Solving Non-smooth Constrained Programs with Lower Complexity than $\mathcal{O}(1/\varepsilon)$: A Primal-Dual Homotopy Smoothing Approach

Xiaohan Wei
Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA, 90089
xiaohanw@usc.edu

Hao Yu
Alibaba Group (U.S.) Inc.
Bellevue, WA, USA, 98004
hao.yu@alibaba-inc.com

Qing Ling
School of Data and Computer Science
Sun Yat-Sen University
Guangzhou, China, 510006
lingqing556@mail.sysu.edu.cn

Michael J. Neely
Department of Electrical Engineering
University of Southern California
Los Angeles, CA, USA, 90089
mikejneely@gmail.com

Abstract

We propose a new primal-dual homotopy smoothing algorithm for a linearly constrained convex program, where neither the primal nor the dual function has to be smooth or strongly convex. The best known iteration complexity solving such a non-smooth problem is $\mathcal{O}(\varepsilon^{-1})$. In this paper, we show that by leveraging a local error bound condition on the dual function, the proposed algorithm can achieve a better primal convergence time of $\mathcal{O}\left(\varepsilon^{-2/(2+\beta)}\log_2(\varepsilon^{-1})\right)$, where $\beta \in (0,1]$ is a local error bound parameter. As an example application of the general algorithm, we show that the distributed geometric median problem, which can be formulated as a constrained convex program, has its dual function non-smooth but satisfying the aforementioned local error bound condition with $\beta = 1/2$, therefore enjoying a convergence time of $\mathcal{O}\left(\varepsilon^{-4/5}\log_2(\varepsilon^{-1})\right)$. This result improves upon the $\mathcal{O}(\varepsilon^{-1})$ convergence time bound achieved by existing distributed optimization algorithms. Simulation experiments also demonstrate the performance of our proposed algorithm.

1 Introduction

We consider the following linearly constrained convex optimization problem:

$$\min \ f(\mathbf{x}) \tag{1}$$
$$\text{s.t. } \mathbf{A}\mathbf{x} - \mathbf{b} = 0, \ \ \mathbf{x} \in \mathcal{X}, \tag{2}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact convex set, $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function, $\mathbf{A} \in \mathbb{R}^{N\times d}$, $\mathbf{b} \in \mathbb{R}^N$. Such an optimization problem has been studied in numerous works under various application scenarios such as machine learning (Yurtsever et al. (2015)), signal processing (Ling and Tian (2010)) and communication networks (Yu and Neely (2017a)). The goal of this work is to design new algorithms for (1-2) achieving an ε approximation with better convergence time than $\mathcal{O}(1/\varepsilon)$.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

1.1 Optimization algorithms related to constrained convex program

Since enforcing the constraint $\mathbf{A}\mathbf{x} - \mathbf{b} = 0$ generally requires a significant amount of computation in large scale systems, the majority of the scalable algorithms solving problem (1-2) are of primal-dual type. Generally, the efficiency of these algorithms depends on two key properties of the dual function of (1-2), namely, the Lipschitz gradient and strong convexity. When the dual function of (1-2) is smooth, primal-dual type algorithms with Nesterov's acceleration on the dual of (1)-(2) can achieve a convergence time of $\mathcal{O}(1/\sqrt{\varepsilon})$ (e.g. Yurtsever et al. (2015); Tran-Dinh et al. (2018))[1]. When the dual function has both the Lipschitz continuous gradient and the strongly convex property, algorithms such as dual subgradient and ADMM enjoy a linear convergence $\mathcal{O}(\log(1/\varepsilon))$ (e.g. Yu and Neely (2018); Deng and Yin (2016)). However, when neither of the properties is assumed, the basic dual-subgradient type algorithm gives a relatively worse $\mathcal{O}(1/\varepsilon^2)$ convergence time (e.g. Wei et al. (2015); Wei and Neely (2018)), while its improved variants yield a convergence time of $\mathcal{O}(1/\varepsilon)$ (e.g. Lan and Monteiro (2013); Deng et al. (2017); Yu and Neely (2017b); Yurtsever et al. (2018); Gidel et al. (2018)).

More recently, several works seek to achieve a better convergence time than $\mathcal{O}(1/\varepsilon)$ under weaker assumptions than Lipschitz gradient and strong convexity of the dual function. Specifically, building upon the recent progress on the gradient type methods for optimization with Hölder continuous gradient (e.g. Nesterov (2015a,b)), the work Yurtsever et al. (2015) develops a primal-dual gradient method solving (1-2), which achieves a convergence time of $\mathcal{O}(1/\varepsilon^{\frac{1+\nu}{1+3\nu}})$, where ν is the modulus of Hölder continuity on the gradient of the dual function of the formulation (1-2).[2] On the other hand, the work Yu and Neely (2018) shows that when the dual function has Lipschitz continuous gradient and satisfies a locally quadratic property (i.e. a local error bound with $\beta = 1/2$, see Definition 2.1 for details), which is weaker than strong convexity, one can still obtain a linear convergence with a dual subgradient algorithm. A similar result has also been proved for ADMM in Han et al. (2015).

In the current work, we aim to address the following question: *Can one design a scalable algorithm with lower complexity than $\mathcal{O}(1/\varepsilon)$ solving (1-2), when both the primal and the dual functions are possibly non-smooth?* More specifically, we look at a class of problems with dual functions satisfying only a local error bound, and show that indeed one is able to obtain a faster primal convergence via a *primal-dual homotopy smoothing method* under a local error bound condition on the dual function.

Homotopy methods were first developed in the statistics literature in relation to the model selection problem for LASSO, where, instead of computing a single solution for LASSO, one computes a complete solution path by varying the regularization parameter from large to small (e.g. Osborne et al. (2000); Xiao and Zhang (2013)).[3] On the other hand, the smoothing technique for minimizing a non-smooth convex function of the following form was first considered in Nesterov (2005):

$$\Psi(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}),\ \mathbf{x} \in \Omega_1 \tag{3}$$

where $\Omega_1 \subseteq \mathbb{R}^d$ is a closed convex set, $h(\mathbf{x})$ is a convex smooth function, and $g(\mathbf{x})$ can be explicitly written as

$$g(\mathbf{x}) = \max_{\mathbf{u}\in\Omega_2} \langle \mathbf{A}\mathbf{x}, \mathbf{u}\rangle - \phi(\mathbf{u}), \tag{4}$$

where for any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, $\langle \mathbf{a}, \mathbf{b}\rangle = \mathbf{a}^T\mathbf{b}$, $\Omega_1 \subseteq \mathbb{R}^d$ is a closed convex set, and $\phi(\mathbf{u})$ is a convex function. By adding a strongly concave proximal function of $\mathbf{u}$ with a smoothing parameter $\mu > 0$ into the definition of $g(\mathbf{x})$, one can obtain a smoothed approximation of $\Psi(\mathbf{x})$ with smooth modulus μ. Then, Nesterov (2005) employs the accelerated gradient method on the smoothed approximation (which delivers a $\mathcal{O}(1/\sqrt{\varepsilon})$ convergence time for the approximation), and sets the parameter to be $\mu = \mathcal{O}(\varepsilon)$, which gives an overall convergence time of $\mathcal{O}(1/\varepsilon)$. An important follow-up question is that whether or not such a smoothing technique can also be applied to solve

[1] Our convergence time to achieve within ε of optimality is in terms of *number of (unconstrained) maximization steps* $\arg\max_{\mathbf{x}\in\mathcal{X}}[\lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b}) - f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}-\tilde{\mathbf{x}}\|^2]$ where constants $\lambda, A, \tilde{x}, \mu$ are known. This is a standard measure of convergence time for Lagrangian-type algorithms that turn a constrained problem into a sequence of unconstrained problems.

[2] The gradient of function $g(\cdot)$ is Hölder continuous with modulus $\nu \in (0, 1]$ on a set $\mathcal{X}$ if $\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \leq L_\nu\|\mathbf{x}-\mathbf{y}\|^\nu,\ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$, where $\|\cdot\|$ is the vector 2-norm and L_ν is a constant depending on ν.

[3] The word "homotopy", which was adopted in Osborne et al. (2000), refers to the fact that the mapping from regularization parameters to the set of solutions of the LASSO problem is a continuous piece-wise linear function.

(1-2) with the same primal convergence time. This question is answered in subsequent works Necoara and Suykens (2008); Li et al. (2016); Tran-Dinh et al. (2018), where they show that indeed one can also obtain an $\mathcal{O}(1/\varepsilon)$ primal convergence time for the problem (1-2) via smoothing.

