Gossip Algorithms over Directed Graphs using Preconditioned Gradient Descent

Rustem Islamov[*]

August 25, 2022

1 Introduction

The size of modern Machine Learning models is growing like a lamplighter. Thus, to guarantee adequate performance in practice such models require a huge amount of data for training. However, existing software does not allow to store large amounts of training samples in one device. As a consequence, data is split across multiple machines that cooperate with each other through a communication network. This is the place where decentralized optimization comes to play.

In this work we work in decentralized setting where clients are connected through communication graph that represents possible communication links between nodes. Modern theory of decentralized optimization usually consider undirected communication graphs. However, we focus on directed case in order to build a more general and flexible theory of gossip algorithms. On top of that, there are real world problems that should be solved respecting directionality. For instance, in wireless sensor networks, sensors broadcasting at different power levels will finally lead to directed communication [Xin and Khan, 2018]. Besides, removing slow communication links in an undirected network leads to directional communication [Xi et al., 2016].

We consider standard Empirical Minimization Problem of the form

$$\min_{\theta\in\mathbb{R}^d}\left[f(\theta) := \sum_{i=1}^{n} f_i(\theta)\right], \tag{1}$$

where θ is a model parameters; f_i represents the local loss associated with data owned by device $i \in [n] := \{1, 2 \ldots, n\}$ only. This formulation aims to train a single machine learning model $x \in \mathbb{R}^d$ composed of d parameters by minimizing empirical loss $f(x)$ using all n clients' data with communication restrictions in mind.

2 Literature overview

In most cases (1) is solved using gossip algorithms [Koloskova et al., 2019, Scaman et al., 2017]. The communication step in algorithms of this type is represented as a multiplication by gossip matrix which encodes communication restrictions of the topology. Yet, most of existing works consider

[*] Institut Polytechnique de Paris (IP Paris), Palaiseau, France. This work was done while Rustem Islamov was an intern at Machine Learning and Optimization Lab, EPFL.

undirected networks, or equivalently symmetric gossip matrices. But working in a more general setting of directed communication graph (or asymmetric gossip matrix) brings additional challenges and requires much deeper understanding of the problem.

The first series of works in the direction of asymmetric gossip matrix analyze different variations of push-sum algorithm [Nedić and Olshevsky, 2015, Tsianos et al., 2012]. Push-Sum methods can be applied not only on fixed communication graphs, but also on time-varying ones. [Assran et al.] studies the stochastic Push-Sum protocol for averaging and derive theoretical guarantees. The key idea of Push-Sum protocol is to jointly learn the optimal parameters of a model and normalization factors. However, this makes the analysis of push-sum method extremely complicated, and as a consequence, it is necessary to make strong assumptions, such as gradient boundedness, to derive convergence rates. In addition, the convergence rates are usually suboptimal.

Another approach has been done by [Kovalev et al., 2021a,b]. The authors introduce ADOM and ADOM + methods that can be applied on time-varying graphs and achieve optimal convergence guarantees in undirected setting. Nevertheless, these two works still have limitations. The first one is the impossibility of stochastic updates due to the assumptions they make. Another problem is that the communication acceleration within ADOM and ADOM + relies on Chebyshev acceleration that provably works only on undirected graphs. There is still no work that claims that Chebyshev technique works in directed case. Hence, it is an open question if these two proposed methods can be accelerated in directed case.

Dual approach has been also investigated in undirected setting. [Scaman et al., 2017] provides the optimal rates and creates algorithms those convergence rates match the lower bounds. [Hendrikx et al., 2019] introduces a stochastic method with variance reduction that handles "sum of sums" case and matches optimal rates. Although dual approach has been successful for undirected graphs, it has never been leveraged for directed ones.

Due to the limitations that existing works have the goal of this work is to develop and analyze a simple gossip method that handles all aforementioned issues.

3 Dual reformulation

We assume that network is represented as a graph $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is a set of nodes i of a size n, and $\mathcal{E}$ is a set of directed edges (i, j). We assume that a graph $\mathcal{G}$ is strongly connected, there is at least one path between any pair of nodes i and j. Introducing additional parameters problem (1) can be equivalently written as follows

$$\min_{\Theta \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} f_i(\theta_i), \quad \theta_i = \theta_j \text{ if } (i, j) \in \mathcal{E}. \tag{2}$$

Here the i-th row of Θ is θ_i. Linear constraints $\theta_i = \theta_j$ might be combined in one which is written as $\mathbf{A}^\top \Theta = \mathbf{0}$ where $\mathrm{Ker}(\mathbf{A}^\top) = \mathrm{Span}(\mathbb{1})$. Since we have a minimization problem with linear constraints, then the Lagrangian reformulation is written in the following way

$$\min_{\Theta \in \mathbb{R}^{n \times d}} \max_{\Lambda \in \mathbb{R}^{n \times d}} \sum_{i=1}^{n} f_i(\theta_i) - \Lambda^\top \mathbf{A}^\top \Theta. \tag{3}$$

Finally, minimizing in Θ we derive the dual reformulation of the problem (1)

$$\max_{\Lambda} - \sum_{i=1}^{n} f_i^*([\mathbf{A}\Lambda]_i), \tag{4}$$

where f_i^* is a Fenchel conjugate of f_i. In the rest of the report we work with dual reformulation. We assume that we have an access to all f_i^* and their gradients. We note that the dual reformulation is done over nodes, but it is usually formulated on edges [Hendrikx et al., 2019].

3.1 Possible choice of gossip matrix

There are a lot of ways how we can define a gossip matrix. We present one of the possible ways for that. But we need to create a gossip matrix in a such way that in the undirected case it will be transformed into a gossip matrix of corresponding undirected graph. For example, we can set elements of gossip matrix as follows

$$[\mathbf{W}]_{ij} = w_{ij}, \quad w_{ij} = \begin{cases} -\frac{1}{\max\{d_i^{\text{in}}, d_j^{\text{out}}\}} & \text{if } (i,j) \in \mathcal{E} \text{ and } i \neq j, \\ -\sum\limits_{i \to j, i \neq j} w_{ij} & \text{if } i = j, \end{cases} \tag{5}$$

where d_i^{in} and d_i^{out} are in-degree and out-degree (node itself is always lies in in- and out-neighbourhood) of vertex i. Note that if a graph is undirected, then $d_i^{\text{in}} = d_i^{\text{out}} = d_i$, and we get one of the standard gossip matrices which is used in the theory of undirected graphs. We assume that any gossip matrix should have a non-zero entry w_{ij} iff an edge $(i,j) \in \mathcal{E}$.

4 Preconditioning is a possible solution

4.1 Pure averaging case

Let us now consider the simplest case of quadratics, i.e. each f_i is equal to $\frac{1}{2}\|x\|^2$. In this instance, the dual problem is written as

$$-\min_{\Lambda} \frac{1}{2}\text{Tr}(\Lambda^\top \mathbf{A}^\top \mathbf{A} \Lambda). \tag{6}$$

The above problem can be solved using Gradient Descent

$$\Lambda^{k+1} = \Lambda^k - \eta \mathbf{A}^\top \mathbf{A} \Lambda^k. \tag{7}$$

Unfortunately, Gradient Descent in this form can not be implemented for directed graphs. Indeed, our desire is to develop a gossip method with asymmetric gossip matrix. However, $\mathbf{A}^\top \mathbf{A}$ is symmetric, and thus, it could not be used as a gossip matrix. We modify the method by multiplying the update by properly chosen preconditioner $\mathbf{P}$

$$\Lambda^{k+1} = \Lambda^k - \eta \mathbf{P} \mathbf{A}^\top \mathbf{A} \Lambda^k, \quad \mathbf{P} = (\mathbf{A}^\top)^\dagger. \tag{8}$$

This method transforms into the next one

$$\Lambda^{k+1} = \Lambda^k - \eta \mathbf{A} \Lambda^k. \tag{9}$$

Now the last term involves only $\mathbf{A}$ that may play the role of gossip matrix. Indeed, the constraints $\mathbf{A}^\top\Theta = \mathbf{0}$ can be written in a such way that $[\mathbf{A}]_{ij} \neq 0$ if and only if $(i,j) \in \mathcal{E}$. Additionally assuming $\mathrm{Ker}(\mathbf{A}) = \mathrm{Ker}(\mathbb{1})$, we are now able to use $\mathbf{A}$ as a gossip matrix. We give **Di**rected **P**reconditioned **G**radient **D**escent name to the proposed method.

Yet, we make three assumptions on $\mathbf{A}$. First two ones are $\mathrm{Ker}(\mathbf{A}^\top) = \mathrm{Ker}(\mathbf{A}) = \mathrm{Span}(\mathbb{1})$. This assumption means the graph we work with is balanced: sum of weights of incoming edges is equal to weights of outgoing edges. This assumption does not restrict the class of graphs we are able to work with. Indeed, assume that we have a matrix $\mathbf{A}$ such that its columns are summed up to 0. Since $\mathbf{A}$ is not of full rank, then we can always rescale the columns of $\mathbf{A}$ in a such way that rows are also summed up to 0. However, such rescaling may change spectral properties of $\mathbf{A}$.

Remark 4.1. *We may consider more general setting when $f_i(x) = \frac{1}{2}\|x - c_i\|^2$. In this case we are still able to apply the same trick with preconditioning, but with change of variables*

$$(\mathbf{A}\Lambda^{k+1} + \mathbf{C}) = (\mathbf{A}\Lambda^k + \mathbf{C}) - \eta\mathbf{A}\mathbf{P}\mathbf{A}^\top(\mathbf{A}\Lambda^k + \mathbf{C}) = (\mathbf{A}\Lambda^k + \mathbf{C}) - \eta\mathbf{A}(\mathbf{A}\Lambda^k + \mathbf{C}). \tag{10}$$

In other words, we need to work with $\mathbf{A}\Lambda^k + \mathbf{C}$ instead of Λ^k.

4.2 Convergence analysis of DiPGD in quadratic case

To derive the linear convergence of Gradient Descent applied on the minimization problem of a function f we may assume strong convexity and co-coercivity of the objective. These two assumptions might be written in the following way:

$$\begin{aligned} \text{Strong monotonicity:} \quad & (x-y)^\top(\nabla f(x) - \nabla f(y)) \geq \mu\|x-y\|^2, \\ \text{Co-coercivity:} \quad & \|\nabla f(x) - \nabla f(y)\|^2 \leq L(x-y)^\top(\nabla f(x) - \nabla f(y)). \end{aligned} \tag{11}$$

We make similar assumptions to analyze DiPGD described above

$$\begin{aligned} \text{Strong monotonicity:} \quad & (x-y)^\top\mathbf{P}(\nabla f(x) - \nabla f(y)) \geq \mu\|x-y\|_{\mathbf{M}}^2, \\ \text{Co-coercivity:} \quad & \|\mathbf{P}(\nabla f(x) - \nabla f(y))\|_{\mathbf{M}}^2 \leq L(x-y)^\top\mathbf{P}(\nabla f(x) - \nabla f(y)). \end{aligned} \tag{12}$$

Here $\mathbf{M} := \mathbf{P}\mathbf{P}^\dagger$ is a matrix that restricts all necessary properties on proper space $\mathrm{Im}(\mathbf{A})$. If we consider the previous problem (6), then (12) transform to the next ones.

Assumption 4.2. *We assume that gossip matrix $\mathbf{A}$ satisfies*

$$\begin{aligned} \text{Strong convexity :} \quad & (X-Y)^\top\mathbf{A}(X-Y) \geq \mu\|X-Y\|_{\mathbf{M}}^2, \\ \text{Co-coercivity :} \quad & (X-Y)^\top\mathbf{A}^\top\mathbf{A}(X-Y) \leq L(X-Y)^\top\mathbf{A}(X-Y) \end{aligned} \tag{13}$$

for any X, Y.

