On Decentralized Nonsmooth Optimization[*]

Savelii Chezhegov[1][0009−0003−7378−3210], Alexander
Rogozin[1][0000−0003−3435−2680], and Alexander
Gasnikov[1,2,3][0000−0002−7386−039X]

[1] Moscow Institute of Physics and Technology, Moscow, Russia
[2] Institute for Information Transportation Problems, Moscow, Russia
[3] Caucasus Mathematic Center of Adygh State University, Moscow, Russia

Abstract. In decentralized optimization, several nodes connected by a
network collaboratively minimize some objective function. For minimization of Lipschitz functions lower bounds are known along with optimal
algorithms. We study a specific class of problems: linear models with nonsmooth loss functions. Our algorithm combines regularization and dual
reformulation to get an effective optimization method with complexity
better than the lower bounds.

Keywords: convex optimization, distributed optimization

1 Introduction

The focus of this work is a particular class of problems in decentralized nonsmooth optimization. We assume that each of computational agents, or nodes,
holds a part of a common optimization problem and the agents are connected
by a network. Each node may communicate with its immediate neighbors, and
the agents aim to collaboratively solve an optimization problem.

On the class of smooth (strongly) convex functions endowed with a first-order
oracle, decentralized optimization can be called a theoretically well-developed
area of research. For this setting, [13] proposed lower bounds and optimal dual
algorithms. After that, optimal gradient methods with primal oracle were developed in [8]. Even if the network is allowed to change, lower bounds and optimal
algorithms are known and established in a series of works [9,7,11].

However, the case when local functions are non-smooth is not that well
studied. Algorithm proposed in [14] uses a gradient approximation via Gaussian smoothing. Such a technique results in additional factor of dimension. Distributed subgradient methods [12] are not optimal and only converge to a neighborhood of the solution if used with a constant step-size. In other words, development of an optimal decentralized algorithm for networks is an open research
question.

As noted below, we restrict our attention to a particular class of decentralized non-smooth optimization problems. Namely, we study linear models with non-smooth loss functions and an entropic regularizer. Problems of such type arise in traffic demands matrix calculation [1,15], optimal transport [10] and distributed training with decomposition over features [2].

Traffic problems. Following the arguments in [1], for example, one seeks to minimize $g(x)$ subject to constraints $Ax = b$. Here function $g(x)$ may be interpreted as some similarity measure between x and a supposed solution. Moving the constraint $Ax = b$ as a penalty into the objective, we obtain a problem of type

$$\min_{x \in \mathbb{R}^d} \ g(x) + \lambda \left\| Ax - b \right\|,$$

where $\|Ax - b\|$ denotes some norm. If the g represents a similarity measure given by KL divergence, we obtain an optimization problem for linear model with entropic regularizer.

Optimal transport. Another example is entropy-regularized optimal transport [10]. In paper [10] the authors show that an optimal transportation problem can be rewritten as

$$\min_{\mathbf{x} \in \Delta_n^n} \min_{\mathbf{y} \in \Delta_2^n} \ \langle \mathbf{x}, \mathbf{a} \rangle - \langle \mathbf{y}, \mathbf{b} \rangle + \langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle + \lambda_{\mathbf{x}} \langle \mathbf{x}, \log \mathbf{x} \rangle - \lambda_{\mathbf{y}} \langle \mathbf{y}, \log \mathbf{y} \rangle,$$

where Δ_n denotes denotes a unit simplex of dimension n. This illustrated that entropy-regularized linear models can arise in saddle-point optimization, as well.

Distributed ML. In distributed statistical inference and machine learning one may want to train a model in a distributed way [2]. Consider a dataset with a moderate number of training examples and a large number of features. Let the dataset be split between the nodes not by samples but by features. Let ℓ be the common loss function, and for each agent i introduce its local dataset (A_i, b_i) and the corresponding regularizer $r_i(x)$. That leads to a fitting problem

$$\min_{x_1, \ldots, x_m} \ \ell \left(\sum_{i=1}^{m} A_i x_i - b_i \right) + \sum_{i=1}^{m} r_i(x_i).$$

Our contribution. In our work we propose a dual algorithm for non-smooth decentralized optimization. The dual problem is smooth although the initial one is non-smooth, but also subject to constraints. The constraints can be equivalently rewritten as a regularizer. We show that a resulting regularized problem can be solved by an accelerated proximal primal-dual gradient method.

We study a specific class of problems, and our approach allows to break the lower bounds in [14]. Omitting problem parameters, the iteration and communication complexities of our algorithm are $O(\sqrt{1/\varepsilon})$, while lower bounds suggest that at least $\Omega(1/\varepsilon)$ communication rounds and at least $\Omega(1/\varepsilon^2)$ local computations at each node are required.

1.1   Notation

Let $\otimes$ denote the Kronecker product. Let $\Delta_d = \{x \in \mathbb{R}^d : \sum_{i=1}^m x_i = 1, \ x_i \geq 0, \ i = 1, \ldots, m\}$ be a unit simplex in $\mathbb{R}^d$. By Δ_d^m we understand a product of m simplices that is a set in $\mathbb{R}^{md}$. For $p \geq 0$, let $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$ denote the p-norm in $\mathbb{R}^d$. By $\langle a, b \rangle$ we denote a scalar product of vectors. Also let $\mathbf{x} = \mathrm{col}[x_1, \ldots, x_m] = (x_1^\top \ldots x_m^\top)^\top \in \mathbb{R}^{md}$ denote a column vector stacked of $x_1, \ldots, x_m \in \mathbb{R}^d$. Similarly for matrices $C_1, \ldots, C_m \in \mathbb{R}^{n \times d}$, introduce $\mathrm{col}[C_1, \ldots, C_m] = (C_1^\top \ldots C_m^\top)^\top \in \mathbb{R}^{mn \times d}$. Moreover, let $\mathrm{diag}[C_1, \ldots, C_m]$ denote a block matrix with blocks $C_1, \ldots, C_m$ at the diagonal. For $x \in \mathbb{R}^d$, let $\log x$ denote a natural logarithm function applied component-wise. We define $\mathbf{1}_n$ to be a vector of all ones of length n and $\mathbf{I}_n$ to be an identity matrix of size $n \times n$. Also denote the i-th coordinate vector of $\mathbb{R}^n$ as $e_i^{(n)}$. Let $\lambda_{\max}(C)$ and $\lambda_{\min}^+(C)$ denote the maximal and minimal nonzero eigenvalue of matrix C. Let $\sigma_{\max}(C)$ and $\sigma_{\min}^+(C)$ denote the maximal and minimal nonzero singular values of C.

