}_{\mathbf{X}} \times \mathcal{B}_{\mathbf{S}}}(\hat{\mathbf{X}}, \hat{\mathbf{S}}) \triangleq \max_{\hat{\mathbf{S}}' \in \mathcal{B}_{\mathbf{S}}} \mathcal{L}(\hat{\mathbf{X}}, \hat{\mathbf{S}}') - \min_{\hat{\mathbf{X}}' \in \mathcal{B}_{\mathbf{X}}} \mathcal{L}(\hat{\mathbf{X}}', \hat{\mathbf{S}}).$$

  - $\mathcal{G}_{\mathcal{B}_{\mathbf{X}} \times \mathcal{B}_{\mathbf{S}}}(\hat{\mathbf{X}}, \hat{\mathbf{S}}) \geq 0$ for any $(\hat{\mathbf{X}}, \hat{\mathbf{S}}) \in \mathtt{dom}(\mathcal{L})$;
  - If $(\hat{\mathbf{X}}, \hat{\mathbf{S}})$ lies in the interior of $\mathcal{B}_{\mathbf{X}} \times \mathcal{B}_{\mathbf{S}}$, then $\mathcal{G}_{\mathcal{B}_{\mathbf{X}} \times \mathcal{B}_{\mathbf{S}}}(\hat{\mathbf{X}}, \hat{\mathbf{S}}) = 0$ indicates $(\hat{\mathbf{X}}, \hat{\mathbf{S}})$ is a saddle point of $\mathcal{L}$ [Chambolle and Pock, 2011].
- Ergodic convergence: We define the measurement of convergence on ergidic sequences

$$\mathcal{G}^k := \mathcal{G}_{\mathcal{B}_{\mathbf{X}} \times \mathcal{B}_{\mathbf{S}}}(\overline{\mathbf{A}}^k, \overline{\mathbf{S}}^k),$$

where

$$\overline{\mathbf{A}}^k := \frac{1}{\theta^k} \sum_{t=1}^{k} \gamma^{t-1} \hat{\mathbf{A}}^t, \quad \overline{\mathbf{S}}^k := \frac{1}{\theta^k} \sum_{t=1}^{k} \gamma^{t-1} \hat{\mathbf{S}}^t, \quad \theta^k := \sum_{t=1}^{k} \gamma^{t-1}.$$

# References I


Aldana-López, R., Macchelli, A., Notarstefano, G., Aragüés, R., and Sagüés, C. (2024).
Towards parameter-free distributed optimization: a port-hamiltonian approach.
*arXiv preprint arXiv:2404.13529*.

Atenas, F., Dao, M. N., and Tam, M. K. (2024).
A distributed proximal splitting method with linesearch for locally lipschitz gradients.
*arXiv preprint arXiv:2410.15583*.

Chambolle, A. and Pock, T. (2011).
A first-order primal-dual algorithm for convex problems with applications to imaging.
*Journal of mathematical imaging and vision*, 40(1):120–145.

Chen, A. I.-A. (2012).
*Fast distributed first-order methods*.
PhD thesis, Massachusetts Institute of Technology.

Chen, X., Karimi, B., Zhao, W., and Li, P. (2023).
On the convergence of decentralized adaptive gradient methods.
In *Asian Conference on Machine Learning*, pages 217–232. PMLR.

Davis, D. and Yin, W. (2017).
A three-operator splitting scheme and its optimization applications.
*Set-valued and variational analysis*, 25(4):829–858.

Kuruzov, I., Chen, X., Scutari, G., and Gasnikov, A. (2025).
Adaptive stepsize selection in decentralized convex optimization.
*arXiv preprint arXiv:2507.23725*.

# References II


Kuruzov, I., Scutari, G., and Gasnikov, A. (2024).
Achieving linear convergence with parameter-free algorithms in decentralized optimization.
*Advances in Neural Information Processing Systems*, 37:96011–96044.

Latafat, P., Themelis, A., Stella, L., and Patrinos, P. (2024).
Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient.
*Mathematical Programming*, pages 1–39.

Li, J., Chen, X., Ma, S., and Hong, M. (2024).
Problem-parameter-free decentralized nonconvex stochastic optimization.
*arXiv preprint arXiv:2402.08821*.

Malitsky, Y. and Pock, T. (2018).
A first-order primal-dual algorithm with linesearch.
*SIAM Journal on Optimization*, 28(1):411–432.

Nazari, P., Tarzanagh, D. A., and Michailidis, G. (2022).
Dadam: A consensus-based distributed adaptive gradient method for online optimization.
*IEEE Transactions on Signal Processing*, 70:6065–6079.

Nedić, A. and Olshevsky, A. (2014).
Distributed optimization over time-varying directed graphs.
*IEEE Transactions on Automatic Control*, 60(3):601–615.

Sundhar Ram, S., Nedić, A., and Veeravalli, V. V. (2010).
Distributed stochastic subgradient projection algorithms for convex optimization.
*Journal of optimization theory and applications*, 147(3):516–545.