In other words, we need to assume that $\mathbf{A} \succeq \mu\mathbf{M}$ and $L\mathbf{A} \succeq \mathbf{A}^\top\mathbf{A}$. These two assumptions could be verified; see remarks A.2 A.3 in the Appendix. Even if we do not work with proper gradients, but we are still able to apply monotone operators theory. These assumptions are reasonable since directed gossip matrices are positive. The simpliest example when they are verified is directed cyclic graph.

After clarifying the intuition behind the assumptions we make we are ready to establish the convergence guarantees for DiPGD.

Theorem 4.3. *Let the matrix* $\mathbf{A}$ *satisfies the conditions* $\mathrm{Ker}(\mathbf{A}^\top) = \mathrm{Ker}(\mathbf{A}) = \mathrm{Span}(\mathbb{1})$ *and* $\mathbf{A}_{ij} \neq 0$ *iff* $(i, j) \in \mathcal{E}$*. Let* (13) *assumptions hold with constants* μ *and* L*, and the stepsize satisfies* $\eta \leq {}^1/_L$*. Then the iterates of* DiPGD *converge linearly*

$$\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 \leq (1 - \eta\mu)^k \left\|\Lambda^0 - \Lambda^*\right\|_{\mathbf{M}}^2. \tag{14}$$

Proof. The proof of this theorem is almost identical to that of standard Gradient Descent. The detailed proof can be find in the Appendix A. □

4.3 Accelerated DiPGD

In this section we describe possible way for acceleration in asymmetric case. It is based on standard existing approaches such as three-point acceleration [Taylor and Bach, 2019], Nesterov acceleration [Nesterov, 2013], and Halpern-iteration [Lieder, 2021].

4.3.1 Halpern-iteration

First, we start with accelerated DiPGD via Halpern-iteration. Its iterates might recurrently be defined as follows

$$\Lambda^{k+1} = \alpha_k \Lambda^0 + (1 - \alpha_k) T(\Lambda^k), \quad \text{where } T(\Lambda) := \Lambda - \eta \mathbf{P}\mathbf{A}^\top \mathbf{A}\Lambda = \Lambda - \eta\mathbf{A}\Lambda \text{ and } \alpha_k = \frac{1}{k+2}. \tag{15}$$

It is a standard way to accelerate co-coercive operators. Theorem 2.1 of [Lieder, 2021] states that it takes $\mathcal{O}({}^1/_{k^2})$ iterations to find a fixed point of co-coercive operator. We adjust the proof of this theorem in our case to get $\mathcal{O}({}^1/_{k^2})$ convergence if the second assumption of 13 holds.

Theorem 4.4. *Let the matrix* $\mathbf{A}$ *satisfies the conditions* $\mathrm{Ker}(\mathbf{A}^\top) = \mathrm{Ker}(\mathbf{A}) = \mathrm{Span}(\mathbb{1})$ *and* $\mathbf{A}_{ij} \neq 0$ *iff* $(i, j) \in \mathcal{E}$*. Let the second assumption of* (13) *holds with constant* L*, and the stepsize satisfies* $\eta \leq {}^1/_L$*. Then the iterates of accelerated* DiPGD *through Halpern-iteration converge sublinearly*

$$\frac{1}{2}\left\|\Lambda^k - T(\Lambda^k)\right\|^2 \leq \frac{\left\|\Lambda^0 - \Lambda^*\right\|^2}{(k+1)^2}. \tag{16}$$

Proof. The proof of this theorem follows the proof of [Lieder, 2021] and is given in Appendix C. □

Remark 4.5. *Recent work of Park and Ryu [2022] provides optimal convergence guarantees if we additionally assume strong monotonicity. The method in our case has the form*

$$\Lambda^{k+1} = \frac{1}{\varphi_k}\Lambda^0 + \left(1 - \frac{1}{\varphi_k}\right) T(\Lambda^k), \tag{17}$$

where $\varphi_k := \sum_{i=0}^{k} \left(\frac{L/\mu}{L/\mu - 1}\right)^{2i}$*. This method achieves* $\mathcal{O}(e^{-4\mu N/L})$ *convergence rate.*

Rustem: check

4.3.2 Three-point acceleration

Next, we try to accelerate DiPGD through three-point acceleration [Taylor and Bach, 2019] where iterates satisfy

$$\begin{aligned} X^{k+1} &= \alpha Z^k + (1-\alpha) Y^k, \\ Y^{k+1} &= X^k - \eta \mathbf{A} X^k, \\ Z^{k+1} &= \beta Z^k + (1-\beta) X^k - \gamma \mathbf{A} X^k. \end{aligned} \tag{18}$$

When we apply three-point acceleration on standard minimization problem, the analysis requires the following assumptions:

1. smoothness of the objective between y^{k+1} and x^k;

2. strong convexity of the objective;

3. convexity of the objective.

First two assumptions are discussed in Section 4.2 and hold for different types of asymmetric graphs (e.g. ring graph). The first issue that we face during the analysis is that all existing prooves of three-point acceleration involve function values in it. In our case $\mathbf{A}X$ with asymmetric $\mathbf{A}$ does not correspond to any function, hence, it is not clear what quantity should replace function value in the analysis. Next,

4.3.3 Nesterov's acceleration

A special case of three-point acceleration is Nesterov's acceleration Nesterov [2013]. It has the form

$$\begin{aligned} X^{k+1} &= Y^k - \eta \mathbf{A} Y^k \\ Y^{k+1} &= X^{k+1} + \beta (X^{k+1} - X^k). \end{aligned} \tag{19}$$

It was analyzed in [Hong and Yavneh, 2017] in asymmetric case, but the authors do not provide with the rate of convergence. They give only the optimal value for parameter β. Nevertheless, it is not clear to what rate of convergence the optimal choice of parameters leads.

4.4 Stochastic updates

In practice, only a few nodes are active at the same time. Thus, we can not perform the full step of DiPGD. Our desire is to perform a step over a random set of nodes. We randomly select a cycle of nodes and perform a step with gossip matrix corresponding to selected cycle. We would like to point out the importance of selecting cycle. Indeed, if we pick, for example, only two nodes connected with directed edge, then we either do not keep the mean or introduce a variance to the optimization process.

Let us briefly describe stochastic version of DiPGD. We assume that $\mathcal{S} = \{c_1, \ldots, c_m\}$ is a set of all cycles for given graph $(\mathcal{V}, \mathcal{E})$. For any cycle $c \in \mathcal{S}$ we are able to construct a gossip matrix $\mathbf{A}_c$. We select a cycle c_k from $\mathcal{S}$ with corresponding gossip matrix $\mathbf{A}_{c_k}$ on the iteration k. Then we perform a step of SDiPGD

$$\Lambda^{k+1} = \Lambda^k - \eta m \mathbf{A}_{c_k} \Lambda^k. \tag{20}$$

We multiply a stepsize η by m to make $m\mathbf{A}_{c_k}$ to be unbiased estimator of full gossip matrix $\mathbf{A}$.

4.4.1 Convergence theory of SDiPGD

We make the following assumption similar to 13.

Assumption 4.6. *For any cycle $c \in \mathcal{S}$ let us assume that corresponding gossip matrix $\mathbf{A}_c \in \mathbb{R}^{n\times n}$ satisfies*

$$\begin{aligned}
\text{Consensus property}: &\quad \mathrm{Ker}(\mathbf{A}_c) = \mathrm{Ker}(\mathbf{A}_c^\top) = \mathrm{Span}(\mathbb{1}_{\mathbb{c}}), \quad [\mathbb{1}_{\mathbb{c}}]_i = 1 \text{ iff } i \in c,\\
\text{Strong convexity}: &\quad (X-Y)^\top \mathbf{A}_c (X-Y) \geq \mu \|X-Y\|^2_{\mathbf{M}},\\
\text{Co-coercivity}: &\quad \|\mathbf{A}_c(X-Y)\|^2_{\mathbf{M}} \leq \frac{\overline{L}}{m}(X-Y)^\top \mathbf{M}\mathbf{A}_c(X-Y).
\end{aligned} \tag{21}$$

for any X, Y.

Now we are ready to establish a convergence theorem of SDiPGD.

Theorem 4.7. *Let* (21) *assumptions hold with constants μ and $\overline{L}$, and the stepsize satisfies $\eta \leq 1/\overline{L}$. Then the iterates of* SDiPGD *converge linearly*

$$\mathbb{E}\left[\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}}\right] \leq (1-\eta\mu)^k \left\|\Lambda^0 - \Lambda^*\right\|^2_{\mathbf{M}}. \tag{22}$$

Proof. The proof of this theorem is almost identical to that of standard Gradient Descent. The detailed proof can be find in the Appendix B. □

The main difficulty of stochastic updates in the directed case is a necessity of cyclic updates. In order to handle that, Push-Sum [Tsianos et al., 2012] protocol was created; see Section 2. The idea of this approach is to introduce normalization variables to overcome problems coming from directionality. If we do not normalize properly or do not use another trick, we either introduce variance to optimization process (i.e. do not converge to the exact optimum) or do not keep the mean.

5 General problem

So far we analyze only quadratic case, i.e. pure averaging. In this section we propose the way to handle general finite sum problem. We consider standard ERM problem of the form

$$\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \left[f_i(\theta) + \frac{\sigma}{2}\|\theta\|^2\right], \tag{23}$$

where $f_i(x) + \frac{\sigma}{2}\|x\|^2$ represents local loss/risk on i-th device. We assume that each function f_i is convex and L-smooth. This implies that the Fenchel conjugate f_i^* is L^{-1}-convex[1]. As a consequence, the proximity operator of f_i^* is well defined. However, we do not make assumptions on strong convexity of f_i. Thus, we are not able to say anything about smoothness of f_i^*.

The idea how to work with such kind of problem is inspired by Hendrikx et al. [2019]. We first rewrite (23) as

$$\min_{\theta \in \mathbb{R}^{n\times d}} \sum_{i=1}^n f_i(\theta_i) + \frac{\sigma}{2}\|\theta_i\|^2 \quad \theta_i = \theta_j \text{ if } j \in \mathcal{N}^{\mathrm{in}}(i), \tag{24}$$

[1] In fact, values of σ and L may differ for different nodes, but we assume they are equal for simplicity.