Given a convex closed set Q, let Π_Q denote a projection operator on it and denote its interior $\mathrm{int}\, Q$. For a closed proper function $h(x) : Q \to \mathbb{R}$ and a scalar $\gamma > 0$, define proximal operator as

$$\mathrm{prox}_{\gamma h}(x) = \arg\min_{y \in Q} \left(h(x) + \frac{1}{2\gamma} \|y - x\|_2^2 \right).$$

2   Problem and assumptions

Consider m independent computational entities, or agents. Agent i locally holds a dataset consisting of matrix A_i and labels b_i. Let $A = \mathrm{col}[A_1, \ldots, A_m] \in \mathbb{R}^{n \times d}$ be the training samples and $\mathbf{b} = \mathrm{col}[b_1, \ldots, b_m]$ be the labels. The whole training dataset $(A, \mathbf{b})$ is distributed between m different machines. We consider p-norm minimization over unit simplex with entropy regularizer.

$$\min_{x \in \Delta_d} \frac{1}{m} \|Ax - \mathbf{b}\|_p + \theta \langle x, \log x \rangle, \tag{1}$$

where $\theta > 0$ is a regularization parameter.

The agents can communicate information through a communication network. We assume that each machine is a node in the network that is represented by a connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The nodes can communicate if and only if they are connected by an edge.

With graph $\mathcal{G}$ we associate a communication matrix W that has the following properties.

Assumption 1
1. (Network compatibility) $[W]_{ij} = 0$ if $i \neq j$ and $(i, j) \notin \mathcal{E}$.
2. (Positive semi-definiteness and symmetry) $W \succeq 0$, $W^\top = W$.
3. (Kernel property) $W x = 0$ if and only if $x_1 = \ldots = x_m$.

We also introduce the condition number of the communication matrix.

$$\chi = \frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)}. \tag{2}$$

In order to get a distributed formulation, assign each agent i in the network a local copy of the solution vector x_i. Define $\mathbf{x} = \text{col}[x_1, \ldots, x_m]$, $\mathbf{A} = \text{diag}[A_1, \ldots, A_m]$ and introduce $\mathbf{y} = \mathbf{A}\mathbf{x}$.

$$\min_{\mathbf{x} \in \Delta_m^d} \|\mathbf{y} - \mathbf{b}\|_p + \theta \langle \mathbf{x}, \log \mathbf{x} \rangle \tag{3}$$

$$\text{s.t. } \mathbf{W}\mathbf{x} = 0$$

$$\mathbf{y} = \mathbf{A}\mathbf{x}$$

The complexity of distributed methods typically depends on the condition number of the communication matrix (it is χ defined in (2)) and on condition numbers of objective functions. For brevity we introduce

$$\sigma_{\max}(\mathcal{A}) = \max_{i=1,\ldots,m} \left(\sigma_{\max}(A_i) \right), \ \sigma_{\min}^+(\mathcal{A}) = \min_{i=1,\ldots,m} \sigma_{\min}^+(A_i). \tag{4}$$

3   Dual problem

Let us derive a dual problem to (3). It is convenient to introduce $F(\mathbf{y}) = \|\mathbf{y} - \mathbf{b}\|_p$, $G(\mathbf{x}) = \theta \langle \mathbf{x}, \log \mathbf{x} \rangle$.

3.1   Conjugate functions

Let us derive the conjugate functions F^* and G^*. Let $q \geq 1$ be such that $\frac{1}{p} + \frac{1}{q} = 1$.

$$F^*(\mathbf{t}) = \sup_{\mathbf{y} \in \mathbb{R}^{mn}} \left(\langle \mathbf{t}, \mathbf{y} \rangle - F(\mathbf{y}) \right) = \sup_{\mathbf{y} \in \mathbb{R}^{mn}} \left(\langle \mathbf{t}, \mathbf{y} - \mathbf{b} \rangle - \|\mathbf{y} - \mathbf{b}\|_p \right) + \langle \mathbf{t}, \mathbf{b} \rangle$$

$$= \sup_{\mathbf{r} \in \mathbb{R}^{mn}} \left(\langle \mathbf{t}, \mathbf{r} \rangle - \|\mathbf{r}\|_p \right) + \langle \mathbf{t}, \mathbf{b} \rangle$$

$$= \begin{cases} \langle \mathbf{t}, \mathbf{b} \rangle, & \|\mathbf{t}\|_q \leq 1 \\ +\infty, & \text{otherwise} \end{cases}$$

Last equation is a result of conjugate function for $\|x\|_p$, which is taken from a classical book by Boyd [3], Chapter 5.

In order to compute G^*, introduce $g(x) = \theta \langle x, \log x \rangle : \ \mathbb{R}^d \to \mathbb{R}$ and note that $G(\mathbf{x}) = \sum_{i=1}^m g_i(x_i)$.

$$g^*(t) = \sup_{x \in \Delta_d} \left(\langle t, x \rangle - \theta \langle x, \log(x) \rangle \right)$$

Writing a Lagrange function:

$$L(t, x) = \langle t, x \rangle - \theta \langle x, \log(x) \rangle + \lambda \left(\mathbf{1}_d^\top x - 1 \right)$$

$$\nabla_x L(t, x) = t - \theta \log x - \theta \mathbf{1}_d + \lambda \mathbf{1}_d = 0 \Rightarrow x = \exp \left(\frac{t}{\theta} + \mathbf{1}_d \left(\frac{\lambda}{\theta} - 1 \right) \right)$$

$$\mathbf{1}_d^\top x = 1 \Rightarrow \exp \left(\frac{\lambda}{\theta} - 1 \right) \mathbf{1}_d^\top \exp \left(\frac{t}{\theta} \right) = 1 \Rightarrow \exp \left(\frac{\lambda}{\theta} - 1 \right) = \frac{1}{\mathbf{1}_d^\top \exp \left(\frac{t}{\theta} \right)}$$