Combining the homotopy method with a smoothing technique to solve problems of the form (3) has been considered by a series of works including Yang and Lin (2015), Xu et al. (2016) and Xu et al. (2017). Specifically, the works Yang and Lin (2015) and Xu et al. (2016) consider a multi-stage algorithm which starts from a large smoothing parameter μ and then decreases this parameter over time. They show that when the function $\Psi(\mathbf{x})$ satisfies a local error bound with parameter $\beta \in (0, 1]$, such a combination gives an improved convergence time of $\mathcal{O}(\log(1/\varepsilon)/\varepsilon^{1-\beta})$ minimizing the unconstrained problem (3). The work Xu et al. (2017) shows that the homotopy method can also be combined with ADMM to achieve a faster convergence solving problems of the form

$$\min_{\mathbf{x}\in\Omega_1} f(\mathbf{x}) + \psi(\mathbf{A}\mathbf{x} - \mathbf{b}),$$

where Ω_1 is a closed convex set, f, ψ are both convex functions with $f(\mathbf{x}) + \psi(\mathbf{A}\mathbf{x} - \mathbf{b})$ satisfying the local error bound, and the proximal operator of $\psi(\cdot)$ can be easily computed. However, due to the restrictions on the function ψ in the paper, it *cannot* be extended to handle problems of the form (1-2).[4]

Contributions: In the current work, we show a multi-stage homotopy smoothing method enjoys a primal convergence time $\mathcal{O}\left(\varepsilon^{-2/(2+\beta)}\log_2(\varepsilon^{-1})\right)$ solving (1-2) when the dual function satisfies a local error bound condition with $\beta \in (0, 1]$. Our convergence time to achieve within ε of optimality is in terms of *number of (unconstrained) maximization steps* $\arg\max_{x\in\mathcal{X}}[\lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b}) - f(\mathbf{x}) - \frac{\mu}{2}||\mathbf{x} - \widetilde{\mathbf{x}}||^2]$, where constants $\lambda, \mathbf{A}, \widetilde{\mathbf{x}}, \mu$ are known, which is a standard measure of convergence time for Lagrangian-type algorithms that turn a constrained problem into a sequence of unconstrained problems. The algorithm essentially restarts a weighted primal averaging process at each stage using the last Lagrange multiplier computed. This result improves upon the earlier $\mathcal{O}(1/\varepsilon)$ result by (Necoara and Suykens (2008); Li et al. (2016)) and at the same time extends the scope of homotopy smoothing method to solve a new class of problems involving constraints (1-2). It is worth mentioning that a similar restarted smoothing strategy is proposed in a recent work Tran-Dinh et al. (2018) to solve problems including (1-2), where they show that, empirically, restarting the algorithm from the Lagrange multiplier computed from the last stage improves the convergence time. Here, we give one theoretical justification of such an improvement.

1.2 The distributed geometric median problem

The geometric median problem, also known as the Fermat-Weber problem, has a long history (e.g. see Weiszfeld and Plastria (2009) for more details). Given a set of n points $\mathbf{b}_1,\ \mathbf{b}_2,\ \cdots,\ \mathbf{b}_n \in \mathbb{R}^d$, we aim to find one point $\mathbf{x}^* \in \mathbb{R}^d$ so as to minimize the sum of the Euclidean distance, i.e.

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^{n} \|\mathbf{x} - \mathbf{b}_i\|, \tag{5}$$

which is a *non-smooth* convex optimization problem. It can be shown that the solution to this problem is *unique* as long as $\mathbf{b}_1,\ \mathbf{b}_2,\ \cdots,\ \mathbf{b}_n \in \mathbb{R}^d$ are not co-linear. Linear convergence time algorithms solving (5) have also been developed in several works (e.g. Xue and Ye (1997), Parrilo and Sturmfels (2003), Cohen et al. (2016)). Our motivation of studying this problem is driven by its recent application in distributed statistical estimation, in which data are assumed to be randomly spreaded to multiple connected computational agents that produce intermediate estimators, and then, these intermediate estimators are aggregated in order to compute some statistics of the whole data set. Arguably one of the most widely used aggregation procedures is computing the *geometric median* of the local estimators (see, for example, Duchi et al. (2014), Minsker et al. (2014), Minsker and Strawn (2017), Yin et al. (2018)). It can be shown that the geometric median is robust against arbitrary corruptions of local estimators in the sense that the final estimator is stable as long as at least half of the nodes in the system perform as expected.

[4]The result in Xu et al. (2017) heavily depends on the assumption that the subgradient of $\psi(\cdot)$ is defined everywhere over the set Ω_1 and uniformly bound by some constant ρ, which excludes the choice of indicator functions necessary to deal with constraints in the ADMM framework.

Contributions: As an example application of our general algorithm, we look at the problem of computing the solution to (5) in a distributed scenario over a network of n agents without any central controller, where each agent holds a local vector $\mathbf{b}_i$. Remarkably, we show theoretically that such a problem, when formulated as (1-2), has its dual function *non-smooth but locally quadratic*. Therefore, applying our proposed primal-dual homotopy smoothing method gives a convergence time of $\mathcal{O}\left(\varepsilon^{-4/5}\log_2(\varepsilon^{-1})\right)$. This result improves upon the performance bounds of the previously known decentralized optimization algorithms (e.g. PG-EXTRA Shi et al. (2015) and decentralized ADMM Shi et al. (2014)), which do not take into account the special structure of the problem and only obtain a convergence time of $\mathcal{O}\left(1/\varepsilon\right)$. Simulation experiments also demonstrate the superior ergodic convergence time of our algorithm compared to other algorithms.

2 Primal-dual Homotopy Smoothing

2.1 Preliminaries

The Lagrange dual function of (1-2) is defined as follows:[5]

$$F(\lambda) := \max_{\mathbf{x}\in\mathcal{X}} \quad \{-\langle\lambda, \mathbf{A}\mathbf{x}-\mathbf{b}\rangle - f(\mathbf{x})\}, \tag{6}$$

where $\lambda \in \mathbb{R}^N$ is the dual variable, $\mathcal{X}$ is a compact convex set and the minimum of the dual function is $F^* := \min_{\lambda\in\mathbb{R}^N} F(\lambda)$. For any closed set $\mathcal{K} \subseteq \mathbb{R}^d$ and $\mathbf{x} \in \mathbb{R}^d$, define the distance function of $\mathbf{x}$ to the set $\mathcal{K}$ as

$$\text{dist}(\mathbf{x}, \mathcal{K}) := \min_{\mathbf{y}\in\mathcal{K}} \|\mathbf{x}-\mathbf{y}\|,$$

where $\|\mathbf{x}\| := \sqrt{\sum_{i=1}^d x_i^2}$. For a convex function $F(\lambda)$, the δ-sublevel set $\mathcal{S}_\delta$ is defined as

$$\mathcal{S}_\delta := \{\lambda \in \mathbb{R}^N :\ F(\lambda) - F^* \le \delta\}. \tag{7}$$

Furthermore, for any matrix $\mathbf{A} \in \mathbb{R}^{N\times d}$, we use $\sigma_{\max}(\mathbf{A}^T\mathbf{A})$ to denote the largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. Let

$$\Lambda^* := \left\{\lambda^* \in \mathbb{R}^N :\ F(\lambda^*) \le F(\lambda),\ \forall\lambda \in \mathbb{R}^N\right\} \tag{8}$$

be the set of optimal Lagrange multipliers. Note that if the constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$ is feasible, then $\lambda^* \in \Lambda^*$ implies $\lambda^* + \mathbf{v} \in \Lambda^*$ for any $\mathbf{v}$ that satisfies $\mathbf{A}^T\mathbf{v} = 0$. The following definition introduces the notion of local error bound.

Definition 2.1. *Let $F(\lambda)$ be a convex function over $\lambda \in \mathbb{R}^N$. Suppose Λ^* is non-empty. The function $F(\lambda)$ is said to satisfy the local error bound with parameter $\beta \in (0, 1]$ if $\exists\delta > 0$ such that for any $\lambda \in \mathcal{S}_\delta$,*

$$\textit{dist}(\lambda, \Lambda^*) \le C_\delta(F(\lambda) - F^*)^\beta, \tag{9}$$

where C_δ is a positive constant possibly depending on δ. In particular, when $\beta = 1/2$, $F(\lambda)$ is said to be locally quadratic and when $\beta = 1$, it is said to be locally linear.

Remark 2.1. *Indeed, a wide range of popular optimization problems satisfy the local error bound condition. The work Tseng (2010) shows that if $\mathcal{X}$ is a polyhedron, $f(\cdot)$ has Lipschitz continuous gradient and is strongly convex, then the dual function of (1-2) is locally linear. The work Burke and Tseng (1996) shows that when the objective is linear and $\mathcal{X}$ is a convex cone, the dual function is also locally linear. The values of β have also been computed for several other problems (e.g. Pang (1997); Yang and Lin (2015)).*

Definition 2.2. *Given an accuracy level $\varepsilon > 0$, a vector $\mathbf{x}_0 \in \mathcal{X}$ is said to achieve an ε approximate solution regarding problem (1-2) if*

$$f(\mathbf{x}_0) - f^* \le \mathcal{O}(\varepsilon),\ \ \|\mathbf{A}\mathbf{x}_0 - \mathbf{b}\| \le \mathcal{O}(\varepsilon),$$

where f^* is the optimal primal objective of (1-2).

Throughout the paper, we adopt the following assumptions:

[5] Usually, the Lagrange dual is defined as $\min_{\mathbf{x}\in\mathcal{X}}\ \langle\lambda, \mathbf{A}\mathbf{x}-\mathbf{b}\rangle + f(\mathbf{x})$. Here, we flip the sign and take the maximum for no reason other than being consistent with the form (4).