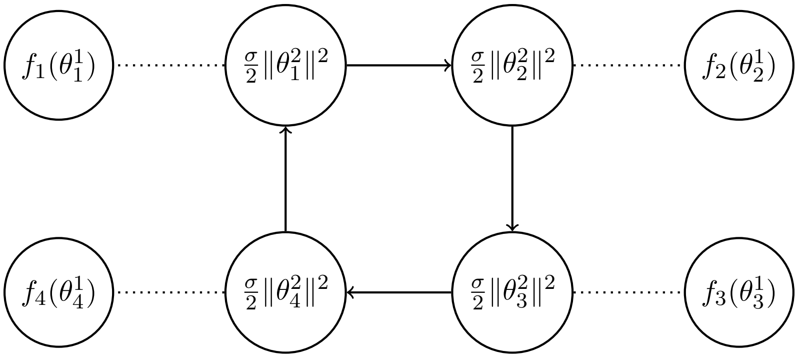

$f_1(\theta_1^1)$
$\frac{\sigma}{2}\|\theta_1^2\|^2$
$\frac{\sigma}{2}\|\theta_2^2\|^2$
$f_2(\theta_2^1)$
$f_4(\theta_4^1)$
$\frac{\sigma}{2}\|\theta_4^2\|^2$
$\frac{\sigma}{2}\|\theta_3^2\|^2$
$f_3(\theta_3^1)$

Figure 1: The example of augmented graph applied on a ring graph.

where $\mathcal{N}^{\text{in}}(i)$ is in-neighbourhood of node i. Next, we copy parameters of i-th node into two parts: communication central node with local loss function $\frac{\sigma}{2}\|\theta_i^2\|^2$ and computation virtual node with local loss function $f_i(\theta_i^1)$; see example in Figure 1. Thus, we transform (24) as follows

$$\min_{\theta\in\mathbb{R}^{2n\times d}} \sum_{i=1}^{n} f_i(\theta_i^1) + \frac{\sigma}{2}\|\theta_i^2\|^2 \quad \theta_i^2 = \theta_j^2 \text{ if } j \in \mathcal{N}^{\text{in}}(i) \text{ and } \theta_i^2 = \theta_i^2\ \forall\ i \in [n]. \tag{25}$$

Note that dotted edges in an augmented graph are virtual undirected, i.e. we do not need to communicate via these edges. The only edges we communicate through are solid directed edges. In the rest of the draft upper index 1 refers to computation nodes (first n rows in $\theta \in \mathbb{R}^{2n\times d}$), and upper index 2 refers to communication nodes (last n rows in θ). The constraints in (25) can be rewritten as $\mathbf{A}^\top\theta = 0$ where $\mathbf{A} \in \mathbb{R}^{2n\times 2n}$ and $\mathrm{Ker}(\mathbf{A}^\top) = \mathbb{1}$ (vector of ones). In more details,

$$\mathbf{A} = \begin{pmatrix} \mu\mathbf{I} & \mathbf{0} \\ -\mu\mathbf{I} & \tilde{\mathbf{A}} \end{pmatrix} = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ -\mathbf{D} & \tilde{\mathbf{A}} \end{pmatrix} := \begin{pmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{pmatrix},$$

where $\mu > 0$ is some constant, and $\mathbf{D} := \mu\mathbf{I}$. We assume that $[\tilde{\mathbf{A}}]_{p,q} \neq$ iff (p, q) is an edge in the original graph.

The constraction of matrix $\mathbf{A}$ is intuitively clear since i-th computation node is connected only to i-th communication node. Note that this implies

$$[\mathbf{A}\lambda]_i^1 = [\mathbf{A}_1\lambda]_i = [\mathbf{D}\lambda^1]_i = \mu\lambda_i^1.$$

The primal-dual reformulation of this problem is of the form

$$\min_{\Theta\in\mathbb{R}^{2n\times d}} \max_{\Lambda\in\mathbb{R}^{2n\times d}} \sum_{i=1}^{n} f_i(\theta_i^1) + \frac{\sigma}{2}\|\theta_i^2\|^2 - \left\langle \Lambda, \mathbf{A}^\top\Theta \right\rangle. \tag{26}$$

Hence, the dual formulation of (26) is defined as

$$\max_{\Lambda\in\mathbb{R}^{2n\times d}} -\sum_{i=1}^{n} f_i^*([\mathbf{A}_1\Lambda)]_i) + \frac{1}{2\sigma}\|[\mathbf{A}_2\Lambda]_i\|^2 = \max_{\Lambda\in\mathbb{R}^{2n\times d}} -\sum_{i=1}^{n} f_i^*(\mu\lambda_i^1) + \mathrm{tr}\left(\frac{1}{2}\Lambda^\top\mathbf{A}_2^\top\Sigma_2^{-1}\mathbf{A}_2\Lambda\right), \tag{27}$$

where $\Sigma_2 := \sigma\mathbf{I} \in \mathbb{R}^{n\times n}$. By the assumption we make, each Fenchel conjugate f_i^* is L^{-1}-convex. As it is stated earlier, we do not assume strong convexity of f_i. This implies that f_i^* could be potentially non-smooth. To overcome this issue we use the proximity operator of the sum of Fenchel conjugates. In order to make the second term in (26) to be strongly convex, we transfer all strong convexity part of the first term to the second one. Then dual problem is written as

$$\max_{\Lambda\in\mathbb{R}^{2n\times d}} -\sum_{i=1}^{n}\tilde{f}_i^*(\lambda_i^1) + \operatorname{tr}\left(\frac{1}{2}\Lambda^\top \mathbf{A}_2^\top \Sigma \mathbf{A}_2 \Lambda\right) = -\min_{\Lambda\in\mathbb{R}^{2n\times d}} \underbrace{\sum_{i=1}^{n}\tilde{f}_i^*(\lambda_i^1)}_{:=\,\psi(\Lambda)} + \underbrace{\operatorname{tr}\left(\frac{1}{2}\Lambda^\top \mathbf{A}^\top \Sigma^{-1}\mathbf{A}\Lambda\right)}_{:=\,g(\Lambda)}, \quad (28)$$

where

$$\Sigma := \begin{pmatrix} L\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix} = \begin{pmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \Sigma_2 \end{pmatrix}$$

and $\tilde{f}_i^* : \nu \mapsto f_i^*(\mu\nu) - \frac{\mu^2}{2L}\|\nu\|^2$.

5.1 DiP2GD is an extension of DiPGD

We solve (28) applying **P**recondition **P**roximal **G**radient **D**escent of the form

$$\Lambda^{k+1} = \operatorname{prox}_{\eta\psi}\left[\Lambda^k - \eta\mathbf{P}\nabla g(\Lambda^k)\right] = \operatorname{prox}_{\eta\psi}\left[\Lambda^k - \eta\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}\Lambda^k\right] \quad (29)$$

with properly chosen preconditioner $\mathbf{P}$ and positive stepsize η.

The proximity operator of $\eta\psi$ can be computed using Moreau identity and separability of ψ as a function of λ_i^1. One can derive that

$$\operatorname{prox}_{\eta\psi}(\Phi) = \begin{pmatrix} \operatorname{prox}_{\eta\tilde{\psi}}(\Phi^1) \\ \Phi^2 \end{pmatrix} = \begin{pmatrix} \operatorname{prox}_{\eta\tilde{f}_i^*}(\varphi_i^1) \\ \vdots \\ \operatorname{prox}_{\eta\tilde{f}_n^*}(\varphi_n^1) \\ \Phi^2 \end{pmatrix}$$

where $\tilde{\psi}(\varphi^1) := \sum_{i=1}^n \tilde{f}_i^*(\varphi_i^1)$. For more details, we refer to the Appendix D.1.

We choose $\mathbf{P}$ equal to

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & (\tilde{\mathbf{A}}^\top)^\dagger \end{pmatrix} \quad (30)$$

and define a matrix $\mathbf{M} = \mathbf{P}\mathbf{P}^\dagger$. Such choice of preconditioning does not break the fixed point and non-expansiveness of the proximity operator. For more details see Appendix D.2.

5.1.1 Practical implementation

In the described setting a matrix $\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}$ plays the role of gossip matrix. However, it has the form

$$\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A} = \begin{pmatrix} \mu^2\left(\sigma^{-1}+L^{-1}\right)\mathbf{I} & -\mu\sigma^{-1}\tilde{\mathbf{A}} \\ -\mu\sigma^{-1}\mathrm{Proj} & \sigma^{-1}\tilde{\mathbf{A}} \end{pmatrix}.$$

where $\mathrm{Proj} = (\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})^\dagger$. This method could not be implemented in practice due to this projection part. Thus, in practice we work with other variables. We perfrom a change of variables

$$Y^k = \hat{\mathbf{A}}\Lambda^k$$

where

$$\hat{\mathbf{A}} := \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{A}}. \end{pmatrix}.$$

Then the algorithm is written in the form

$$Y^{k+1} = \operatorname{prox}_{\eta\psi}\left[Y^k - \eta\hat{\mathbf{A}}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{B}Y^k\right] \tag{31}$$

where

$$\mathbf{B} := \begin{pmatrix} \mu\mathbf{I} & \mathbf{0} \\ -\mu\mathbf{I} & \mathbf{I} \end{pmatrix}.$$

is a basis matrix. In other words, instead of variables Λ^1 and Λ^2 we work with $Y^1 = \Lambda^1$ and $Y^2 = \tilde{\mathbf{A}}\Lambda^2$. We point out that such change of variabels does not affect properties of the proximity operator of ψ. Indeed, we change only communication variables while ψ influences only computation variabels.

Gossip matrix in the form (31) $\hat{\mathbf{A}}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{B}$ can be explicitly written as

$$\hat{\mathbf{A}}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{B} = \begin{pmatrix} \mu^2\left(\sigma^{-1} + L^{-1}\right)\mathbf{I} & -\mu\sigma^{-1}\mathbf{I} \\ -\mu\sigma^{-1}\tilde{\mathbf{A}} & \sigma^{-1}\tilde{\mathbf{A}} \end{pmatrix}.$$

Now gossip matrix does not involve projection part, and as a consequence, can be implemented in practice.

5.1.2 Convergence theory of DiP2GD

We should assume that g satisfies assumptions similar to assumptions 13. We again define $\mathbf{M} = := \mathbf{P}\mathbf{P}^\dagger$.

Assumption 5.1. *Let $\mathbf{P}$ be a possibly asymmetric matrix and $\mathbf{P}^\dagger$ be its pseudo-inverse. Let α and β be positive constants. We say that a function g defined above is α-strongly convex and β-co-coercive in the norm w.r.t. a matrix $\mathbf{M}$ if it satisfies for all $x, y \in \mathbb{R}^d$*

$$\text{strong monotonicity :} \quad (X-Y)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(X-Y) \geq \alpha\left\|X-Y\right\|^2_{\mathbf{M}},$$

$$\text{co-coercivity :} \quad \left\|\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(X-Y)\right\|^2_{\mathbf{M}} \leq \beta(X-Y)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(X-Y) \tag{32}$$

for any X, Y.

Remark 5.2. *These assumptions may hold only for small condition number in practice. The reason why it is so beacause in the general case a matrix $\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}$ might have negative eigenvalues (to be more precise, symmetrized version of this matrix), i.e could be non-positive definite. Analysis of the spectrum of $\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}$ is provided in Appendix G.*

Then the analysis is almost the same as for the quadratic case combining the fact that the proximity operator is non-expansive. This leads us to the following theorem.

Theorem 5.3. *Let a function g satisfies assumptions (32) with constants α and β. Let $\operatorname{Ker}(\tilde{\mathbf{A}}) = \operatorname{Ker}(\tilde{\mathbf{A}}^\top) = \operatorname{Span}(\mathbb{1})$. Then the iterates of DiP2GD with the stepsize $\eta \leq \frac{1}{\beta}$ satisfy*

$$\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} \leq (1-\eta\alpha)\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}}.$$

Proof. It follows the proof of theorem 4.4 combining properties of the proximity operator. See Appendix E. □

5.2 Split communication and computation steps of DiP2GD

Let us write one step of DiP2GD in the form (31)

$$\begin{aligned}\text{computation step :} \qquad & (Y^1)^{k+1} = \mathrm{prox}_{\eta\tilde{\psi}}\left[(Y^1)^k - \eta\left(\mu^2(L^{-1}+\sigma^{-1})(Y^1)^k - \mu\sigma^{-1}(Y^2)^k\right)\right] \\ \text{communication step :} \qquad & (Y^2)^{k+1} = (Y^2)^k - \eta\sigma^{-1}\tilde{\mathbf{A}}\left(-\mu(Y^1)^k + (Y^2)^k\right). \qquad (33)\end{aligned}$$

As we observe, one step may be split into two parts: computation and communication. In computation part calculations are done locally on nodes (i.e. no need to communicate). In communication step we gossip with matrix $\tilde{\mathbf{A}}$. If one part costs much more than another it is better to perform more cheap steps. In other words, we perform computation step with probability p and communication step with probability $(1-p)$, where the value of p is adjusted according to the cost of each step.