As a consequence

$$x = \frac{\exp \left(\frac{t}{\theta} \right)}{\mathbf{1}_d^\top \exp \left(\frac{t}{\theta} \right)}$$

Using equation to x,

$$g^*(t) = \theta \log \left(\mathbf{1}_d^\top \exp \left(\frac{t}{\theta} \right) \right)$$

As noted above, $G(\mathbf{x})$ is separable, i.e. $G(\mathbf{x}) = \sum_{i=1}^m g(x_i)$. Therefore,

$$G^*(\mathbf{t}) = \sup_{\mathbf{x} \in \Delta_d^m} \left(\langle \mathbf{t}, \mathbf{x} \rangle - \sum_{i=1}^m g(x_i) \right) = \sum_{i=1}^m \sup_{x \in \Delta_d} \left(\langle t_i, x \rangle - g(x) \right) = \sum_{i=1}^m g^*(t_i).$$

It is convenient to express t_i through $\mathbf{t}$. Introduce matrix

$$\mathbf{E}_i = \left(e_i^{(m)} \right)^\top \otimes \mathbf{I} = [0 \ldots 0 \ \mathbf{I} \ 0 \ldots 0]. \tag{5}$$

Then $t_i = \mathbf{E}_i \mathbf{t}$. It holds

$$G^*(\mathbf{t}) = \sum_{i=1}^m g^*(\mathbf{E}_i \mathbf{t}).$$

3.2   Dual problem formulation

Let us derive a dual problem to (3). It is convenient to denote $F(\mathbf{y}) = \|\mathbf{y} - \mathbf{b}\|_p$, $G(\mathbf{x}) = \theta \langle \mathbf{x}, \log \mathbf{x} \rangle$. Introduce dual function

$$\begin{aligned}
\Phi(\mathbf{z}, \mathbf{s}) &= \inf_{\mathbf{x} \in \Delta_d^m, \mathbf{y} \in \mathbb{R}^n} \left[F(\mathbf{y}) + G(\mathbf{x}) + \langle \mathbf{z}, \mathbf{W}\mathbf{x} \rangle + \langle \mathbf{s}, \mathbf{A}\mathbf{x} - \mathbf{y} \rangle \right] \\
&= \inf_{\mathbf{y} \in \mathbb{R}^{mn}} \left[F(\mathbf{y}) - \langle \mathbf{s}, \mathbf{y} \rangle \right] + \inf_{\mathbf{x} \in \Delta_d^m} \left[G(\mathbf{x}) + \langle \mathbf{W}\mathbf{z} + \mathbf{A}^\top \mathbf{s}, \mathbf{x} \rangle \right] \\
&= - \sup_{\mathbf{y} \in \mathbb{R}^{mn}} \left[\langle \mathbf{s}, \mathbf{y} \rangle - F(\mathbf{y}) \right] - \sup_{\mathbf{x} \in \Delta_d^m} \left[\langle -\mathbf{W}\mathbf{z} - \mathbf{A}^\top \mathbf{s}, \mathbf{x} \rangle - G(\mathbf{x}) \right] \\
&= -F^*(\mathbf{s}) - G^*(-\mathbf{W}\mathbf{z} - \mathbf{A}^\top \mathbf{s})
\end{aligned}$$

As a consequence, dual problem can be formulated as

$$\min_{\mathbf{z} \in \mathbb{R}^{md}, \mathbf{s} \in \mathbb{R}^{mn}} F^*(\mathbf{s}) + G^*(-\mathbf{W}\mathbf{z} - \mathbf{A}^\top \mathbf{s}).$$

Results from 3.2 and 3.1 leads us to final dual problem formulation

$$\min_{\mathbf{z},\mathbf{s}:\|\mathbf{s}\|_q\leq 1} \langle\mathbf{s},\mathbf{b}\rangle + \sum_{i=1}^{m}\theta\log\left(\mathbf{1}_d^\top\exp\left(-\frac{1}{\theta}\mathbf{E}_i\left(\mathbf{Wz}+\mathbf{A}^\top\mathbf{s}\right)\right)\right) \qquad (6)$$

The constrained problem above is equivalent to a regularized problem

$$\min_{\mathbf{z},\mathbf{s}} \ \langle\mathbf{s},\mathbf{b}\rangle + \sum_{i=1}^{m}\theta\log\left(\mathbf{1}_d^\top\exp\left(-\frac{1}{\theta}\mathbf{E}_i(\mathbf{Wz}+\mathbf{A}^\top\mathbf{s})\right)\right) + \nu\|\mathbf{s}\|_q^q, \qquad (7)$$

where $\nu > 0$ is a scalar.
As a result, the dual problem writes as

$$\min_{\mathbf{q}} \ H(\mathbf{z},\mathbf{s}) + R(\mathbf{z},\mathbf{s}) \qquad (8)$$

$$H(\mathbf{z},\mathbf{s}) = \langle\mathbf{s},\mathbf{b}\rangle + \sum_{i=1}^{m}\theta\log\left(\mathbf{1}_d^\top\exp\left(-\frac{1}{\theta}\mathbf{E}_i(\mathbf{Wz}+\mathbf{A}^\top\mathbf{s})\right)\right)$$

$$R(\mathbf{z},\mathbf{s}) = \nu\|\mathbf{s}\|_q^q\,.$$

Recall problem (8) and denote $\mathbf{B} = (-\mathbf{W} \ -\mathbf{A}^\top)$, $\mathbf{q} = \mathrm{col}[\mathbf{z},\mathbf{s}]$, $\mathbf{p} = \mathrm{col}[0,\mathbf{b}]$. With slight abuse of notation we write $H(\mathbf{q}) = H(\mathbf{z},\mathbf{s})$ and $R(\mathbf{q}) = R(\mathbf{z},\mathbf{s})$. Problem (8) takes the form

$$\min_{\mathbf{q}} \ H(\mathbf{q}) + R(\mathbf{q}).$$

Here H is a differentiable function and R is a regularizer, or composite term. Problems of such type are typically solved by proximal optimization methods.