Assumption 2.1. *(a) The feasible set* $\{\mathbf{x} \in \mathcal{X} : \mathbf{A}\mathbf{x} - \mathbf{b} = 0\}$ *is nonempty and non-singleton.*
(b) The set $\mathcal{X}$ *is bounded, i.e.* $\sup_{\mathbf{x},\mathbf{y}\in\mathcal{X}} \|\mathbf{x} - \mathbf{y}\| \leq D$*, for some positive constant* D*. Furthermore, the function* $f(\mathbf{x})$ *is also bounded, i.e.* $\max_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})| \leq M$*, for some positive constant* M*.*
(c) The dual function defined in (6) *satisfies the local error bound for some parameter* $\beta \in (0, 1]$ *and some level* $\delta > 0$*.*
(d) Let $\mathcal{P}_{\mathbf{A}}$ *be the projection operator onto the column space of* $\mathbf{A}$*. There exists a unique vector* $\nu^* \in \mathbb{R}^N$ *such that for any* $\lambda^* \in \Lambda^*$*,* $\mathcal{P}_{\mathbf{A}}\lambda^* = \nu^*$*, i.e.* $\Lambda^* = \left\{\lambda^* \in \mathbb{R}^N : \mathcal{P}_{\mathbf{A}}\lambda^* = \nu^*\right\}$*.*

Note that assumption (a) and (b) are very mild and quite standard. For most applications, it is enough to check (c) and (d). We will show, for example, in Section 4 that the distributed geometric median problem satisfies all the assumptions. Finally, we say a function $g : \mathcal{X} \to \mathbb{R}$ is smooth with modulus $L > 0$ if

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

2.2 Primal-dual homotopy smoothing algorithm

This section introduces our proposed algorithm for optimization problem (1-2) satisfying Assumption 2.1. The idea of smoothing is to introduce a smoothed Lagrange dual function $F_\mu(\lambda)$ that approximates the original possibly non-smooth dual function $F(\lambda)$ defined in (6).

For any constant $\mu > 0$, define

$$f_\mu(\mathbf{x}) = f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2, \tag{10}$$

where $\widetilde{\mathbf{x}}$ is an arbitrary fixed point in $\mathcal{X}$. For simplicity of notation, we drop the dependency on $\widetilde{\mathbf{x}}$ in the definition of $f_\mu(\mathbf{x})$. Then, by the boundedness assumption of $\mathcal{X}$, we have $f(\mathbf{x}) \leq f_\mu(\mathbf{x}) \leq f(\mathbf{x}) + \frac{\mu}{2}D^2, \ \forall \mathbf{x} \in \mathcal{X}$. For any $\lambda \in \mathbb{R}^N$, define

$$F_\mu(\lambda) = \max_{\mathbf{x}\in\mathcal{X}} \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle - f_\mu(\mathbf{x}) \tag{11}$$

as the smoothed dual function. The fact that $F_\mu(\lambda)$ is indeed smooth with modulus μ follows from Lemma 6.1 in the Supplement. Thus, one is able to apply an accelerated gradient descent algorithm on this modified Lagrange dual function, which is detailed in Algorithm 1 below, starting from an initial primal-dual pair $(\widetilde{\mathbf{x}}, \widetilde{\lambda}) \in \mathbb{R}^d \times \mathbb{R}^N$.

Algorithm 1 Primal-Dual Smoothing: PDS$\left(\widetilde{\lambda}, \widetilde{\mathbf{x}}, \mu, T\right)$

Let $\lambda_0 = \lambda_{-1} = \widetilde{\lambda}$ and $\theta_0 = \theta_{-1} = 1$.
For $t = 0$ to $T - 1$ **do**

- Compute a tentative dual multiplier: $\widehat{\lambda}_t = \lambda_t + \theta_t(\theta_{t-1}^{-1} - 1)(\lambda_t - \lambda_{t-1})$,
- Compute the primal update: $\mathbf{x}(\widehat{\lambda}_t) = \text{argmax}_{\mathbf{x}\in\mathcal{X}} \ -\left\langle \widehat{\lambda}_t, \mathbf{A}\mathbf{x} - \mathbf{b}\right\rangle - f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x} - \widetilde{\mathbf{x}}\|^2$.
- Compute the dual update: $\lambda_{t+1} = \widehat{\lambda}_t + \mu(\mathbf{A}\mathbf{x}(\widehat{\lambda}_t) - \mathbf{b})$.
- Update the stepsize: $\theta_{t+1} = \frac{\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2}{2}$.

end for
Output: $\overline{\mathbf{x}}_T = \frac{1}{S_T}\sum_{t=0}^{T-1} \frac{1}{\theta_t}\mathbf{x}(\widehat{\lambda}_t)$ and λ_T, where $S_T = \sum_{t=0}^{T-1} \frac{1}{\theta_t}$.

Our proposed algorithm runs Algorithm 1 in multiple stages, which is detailed in Algorithm 2 below.

3 Convergence Time Results

We start by defining the set of optimal Lagrange multipliers for the smoothed problem:[6]

$$\Lambda_\mu^* := \left\{\lambda_\mu^* \in \mathbb{R}^N : F_\mu(\lambda_\mu^*) \leq F_\mu(\lambda), \ \forall \lambda \in \mathbb{R}^N\right\} \tag{12}$$

[6]By Assumption 2.1(a) and Farkas' Lemma, this is non-empty.

Algorithm 2 Homotopy Method:

Let ε_0 be a fixed constant and $\varepsilon < \varepsilon_0$ be the desired accuracy. Set $\mu_0 = \frac{\varepsilon_0}{D^2}$, $\lambda^{(0)} = 0$, $\overline{\mathbf{x}}^{(0)} \in \mathcal{X}$, the number of stages $K \geq \lceil \log_2(\varepsilon_0/\varepsilon) \rceil + 1$, and the time horizon during each stage $T \geq 1$.
For $k = 1$ to K **do**

- Let $\mu_k = \mu_{k-1}/2$.
- Run the primal-dual smoothing algorithm $(\lambda^{(k)}, \overline{\mathbf{x}}^{(k)}) = \text{PDS}\left(\lambda^{(k-1)}, \overline{\mathbf{x}}^{(k-1)}, \mu_k, T\right)$.

end for
Output: $\overline{\mathbf{x}}^{(K)}$.

Our convergence time analysis involves two steps. The first step is to derive a primal convergence time bound for Algorithm 1, which involves the location information of the initial Lagrange multiplier at the beginning of this stage. The details are given in Supplement 6.2.

Theorem 3.1. *Suppose Assumption 2.1(a)(b) holds. For any $T \geq 1$ and any initial vector $(\widetilde{\mathbf{x}}, \widetilde{\lambda}) \in \mathbb{R}^d \times \mathbb{R}^N$, we have the following performance bound regarding Algorithm 1,*

$$f\left(\overline{\mathbf{x}}_T\right) - f^* \leq \|\mathcal{P}_{\mathbf{A}}\widetilde{\lambda}^*\| \cdot \|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\| + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\left\|\widetilde{\lambda}^* - \widetilde{\lambda}\right\|^2 + \frac{\mu D^2}{2}, \tag{13}$$

$$\|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\| \leq \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu S_T}\left(\left\|\widetilde{\lambda}^* - \widetilde{\lambda}\right\| + \textit{dist}(\lambda_\mu^*, \Lambda^*)\right), \tag{14}$$

where $\widetilde{\lambda}^* \in \textit{argmin}_{\lambda^* \in \Lambda^*}\|\lambda^* - \widetilde{\lambda}\|$, $\overline{\mathbf{x}}_T := \frac{1}{S_T}\sum_{t=0}^{T-1}\frac{\mathbf{x}(\widehat{\lambda}_t)}{\theta_t}$, $S_T = \sum_{t=0}^{T-1}\frac{1}{\theta_t}$ and λ_μ^* is any point in Λ_μ^* defined in (12).

An inductive argument shows that $\theta_t \leq 2/(t+2)\ \forall t \geq 0$. Thus, Theorem 3.1 already gives an $\mathcal{O}(1/\varepsilon)$ convergence time by setting $\mu = \varepsilon$ and $T = 1/\varepsilon$. Note that this is the best trade-off we can get from Theorem 3.1 when simply bounding the terms $\|\widetilde{\lambda}^* - \widetilde{\lambda}\|$ and dist$(\lambda_\mu^*, \Lambda^*)$ by constants. To see how this bound leads to an improved convergence time when running in multiple rounds, suppose the computation from the last round gives a $\widetilde{\lambda}$ that is close enough to the optimal set Λ^*, then, $\|\widetilde{\lambda}^* - \widetilde{\lambda}\|$ would be small. When the local error bound condition holds, one can show that dist$(\lambda_\mu^*, \Lambda^*) \leq \mathcal{O}(\mu^\beta)$. As a consequence, one is able to choose μ smaller than ε and get a better trade-off. Formally, we have the following overall performance bound. The proof is given in Supplement 6.3.