In order to make the step to be unbiased (i.e. expected step should be equal to deterministic step) stochastic steps have the form

$$\begin{aligned}\text{computation step :} \qquad & (Y^1)^{k+1} = \mathrm{prox}_{\eta\tilde{\psi}}\left[(Y^1)^k - \frac{1}{p}\eta\left(\mu^2(L^{-1}+\sigma^{-1})(Y^1)^k - \mu\sigma^{-1}(Y^2)^k\right)\right] \\ \text{communication step :} \qquad & (Y^2)^{k+1} = (Y^2)^k - \frac{1}{1-p}\eta\sigma^{-1}\tilde{\mathbf{A}}\left(-\mu(Y^1)^k + (Y^2)^k\right). \qquad (34)\end{aligned}$$

5.2.1 Convergence of SDiP2GD

In stochastic case we have two types of steps: communication and computation (33). It is equivalent to say that we gossip with matrix $\tilde{\mathbf{G}}$ which is equal to either $p^{-1}\mathbf{G}_1$ with probability p or to $(1-p)^{-1}\mathbf{G}_2$ with probability $(1-p)$ where

$$\mathbf{G}_1 := \begin{pmatrix}\mu^2\left(\sigma^{-1}+L^{-1}\right)\mathbf{I} & -\mu\sigma^{-1}\mathbf{I} \\ \mathbf{0} & \mathbf{0}\end{pmatrix}, \quad \mathbf{G}_2 := \begin{pmatrix}\mathbf{0} & \mathbf{0} \\ -\mu\sigma^{-1}\tilde{\mathbf{A}} & \sigma^{-1}\tilde{\mathbf{A}}\end{pmatrix}.$$

This means that in expectation $\mathbb{E}[\tilde{\mathbf{G}}] = \mathbf{G}$ where $\mathbf{G} := \hat{\mathbf{A}}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{B}$ is a deterministic gossip matrix.

Again, we should assume that g satisfies assumptions similar to assumptions 32, but co-coercivity assumption should hold in expectaion

Assumption 5.4. *Let $\mathbf{P}$ be a possibly asymmetric matrix and $\mathbf{P}^\dagger$ be its pseudo-inverse. Let α and β be positive constants. We say that a function g defined above is α-strongly convex and β-co-coercive in the norm w.r.t. a matrix $\mathbf{M}$ if it satisfies for all $x, y \in \mathbb{R}^d$*

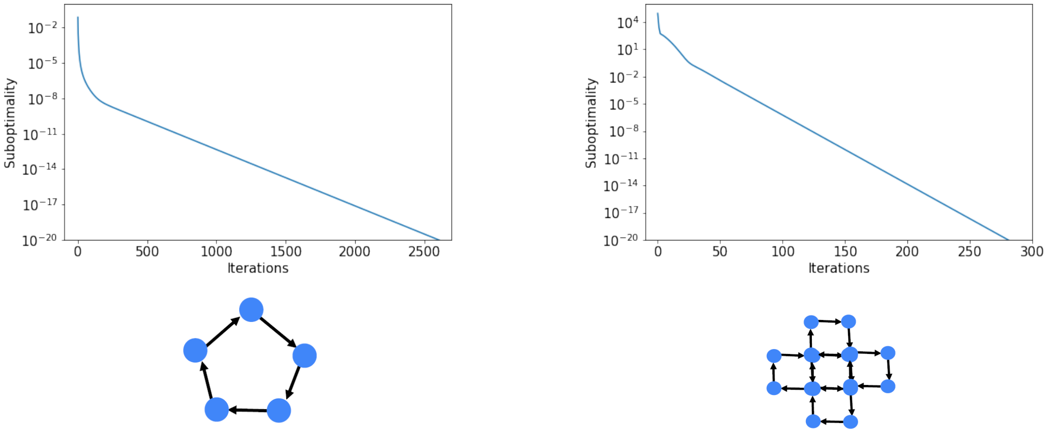
$$\begin{aligned}\text{strong convexity :} \qquad & (X-Y)^\top\mathbf{G}(X-Y) \geq \alpha\left\|X-Y\right\|^2_{\mathbf{M}}, \\ \text{expected co-coercivity :} \qquad & \mathbb{E}\left[\left\|\tilde{\mathbf{G}}Y - \mathbf{G}Y^*\right\|^2_{\mathbf{M}}\right] \leq \overline{\beta}(Y-Y^*)^\top\mathbf{G}(Y-Y^*), \qquad (35)\end{aligned}$$

for any X, Y where the expectation is taken w.r.t randomness coming from $\tilde{\mathbf{G}}$.

As we can see, strong convexity assumption remains unchanged, but co-coercivity assumption swithces from deterministic to expected case. We establish the convergence theorem of SDiP2GD.

Theorem 5.5. *Let a function g satisfies assumptions* (35) *with constants α and $\overline{\beta}$. Let* $\mathrm{Ker}(\tilde{\mathbf{A}}) = \mathrm{Ker}(\tilde{\mathbf{A}}^\top) = \mathrm{Span}(\mathbb{1})$. *Then the iterates of* SDiP2GD *with the stepsize $\eta \leq \frac{1}{\overline{\beta}}$ satisfy*

$$\mathbb{E}\left[\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}}\right] \leq (1-\eta\alpha)\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}}.$$



Ring: $n = 30, d = 20, \mu = 0.0072, L = 0.067$ Grid: $n = 12, d = 30, \mu = 0.0329, L = 0.500$

Figure 2: The performance of DiPGD on ring and grid graphs and illustrations of graphs used in the experiments.

Proof. The proof is given in Appendix F □

6 Experiments

6.1 Pure averaging

6.1.1 Convergence of DiPGD

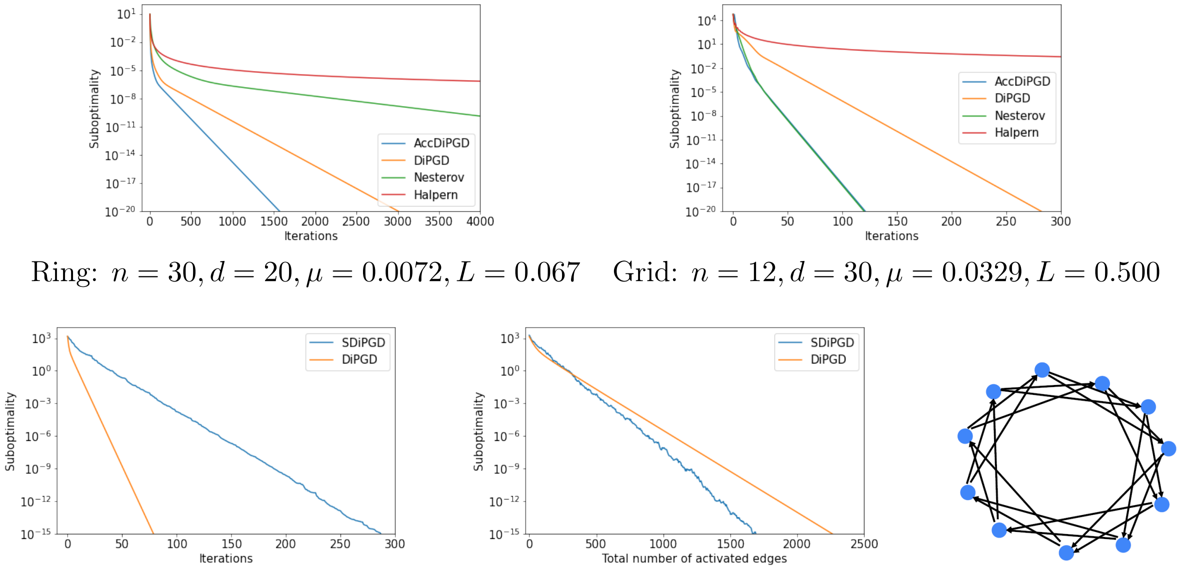
First, we check the convergence of DiPGD for two types of graphs: ring and grid. Both of these graphs satisfy assumptions 13 with constants given in Figure 2. Stepsizes are set according to the theory. The convergence plots are also presented in Figure 2. As we observe, DiPGD converges linearly as we expect.

6.1.2 Acceleration of DiPGD

Now we test accelerated versions of DiPGD that are presented in Section 4.3. We run simulations on the same graphs: ring and grid. Stepsizes for AccDiPGD and Nesterov are fine-tuned. According to the results presented in Figure 3, we state that AccDiPGD performs better than others in all cases. However, it is expected that Nesterov should have almost the same performance as AccDiPGD since it is a special case of three-point acceleration. Besides, Nesterov could be both faster and slower than DiPGD. Probably these two observations can be explained due to the fact stepsizes might be fine-tuned badly in the case of ring. Moreover, we observe that Halpern demonstrates only sublinear convergence, not linear. But this fact is expected since α_k decays as $1/k$.

6.1.3 Performance of SDiPGD

In our next experiment, we investigate empirical behavior of SDiPGD. The graph we use in the experiments as well as convergence plot are presented in Figure 3. Stepsizes are chosen according to



Ring: $n = 30, d = 20, \mu = 0.0072, L = 0.067$ Grid: $n = 12, d = 30, \mu = 0.0329, L = 0.500$

Figure 3: **First line:** The performance of accelerated DiPGD through three-point acceeration (AccDiPGD), Nesterov acceleration (Nesterov), and Halpern-iteration (Halpern). In addition, we add plots of convergence of DiPGD for comparison. **Second line:** The performance of SDiPGD and the graph illustraion used in the experiments.

the theory. As we can see, SDiPGD converges linearly, but has small fluctuations. It performs slower than DiPGD in terms of iterations since in each iteration only a subset of clients communicate with each other. However, the number of edge activations(or equivalently the number of communications) required for convergence is smaller than that of DiPGD.

6.2 Linear regression problem

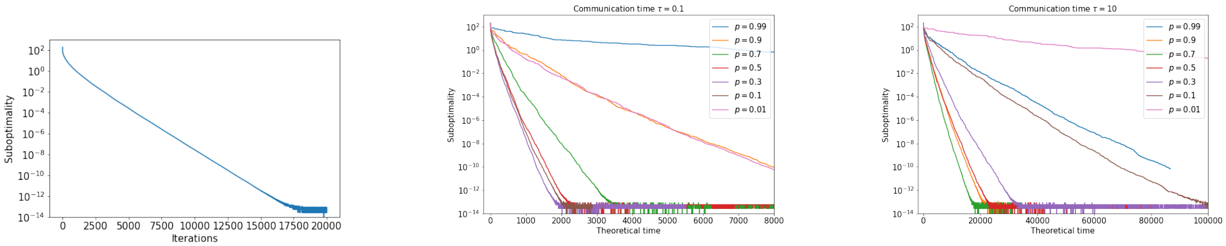
All experiments are done on linear regression problem of the form

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \sum_{i=1}^{n} f_i(x)\right], \quad f_i(x) = \frac{1}{2}\|\mathbf{A}_i x - b_i\|^2, \tag{36}$$

where $\mathbf{A}_i \in \mathbb{R}^{p\times d}$ is a feature matrix of node i, and b_i is corresponding labels. We work with matrices $\mathbf{A}_i$ such that $\mathbf{A}_i^\top \mathbf{A}_i$ is invertable. It is straightforward to check that these functions satisfy all assumption we make on f_i.

6.2.1 Convergence of DiP2GD

Next, we simulate the performance of DiP2GD on a ring graph. We expect to obtain a linear convergence. This is indeed demonstrated in Figure 4 (left plot).