4   Algorithms and Complexities

4.1   Similar Triangles Method

We apply an accelerated primal-dual algorithm called Similar Triangles Method (STM) [4].

Algorithm 1 Similar triangles method(STM)

Require: $A_0 = \alpha_0 = 0,\ \mathbf{q}^0 = \mathbf{u}^0 = \mathbf{y}^0$.
1: **for** $k = 0, \ldots, N-1$ **do**
2:    Find α_{k+1} from equality $(A_k + \alpha_k)(1 + A_k\mu) = L\alpha_{k+1}$ and put $A_{k+1} = A_k + \alpha_{k+1}$.

3:    Introduce

$$\phi_{k+1}(\mathbf{x}) = \alpha_{k+1}\left(\left\langle \nabla H(\mathbf{y}^{k+1}), \mathbf{x}\right\rangle + R(\mathbf{x})\right) + \frac{1+A_k\mu}{2}\left\|\mathbf{x} - \mathbf{u}^k\right\|_2^2 + \frac{\mu\alpha_{k+1}}{2}\left\|\mathbf{x} - \mathbf{y}^{k+1}\right\|_2^2$$

4:    $\mathbf{y}^{k+1} = \frac{\alpha_{k+1}\mathbf{u}^k + A_k\mathbf{q}^k}{A_{k+1}}$
5:    $\mathbf{u}^{k+1} = \arg\min_{\mathbf{q}}[\phi_{k+1}(\mathbf{x})]$
6:    $\mathbf{q}^{k+1} = \frac{\alpha^{k+1}\mathbf{u}^{k+1} + A_k\mathbf{q}^k}{A_{k+1}}$
7: **end for**

First, note that line 5 of Algorithm 1 can be decomposed into a gradient step and computation of proximal operator of R.

$$\mathbf{u}^{k+1} = \arg\min_{\mathbf{q}}[\phi_{k+1}(\mathbf{x})] = \mathrm{prox}_{\gamma_k R}\left[\mu\gamma_k\mathbf{y}^{k+1} + (1 - \mu\gamma_k)\mathbf{u}^k - \gamma_k\nabla F(\mathbf{y}^{k+1})\right],$$

where $\gamma_k = \frac{\alpha_{k+1}}{1+\mu A_{k+1}}$. Let us show that this operator can be easily computed. Let $\mathbf{t} = \mathrm{col}[\mathbf{t_z}, \mathbf{t_s}]$. By definition of proximal operator we have

$$\mathrm{prox}_{\gamma_k R}(\mathbf{t}) = \arg\min_{\mathbf{s}}\left(\frac{1}{2\gamma_k}\|\mathbf{t} - \mathbf{s}\|_2^2 + R(\mathbf{q})\right)$$

$$= \arg\min_{\mathbf{z},\mathbf{s}}\left(\frac{1}{2\gamma_k}(\|\mathbf{t_s} - \mathbf{s}\|_2^2 + \|\mathbf{t_z} - \mathbf{z}\|_2^2) + \nu\|\mathbf{s}\|_q^q\right).$$

Let $\tilde{\mathbf{q}} = \mathrm{col}[\tilde{\mathbf{z}}, \tilde{\mathbf{s}}] = \mathrm{prox}_{\gamma_k R}(\mathbf{t})$. We have $\tilde{\mathbf{z}} = \mathbf{t_z}$. Let $\tilde{s}_i$ denote the i-th component of $\tilde{\mathbf{s}}$ and t_i denote the i-th component of $\mathbf{t_s}$; then $\tilde{s}_i$ can be found from equation

$$t_i - \tilde{s}_i + \gamma_k q\nu|\tilde{s}_i|^{q-1} = 0.$$

We assume that the equation above can be efficiently numerically solved w.r.t. $\tilde{s}_i$. For example, it can be done by solution localization methods such as binary search. As a result, we see that the proximal operator of R can be computed cheaply.

Let us formulate the theorem on convergence of Algorithm 1 for the problem (8).

Theorem 2. *Algorithm 1 requires*

$$O\left(\left(\frac{\theta m\left(\sigma_{\max}^2(\mathcal{A}) + \sigma_{\max}^2(W)\right)\|\log x^* + \mathbf{1}_d\|_2^2}{\min((\sigma_{min}^+(\mathcal{A}))^2, (\lambda_{min}^+(W))^2)\varepsilon}\right)^{1/2}\right)$$

iterations to reach ε-accuracy for the problem (8)

Before we prove the above result, we need to formulate some lemmas.
We need to find Lipschitz constant for dual problem. Namely, let us find the
Lipschitz constant for function $G^*(-\mathbf{W}\mathbf{z} - \mathbf{A}^\top \mathbf{s})$ as a function of $\mathbf{q} = \mathrm{col}[\mathbf{z}, \mathbf{s}]$.

Lemma 1. *Function $H(\mathbf{q})$ has a Lipschitz gradient with constant*

$$L_H = \frac{m\left(\sigma_{\max}^2(\mathcal{A}) + \sigma_{\max}^2(W)\right)}{\theta}$$

Proof. According to [6], if a function is μ-strongly convex in norm $\|\cdot\|_2$, then its
conjugate function $h^*(y)$ has a $\frac{1}{\mu}$-Lipschitz gradient in $\|\cdot\|_2$.
Using the fact from [3], Chapter 3, we obtain that the conjugate function of

$$h(x) = \log\left(\sum_{i=1}^d \exp(x_i)\right)$$

is

$$h^*(y) = \begin{cases} \langle y, \log y \rangle, & y \in \Delta_d \\ \infty, & \text{otherwise} \end{cases}$$

To have a constant of strongly convexity, we can find a minimal eigenvalue of
Hessian of $h^*(y)$

$$\nabla^2 h^*(y) = \mathrm{diag}\left(\frac{1}{y_1}, \ldots, \frac{1}{y_d}\right).$$