Theorem 3.2. *Suppose Assumption 2.1 holds, $\varepsilon_0 \geq \max\{2M, 1\}$, $0 < \varepsilon \leq \min\{\delta/2, 2M, 1\}$, $T \geq \frac{2DC_\delta\sqrt{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}(2M)^{\beta/2}}{\varepsilon^{2/(2+\beta)}}$. The proposed homotopy method achieves the following objective and constraint violation bound:*

$$f(\overline{\mathbf{x}}^{(K)}) - f^* \leq \left(\frac{24\|\mathcal{P}_{\mathbf{A}}\lambda_*\|(1+C_\delta)}{C_\delta^2(2M)^{2\beta}} + \frac{6}{C_\delta^2(2M)^{2\beta}} + \frac{1}{4}\right)\varepsilon,$$

$$\|\mathbf{A}\overline{\mathbf{x}}^{(K)} - \mathbf{b}\| \leq \frac{24(1+C_\delta)}{C_\delta^2(2M)^\beta}\varepsilon,$$

with running time $\frac{2DC_\delta\sqrt{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}(2M)^{\beta/2}}{\varepsilon^{2/(2+\beta)}}(\lceil\log_2(\varepsilon_0/\varepsilon)\rceil + 1)$, i.e. the algorithm achieves an ε approximation with convergence time $\mathcal{O}\left(\varepsilon^{-2/(2+\beta)}\log_2(\varepsilon^{-1})\right)$.

4 Distributed Geometric Median

Consider the problem of computing the geometric median over a connected network $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \cdots, n\}$ is a set of n nodes, $\mathcal{E} = \{e_{ij}\}_{i,j\in\mathcal{V}}$ is a collection of undirected edges, $e_{ij} = 1$ if there exists an undirected edge between node i and node j, and $e_{ij} = 0$ otherwise. Furthermore, $e_{ii} = 1,\ \forall i \in \{1, 2, \cdots, n\}$.Furthermore, since the graph is undirected, we always have $e_{ij} = e_{ji},\ \forall i, j \in \{1, 2, \cdots, n\}$. Two nodes i and j are said to be neighbors of each other if $e_{ij} = 1$. Each node i holds a local vector $\mathbf{b}_i \in \mathbb{R}^d$, and the goal is to compute the solution to (5) without having a central controller, i.e. each node can only communicate with its neighbors.

Computing geometric median over a network has been considered in several works previously and various distributed algorithms have been developed such as decentralized subgradient methd (DSM, Nedic and Ozdaglar (2009); Yuan et al. (2016)), PG-EXTRA (Shi et al. (2015)) and ADMM (Shi et al. (2014); Deng et al. (2017)). The best known convergence time for this problem is $\mathcal{O}(1/\varepsilon)$. In this section, we will show that it can be written in the form of problem (1-2), has its Lagrange dual function *locally quadratic* and optimal Lagrange multiplier unique up to the null space of $\mathbf{A}$, thereby satisfying Assumption 2.1.

Throughout this section, we assume that $n \geq 3$, $\mathbf{b}_1,\ \mathbf{b}_2,\ \cdots,\ \mathbf{b}_n \in \mathbb{R}^d$ are not co-linear and they are distinct (i.e. $\mathbf{b}_i \neq \mathbf{b}_j$ if $i \neq j$). We start by defining a mixing matrix $\widetilde{\mathbf{W}} \in \mathbb{R}^{n\times n}$ with respect to this network. The mixing matrix will have the following properties:

1. Decentralization: The (i,j)-th entry $\widetilde{w}_{ij} = 0$ if $e_{ij} = 0$.
2. Symmetry: $\widetilde{\mathbf{W}} = \widetilde{\mathbf{W}}^T$.
3. The null space of $\mathbf{I}_{n\times n} - \widetilde{\mathbf{W}}$ satisfies $\mathcal{N}(\mathbf{I}_{n\times n} - \widetilde{\mathbf{W}}) = \{c\mathbf{1},\ c \in \mathbb{R}\}$, where $\mathbf{1}$ is an all 1 vector in $\mathbb{R}^n$.

These conditions are rather mild and satisfied by most doubly stochastic mixing matrices used in practice. Some specific examples are Markov transition matrices of max-degree chain and Metropolis-Hastings chain (see Boyd et al. (2004) for detailed discussions). Let $\mathbf{x}_i \in \mathbb{R}^d$ be the local variable on the node i. Define

$$\mathbf{x} := \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \in \mathbb{R}^{nd},\ \ \mathbf{b} := \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_n \end{bmatrix} \in \mathbb{R}^{nd},\ \ \mathbf{A} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{n1} & \cdots & \mathbf{W}_{nn} \end{bmatrix} \in \mathbb{R}^{(nd)\times(nd)},$$

where

$$\mathbf{W}_{ij} = \begin{cases} (1-\widetilde{w}_{ij})\mathbf{I}_{d\times d}, & \text{if } i = j \\ -\widetilde{w}_{ij}\mathbf{I}_{d\times d}, & \text{if } i \neq j \end{cases},$$

and $\widetilde{w}_{ij}$ is ij-th entry of the mixing matrix $\widetilde{\mathbf{W}}$. By the aforementioned null space property of the mixing matrix $\widetilde{\mathbf{W}}$, it is easy to see that the null space of the matrix $\mathbf{A}$ is

$$\mathcal{N}(\mathbf{A}) = \left\{\mathbf{u} \in \mathbb{R}^{nd} :\ \mathbf{u} = [\mathbf{u}_1^T, \cdots, \mathbf{u}_n^T]^T,\ \mathbf{u}_1 = \mathbf{u}_2 = \cdots = \mathbf{u}_n\right\}, \tag{15}$$

Then, because of the null space property (15), one can equivalently write problem (5) in a "distributed fashion" as follows:

$$\min\ \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{b}_i\| \tag{16}$$

$$s.t.\ \ \mathbf{A}\mathbf{x} = 0, \|\mathbf{x}_i - \mathbf{b}_i\| \leq D,\ i = 1, 2, \cdots, n, \tag{17}$$

where we set the constant D to be large enough so that the solution belongs to the set $\mathcal{X} := \left\{\mathbf{x} \in \mathbb{R}^{nd} :\ \|\mathbf{x}_i - \mathbf{b}_i\| \leq D, i = 1, 2, \cdots, n\right\}$. This is in the same form as (1-2) with $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^{nd} :\ \|\mathbf{x}_i - \mathbf{b}_i\| \leq D,\ i = 1, 2, \cdots, n\}$.

4.1 Distributed implementation

In this section, we show how to implement the proposed algorithm to solve (16-17) in a distributed way. Let $\lambda_t = [\lambda_{t,1}^T,\ \lambda_{t,2}^T, \cdots,\ \lambda_{t,n}^T] \in \mathbb{R}^{nd}$, $\widehat{\lambda}_t = [\widehat{\lambda}_{t,1}^T,\ \widehat{\lambda}_{t,2}^T, \cdots,\ \widehat{\lambda}_{t,n}^T] \in \mathbb{R}^{nd}$ be the vectors of Lagrange multipliers defined in Algorithm 1, where each $\lambda_{t,i},\ \widehat{\lambda}_{t,i} \in \mathbb{R}^d$. Then, each agent $i \in \{1, 2, \cdots, n\}$ in the network is responsible for updating the corresponding Lagrange multipliers $\lambda_{t,i}$ and $\widehat{\lambda}_{t,i}$ according to Algorithm 1, which has the initial values $\lambda_{0,i} = \lambda_{-1,i} = \widetilde{\lambda}_i$. Note that the first, third and fourth steps in Algorithm 1 are naturally separable regarding each agent. It remains to check if the second step can be implemented in a distributed way.

Note that in the second step, we obtain the primal update $\mathbf{x}(\widehat{\lambda}_t) = [\mathbf{x}_1(\widehat{\lambda}_t)^T, \cdots, \mathbf{x}_n(\widehat{\lambda}_t)^T] \in \mathbb{R}^{nd}$ by solving the following problem:

$$\mathbf{x}(\widehat{\lambda}_t) = \text{argmax}_{\mathbf{x}:\|\mathbf{x}_i-\mathbf{b}_i\|\leq D,\ i=1,2,\cdots,n}\ -\left\langle \widehat{\lambda}_t, \mathbf{A}\mathbf{x}\right\rangle - \sum_{i=1}^{n}\left(\|\mathbf{x}_i - \mathbf{b}_i\| + \frac{\mu}{2}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2\right),$$

where $\widetilde{\mathbf{x}}_i \in \mathbb{R}^d$ is a fixed point in the feasible set. We separate the maximization according to different agent $i \in \{1, 2, \cdots, n\}$:

$$\mathbf{x}_i(\widehat{\lambda}_t) = \text{argmax}_{\mathbf{x}_i:\|\mathbf{x}_i-\mathbf{b}_i\|\leq D} \ -\sum_{j=1}^{n} \left\langle \widehat{\lambda}_{t,j}, \mathbf{W}_{ji}\mathbf{x}_i \right\rangle - \|\mathbf{x}_i - \mathbf{b}_i\| - \frac{\mu}{2}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2.$$