Ring: $n = 10, d = 20, \mathbf{A}_i \in \mathbb{R}^{30\times 20}$

Figure 4: **Left plot:** The performance of DiP2GD. **Central and right plots:** The performance of SDiP2GD in two scenarious: when communication is cheaper than computation ($\tau = 0.1$, central plot) and vice versa ($\tau = 10$, right plot).

6.2.2 Behavior of SDiP2GD

Finally, we demonstrate how SDiP2GD behaves in two scenarious: when communication is expensive or cheap comparing with computation. We assume that computation costs 1 unit of time, and communication — τ units of time. We test SDiP2GD in two cases when τ is larger than 1 and smaller than 1. We choose different values of probability p and plot the theoretical time of convergence in Figure 4. Observing the results, we may claim that in the case when communication is cheap ($\tau = 0.1$, left plot in Figure 4) it is beneficial to perform more communication steps (i.e. $p < 0.5$). Conversely, if communication is expensive ($\tau = 10$, right plot in Figure 4) it is efficient to perform more computation steps ($p > 0.5$). As a consequence, there is a trade-off to choose the optimal value of p depending on the cost ratio of communication and computation. Nevertheless, we claim that the performance of SDiP2GD with $p = 0.5$ is close to the best one in both cases (small and large τ). According to the theory of undirected graphs, there is no such trade-off, and values of optimal parameters, including the value of p, are set according to smoothness and other constants depending on the problem only.

7 Future work

We made a step towards understanding of gradient-type methods for directed decentralized optimization. We create algorithms that supports acceleration, stochasticity and other important features. However, the analysis relies on strong assumptions that may work for small range of the condition number or do not hold at all in practice. Thus, much more work should be done in this direction to improve the analysis. Finally, following [Hendrikx et al., 2019] we are able to construct a modification of DiP2GD that additionally supports variance reduction mechanism in the case when f_i is also a finite sum of functions for all i. However, the analysis still should be impoved.

A Convergence of DiPGD

First of all, we would like to point out that the choice of preconditioner allows us to simplify the iterative process of DiPGD. Indeed, we have

$$\begin{aligned}\mathbf{M}\mathbf{P}\mathbf{A}^\top\mathbf{A} &= \mathbf{P}\mathbf{P}^\dagger\mathbf{P}\mathbf{A}^\top\mathbf{A}\\ &= \mathbf{U}\mathbf{D}^\dagger\mathbf{V}^\top\cdot\mathbf{V}\mathbf{D}\mathbf{U}^\top\cdot\mathbf{U}\mathbf{D}^\dagger\mathbf{V}^\top\cdot\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\\ &= \mathbf{U}\mathbf{D}\mathbf{V}^\top = \mathbf{A}.\end{aligned}$$

To repeat, we assume that Assumptions 13 hold. Then we are able to establish the convergence theorem for DiPGD that was previously stated in Section 4.2.

Theorem A.1. *Let the matrix* $\mathbf{A}$ *satisfies the conditions* $\mathrm{Ker}(\mathbf{A}^\top) = \mathrm{Ker}(\mathbf{A}) = \mathrm{Span}(\mathbb{1})$ *and* $\mathbf{A}_{ij} \neq 0$ *iff* $(i,j) \in \mathcal{E}$. *Let* (13) *assumptions hold with constants* μ *and* L, *and the stepsize satisfies* $\eta \leq 1/L$. *Then the iterates of* DiPGD *converge linearly*

$$\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 \leq (1-\eta\mu)^k \left\|\Lambda^0 - \Lambda^*\right\|_{\mathbf{M}}^2. \tag{37}$$

Proof. Necessary optimality condition gives us the fact that $\mathbf{A}\Lambda^* = 0$. Then we need to follow the convergence proof of Gradient Descent

$$\begin{aligned}\left\|\Lambda^{k+1} - \Lambda^*\right\|_{\mathbf{M}}^2 &= \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 + \eta^2\left\|\mathbf{A}(\Lambda^k - \Lambda^*)\right\|_{\mathbf{M}}^2 - 2\eta(\Lambda^k - \Lambda^*)^\top\mathbf{M}\mathbf{A}(\Lambda^k - \Lambda^*)\\ &= \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 + \eta^2(\Lambda^k - \Lambda^*)^\top\mathbf{A}^\top\mathbf{A}(\Lambda^k - \Lambda^*)\\ &\quad -2\eta(\Lambda^k - \Lambda^*)^\top\mathbf{A}(\Lambda^k - \Lambda^*)\\ &\overset{\text{co-coer.}}{\leq} \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 + L\eta^2(\Lambda^k - \Lambda^*)^\top\mathbf{A}(\Lambda^k - \Lambda^*)\\ &\quad -2\eta(\Lambda^k - \Lambda^*)^\top\mathbf{A}(\Lambda^k - \Lambda^*)\\ &= \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - \eta(2-\eta L)(\Lambda^k - \Lambda^*)^\top\mathbf{A}(\Lambda^k - \Lambda^*).\end{aligned}$$

Note that using the restriction on the stepsize η we have $2 - \eta L \geq 1$. Since we assume $\mathbf{A}$ to be positive semidefinite, then we can continue the chain of inequalities as follows

$$\left\|\Lambda^{k+1} - \Lambda^*\right\|_{\mathbf{M}}^2 \overset{\text{str.conv}}{\leq} \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - \alpha\mu\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 = (1-\alpha\mu)\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2$$

that finalises the proof. □

Remark A.2. *We need to understand when Assumptions 13 hold. For that we need to check if* $\mathbf{A} \succeq \mu\mathbf{M}$ *is true for some positive* μ. *Note that if* $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, *then* $\mathbf{M} = \mathbf{U}\mathbf{D}_p\mathbf{U}^\top$ *where* $\mathbf{D}_p = \mathbf{D}\mathbf{D}^\dagger$. *Then strong convexity assumption is verified. Indeed,*

$$\begin{aligned}\mu x^\top\mathbf{U}\mathbf{D}_p\mathbf{U}^\top x &= \mu\sum_{i=1}^{n-1}[\mathbf{U}^\top x]_i^2\\ &\leq \mu\sum_{i=1}^{n}[\mathbf{U}^\top x]_i^2\\ &= \mu x^\top\mathbf{U}\mathbf{U}^\top x.\end{aligned}$$

Since $\mathbf{U}^\top \mathbf{1} = 0$*, then* $x^\top \mathbf{A} x \geq \mu x^\top \mathbf{U}\mathbf{D}_p\mathbf{U}^\top x$ *with* $\mu = \lambda^+_{\min}(\mathbf{A})$*.*

Remark A.3. *In the similar way we can derive that for* L*-co-coercivity we need to check if* $L\mathbf{A} \succeq \mathbf{A}^\top\mathbf{A}$ *or* $L\frac{\mathbf{A}+\mathbf{A}^\top}{2} \succeq \mathbf{A}^\top\mathbf{A}$*. This could be satisfied if we upper bound* $\mathbf{A}^\top\mathbf{A}$ *by* $\lambda_{\max}(\mathbf{A}^\top\mathbf{A})\mathbf{I}$ *and lower bound* $\frac{\mathbf{A}+\mathbf{A}^\top}{2}$ *by* $\lambda^+_{\min}\left(\frac{1}{2}(\mathbf{A}+\mathbf{A}^\top)\right)\mathbf{I}$*. Here we highlight that all the above works because we work in the range of* $\mathbf{A}$*.*

B Convergence on SDiPGD

Theorem B.1. *Let* (21) *assumptions hold with constants* μ *and* $\overline{L}$*, and the stepsize satisfies* $\eta \leq 1/\overline{L}$*. Then the iterates of* SDiPGD *converge linearly*

$$\mathbb{E}\left[\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}}\right] \leq (1-\eta\mu)^k \left\|\Lambda^0 - \Lambda^*\right\|^2_{\mathbf{M}}. \tag{38}$$

Proof. Necessary optimality condition gives us the fact that $\mathbf{A}_k\Lambda^* = 0$. Then we need to follow the convergence proof of Gradient Descent. Here we use the expression $\mathbb{E}_k[\cdot]$ for the expectation w.r.t. to iterate Λ^k. We have

$$\begin{aligned}
\mathbb{E}_k\left[\left\|\Lambda^{k+1} - \Lambda^*\right\|^2_{\mathbf{M}}\right] &= \left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} + \eta^2 m^2 \mathbb{E}_k\left[\left\|\mathbf{A}_k(\Lambda^k - \Lambda^*)\right\|^2_{\mathbf{M}}\right] \\
&\quad -2\eta(\Lambda^k - \Lambda^*)^\top \mathbf{M}\mathbf{A}(\Lambda^k - \Lambda^*) \\
&\overset{\text{co-coer.}}{\leq} \left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} + \overline{L}\eta^2 m \mathbb{E}_k\left[(\Lambda^k - \Lambda^*)^\top \mathbf{M}\mathbf{A}_k(\Lambda^k - \Lambda^*)\right] \\
&\quad -2\eta(\Lambda^k - \Lambda^*)^\top \mathbf{A}(\Lambda^k - \Lambda^*) \\
&= \left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} + \overline{L}\eta^2 (\Lambda^k - \Lambda^*)^\top \mathbf{M}\mathbf{A}(\Lambda^k - \Lambda^*) \\
&\quad -2\eta(\Lambda^k - \Lambda^*)^\top \mathbf{A}(\Lambda^k - \Lambda^*) \\
&= \left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} - \eta(2-\eta\overline{L})(\Lambda^k - \Lambda^*)^\top \mathbf{A}(\Lambda^k - \Lambda^*).
\end{aligned}$$

Note that using the restriction on the stepsize η we have $2 - \eta\overline{L} \geq 1$. Since we assume $\mathbf{A}$ to be positive semidefinite, then we can continue the chain of inequalities as follows

$$\mathbb{E}_k\left[\left\|\Lambda^{k+1} - \Lambda^*\right\|^2_{\mathbf{M}}\right] \overset{\text{str. cvx.}}{\leq} \left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} - \alpha\mu\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}} = (1-\alpha\mu)\left\|\Lambda^k - \Lambda^*\right\|^2_{\mathbf{M}}.$$

Unrolling this inequality till the initial iterate, we obtain the statement of the theorem. □

C Accelerated DiPGD via Halpern-iteration

We start with the lemma that connects non-expansive and co-coercive operators.