For any $y \in \mathrm{int}\Delta_d$, we have that $1/y_i \geq 1$, $i = 1, \ldots, d$. Therefore, we have
$\lambda_{\min}(\nabla^2 h^*(y)) \geq 1$, i.e. $\mu_{h^*} \geq 1$.
As a consequence, for function

$$h(x) = \log\left(\sum_{i=1}^d \exp(x_i)\right)$$

Lipschitz constant is equal to $L_h = 1$.
Therefore, for a function

$$g^*(x) = \theta h\left(\frac{x}{\theta}\right)$$

Lipschitz constant is equal to $L_g = 1/\theta$.
Introduce $\mathbf{B} = \left(-\mathbf{W}, -\mathbf{A}^\top\right)$. We have

$$H(\mathbf{q}) = G^*(\mathbf{B}\mathbf{q}) + \langle \mathbf{s}, \mathbf{b} \rangle = \sum_{i=1}^m g^*(\mathbf{E}_i \mathbf{B}\mathbf{q}) + \langle \mathbf{s}, \mathbf{b} \rangle.$$

It holds

$$\|\nabla H(\mathbf{q}_2) - \nabla H(\mathbf{q}_2)\|_2 = \|\mathbf{B}^\top \nabla G^*(\mathbf{Bq}_2) - \mathbf{B}^\top \nabla G^*(\mathbf{Bq}_2)\|_2$$

$$\leq \sigma_{\max}(\mathbf{B}) \|\nabla G^*(\mathbf{Bq}_2) - \nabla G^*(\mathbf{Bq}_1)\| \leq \sigma_{\max}(\mathbf{B}) \sum_{i=1}^{m} \|\nabla g^*(\mathbf{E}_i \mathbf{Bq}_2) - \nabla g^*(\mathbf{E}_i \mathbf{Bq}_1)\|_2$$

$$\leq \sigma_{\max}(\mathbf{B}) \sum_{i=1}^{m} \frac{\sigma_{\max}(\mathbf{E}_i \mathbf{B})}{\theta} \|\mathbf{q}_2 - \mathbf{q}_1\|_2 \leq \frac{m\sigma_{\max}^2(\mathbf{B})}{\theta} \|\mathbf{q}_2 - \mathbf{q}_1\|_2$$

$$= \frac{m(\sigma_{\max}^2(\mathbf{A}) + \sigma_{\max}^2(\mathbf{W}))}{\theta} \|\mathbf{q}_2 - \mathbf{q}_1\|_2$$

$$\leq \frac{m\left(\max\limits_{i=1,\ldots,m}(\sigma_{\max}^2(A_1),\ldots,\sigma_{\max}^2(A_m)) + \sigma_{\max}^2(W)\right)}{\theta} \|\mathbf{q}_2 - \mathbf{q}_1\|_2$$

$$= L_H \|\mathbf{q}_2 - \mathbf{q}_1\|_2\,,$$

which finishes the proof of lemma.

For writing a complexity of solver for our problem, we also need to bound the dual distance.

Lemma 2. *Let* $\mathbf{q}^* = \mathrm{col}[\mathbf{z}^*, \mathbf{s}^*]$ *be the solution of dual problem* (8) *and let* x^* *be a solution of* (1). *It holds*

$$\|\mathbf{q}^*\|_2^2 \leq R_{dual}^2 = \frac{\theta^2 m \|\log x^* + \mathbf{1}_d\|_2^2}{\min((\sigma_{min}^+(\mathcal{A}))^2, (\lambda_{min}^+(W))^2)}.$$

Proof. Let $(\mathbf{x}^*, \mathbf{y}^*)$ be a solution to primal problem (3). In particular, we have $\mathbf{x}^* = \mathbf{1}_m \otimes x^*$. Then $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*, \mathbf{s}^*)$ is a saddle point of Lagrange function. For any $\mathbf{x} \in \Delta_m^d$, $\mathbf{y} \in \mathbb{R}^{md}$, $\mathbf{z} \in \mathbb{R}^{md}, \mathbf{s} \in \mathbb{R}^{mn}$ it holds

$$F(\mathbf{y}^*) + G(\mathbf{x}^*) + \langle \mathbf{z}, \mathbf{Wx}^* \rangle + \langle \mathbf{s}, \mathbf{Ax}^* - \mathbf{y}^* \rangle$$
$$\leq F(\mathbf{y}^*) + G(\mathbf{x}^*) + \langle \mathbf{z}^*, \mathbf{Wx}^* \rangle + \langle \mathbf{s}^*, \mathbf{Ax}^* - \mathbf{y}^* \rangle$$
$$\leq F(\mathbf{y}) + G(\mathbf{x}) + \langle \mathbf{z}^*, \mathbf{Wx} \rangle + \langle \mathbf{s}^*, \mathbf{Ax} - \mathbf{y} \rangle$$

Substituting $\mathbf{y} = \mathbf{y}^*$ we obtain

$$G(\mathbf{x}) \geq G(\mathbf{x}^*) + \langle -\mathbf{Wz}^* - \mathbf{A}^\top \mathbf{s}^*, \mathbf{x} - \mathbf{x}^* \rangle$$
$$-\mathbf{Wz}^* - \mathbf{A}^\top \mathbf{s}^* = \nabla G(\mathbf{x}^*)$$

Recalling that $\mathbf{B} = (-\mathbf{W}, -\mathbf{A}^\top)$ we derive

$$\mathbf{Bq}^* = \nabla G(\mathbf{x}^*)$$
$$\langle \mathbf{B}^\top \mathbf{Bq}^*, \mathbf{q}^* \rangle = \|\nabla G(\mathbf{x}^*)\|_2^2$$
$$\lambda_{min}^+(\mathbf{B}^\top \mathbf{B}) \|\mathbf{q}^*\|_2^2 \leq \|\nabla G(\mathbf{x}^*)\|_2^2$$
$$\|\mathbf{q}^*\|_2^2 \leq \frac{\|\nabla G(\mathbf{x}^*)\|_2^2}{\lambda_{min}^+(\mathbf{B}^\top \mathbf{B})}$$