Note that according to the definition of $\mathbf{W}_{ji}$, it is equal to 0 if agent j is not the neighbor of agent i. More specifically, Let $\mathcal{N}_i$ be the set of neighbors of agent i (including the agent i itself), then, the above maximization problem can be equivalently written as

$$\begin{aligned}
&\text{argmax}_{\mathbf{x}_i:\|\mathbf{x}_i-\mathbf{b}_i\|\leq D} - \sum_{j\in\mathcal{N}_i} \left\langle \widehat{\lambda}_{t,j}, \mathbf{W}_{ji}\mathbf{x}_i \right\rangle - \|\mathbf{x}_i - \mathbf{b}_i\| - \frac{\mu}{2}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2 \\
=&\text{argmax}_{\mathbf{x}_i:\|\mathbf{x}_i-\mathbf{b}_i\|\leq D} - \left\langle \sum_{j\in\mathcal{N}_i} \mathbf{W}_{ji}\widehat{\lambda}_{t,j}, \mathbf{x}_i \right\rangle - \|\mathbf{x}_i - \mathbf{b}_i\| - \frac{\mu}{2}\|\mathbf{x}_i - \widetilde{\mathbf{x}}_i\|^2 \ \ i \in \{1, 2, \cdots, n\},
\end{aligned}$$

where we used the fact that $\mathbf{W}_{ji}^T = \mathbf{W}_{ji}$. Solving this problem only requires the local information from each agent. Completing the squares gives

$$\mathbf{x}_i(\widehat{\lambda}_t) = \text{argmax}_{\|\mathbf{x}_i-\mathbf{b}_i\|\leq D} - \frac{\mu}{2}\left\|\mathbf{x}_i - \left(\widetilde{\mathbf{x}}_i - \frac{1}{\mu}\sum_{j\in\mathcal{N}_i}\mathbf{W}_{ji}\widehat{\lambda}_{t,j}\right)\right\|^2 - \|\mathbf{x}_i - \mathbf{b}_i\|. \tag{18}$$

The solution to such a subproblem has a closed form, as is shown in the following lemma (the proof is given in Supplement 6.4):

Lemma 4.1. *Let* $\mathbf{a}_i = \widetilde{\mathbf{x}}_i - \frac{1}{\mu}\sum_{j\in\mathcal{N}_i}\mathbf{W}_{ji}\widehat{\lambda}_{t,j}$*, then, the solution to* (18) *has the following closed form:*

$$\mathbf{x}_i(\widehat{\lambda}_t) = \begin{cases} \mathbf{b}_i, & \text{if } \|\mathbf{b}_i - \mathbf{a}_i\| \leq 1/\mu, \\ \mathbf{b}_i - \frac{\mathbf{b}_i-\mathbf{a}_i}{\|\mathbf{b}_i-\mathbf{a}_i\|}\left(\|\mathbf{b}_i - \mathbf{a}_i\| - \frac{1}{\mu}\right), & \text{if } \frac{1}{\mu} < \|\mathbf{b}_i - \mathbf{a}_i\| \leq \frac{1}{\mu} + D, \\ \mathbf{b}_i - \frac{\mathbf{b}_i-\mathbf{a}_i}{\|\mathbf{b}_i-\mathbf{a}_i\|}D, & \text{otherwise}. \end{cases}$$

4.2 Local error bound condition

The proof of the this theorem is given in Supplement 6.5.

Theorem 4.1. *The Lagrange dual function of (16-17) is non-smooth and given by the following*

$$F(\lambda) = -\left\langle \mathbf{A}^T\lambda, \mathbf{b} \right\rangle + D\sum_{i=1}^{n}(\|\mathbf{A}_{[i]}^T\lambda\| - 1)\cdot I\left(\|\mathbf{A}_{[i]}^T\lambda\| > 1\right),$$

where $\mathbf{A}_{[i]} = [\mathbf{W}_{1i}\ \mathbf{W}_{2i}\ \cdots\ \mathbf{W}_{ni}]^T$ *is the* i*-th column block of the matrix* $\mathbf{A}$*,* $I\left(\|\mathbf{A}_{[i]}^T\lambda\| > 1\right)$ *is the indicator function which takes 1 if* $\|\mathbf{A}_{[i]}^T\lambda\| > 1$ *and 0 otherwise. Let* Λ^* *be the set of optimal Lagrange multipliers defined according to* (8)*. Suppose* $D \geq 2n \cdot \max_{i,j\in\mathcal{V}}\|\mathbf{b}_i - \mathbf{b}_j\|$*, then, for any* $\delta > 0$*, there exists a* $C_\delta > 0$ *such that*

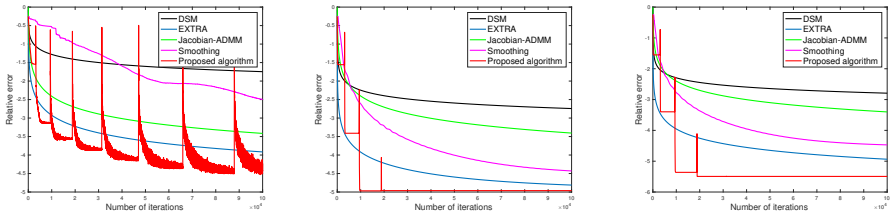
$$dist(\lambda, \Lambda^*) \leq C_\delta(F(\lambda) - F^*)^{1/2}, \ \forall\lambda \in \mathcal{S}_\delta.$$

Furthermore, there exists a unique vector $\nu^* \in \mathbb{R}^{nd}$ *s.t.* $\mathcal{P}_{\mathbf{A}}\lambda^* = \nu^*$*,* $\forall\lambda^* \in \Lambda^*$*, i.e. Assumption 2.1(d) holds. Thus, applying the proposed method gives the convergence time* $\mathcal{O}\left(\varepsilon^{-4/5}\log_2(\varepsilon^{-1})\right)$.

5 Simulation Experiments

In this section, we conduct simulation experiments on the distributed geometric median problem. Each vector $\mathbf{b}_i \in \mathbb{R}^{100}$, $i \in \{1, 2, \cdots, n\}$ is sampled from the uniform distribution in $[0, 10]^{100}$, i.e. each entry of $\mathbf{b}_i$ is independently sampled from uniform distribution on $[0, 10]$. We compare our algorithm with DSM (Nedic and Ozdaglar (2009)), P-EXTRA (Shi et al. (2015)), Jacobian parallel ADMM (Deng et al. (2017)) and Smoothing (Necoara and Suykens (2008)) under different network

sizes ($n = 20, 50, 100$). Each network is randomly generated with a particular connectivity ratio[7], and the mixing matrix is chosen to be the Metropolis-Hastings Chain (Boyd et al. (2004)), which can be computed in a distributed manner. We use the relative error as the performance metric, which is defined as $\|\overline{\mathbf{x}}_t - \mathbf{x}^*\|/\|\mathbf{x}_0 - \mathbf{x}^*\|$ for each iteration t. The vector $\mathbf{x}_0 \in \mathbb{R}^{nd}$ is the initial primal variable. The vector $\mathbf{x}^* \in \mathbb{R}^{nd}$ is the optimal solution computed by CVX Grant et al. (2008). For our proposed algorithm, $\overline{\mathbf{x}}_t$ is the restarted primal average up to the current iteration. For all other algorithms, $\overline{\mathbf{x}}_t$ is the primal average up to the current iteration. The results are shown below. We see in all cases, our proposed algorithm is much better than, if not comparable to, other algorithms. For detailed simulation setups and additional simulation results, see Supplement 6.6.



(a) (b) (c)

Figure 1: Comparison of different algorithms on networks of different sizes. (a) $n = 20$, connectivity ratio=0.15. (b) $n = 50$, connectivity ratio=0.13. (c) $n = 100$, connectivity ratio=0.1.

Acknowledgments

The authors thank Stanislav Minsker and Jason D. Lee for helpful discussions related to the geometric median problem. Qing Ling's research is supported in part by the National Science Foundation China under Grant 61573331 and Guangdong IIET Grant 2017ZT07X355. Qing Ling is also affiliated with Guangdong Province Key Laboratory of Computational Science. Michael J. Neely's research is supported in part by the National Science Foundation under Grant CCF-1718477.

References

Beck, A., A. Nedic, A. Ozdaglar, and M. Teboulle (2014). An $o(1/k)$ gradient method for network resource allocation problems. *IEEE Transactions on Control of Network Systems 1*(1), 64–73.

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific Belmont.

Bertsekas, D. P. (2009). *Convex optimization theory*. Athena Scientific Belmont.

Boyd, S., P. Diaconis, and L. Xiao (2004). Fastest mixing markov chain on a graph. *SIAM Review 46*(4), 667–689.

Burke, J. V. and P. Tseng (1996). A unified analysis of Hoffman's bound via Fenchel duality. *SIAM Journal on Optimization 6*(2), 265–282.

Cohen, M. B., Y. T. Lee, G. Miller, J. Pachocki, and A. Sidford (2016). Geometric median in nearly linear time. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 9–21.

Deng, W., M.-J. Lai, Z. Peng, and W. Yin (2017). Parallel multi-block ADMM with $o(1/k)$ convergence. *Journal of Scientific Computing 71*(2), 712–736.

Deng, W. and W. Yin (2016). On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing 66*(3), 889–916.

Duchi, J. C., M. I. Jordan, M. J. Wainwright, and Y. Zhang (2014). Optimality guarantees for distributed statistical estimation. *arXiv preprint arXiv:1405.0782*.

Gidel, G., F. Pedregosa, and S. Lacoste-Julien (2018). Frank-Wolfe splitting via augmented Lagrangian method. *arXiv preprint arXiv:1804.03176*.