Lemma C.1. *Let the matrix* $\mathbf{A}$ *satisfies the conditions* $\mathrm{Ker}(\mathbf{A}^\top) = \mathrm{Ker}(\mathbf{A}) = \mathrm{Span}(\mathbb{1})$ *and* $\mathbf{A}_{ij} \neq 0$ *iff* $(i,j) \in \mathcal{E}$*. Let the second assumption of* (13) *holds with constant* L*, and the stepsize satisfies* $\eta \leq 1/L$*. Then operator* $T(\Lambda) := \Lambda - \eta\mathbf{A}\Lambda$ *is non-expansive.*

Proof. The proof is straightforward. Indeed,

$$\begin{aligned}
\left\|(\mathbf{X}-\eta\mathbf{A}\mathbf{X})-(\mathbf{Y}-\eta\mathbf{A}\mathbf{Y})\right\|_{\mathbf{M}}^2 &= \|\mathbf{X}-\mathbf{Y}\|_{\mathbf{M}}^2+\eta^2\|\mathbf{A}(\mathbf{X}-\mathbf{Y})\|_{\mathbf{M}}^2-2\eta(\mathbf{X}-\mathbf{Y})^\top\mathbf{A}(\mathbf{X}-\mathbf{Y})\\
&\overset{\text{co}-\text{coer.}}{\leq} \|\mathbf{X}-\mathbf{Y}\|_{\mathbf{M}}^2+L\eta^2(\mathbf{X}-\mathbf{Y})^\top\mathbf{A}(\mathbf{X}-\mathbf{Y})\\
&\qquad -2\eta(\mathbf{X}-\mathbf{Y})^\top\mathbf{A}(\mathbf{X}-\mathbf{Y})\\
&= \|\mathbf{X}-\mathbf{Y}\|_{\mathbf{M}}^2-\eta(2-L\eta)(\mathbf{X}-\mathbf{Y})^\top\mathbf{A}(\mathbf{X}-\mathbf{Y})\\
&\leq (1-\eta\mu)\|\mathbf{X}-\mathbf{Y}\|_{\mathbf{M}}^2 \qquad (39)
\end{aligned}$$

where the last inequality holds due to stepsize restriction. □

Theorem C.2. *Let the matrix* $\mathbf{A}$ *satisfies the conditions* $\mathrm{Ker}(\mathbf{A}^\top)=\mathrm{Ker}(\mathbf{A})=\mathrm{Span}(\mathbb{1})$ *and* $\mathbf{A}_{ij}\neq 0$ *iff* $(i,j)\in\mathcal{E}$. *Let the second assumption of* (13) *holds with constant* L, *and the stepsize satisfies* $\eta\leq {}^1\!/_L$. *Then the iterates of accelerated* DiPGD *through Halpern-iteration converge sublinearly*

$$\frac{1}{2}\left\|\Lambda^k-T(\Lambda^k)\right\|_{\mathbf{M}}^2\leq\frac{\|\Lambda^0-\Lambda^*\|_{\mathbf{M}}^2}{(k+1)^2}. \qquad (40)$$

Proof. The recurrent formula of Halpern-iteration acceleration implies

$$\Lambda^j=\frac{1}{j+1}\Lambda^0+\frac{j}{j+1}T(\Lambda^{j-1}) \quad \text{or } T(\Lambda^{j-1})=\frac{j+1}{j}\Lambda^j-\frac{1}{j}\Lambda^0. \qquad (41)$$

By non-expansiveness of operator T which follows from Lemma (C.1) we get

$$\left\|T(\Lambda^k)-T(\Lambda^*)\right\|_{\mathbf{M}}^2\leq\left\|\Lambda^k-\Lambda^*\right\|_{\mathbf{M}}^2, \qquad (42)$$

and

$$\left\|T(\Lambda^j)-T(\Lambda^{j-1})\right\|_{\mathbf{M}}^2\leq\left\|\Lambda^j-\Lambda^{j-1}\right\|_{\mathbf{M}}^2 \quad \text{for } j=1,\dots,k. \qquad (43)$$

Below we reformulate the following weighted sum of (43)

$$0\geq\sum_{j=1}^k j(j+1)\left(\left\|T(\Lambda^j)-T(\Lambda^{j-1})\right\|_{\mathbf{M}}^2-\left\|\Lambda^j-\Lambda^{j-1}\right\|_{\mathbf{M}}^2\right). \qquad (44)$$

Using the second relation in (41) the first terms in the summation (44) are

$$\begin{aligned}
j(j+1)\left\|T(\Lambda^j)-T(\Lambda^{j-1})\right\|_{\mathbf{M}}^2 &= j(j+1)\left\|\Lambda^j-T(\Lambda^j)+\frac{1}{j}(\Lambda^j-\Lambda^0)\right\|_{\mathbf{M}}^2\\
&= j(j+1)\left\|\Lambda^j-T(\Lambda^j)\right\|_{\mathbf{M}}^2+2(j+1)(\Lambda^j-T(\Lambda^j))^\top\mathbf{M}(\Lambda^j-\Lambda^0)\\
&\qquad \frac{j+1}{j}\left\|\Lambda^j-\Lambda^0\right\|_{\mathbf{M}}^2, \qquad (45)
\end{aligned}$$

and using the first relation in (41) it follows for the second terms in (44)

$$\begin{aligned}
-j(j+1)\left\|\Lambda^j-\Lambda^{j-1}\right\|_{\mathbf{M}}^2 &= -j(j+1)\left\|\frac{1}{j}(\Lambda^0-T(\Lambda^{j-1}))+T(\Lambda^{j-1})-\Lambda^{j-1}\right\|_{\mathbf{M}}^2\\
&= j(j+1)\left\|\Lambda^0-T(\Lambda^{j-1})\right\|_{\mathbf{M}}^2\\
&\qquad -2j(\Lambda^0-T(\Lambda^{j-1}))^\top\mathbf{M}(T(\Lambda^{j-1})-\Lambda^{j-1})\\
&\qquad -j(j+1)\left\|T(\Lambda^{j-1})-\Lambda^{j-1}\right\|_{\mathbf{M}}^2. \qquad (46)
\end{aligned}$$

Observe [using again the second relation in (41)] that the first term in (46)

$$\begin{aligned} -\frac{j}{j+1}\left\|\Lambda^0 - T(\Lambda^{j-1})\right\|_{\mathbf{M}}^2 &= -\frac{j+1}{j}\left\|\frac{j}{j+1}\Lambda^0 - \frac{j+1}{j}\Lambda^j\right\|_{\mathbf{M}}^2 \\ &= -\frac{j+1}{j}\left\|\Lambda^0 - \Lambda^j\right\|_{\mathbf{M}}^2 \end{aligned} \tag{47}$$

cancels the third term in (45). Summing up the second terms in (46) for $j = 1, \ldots, k$ we shift the summation index,

$$-\sum_{j=1}^{k} 2j(\Lambda^0 - T(\Lambda^{j-1}))^\top \mathbf{M}(T(\Lambda^{j-1}) - \Lambda^{j-1}) = \sum_{j=0}^{k-1} 2(j+1)(\Lambda^j - T(\Lambda^j))\mathbf{M}(\Lambda^0 - T(\Lambda^j)),$$

so that summing up the second terms in (45) and in (46) for $j = 1, \ldots, k$ results in

$$2(k+1)(\Lambda^k - T(\Lambda^k))^\top \mathbf{M}(\Lambda^k - \Lambda^0) + 2\sum_{j=1}^{k-1}(j+1)(\Lambda^j - T(\Lambda^j))^\top \mathbf{M}(\Lambda^j - T(\Lambda^j)) + 2\left\|\Lambda^0 - T(\Lambda^0)\right\|_{\mathbf{M}}^2. \tag{48}$$

Shifting again the index in the summation of the third term in (46)

$$-\sum_{j=1}^{k} j(j+1)\left\|\Lambda^{j-1} - T(\Lambda^{j-1}))\right\|_{\mathbf{M}}^2 = -\sum_{j=0}^{k-1}(j+1)(j+2)\left\|\Lambda^j - T(\Lambda^j)\right\|_{\mathbf{M}}^2$$

and summing up the first terms in (45) and the third terms in (46) for $j = 1, \ldots, k$ gives

$$k(k+1)\left\|\Lambda^k - T(\Lambda^k)\right\|_{\mathbf{M}}^2 - 2\sum_{j=1}^{k-1}(j+1)\left\|\Lambda^j - T(\Lambda^j)\right\|_{\mathbf{M}}^2 - 2\left\|\Lambda^0 - T(\Lambda^0)\right\|_{\mathbf{M}}^2 \tag{49}$$

where the sum in the middle cancels the sum in the middle of (48) and the terms $2\left\|\Lambda^0 - T(\Lambda^0)\right\|_{\mathbf{M}}^2$ cancel as well. The only remaining terms are the first terms in (48) and (49).

Thus, inserting (47), (48) and (49) in (48) leads to

$$0 \geq k(k+1)\left\|\Lambda^k - T(\Lambda^k)\right\|_{\mathbf{M}}^2 + 2(k+1)(\Lambda^k - T(\Lambda^k))^\top \mathbf{M}(\Lambda^k - \Lambda^0). \tag{50}$$

Applying Cauchy-Schwartz inequality to the second term in (50) leads to

$$\frac{1}{2}\left\|\Lambda^k - T(\Lambda^k)\right\|_{\mathbf{M}} \leq \frac{1}{k}\left\|\Lambda^k - \Lambda^0\right\|.$$

To prove the theorem, (50) is divided by $(k+1)$ and then (42) is added:

$$\begin{aligned} 0 &\geq k\left\|\Lambda^k - T(\Lambda^k)\right\|_{\mathbf{M}}^2 + 2(\Lambda^k - T(\Lambda^k))^\top \mathbf{M}(\Lambda^k - \Lambda^0) + \left\|T(\Lambda^k) - \Lambda^*\right\|_{\mathbf{M}}^2 - \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 \\ &= \frac{k+1}{2}\left\|\Lambda^k - T(\Lambda^k)\right\|_{\mathbf{M}}^2 - \frac{2}{k+1}\left\|\Lambda^0 - \Lambda^*\right\|_{\mathbf{M}}^2 \\ &\qquad + \frac{2}{k+1}\left\|\Lambda^0 - \Lambda^* - \frac{k+1}{2}(\Lambda^k - T(\Lambda^k))\right\|_{\mathbf{M}}^2. \end{aligned} \tag{51}$$

To see the last equation, the last two terms in (51) can be combined, and then a straightforward multiplication of the terms $a := \Lambda^k - T(\Lambda^k)$, $b := \Lambda^k - \Lambda^0$, $c := T(\Lambda^k) - \Lambda^*$, and $a + c - b = \Lambda^0 - \Lambda^*$ reveals the identity. Omitting the last term in (51) one obtains

$$\left\|\Lambda^k - T(\Lambda^k)\right\|_{\mathbf{M}}^2 \leq \left(\frac{2}{k+1}\right)^2 \left\|\Lambda^0 - \Lambda^*\right\|_{\mathbf{M}}^2,$$

which proves the theorem. □

D Detailed explanation of DiP2GD

In this section we clarify the explanation of how to compute proximity operator and other aspects that are discussed in Section 5.1.

D.1 Proximity operator

We need to understand how to compute the proximity operator of $\eta\psi$. Since ψ is a separable function of λ^1 w.r.t. λ_i^1, then

$$\begin{aligned}
\operatorname{prox}_{\eta\psi}(\varphi) &= \arg\min_{t^1,t^2} \left\{ \sum_{i=1}^n \eta \tilde{f}_i^*(t_i^1) + \frac{1}{2}\|\varphi - t\|^2 \right\} \\
&= \begin{pmatrix} \vdots \\ \arg\min_{t_i^1} \eta \tilde{f}_i^*(t_i^1) + \frac{1}{2}\|\varphi_i^1 - t_i^1\|^2 \\ \vdots \\ \varphi^2 \end{pmatrix} \\
&= \begin{pmatrix} \operatorname{prox}_{\eta\tilde{\psi}}(\varphi^1) \\ \varphi^2 \end{pmatrix},
\end{aligned}$$

where $\tilde{\psi}(\varphi^1) := \sum_{i=1}^n \tilde{f}_i^*(\varphi_i^1)$.