We have

$$\lambda_{\min}^{+}(\mathbf{B}^{\top}\mathbf{B}) = \lambda_{\min}^{+}(\mathbf{B}\mathbf{B}^{\top}) = \lambda_{\min}^{+}(\mathbf{W}^2 + \mathbf{A}^{\top}\mathbf{A}) = \lambda_{\min}^{+}(\mathbf{W}^2 \otimes \mathbf{I}_d + \mathbf{I}_m \otimes \mathbf{A}^{\top}\mathbf{A})$$
$$= \min((\lambda_{\min}^{+}(\mathbf{W}))^2, (\sigma_{\min}^{+}(\mathbf{A}))^2) = \min\left((\lambda_{\min}^{+}(W))^2, (\sigma_{\min}^{+}(\mathcal{A}))^2\right)$$

and

$$\|\nabla G(\mathbf{x}^*)\|_2^2 = \theta^2 \|\log \mathbf{x}^* + \mathbf{1}_{md}\|_2^2 = \theta^2 m \|\log x^* + \mathbf{1}_d\|_2^2.$$

As a result, we obtain

$$R_{dual}^2 = \frac{\theta^2 m \|\log x^* + \mathbf{1}_d\|_2^2}{\min((\sigma_{min}^{+}(\mathcal{A}))^2, (\lambda_{min}^{+}(W))^2)}.$$

Now we prove the theorem about complexity of Similar Triangles Method.

Proof of Theorem 2

Proof. First, note that solution accuracy ε for problem (1) is equivalent to accuracy $m\varepsilon$ for problem (8). STM requires $O((L_H R_{dual}^2/(m\varepsilon))^{1/2})$ iterations to reach ε-accuracy. Combining the results from lemmas 1 and 2 we obtain the final complexity.

4.2   Accelerated block-coordinate method

In previous section, our approach was based on a way where we apply a first-order method without separating the variables. But we can treat variable blocks $\mathbf{z}$ and $\mathbf{s}$ separately and get a better convergence bound. We apply an accelerated method ACRCD (Accelerated by Coupling Randomized Coordinate Descent) from [5]. We describe the result only for the case $p = 1$. In this case, we apply ACRCD not to regularized dual problem (8), but to constrained version of dual problem (6). We also note that ACRCD is primal-dual, so solving the dual problem with accuracy ε is sufficient to restore the solution of the primal with accuracy ε.

Algorithm 2 ACRCD

Require: Define coefficients $\alpha_{k+1} = \frac{k+2}{8}$, $\tau_k = \frac{2}{k+2}$. Choose stepsizes $L_{\mathbf{z}}$, $L_{\mathbf{s}}$. Put $\overline{\mathbf{z}}^0 = \underline{\mathbf{z}}^0 = \mathbf{z}^0$, $\overline{\mathbf{s}}^0 = \underline{\mathbf{s}}^0 = \mathbf{s}^0$.

1: **for** $k = 0, 1, \ldots, N-1$ **do**
2:     $\mathbf{z}^{k+1} = \tau_k \underline{\mathbf{z}}^k + (1 - \tau_k)\overline{\mathbf{z}}^k$
3:     $\mathbf{s}^{k+1} = \tau_k \underline{\mathbf{s}}^k + (1 - \tau_k)\overline{\mathbf{s}}^k$
4:     Put $\xi_i = 1$ with probability η and $\xi = 0$ with probability $(1 - \eta)$, where $\eta = \frac{\lambda_{\max}(W)}{\lambda_{\max}(W) + \sigma_{\max}(\mathcal{A})}$
5:     **if** $\xi_i = 1$ **then**
6:         $\overline{\mathbf{z}}^{k+1} = \mathbf{z}^{k+1} - \frac{1}{L_{\mathbf{z}}}\nabla H_{\mathbf{z}}(\mathbf{z}^{k+1}, \mathbf{s}^{k+1})$
7:         $\underline{\mathbf{z}}^{k+1} = \underline{\mathbf{z}}^k - \frac{2\alpha_{k+1}}{L_{\mathbf{z}}}\nabla H_{\mathbf{z}}(\mathbf{z}^{k+1}, \mathbf{s}^{k+1})$
8:     **else**
9:         $\overline{\mathbf{s}}^{k+1} = \Pi_{[-1,1]^{mn}}\left[\mathbf{s}^{k+1} - \frac{1}{L_{\mathbf{s}}}\nabla H_{\mathbf{s}}(\mathbf{z}^{k+1}, \mathbf{s}^{k+1})\right]$
10:        $\underline{\mathbf{s}}^{k+1} = \Pi_{[-1,1]^{mn}}\left[\underline{\mathbf{s}}^k - \frac{2\alpha_{k+1}}{L_{\mathbf{s}}}\nabla H_{\mathbf{s}}(\mathbf{z}^{k+1}, \mathbf{s}^{k+1})\right]$
11:    **end if**
12: **end for**

Theorem 3. *To reach accuracy ε with probability at least $(1 - \delta)$, Algorithm 2 requires N_{comm} communication rounds and N_{comp} local computations, where*

$$N_{comm} = \frac{m^{1/4}}{\sqrt{\theta}\varepsilon} \frac{\lambda_{\max}(W)}{\lambda_{\min}^+(W)} \left(2\theta^2 \|\log \mathbf{x}^* + \mathbf{1}_d\|_2^2 + 2n\sigma_{\max}^2(\mathcal{A}) + n(\lambda_{\min}^+(W))^2\right)^{1/2} \log\left(\frac{1}{\delta}\right),$$

$$N_{comp} = \frac{m^{1/4}}{\sqrt{\theta}\varepsilon} \frac{\sigma_{\max}(\mathcal{A})}{\lambda_{\min}^+(W)} \left(2\theta^2 \|\log \mathbf{x}^* + \mathbf{1}_d\|_2^2 + 2n\sigma_{\max}^2(\mathcal{A}) + n(\lambda_{\min}^+(W))^2\right)^{1/2} \log\left(\frac{1}{\delta}\right).$$

First, we need to estimate Lipschitz constants for gradients of each block of variables. If we consider the function H as a function of two blocks of variables, the next result follows.