Grant, M., S. Boyd, and Y. Ye (2008). CVX: Matlab software for disciplined convex programming.

[7]The connectivity ratio is defined as the number of edges divided by the total number of possible edges $n(n+1)/2$.

Han, D., D. Sun, and L. Zhang (2015). Linear rate convergence of the alternating direction method of multipliers for convex composite quadratic and semi-definite programming. *arXiv preprint arXiv:1508.02134*.

Lan, G. and R. D. Monteiro (2013). Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming 138*(1-2), 115–139.

Li, J., G. Chen, Z. Dong, and Z. Wu (2016). A fast dual proximal-gradient method for separable convex optimization with linear coupled constraints. *Computational Optimization and Applications 64*(3), 671–697.

Ling, Q. and Z. Tian (2010). Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. *IEEE Transactions on Signal Processing 58*(7), 3816–3827.

Luo, X.-D. and Z.-Q. Luo (1994). Extension of hoffman's error bound to polynomial systems. *SIAM Journal on Optimization 4*(2), 383–392.

Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2014). Robust and scalable bayes via a median of subset posterior measures. *arXiv preprint arXiv:1403.2660*.

Minsker, S. and N. Strawn (2017). Distributed statistical estimation and rates of convergence in normal approximation. *arXiv preprint arXiv:1704.02658*.

Motzkin, T. (1952). *Contributions to the theory of linear inequalities*. D.R. Fulkerson (Transl.) (Santa Monica: RAND Corporation). RAND Corporation Translation 22.

Necoara, I. and J. A. Suykens (2008). Application of a smoothing technique to decomposition in convex optimization. *IEEE Transactions on Automatic control 53*(11), 2674–2679.

Nedic, A. and A. Ozdaglar (2009). Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control 54*(1), 48–61.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming 103*(1), 127–152.

Nesterov, Y. (2015a). Complexity bounds for primal-dual methods minimizing the model of objective function. *Mathematical Programming*, 1–20.

Nesterov, Y. (2015b). Universal gradient methods for convex optimization problems. *Mathematical Programming 152*(1-2), 381–404.

Osborne, M. R., B. Presnell, and B. A. Turlach (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis 20*(3), 389–403.

Pang, J.-S. (1997, Oct). Error bounds in mathematical programming. *Mathematical Programming 79*(1), 299–332.

Parrilo, P. A. and B. Sturmfels (2003). Minimizing polynomial functions. *Algorithmic and quantitative real algebraic geometry, DIMACS Series in Discrete Mathematics and Theoretical Computer Science 60*, 83–99.

Shi, W., Q. Ling, G. Wu, and W. Yin (2015). A proximal gradient algorithm for decentralized composite optimization. *IEEE Transactions on Signal Processing 63*(22), 6013–6023.

Shi, W., Q. Ling, K. Yuan, G. Wu, and W. Yin (2014). On the linear convergence of the admm in decentralized consensus optimization. *IEEE Trans. Signal Processing 62*(7), 1750–1761.

Tran-Dinh, Q., O. Fercoq, and V. Cevher (2018). A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *SIAM Journal on Optimization 28*(1), 96–134.

Tseng, P. (2010). Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming 125*(2), 263–295.

Wang, T. and J.-S. Pang (1994). Global error bounds for convex quadratic inequality systems. *Optimization 31*(1), 1–12.

Wei, X. and M. J. Neely (2018). Primal-dual Frank-Wolfe for constrained stochastic programs with convex and non-convex objectives. *arXiv preprint arXiv:1806.00709*.

Wei, X., H. Yu, and M. J. Neely (2015). A probabilistic sample path convergence time analysis of drift-plus-penalty algorithm for stochastic optimization. *arXiv preprint arXiv:1510.02973*.

Weiszfeld, E. and F. Plastria (2009). On the point for which the sum of the distances to n given points is minimum. *Annals of Operations Research 167*(1), 7–41.

Xiao, L. and T. Zhang (2013). A proximal-gradient homotopy method for the sparse least-squares problem. *SIAM Journal on Optimization 23*(2), 1062–1091.

Xu, Y., M. Liu, Q. Lin, and T. Yang (2017). ADMM without a fixed penalty parameter: Faster convergence with new adaptive penalization. In *Advances in Neural Information Processing Systems*, pp. 1267–1277.

Xu, Y., Y. Yan, Q. Lin, and T. Yang (2016). Homotopy smoothing for non-smooth problems with lower complexity than $o(1/\epsilon)$. In *Advances In Neural Information Processing Systems*, pp. 1208–1216.

Xue, G. and Y. Ye (1997). An efficient algorithm for minimizing a sum of euclidean norms with applications. *SIAM Journal on Optimization 7*(4), 1017–1036.

Yang, T. and Q. Lin (2015). Rsg: Beating subgradient method without smoothness and strong convexity. *arXiv preprint arXiv:1512.03107*.

Yin, D., Y. Chen, K. Ramchandran, and P. Bartlett (2018). Byzantine-robust distributed learning: Towards optimal statistical rates. *arXiv preprint arXiv:1803.01498*.

Yu, H. and M. J. Neely (2017a). A new backpressure algorithm for joint rate control and routing with vanishing utility optimality gaps and finite queue lengths. In *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*, pp. 1–9. IEEE.

Yu, H. and M. J. Neely (2017b). A simple parallel algorithm with an $o(1/t)$ convergence rate for general convex programs. *SIAM Journal on Optimization 27*(2), 759–783.

Yu, H. and M. J. Neely (2018). On the convergence time of dual subgradient methods for strongly convex programs. *IEEE Transactions on Automatic Control*.

Yuan, K., Q. Ling, and W. Yin (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization 26*(3), 1835–1854.

Yurtsever, A., Q. T. Dinh, and V. Cevher (2015). A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*, pp. 3150–3158.

Yurtsever, A., O. Fercoq, F. Locatello, and V. Cevher (2018). A conditional gradient framework for composite convex minimization with applications to semidefinite programming. *arXiv preprint arXiv:1804.08544*.

6 Supplement

6.1 Smoothing lemma

In this section, we show that adding the strongly convex term on the primal indeed gives a smoothed dual.

Lemma 6.1. *Let $f_\mu(\mathbf{x})$ be defined as above and let $g_i : \mathcal{X} \to \mathbb{R},\ i = 1, 2, \cdots, N$ be a sequence of G-Lipschitz continuous convex functions, i.e. $\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq G\|\mathbf{x} - \mathbf{y}\|,\ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$, where $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), \ldots, g_N(\mathbf{x})]$. Then, the Lagrange dual function*

$$d_\mu(\lambda) := \max_{\mathbf{x}\in\mathcal{X}} -\langle \lambda, \mathbf{g}(\mathbf{x})\rangle - f_\mu(\mathbf{x}),\ \lambda \in \mathbb{R}^N$$

is smooth with modulus G^2/μ. In particular, if $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$, then, the smooth modulus is equal to $\sigma_{\max}(\mathbf{A}^T\mathbf{A})/\mu$, where $\sigma_{\max}(\mathbf{A}^T\mathbf{A})$ denotes the maximum eigenvalue of $\mathbf{A}^T\mathbf{A}$.

This proof of this lemma is rather standard (see also proof of Lemma 6 of Yu and Neely (2018)) and the special case of $\mathbf{g}(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ can also be derived from Fenchel duality (Beck et al. (2014)).

Proof of Lemma 6.1. First of all, note that the function $h_\lambda(\mathbf{x}) = -\langle \lambda, \mathbf{g}(\mathbf{x})\rangle - f_\mu(\mathbf{x})$ is strongly concave, it follows that there exists a unique minimizer $\mathbf{x}(\lambda) := \text{argmax}_{\mathbf{x}\in\mathcal{X}} h_\lambda(\mathbf{x})$. By Danskin's theorem (see Bertsekas (1999) for details), we have for any $\lambda \in \mathbb{R}^N$,

$$\nabla d_\mu(\lambda) = \mathbf{g}(\mathbf{x}(\lambda)).$$

Now, consider any $\lambda_1, \lambda_2 \in \mathbb{R}^N$, we have

$$\|\nabla d_\mu(\lambda_1) - \nabla d_\mu(\lambda_2)\| = \|\mathbf{g}(\mathbf{x}(\lambda_1)) - \mathbf{g}(\mathbf{x}(\lambda_2))\| \leq G\|x(\lambda_1) - x(\lambda_2)\|. \tag{19}$$

where the equality follows from Danskin's Theorem and the inequality follows from Lipschitz continuity of $\mathbf{g}(\mathbf{x})$. Again, by the fact that $h_\mu(\mathbf{x})$ is strongly concave with modulus μ,

$$h_{\lambda_1}(\mathbf{x}(\lambda_2)) \leq h_{\lambda_1}(\mathbf{x}(\lambda_1)) - \frac{\mu}{2}\|x(\lambda_1) - x(\lambda_2)\|^2,$$
$$h_{\lambda_2}(\mathbf{x}(\lambda_1)) \leq h_{\lambda_2}(\mathbf{x}(\lambda_2)) - \frac{\mu}{2}\|x(\lambda_1) - x(\lambda_2)\|^2,$$

which implies

$$-\langle \lambda_1, \mathbf{g}(\mathbf{x}(\lambda_2))\rangle - f_\mu(\mathbf{x}(\lambda_2)) \leq -\langle \lambda_1, \mathbf{g}(\mathbf{x}(\lambda_1))\rangle - f_\mu(\mathbf{x}(\lambda_1)) - \frac{\mu}{2}\|x(\lambda_1) - x(\lambda_2)\|^2,$$
$$-\langle \lambda_2, \mathbf{g}(\mathbf{x}(\lambda_1))\rangle - f_\mu(\mathbf{x}(\lambda_1)) \leq -\langle \lambda_2, \mathbf{g}(\mathbf{x}(\lambda_2)\rangle - f_\mu(\mathbf{x}(\lambda_2)) - \frac{\mu}{2}\|x(\lambda_1) - x(\lambda_2)\|^2.$$