Using Moreau identity, we can derive the explicit way to compute the proximity operator. We denote $\tilde{\eta} = \eta\mu^2$, and then

$$\begin{aligned}
\operatorname{prox}_{\eta\tilde{f}_i^*}(\varphi_i^1) &= \arg\min_v f_i^*(\mu v) - \frac{\mu^2}{2L}\|v\|^2 + \frac{1}{2\eta}\|v - \varphi_i^1\|^2 \\
&= \frac{1}{\mu} \arg\min_u f_i^*(u) - \frac{1}{2L}\|u\|^2 + \frac{1}{2\eta}\|u/\mu - \varphi_i^1\|^2 \\
&= \frac{1}{\mu} \arg\min_u f_i^*(u) + \frac{1}{2}(\tilde{\eta}^{-1} - L^{-1})\|u\|^2 - \frac{\mu}{\tilde{\eta}} u^\top \varphi_i^1 \\
&= \frac{1}{\mu} \arg\min_u f_i^*(u) + \frac{1}{2}(\tilde{\eta}^{-1} - L_i^{-1}) \left\| u - (\tilde{\eta}^{-1} - L^{-1})^{-1} \frac{\mu}{\tilde{\eta}} \varphi_i^1 \right\|^2 \\
&= \operatorname{prox}_{(\tilde{\eta}^{-1} - L_i^{-1})^{-1} f_i^*} \left((\tilde{\eta}^{-1} - L^{-1})^{-1} \frac{\mu}{\tilde{\eta}} \varphi_i^1 \right).
\end{aligned}$$

Now we use Moreau identity $\operatorname{prox}_{\gamma f^*}(z) = z - \gamma \operatorname{prox}_{\gamma^{-1} f}(\gamma^{-1} z)$ to obtain with

$$\gamma = \left(\frac{1}{\tilde{\eta}} - \frac{1}{L}\right)^{-1},$$

and

$$z = \frac{\mu_i}{\tilde{\eta}} \varphi_i^1 .$$

Then we derive

$$\begin{aligned} \operatorname{prox}_{\eta \tilde{f}_i^*}(\varphi_i^1) &= \operatorname{prox}_{(\tilde{\eta}^{-1} - L^{-1})^{-1} f_i^*} \left((\tilde{\eta}^{-1} - L^{-1})^{-1} \frac{\mu}{\tilde{\eta}} \varphi_i^1 \right) \\ &= \frac{\mu}{\tilde{\eta}} \varphi_i^1 - (\tilde{\eta}^{-1} - L^{-1})^{-1} \operatorname{prox}_{(\tilde{\eta}^{-1} - L^{-1}) f_i} \left(\frac{\varphi_i^1 \mu}{\tilde{\eta}} \left(\frac{1}{\tilde{\eta}} - \frac{1}{L} \right) \right). \qquad (52) \end{aligned}$$

The above could be done if $L > \tilde{\eta}$, i.e. we need to assume that $\eta \mu^2 < L$.

D.2 Choice of matrices does not break properties of proximity operator

Note that $\partial \psi(\Lambda)$ has the form

$$\partial \psi(\Lambda) = \begin{pmatrix} \partial \tilde{f}_1^*(\lambda_i^1) \\ \vdots \\ \partial \tilde{f}_n^*(\lambda_i^n) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

As a consequence, $\mathbf{P} \partial \psi = \partial \psi$.

The second observation is that the first-order necessary condition of λ^* to be a solution to (28) is

$$\begin{aligned} 0 \in \partial \psi(\Lambda^*) + \mathbf{A}^\top \Sigma^{-1} \mathbf{A} \Lambda^* \quad &\Leftrightarrow \\ 0 \in \eta \mathbf{P} \partial \psi(\Lambda^*) + \eta \mathbf{P} \mathbf{A}^\top \Sigma^{-1} \mathbf{A} \Lambda^* \quad &\Leftrightarrow \\ 0 \in \eta \partial \psi(\Lambda^*) + \eta \mathbf{P} \mathbf{A}^\top \Sigma^{-1} \mathbf{A} \Lambda^* \quad &\Leftrightarrow \\ 0 \in \eta \partial \psi(\Lambda^*) + \Lambda^* - \Lambda^* + \eta \mathbf{P} \mathbf{A}^\top \Sigma^{-1} \mathbf{A} \Lambda^* \quad &\Leftrightarrow \\ \Lambda^* = \operatorname{prox}_{\eta \psi} \left[\Lambda^* - \eta \mathbf{P} \mathbf{A}^\top \Sigma^{-1} \mathbf{A} \Lambda^* \right]. \qquad (53) \end{aligned}$$

The above means that even in the case of preconditioning the proper choice of $\mathbf{P}$ does not break the fixed point property of the proximity operator.

On top of that, the non-expansiveness property of the proximity operator also remains unchanged in the $\mathbf{M}$-norm. Indeed, we have

$$\begin{aligned} \left\| \operatorname{prox}_{\eta \psi}(X) - \operatorname{prox}_{\eta \psi}(Y) \right\|_{\mathbf{M}}^2 &= \left\| \begin{pmatrix} \operatorname{prox}_{\eta \tilde{\psi}}(X^1) \\ X^2 \end{pmatrix} - \begin{pmatrix} \operatorname{prox}_{\eta \tilde{\psi}}(Y^1) \\ Y^2 \end{pmatrix} \right\|_{\mathbf{M}}^2 \\ &= \left\| \operatorname{prox}_{\eta \tilde{\psi}}(X^1) - \operatorname{prox}_{\eta \tilde{\psi}}(Y^1) \right\|^2 + \left\| X^2 - Y^2 \right\|_{\mathbf{I}_p}^2 \\ &\leq \left\| X^1 - Y^1 \right\|^2 + \left\| X^2 - Y^2 \right\|_{\mathbf{I}_p}^2 \\ &= \left\| X - Y \right\|_{\mathbf{M}}^2. \qquad (54) \end{aligned}$$

E Convergence of DiP2GD

Theorem E.1. *Let a function g satisfies assumptions* (32) *with constants α and β. Let* $\mathrm{Ker}(\tilde{\mathbf{A}}) = \mathrm{Ker}(\tilde{\mathbf{A}}^\top) = \mathrm{Span}(\mathbb{1})$*. Then the iterates of* DiP2GD *with the stepsize $\eta \le \frac{1}{\beta}$ satisfy*

$$\left\|\Lambda^{k+1} - \Lambda^*\right\|_{\mathbf{M}}^2 \le (1-\eta\alpha)\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2.$$

Proof. It follows the proof of theorem 4.4 combining properties of the proximity operator. First, recall the fixed point property of the proximity operator that we show in Section D.2

$$\Lambda^* = \mathrm{prox}_{\eta\psi}\left[\Lambda^* - \eta\mathbf{P}\nabla g(\Lambda^*)\right].$$

We need to follow the convergence proof of Proximal Gradient Descent

$$\begin{aligned}
\left\|\Lambda^{k+1} - \Lambda^*\right\|_{\mathbf{M}}^2 &\overset{(53)}{=} \left\|\mathrm{prox}_{\eta\psi}\left[\Lambda^k - \eta\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}\Lambda^k\right] - \mathrm{prox}_{\eta\psi}\left[\Lambda^* - \eta\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}\lambda^*\right]\right\|_{\mathbf{M}}^2 \\
&\overset{(54)}{\le} \left\|\Lambda^k - \eta\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}\Lambda^k - \Lambda^* + \eta\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}\Lambda^*\right\|_{\mathbf{M}}^2 \\
&= \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - 2\eta(\Lambda^k - \Lambda^*)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(\Lambda^k - \Lambda^*) \\
&\quad + \eta^2(\Lambda^k - \Lambda^*)^\top\mathbf{A}^\top\Sigma^{-1}\mathbf{A}\mathbf{P}^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(\Lambda^k - \Lambda^*).
\end{aligned}$$

Now we use relative smoothness and strong convexity assumption and obtain

$$\begin{aligned}
\left\|\Lambda^{k+1} - \Lambda^*\right\|_{\mathbf{M}}^2 &\overset{(32)}{\le} \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - 2\eta(\Lambda^k - \Lambda^*)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(\Lambda^k - \Lambda^*) \\
&\quad + \beta\eta^2(\Lambda^k - \Lambda^*)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(\lambda^k - \Lambda^*) \\
&= \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - \eta(2-\eta\beta)(\Lambda^k - \Lambda^*)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(\Lambda^k - \Lambda^*) \\
&\le \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - \eta(\Lambda^k - \Lambda^*)^\top\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}(\Lambda^k - \Lambda^*) \\
&\overset{(32)}{\le} \left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2 - \eta\alpha\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2,
\end{aligned}$$

where in the second inequality we use the restriction on the stepsize η: $2 - \eta L \ge 1$, and the fact that $\mathbf{M}\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}$ that follows from the co-coercivity assumption (32). This finalizes the proof. □

F Convergence of SDiP2GD

Theorem F.1. *Let a function g satisfies assumptions* (35) *with constants α and $\overline{\beta}$. Let* $\mathrm{Ker}(\tilde{\mathbf{A}}) = \mathrm{Ker}(\tilde{\mathbf{A}}^\top) = \mathrm{Span}(\mathbb{1})$*. Then the iterates of* SDiP2GD *with the stepsize $\eta \le \frac{1}{\overline{\beta}}$ satisfy*

$$\left\|\Lambda^{k+1} - \Lambda^*\right\|_{\mathbf{M}}^2 \le (1-\eta\alpha)\left\|\Lambda^k - \Lambda^*\right\|_{\mathbf{M}}^2.$$

Proof. We have

$$\begin{aligned}
\mathbb{E}\left[\left\|Y^{k+1}-Y^*\right\|^2_{\mathbf{M}}\right] &= p\left\|\mathrm{prox}_{\eta\psi}\left[Y^k-\eta p^{-1}\mathbf{G}_1\right]-\mathrm{prox}_{\eta\psi}\left[Y^*-\eta\mathbf{G}Y^*\right]\right\|^2_{\mathbf{M}}\\
&\quad+(1-p)\left\|\mathrm{prox}_{\eta\psi}\left[Y^k-\eta(1-p)^{-1}\mathbf{G}_2\right]-\mathrm{prox}_{\eta\psi}\left[Y^*-\eta\mathbf{G}Y^*\right]\right\|^2_{\mathbf{M}}\\
&\overset{\text{non-exp.}}{\leq} p\left\|\left[Y^k-\eta p^{-1}\mathbf{G}_1\right]-\left[Y^*-\eta\mathbf{G}Y^*\right]\right\|^2_{\mathbf{M}}\\
&\quad+(1-p)\left\|\left[Y^k-\eta(1-p)^{-1}\mathbf{G}_2\right]-\left[Y^*-\eta\mathbf{G}Y^*\right]\right\|^2_{\mathbf{M}}\\
&= \left\|Y^k-Y^*\right\|^2_{\mathbf{M}}-2\eta(Y^k-Y^*)^\top\mathbf{M}(\mathbf{G}_1Y^k+\mathbf{G}_2Y^k-\mathbf{G}Y^*)\\
&\quad+\eta^2p\left\|p^{-1}\mathbf{G}_1Y^k-\mathbf{G}Y^*\right\|^2_{\mathbf{M}}+\eta^2(1-p)\left\|(1-p)^{-1}\mathbf{G}_2Y^k-\mathbf{G}Y^*\right\|^2_{\mathbf{M}}\\
&= \left\|Y^k-Y^*\right\|^2_{\mathbf{M}}-2\eta(Y^k-Y^*)^\top\mathbf{G}(Y^k-Y^*)\\
&\quad+\eta^2p\left\|p^{-1}\mathbf{G}_1Y^k-\mathbf{G}Y^*\right\|^2_{\mathbf{M}}+\eta^2(1-p)\left\|(1-p)^{-1}\mathbf{G}_2Y^k-\mathbf{G}Y^*\right\|^2_{\mathbf{M}}.
\end{aligned}$$