Lemma 3. *Function $H(\mathbf{z}, \mathbf{s})$ has a $L_{\mathbf{z}}$-Lipschitz gradient w.r.t. $\mathbf{z}$ and $L_{\mathbf{s}}$-Lipschitz gradient w.r.t. $\mathbf{s}$, where*

$$L_{\mathbf{z}} = \frac{\sqrt{m}\sigma_{\max}^2(W)}{\theta}, \;\; L_{\mathbf{s}} = \frac{\sqrt{m}\sigma_{\max}^2(\mathcal{A})}{\theta}.$$

Proof. Recall that we denoted $\mathbf{x} = \text{col}[x_1, \ldots, x_m]$ and consider $\mathbf{s}_1, \mathbf{s}_2 \in \mathbb{R}^{nm}$. Also denote $[\mathbf{x}]_i = E_i \mathbf{x} = x_i$.

$$
\begin{aligned}
\|\nabla_{\mathbf{s}} & H(\mathbf{z}, \mathbf{s}_2) - \nabla_{\mathbf{s}} H(\mathbf{z}, \mathbf{s}_1)\|_2 \\
&= \|\mathbf{A}\nabla G^*(-\mathbf{W}\mathbf{z} - \mathbf{A}^\top \mathbf{s}_2) - \mathbf{A}\nabla G^*(-\mathbf{W}\mathbf{z} - \mathbf{A}^\top \mathbf{s}_1)\|_2 \\
&\leq \sum_{i=1}^m \|A_i \nabla g^* \left(-[\mathbf{W}\mathbf{z}]_i - [\mathbf{A}^\top \mathbf{s}_2]_i\right) - A_i \nabla g^* \left(-[\mathbf{W}\mathbf{z}]_i - [\mathbf{A}^\top \mathbf{s}_1]_i\right)\|_2 \\
&\overset{①}{=} \sum_{i=1}^m \|A_i \nabla g^* \left(-[\mathbf{W}\mathbf{z}]_i - A_i^\top [\mathbf{s}_2]_i\right) - A_i \nabla g^* \left(-[\mathbf{W}\mathbf{z}]_i - A_i^\top [\mathbf{s}_1]_i\right)\|_2 \\
&\leq \frac{\sigma_{\max}(\mathcal{A})}{\theta} \sum_{i=1}^m \|A_i^\top [\mathbf{s}_2]_i - A_i^\top [\mathbf{s}_1]_i\|_2 \leq \frac{\sigma_{\max}^2(\mathcal{A})}{\theta} \sum_{i=1}^m \|[\mathbf{s}_2]_i - [\mathbf{s}_1]_i\|_2 \\
&\overset{②}{\leq} \frac{\sqrt{m}\sigma_{\max}^2(\mathcal{A})}{\theta} \|\mathbf{s}_2 - \mathbf{s}_1\|_2,
\end{aligned}
$$

where ① holds due to the structure of $\mathbf{A} = \text{diag}[A_1, \ldots, A_m]$ and ② holds by convexity of the 2-norm.

Now consider the gradient w.r.t. $\mathbf{z}$. Let $[\mathbf{x}]^{(i)} = [x_1^{(i)} \ldots x_m^{(i)}]^\top$ denote a vector consisting of i-th components of $x_1, \ldots, x_m$. We have $[\mathbf{W}\mathbf{x}]_i = W[\mathbf{x}]^{(i)}$ due to the structure of $\mathbf{W} = W \otimes \mathbf{I}_d$.

$$
\begin{aligned}
\|\nabla H_{\mathbf{z}}^* & (\mathbf{z}_2, \mathbf{s}) - \nabla H_{\mathbf{z}}^*(\mathbf{z}_1, \mathbf{s})\|_2 \\
&= \|\mathbf{W}\nabla G^*(-\mathbf{W}\mathbf{z}_2 - \mathbf{A}^\top \mathbf{s}) - \mathbf{W}\nabla G^*(-\mathbf{W}\mathbf{z}_1 - \mathbf{A}^\top \mathbf{s})\|_2 \\
&\leq \sum_{i=1}^m \|W\nabla g^* \left(-W[\mathbf{z}_2]^{(i)} - [\mathbf{A}^\top \mathbf{s}]_i\right) - W\nabla g^* \left(-W[\mathbf{z}_1]^{(i)} - [\mathbf{A}^\top \mathbf{s}]_i\right)\|_2 \\
&\leq \frac{\lambda_{\max}(W)}{\theta} \sum_{i=1}^m \|W([\mathbf{z}_2]^{(i)} - [\mathbf{z}_1]^{(i)})\|_2 \leq \frac{\lambda_{\max}^2(W)}{\theta} \sum_{i=1}^m \|[\mathbf{z}_2]^{(i)} - [\mathbf{z}_1]^{(i)}\|_2 \\
&\overset{①}{\leq} \frac{\sqrt{m}\lambda_{\max}^2(W)}{\theta} \|\mathbf{z}_2 - \mathbf{z}_1\|_2,
\end{aligned}
$$

where ① holds by convexity of the 2-norm.

We need to bound dual distance for each block of variables, but first we need to claim an useful proposition from functional analysis.

Proposition 1. *Let $p > r \geq 1, x \in \mathbb{R}^d$. It holds*

$$
\|x\|_p \leq \|x\|_r \leq d^{\left(\frac{1}{r} - \frac{1}{p}\right)} \|x\|_p
$$

Proof. This is a fairly well-known fact with a simple idea of proof. In fact, it is a direct consequence of Hólder's inequality, what means that constant in an inequality are unimprovable.

Now we derive the bound on the norm of the dual solution. The convergence result only relies on the case $p = 1$ ($q = \infty$), but we derive a bound for any $q \geq 1$.