Adding the two inequalities gives

$$\begin{aligned}\mu\|x(\lambda_1)) - x(\lambda_2))\|^2 \leq& \langle \lambda_1 - \lambda_2, \mathbf{g}(\mathbf{x}(\lambda_1)) - \mathbf{g}(\mathbf{x}(\lambda_2))\rangle \\ \leq& \|\lambda_1 - \lambda_2\| \cdot \|\mathbf{g}(\mathbf{x}(\lambda_1)) - \mathbf{g}(\mathbf{x}(\lambda_2))\| \\ \leq& G\|\lambda_1 - \lambda_2\| \cdot \|x(\lambda_1)) - x(\lambda_2))\|,\end{aligned}$$

where the last inequality follows from Lipschitz continuity of $\mathbf{g}(\mathbf{x})$ again. This implies

$$\|x(\lambda_1) - x(\lambda_2)\| \leq \frac{G}{\mu}\|\lambda_1 - \lambda_2\|.$$

Combining this inequality with (19) gives

$$\|\nabla d_\mu(\lambda_1) - \nabla d_\mu(\lambda_2)\| \leq \frac{G^2}{\mu}\|\lambda_1 - \lambda_2\|,$$

finishing the first part of the proof. The second part of the claim follows easily from the fact that $\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{y}\| \leq \sqrt{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}\|\mathbf{x} - \mathbf{y}\|$. □

6.2 Proof of Theorem 3.1

In this section, we give a convergence time proof of each stage. As a preliminary, we have the following basic lemma which bounds the perturbation of the Lagrange dual due to the primal smoothing.

Lemma 6.2. *Let $F(\lambda)$ and $F_\mu(\lambda)$ be functions defined in* (6) *and* (11)*, respectively. Then, we have for any $\lambda \in \mathbb{R}^N$,*

$$0 \leq F(\lambda) - F_\mu(\lambda) \leq \mu D^2/2$$

and

$$0 \leq F(\lambda^*) - F_\mu(\lambda_\mu^*) \leq \mu D^2/2,$$

for any $\lambda^* \in \Lambda^*$ and $\lambda_\mu^* \in \Lambda_\mu^*$.

Proof of Lemma 6.2. First of all, for any $\lambda \in \mathbb{R}^N$, define

$$\begin{aligned} h(\mathbf{x}) &:= -\langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle - f(\mathbf{x}), \\ h_\mu(\mathbf{x}) &:= -\langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle - f_\mu(\mathbf{x}). \end{aligned}$$

Then, let

$$\begin{aligned} \mathbf{x}(\lambda) &\in \text{argmax}_{\mathbf{x}\in\mathcal{X}}\ h(\mathbf{x}), \\ \mathbf{x}_\mu(\lambda) &\in \text{argmax}_{\mathbf{x}\in\mathcal{X}}\ h_\mu(\mathbf{x}), \end{aligned}$$

and we have for any $\lambda \in \mathbb{R}^N$,

$$\begin{aligned} F(\lambda) - F_\mu(\lambda) =& h(\mathbf{x}(\lambda)) - h_\mu(\mathbf{x}_\mu(\lambda)) \\ =& h(\mathbf{x}(\lambda)) - h_\mu(\mathbf{x}(\lambda)) + h_\mu(\mathbf{x}(\lambda)) - h_\mu(\mathbf{x}_\mu(\lambda)) \\ \leq& h(\mathbf{x}(\lambda)) - h_\mu(\mathbf{x}(\lambda)) \\ =& f_\mu(\mathbf{x}(\lambda)) - f(\mathbf{x}(\lambda)) \leq \mu D^2/2, \end{aligned}$$

where the first inequality follows from the fact that $\mathbf{x}_\mu(\lambda)$ maximizes $h_\mu(\lambda)$. Similarly, we have

$$\begin{aligned} F_\mu(\lambda) - F(\lambda) =& h_\mu(\mathbf{x}_\mu(\lambda)) - h(\mathbf{x}(\lambda)) \\ =& h_\mu(\mathbf{x}_\mu(\lambda)) - h(\mathbf{x}_\mu(\lambda)) + h(\mathbf{x}_\mu(\lambda)) - h(\mathbf{x}(\lambda)) \\ \leq& h_\mu(\mathbf{x}_\mu(\lambda)) - h(\mathbf{x}_\mu(\lambda)) \\ =& f(\mathbf{x}(\lambda)) - f_\mu(\mathbf{x}(\lambda)) \leq 0, \end{aligned}$$

where the first inequality follows from the fact that $\mathbf{x}(\lambda)$ maximizes $h(\lambda)$. Furthermore, we have

$$\begin{aligned} F(\lambda^*) - F_\mu(\lambda_\mu^*) &= F(\lambda^*) - F(\lambda_\mu^*) + F(\lambda_\mu^*) - F_\mu(\lambda_\mu^*) \leq F(\lambda_\mu^*) - F_\mu(\lambda_\mu^*) \leq \mu D^2/2, \\ F_\mu(\lambda_\mu^*) - F(\lambda^*) &= F_\mu(\lambda_\mu^*) - F_\mu(\lambda^*) + F_\mu(\lambda^*) - F(\lambda^*) \leq F_\mu(\lambda^*) - F(\lambda^*) \leq 0, \end{aligned}$$

finishing the proof. □

To prove Theorem 3.1, we start by rewriting the primal-dual smoothing algorithm (Algorithm 1) as the Nesterov's accelerated gradient algorithm on the smoothed dual function $F_\mu(\lambda)$: For any $t = 0, 1, \cdots, T-1$,

$$\begin{cases} \widehat{\lambda}_t = \lambda_t + \theta_t(\theta_{t-1}^{-1} - 1)(\lambda_t - \lambda_{t-1}) \\ \lambda_{t+1} = \widehat{\lambda}_t - \mu\nabla F_\mu(\widehat{\lambda}_t) \\ \theta_{t+1} = \frac{\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2}{2} \end{cases} \tag{20}$$

where we use Danskin's Theorem to claim that $\nabla F_\mu(\widehat{\lambda}_t) = \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t)$. As $t \to \infty$, we have $\frac{\theta_t}{\theta_{t-1}} = \sqrt{1-\theta_t} \to 1$. Classical results on the convergence time of accelerated gradient methods are as follows:

Theorem 6.1 (Theorem 1 of Tseng (2010))**.** *Consider the algorithm* (20) *starting from $\lambda_0 = \lambda_{-1} = \widetilde{\lambda}$. For any $\lambda \in \mathbb{R}^N$, we have*

$$F_\mu(\lambda_t) \leq F_\mu(\lambda) + \theta_{t-1}^2 \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})\|\lambda - \widetilde{\lambda}\|^2}{\mu}, \tag{21}$$

Furthermore, for any slot $t \in \{0, 1, 2, \cdots, T-1\}$,

$$F_\mu(\lambda_{t+1}) \leq (1-\theta_t)\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda_t - \widehat{\lambda}_t \right\rangle\right) + \theta_t \left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle\right) + \frac{\theta_t^2 \sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_t\|^2 - \|\lambda - \mathbf{z}_{t+1}\|^2\right), \quad (22)$$

where $\mathbf{z}_t = -(\theta_t^{-1} - 1)\lambda_t + \theta_t^{-1}\widehat{\lambda}_t$.