Note that the sum of two last terms above is equal to $\eta^2\mathbb{E}\left[\left\|\tilde{\mathbf{G}}Y^k-\mathbf{G}Y^*\right\|^2_{\mathbf{M}}\right]$. Using assumption 35 we can continue

$$\begin{aligned}
\mathbb{E}\left[\left\|Y^{k+1}-Y^*\right\|^2_{\mathbf{M}}\right] &\leq \left\|Y^k-Y^*\right\|^2_{\mathbf{M}}-2\eta(Y^k-Y^*)^\top\mathbf{G}(Y^k-Y^*)\\
&\quad+\overline{\beta}\eta^2(Y^k-Y^*)^\top\mathbf{G}(Y^k-Y^*)\\
&= \left\|Y^k-Y^*\right\|^2_{\mathbf{M}}-\eta(2-\eta\overline{\beta})(Y^k-Y^*)^\top\mathbf{G}(Y^k-Y^*)\\
&\overset{\text{str. cvx.}}{\leq} \left\|Y^k-Y^*\right\|^2_{\mathbf{M}}-\eta\alpha\left\|Y^k-Y^*\right\|^2_{\mathbf{M}}.
\end{aligned} \tag{55}$$

□

G Analysis of the spectrum of gossip matrix

We need to satisfy the condition $\mathbf{PA}^\top\Sigma^{-1}\mathbf{A}\succeq\alpha_0\mathbf{M}$ for some $\alpha>0$. As a reminder, $\mathbf{PA}^\top\Sigma^{-1}\mathbf{A}$ has the form of

$$\mathbf{PA}^\top\Sigma^{-1}\mathbf{A}=\begin{pmatrix}\frac{\mu^2(L+\sigma)}{L\sigma}\mathbf{I} & -\frac{\mu}{\sigma}\tilde{\mathbf{A}}\\ -\frac{\mu}{\sigma}\mathbf{I}_p & \frac{1}{\sigma}\tilde{\mathbf{A}}\end{pmatrix}.$$

We want to clarify that we can always work with $\alpha\in\mathbb{R}$ since the positive definiteness of $\mathbf{PA}^\top\Sigma^{-1}\mathbf{A}$ and its symmetrized version is equivalent.

Now, we write the definition of positive definiteness of $\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}$.

$$\begin{aligned}\begin{pmatrix}x\\y\end{pmatrix}^\top \begin{pmatrix}\frac{\mu^2(L+\sigma)}{L\sigma}\mathbf{I}-\alpha_0\mathbf{I} & -\frac{\mu}{\sigma}\tilde{\mathbf{A}}\\ -\frac{\mu}{\sigma}\mathbf{I}_p & \frac{1}{\sigma}\tilde{\mathbf{A}}-\alpha_0\mathbf{I}_p\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix} &= \left(\frac{\mu^2(L+\sigma)}{L\sigma}-\alpha_0\right)\|x\|^2\\ &\quad -\frac{\mu}{\sigma}x^\top(\tilde{\mathbf{A}}+\mathbf{I}_p)y+y^\top\left(\frac{1}{\sigma}\tilde{\mathbf{A}}-\alpha_0\mathbf{I}_p\right)y\\ &= (\alpha-\alpha_0)\left\|x-\frac{\mu}{2\sigma(\alpha-\alpha_0)}(\tilde{\mathbf{A}}+\mathbf{I}_p)y\right\|^2\\ &\quad -\frac{\mu^2}{4\sigma^2(\alpha-\alpha_0)}\left\|(\tilde{\mathbf{A}}+\mathbf{I}_p)y\right\|^2+y^\top\left(\frac{1}{\sigma}\tilde{\mathbf{A}}-\alpha_0\mathbf{I}_p\right)y\\ &\geq 0,\end{aligned}$$

where $\alpha := \frac{\mu^2(L+\sigma)}{L\sigma}$. We define the condition number $\kappa := \frac{L+\sigma}{\sigma} > 1$ of the problem, and perform the simple change $\alpha_0 := \frac{\xi\mu^2}{\sigma}\frac{\kappa}{\kappa-1}$. From the above we have the first restriction on ξ:

$$\begin{aligned}\alpha-\alpha_0\geq 0 &\quad\Leftrightarrow\\ \frac{\mu^2(L+\sigma)}{L\sigma}-\frac{\xi\mu^2}{\sigma}\frac{\kappa}{\kappa-1}\geq 0 &\quad\Leftrightarrow\\ \kappa\frac{\sigma}{L}-\xi\frac{\kappa}{\kappa-1}\geq 0 &\quad\Leftrightarrow\\ \frac{\kappa}{\kappa-1}-\xi\frac{\kappa}{\kappa-1}\geq 0 &\quad\Leftrightarrow\\ 1\geq\xi. &\end{aligned}$$

Now we want to want to prove that the third term is larger than the second one in the definition of positive definiteness of $\mathbf{B}$. It is equivalent to show that

$$\begin{aligned}\frac{1}{\sigma}\tilde{\mathbf{A}}-\alpha_0\mathbf{I}_p \succeq \frac{\mu^2}{4\sigma^2(\alpha-\alpha_0)}(\tilde{\mathbf{A}}+\mathbf{I}_p)^\top(\tilde{\mathbf{A}}+\mathbf{I}_p) &\quad\Leftrightarrow\\ \forall\ \lambda=\lambda(\tilde{\mathbf{A}})\hookrightarrow \lambda-\alpha_0\sigma\geq\frac{\mu^2}{4\sigma(\alpha-\alpha_0)}(\lambda+1)^2 &\quad\Leftrightarrow\\ \lambda-\xi\mu^2\frac{\kappa}{\kappa-1}\geq\frac{\mu^2}{4\sigma(\alpha-\alpha_0)}(\lambda+1)^2. &\end{aligned}$$

We can continue the above series of inequalities

$$\begin{aligned}\lambda-\xi\mu^2\frac{\kappa}{\kappa-1}\geq\frac{\mu^2(\lambda+1)^2}{4\sigma\left(\frac{\mu^2\kappa}{\kappa-1}-\frac{\xi\mu^2}{\sigma}\frac{\kappa}{\kappa-1}\right)} &\quad\Leftrightarrow\\ \lambda-\xi\mu^2\frac{\kappa}{\kappa-1}\geq\frac{(\kappa-1)(\lambda+1)^2}{4\kappa(1-\xi)} &\quad\Leftrightarrow\\ 4\kappa\lambda(\kappa-1)(1-\xi)-4\xi(1-\xi)\kappa^2\mu^2-(\kappa-1)^2(\lambda+1)^2\geq 0 &\quad\Leftrightarrow\\ \xi^2\cdot 4\kappa^2\mu^2-\xi(4\kappa(\kappa-1)\lambda+4\mu^2\kappa^2)+(4\kappa(\kappa-1)\lambda-(\kappa-1)^2(\lambda+1)^2)\geq 0. &\end{aligned}$$

We derive a quadratic polynomial of ξ, thus, we need to check whether it has solutions. Note that the main coefficient is positive, hence, that is why wew can always choose sufficiently large ξ such that the inequality will hold for any eigenvalue λ. However, we are interested to find as tight value of ξ_0 as possible. Since $\kappa > 1$ we have that the discriminant of a quadratic polynomial is

$$\begin{aligned}
16\kappa^2((\kappa-1)\lambda+\mu^2\kappa)^2 - 4\cdot 4\kappa^2\mu^2(4\kappa(\kappa-1)\lambda-(\kappa-1)^2(\lambda+1)^2) &= \\
16\kappa^2\left[(\kappa-1)^2\lambda^2+\mu^4\kappa^2+2(\kappa-1)\lambda\mu^2\kappa-4\mu^2\kappa(\kappa-1)\lambda+\mu^2(\kappa-1)^2(\lambda+1)^2\right] &= \\
16\kappa^2\left[(\kappa-1)^2\lambda^2+\mu^4\kappa^2-2\mu^2\kappa(\kappa-1)\lambda+\mu^2(\kappa-1)^2(\lambda+1)^2\right] &= \\
16\kappa^2\left[((\kappa-1)\lambda-\mu^2\kappa)^2+\mu^2(\kappa-1)^2(\lambda+1)^2\right] &> 0.
\end{aligned}$$

That is why the quadratic polynomial has two positive solutions because the middle coefficient is also positive. If we choose $\xi < \xi_0$, where ξ_0 is the smallest solution, then the value of the quadratic polynomial will be larger than zero. We observe that if we choose

$$\begin{aligned}
\xi_0 &= \min_{\lambda}\left\{1, \frac{4\mu^2\kappa^2+4\kappa(\kappa-1)\lambda-\sqrt{16\kappa^2\left[((\kappa-1)\lambda-\mu^2\kappa)^2+\mu^2(\kappa-1)^2(\lambda+1)^2\right]}}{8\mu^2\kappa^2}\right\} \\
&= \min_{\lambda}\left\{1, \frac{1}{2}+\frac{(1-{}^1\!/\!_\kappa)\,\lambda-\sqrt{((1-{}^1\!/\!_\kappa)\lambda-\mu^2)^2+\mu^2(1-{}^1\!/\!_\kappa)^2(\lambda+1)^2}}{2\mu^2}\right\}
\end{aligned}$$

then $\mathbf{B}-\alpha_0\mathbf{M}$, where $\alpha_0 = \frac{\mu^2\xi_0}{\sigma}\frac{\kappa}{\kappa-1}$ is defined via ξ_0, is positive definite. Let us assume that $\kappa \gg 1$, then we have

$$\begin{aligned}
\frac{4\mu^2\kappa^2+4\kappa^2\lambda-\sqrt{16\kappa^2\left[(\kappa\lambda-\mu^2\kappa)^2+\mu^2\kappa^2(\lambda+1)^2\right]}}{8\mu^2\kappa^2} &= \\
\frac{4\mu^2\kappa^2+4\kappa^2\lambda-\sqrt{16\kappa^2\left[(\kappa\lambda-\mu^2\kappa)^2+\mu^2\kappa^2(\lambda+1)^2\right]}}{8\mu^2\kappa^2} &= \\
\frac{\mu^2+\lambda-\sqrt{(\lambda-\mu^2)^2+\mu^2(\lambda+1)^2}}{2\mu^2} &= \\
\frac{\mu^2+\lambda-\sqrt{\lambda^2+\mu^4+\mu^2\lambda^2+\mu^2}}{2\mu^2}&.
\end{aligned}$$

This quantity could be negative, hence, $\mathbf{P}\mathbf{A}^\top\Sigma^{-1}\mathbf{A}$ might be non-positive definite.

References

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning.

Hadrien Hendrikx, Francis Bach, and Laurent Massoulié. An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems*, 2019.

Tao Hong and Irad Yavneh. On adapting nesterov's scheme to accelerate iterative methods for linear problems. *Numerical Linear Algebra with Applications*, 2017.

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. 2019.

Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. 2021a.

Dmitry Kovalev, Egor Shulgin, Peter Richtarik, Alexander V Rogozin, and Alexander Gasnikov. Adom: Accelerated decentralized optimization method for time-varying networks. pages 5784–5793, 2021b.

Felix Lieder. On the convergence rate of the halpern-iteration. *Optim Letters 15, 405–418*, 2021.

Angelia Nedić and Alex Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 2015.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Jisun Park and Ernest K Ryu. Exact optimal accelerated complexity for fixed-point iterations. 2022.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. 2017.

Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. *arXiv preprint: arXiv 1902.00947*, 2019.

Konstantinos I. Tsianos, Sean Lawlor, and Michael G. Rabbat. Push-sum distributed dual averaging for convex optimization. 2012.

Chenguang Xi, Van Sy Mai, Ran Xin, Eyad H. Abed, and Usman A. Khan. Linear convergence in optimization over directed graphs with row-stochastic matrices. *IEEE Transactions on Automatic Control*, 2016.

Ran Xin and Usman A. Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2018.