Lemma 4. *Let* $\mathbf{z}^*, \mathbf{s}^*$ *be the solutions of dual problem* (8) *and let* x^* *be a solution of* (1). *It holds*

$$\|\mathbf{z}^*\|_2^2 \le R_{\mathbf{z}}^2 = \frac{2\theta^2 m\|\log x^* + \mathbf{1}_d\|_2^2 + 2\sigma_{\max}^2(\mathcal{A}) \cdot \max\left(1, (mn)^{\left(1 - \frac{2}{q}\right)}\right)}{(\lambda_{min}^+(W))^2}$$

$$\|\mathbf{s}^*\|_2^2 \le R_{\mathbf{s}}^2 = \max\left(1, (mn)^{\left(1 - \frac{2}{q}\right)}\right)$$

Proof. Using that the problems 6 and 7 are equal, that means

$$\|\mathbf{s}^*\|_q \le 1 \tag{9}$$

Using Proposition 1 we have

$$\|\mathbf{s}^*\|_2^2 \le \begin{cases} \|\mathbf{s}^*\|_q^2, & q < 2 \\ (mn)^{\left(1 - \frac{2}{q}\right)} \|\mathbf{s}^*\|_q^2, & q \ge 2 \end{cases} \tag{10}$$

Combininq 9 and 10, we have

$$\|\mathbf{s}^*\|_2^2 \le \begin{cases} 1, & q < 2 \\ (mn)^{\left(1 - \frac{2}{q}\right)}, & q \ge 2 \end{cases} \tag{11}$$

With the fact that $(mn)^{\left(1 - \frac{2}{q}\right)} < 1$ where $q < 2$ we state the claimed result. Using the fact from proof of Lemma 2 such that

$$- \mathbf{W}\mathbf{z}^* - \mathbf{A}^\top \mathbf{s}^* = \nabla G(\mathbf{x}^*)$$

we have

$$\|\mathbf{W}\mathbf{z}^*\|_2^2 \le 2\|\nabla G(\mathbf{x}^*)\|_2^2 + 2\|\mathbf{A}^\top \mathbf{s}^*\|_2^2 \le 2\theta^2 m\|\log x^* + \mathbf{1}_d\|_2^2 + 2\sigma_{\max}^2(\mathcal{A}) \cdot \|\mathbf{s}^*\|_2^2$$

As a result

$$\|\mathbf{z}^*\|_2^2 \le \frac{2\theta^2 m\|\log x^* + \mathbf{1}_d\|_2^2 + 2\sigma_{\max}^2(\mathcal{A}) \cdot \|\mathbf{s}^*\|_2^2}{(\lambda_{min}^+(W))^2} \tag{12}$$

Using 11 into 12, we claim the final result.

Proof (Proof of Theorem 3). The proof is based on results in [5]. We have two blocks of variables: $\mathbf{z}$ and $\mathbf{s}$. Firstly, Remark 3 of [5] shows that a block coordinate method is applicable to constrained problems, provided that the constraint set is separable over variable blocks. Secondly, we apply Remark 6 of the same paper with coefficient $\beta = 1/2$. At each step, we randomly choose one of two variable blocks, and factor β rules the probability distribution. In Algorithm 2, the probability of choosing block $\mathbf{z}$ is $\eta = \sqrt{L_{\mathbf{z}}}/(\sqrt{L_{\mathbf{z}}} + \sqrt{L_{\mathbf{s}}})$, and block $\mathbf{s}$ is chosen with probability $(1 - \eta)$. Recall that for accuracy ε in primal problem (1) we need accuracy $m\varepsilon$ in dual problem (6). Combining the two remarks, we

obtain that a resulting method makes N iterations to reach ε accuracy with probability at least $1 - \delta$, where

$$N = O\left(\left(\sqrt{L_{\mathbf{z}}} + \sqrt{L_{\mathbf{s}}} \right) \sqrt{\frac{R_{\mathbf{z}}^2 + R_{\mathbf{s}}^2}{m\varepsilon}} \log\left(\frac{1}{\delta} \right) \right).$$

Consequently, the expected number of computations of $\nabla H_{\mathbf{z}}$ (that equals the number of communications) is ηN, and the expected number of computations of $\nabla H_{\mathbf{s}}$ (that corresponds to the number of local computations) is $(1 - \eta)N$. Substituting the expressions for N and η, we obtain the desired result.

5   Conclusion

In this paper, we considered a particular class of non-smooth decentralized problems. Due to specific problem structure we obtained methods that have a better dependency on problem complexity than general lower bounds. Our approach is based on passing to the dual problem. Moreover, we proposed two accelerated algorithms. The first algorithm is an accelerated primal-dual gradient method that is directly applied to the problem. The second method is a block-coordinate algorithm that allows to split communication and computation complexities.

References

1. A. Anikin, P. Dvurechensky, A. Gasnikov, A. Golov, A. Gornov, Y. Maximov, M. Mendel, and V. Spokoiny. Modern efficient numerical approaches to regularized regression problems in application to traffic demands matrix calculation from link loads. In *Proceedings of International conference ITAS-2015. Russia, Sochi*, 2015.
2. S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, Jan. 2011.
3. S. Boyd and L. Vandenberghe. *Convex Optimization*. NY Cambridge University Press, 2004.
4. P. Dvurechensky, A. Gasnikov, S. Omelchenko, and A. Tiurin. Adaptive similar triangles method: a stable alternative to Sinkhorn's algorithm for regularized optimal transport. *arXiv:1706.07622*, 2017.
5. A. Gasnikov, P. Dvurechensky, and I. Usmanova. On nontriviality of fast (accelerated) randomized methods. *Proccedings of Moscow Institute of Physics and Technology*, 8(2 (30)):67–100, 2016.
6. S. Kakade, S. Shalev-Shwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09.pdf*, 2(1), 2009.
7. D. Kovalev, E. Gasanov, A. Gasnikov, and P. Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34, 2021.
8. D. Kovalev, A. Salim, and P. Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

9. D. Kovalev, E. Shulgin, P. Richtárik, A. Rogozin, and A. Gasnikov. Adom: Accelerated decentralized optimization method for time-varying networks. *arXiv preprint arXiv:2102.09234*, 2021.
10. G. Li, Y. Chen, Y. Chi, H. V. Poor, and Y. Chen. Fast computation of optimal transport via entropy-regularized extragradient methods. *arXiv preprint arXiv:2301.13006*, 2023.
11. H. Li and Z. Lin. Accelerated gradient tracking over time-varying graphs for decentralized optimization. *arXiv preprint arXiv:2104.02596*, 2021.
12. A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
13. K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3027–3036. JMLR. org, 2017.
14. K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pages 2740–2749, 2018.
15. Y. Zhang, M. Roughan, C. Lund, and D. L. Donoho. Estimating point-to-point and point-to-multipoint traffic matrices: An information-theoretic approach. *IEEE/ACM Trans. Netw.*, 13(5):947–960, Oct. 2005.