This theorem bounds the convergence time of the dual function. Our goal is to pass this dual convergence result to that of primal objective and constraint. Specifically, we aim to show the following primal objective bound and constraint violation:

To prove Theorem 3.1, we start by proving the following bound:

Lemma 6.3. *Consider running Algorithm 1 with a given initial condition $\widetilde{\lambda}$ in $\mathbb{R}^N$. For any $\lambda \in \mathbb{R}^N$, we have*

$$f_\mu(\overline{\mathbf{x}}_T) - \langle \mathbf{b} - \mathbf{A}\overline{\mathbf{x}}_T, \lambda\rangle - f_\mu^* \leq \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\left(\|\lambda - \widetilde{\lambda}\|^2 - \|\lambda - \mathbf{z}_T\|^2\right), \quad (23)$$

where $\mathbf{z}_T$ is defined in Theorem 6.1,

$$f_\mu^* := \min_{\mathbf{A}\mathbf{x}-\mathbf{b}=0,\ \mathbf{x}\in\mathcal{X}} f_\mu(x), \ \ \overline{\mathbf{x}}_T := \frac{1}{S_T}\sum_{t=0}^{T-1}\frac{\mathbf{x}(\widehat{\lambda}_t)}{\theta_t},$$

Proof of Lemma 6.3. First, subtracting $F_\mu(\lambda_\mu^*)$ from both sides of (22) in Theorem 6.1, we have for any $\lambda \in \mathbb{R}^N$ and any $t \in \{0, 1, 2, \cdots, T-1\}$,

$$\begin{aligned}
F_\mu(\lambda_{t+1}) - F_\mu(\lambda_\mu^*) \leq & (1-\theta_t)\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda_t - \widehat{\lambda}_t \right\rangle - F_\mu(\lambda_\mu^*)\right) \\
& + \theta_t\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle - F_\mu(\lambda_\mu^*)\right) \\
& + \frac{\theta_t^2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_t\|^2 - \|\lambda - \mathbf{z}_{t+1}\|^2\right) \\
\leq & (1-\theta_t)\left(F_\mu(\lambda_t) - F_\mu(\lambda_\mu^*)\right) + \theta_t\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle - F_\mu(\lambda_\mu^*)\right) \\
& + \frac{\theta_t^2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_t\|^2 - \|\lambda - \mathbf{z}_{t+1}\|^2\right),
\end{aligned}$$

where the second inequality follows from the convexity of F_μ that $F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda_t - \widehat{\lambda}_t \right\rangle \leq F_\mu(\lambda_t)$. Dividing θ_t^2 from both sides gives $\forall t \geq 1$,

$$\begin{aligned}
\frac{1}{\theta_t^2}\left(F_\mu(\lambda_{t+1}) - F_\mu(\lambda_\mu^*)\right) \leq & \frac{1-\theta_t}{\theta_t^2}\left(F_\mu(\lambda_t) - F_\mu(\lambda_\mu^*)\right) + \frac{1}{\theta_t}\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle - F_\mu(\lambda_\mu^*)\right) \\
& + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_t\|^2 - \|\lambda - \mathbf{z}_{t+1}\|^2\right) \\
= & \frac{1}{\theta_{t-1}^2}\left(F_\mu(\lambda_t) - F_\mu(\lambda_\mu^*)\right) + \frac{1}{\theta_t}\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle - F_\mu(\lambda_\mu^*)\right) \\
& + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_t\|^2 - \|\lambda - \mathbf{z}_{t+1}\|^2\right), \quad (24)
\end{aligned}$$

where the last equality uses the identity $(1-\theta_t)/\theta_t^2 = 1/\theta_{t-1}^2$. On the other hand, applying equation (24) at $t = 0$ and using $\theta_0 = \theta_{-1} = 1$ gives $(1-\theta_0)/\theta_0^2 = 0$ and

$$\begin{aligned}
\frac{1}{\theta_0^2}\left(F_\mu(\lambda_1) - F_\mu(\lambda_\mu^*)\right) \leq & \frac{1}{\theta_0}\left(F_\mu(\widehat{\lambda}_0) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_0 \right\rangle - F_\mu(\lambda_\mu^*)\right) \\
& + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_0\|^2 - \|\lambda - \mathbf{z}_1\|^2\right).
\end{aligned}$$

Taking telescoping sums from both sides from $t = 0$ to $t = T - 1$ gives

$$0 \leq \frac{1}{\theta_{T-1}^2}\left(F_\mu(\lambda_T) - F_\mu(\lambda_\mu^*)\right) \leq \sum_{t=0}^{T-1} \frac{1}{\theta_t}\left(F_\mu(\widehat{\lambda}_t) + \left\langle \nabla F_\mu(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle - F_\mu(\lambda_\mu^*)\right) + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_0\|^2 - \|\lambda - \mathbf{z}_T\|^2\right).$$

By Assumption 2.1(a), the feasible set $\{\mathbf{A}\mathbf{x} - \mathbf{b} = 0\}$ is not empty, and thus, strong duality holds for problem

$$\min_{\mathbf{A}\mathbf{x}-\mathbf{b}=0,\ \mathbf{x}\in\mathcal{X}} f_\mu(x)$$

(See, for example Proposition 5.3.1 of Bertsekas (2009)), and we have $F_\mu(\lambda_\mu^*) = -f_\mu^*$. Since

$$\nabla F_\mu(\widehat{\lambda}_t) = \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t), \quad F_\mu(\widehat{\lambda}_t) = \left\langle \widehat{\lambda}_t, \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t) \right\rangle - f_\mu(\mathbf{x}(\widehat{\lambda}_t)),$$

it follows,

$$\begin{aligned} 0 &\leq \sum_{t=0}^{T-1} \frac{1}{\theta_t}\left(\left\langle \widehat{\lambda}_t, \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t)\right\rangle - f_\mu(\mathbf{x}(\widehat{\lambda}_t)) + \left\langle \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t), \lambda - \widehat{\lambda}_t \right\rangle + f_\mu^*\right) \\ &\quad + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_0\|^2 - \|\lambda - \mathbf{z}_T\|^2\right) \\ &= \sum_{t=0}^{T-1} \frac{1}{\theta_t}\left(-f_\mu(\mathbf{x}(\widehat{\lambda}_t)) + \left\langle \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t), \lambda \right\rangle + f_\mu^*\right) + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu}\left(\|\lambda - \mathbf{z}_0\|^2 - \|\lambda - \mathbf{z}_T\|^2\right) \end{aligned}$$

Rearranging the terms and divding $S_T = \sum_{t=0}^{T-1} \frac{1}{\theta_t}$ from both sides,

$$\frac{1}{S_T}\sum_{t=0}^{T-1} \frac{1}{\theta_t}\left(f_\mu(\mathbf{x}(\widehat{\lambda}_t)) - \left\langle \mathbf{b} - \mathbf{A}\mathbf{x}(\widehat{\lambda}_t), \lambda \right\rangle - f_\mu^*\right) \leq \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\left(\|\lambda - \mathbf{z}_0\|^2 - \|\lambda - \mathbf{z}_T\|^2\right).$$

Note that $\mathbf{z}_0 = \widetilde{\lambda}$ by the definition of $\mathbf{z}_t$. By Jensen's inequality, we can move the weighted average inside the function f_μ and finish the proof. □

Proof of Theorem 3.1. First of all, we have by definition of Λ_μ^* in (12) and strong duality, for any $\lambda_\mu^* \in \Lambda_\mu^*$,

$$f_\mu(\overline{\mathbf{x}}_T) + \left\langle \mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}, \lambda_\mu^* \right\rangle \geq f_\mu^*.$$

Substituting this bound into (23) gives

$$\left\langle \mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}, \lambda - \lambda_\mu^* \right\rangle \leq \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\left(\|\lambda - \widetilde{\lambda}\|^2 - \|\lambda - \mathbf{z}_T\|^2\right) \leq \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\|\lambda - \widetilde{\lambda}\|^2.$$

Since this holds for any $\lambda \in \mathbb{R}^N$, the following holds:

$$\max_{\lambda\in\mathbb{R}^N}\left[\left\langle \mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}, \lambda - \lambda_\mu^* \right\rangle - \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\|\lambda - \widetilde{\lambda}\|^2\right] \leq 0.$$

The maximum is attained at $\lambda = \widetilde{\lambda} + \frac{\mu S_T}{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}\left(\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\right)$, which implies,

$$\begin{aligned} &\left\langle \mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}, \widetilde{\lambda} - \lambda_\mu^* \right\rangle + \frac{\mu S_T}{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}\|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\|^2 \leq 0. \\ \Rightarrow &\left\langle \mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}, \mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda} - \lambda_\mu^*\right) \right\rangle + \frac{\mu S_T}{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}\|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\|^2 \leq 0, \end{aligned}$$

where we used the fact that $\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b} = \mathcal{P}_{\mathbf{A}}(\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b})$ because $\mathbf{b}$ is in the column space of $\mathbf{A}$. By Cauchy-Schwarz inequality, we have

$$\begin{aligned} \frac{\mu S_T}{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}\|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\|^2 &\leq \|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\| \cdot \|\mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda} - \lambda_\mu^*\right)\| \\ \Rightarrow \|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\| &\leq \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu S_T}\|\mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda} - \lambda_\mu^*\right)\|. \end{aligned}$$

Let $\widetilde{\lambda}^* = \text{argmin}_{\lambda^* \in \Lambda^*} \|\lambda^* - \widetilde{\lambda}\|$, by triangle inequality,

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\| \leq & \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu S_T}\left(\|\mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda} - \widetilde{\lambda}^*\right)\| + \|\mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda}^* - \lambda_\mu^*\right)\|\right) \\ \leq & \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu S_T}\left(\|\widetilde{\lambda} - \widetilde{\lambda}^*\| + \|\mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda}^* - \lambda_\mu^*\right)\|\right),\end{aligned}$$

where the second inequality follows from the non-expansiveness of the projection. Now we look at the second term on the right hand side of the above inequality, Using Assumption 2.1(d), there exists a unique vector ν^* such that $\mathcal{P}_{\mathbf{A}}\lambda^* = \nu^*,\ \forall \lambda^* \in \Lambda^*$. Thus,

$$\begin{aligned}\|\mathcal{P}_{\mathbf{A}}\left(\widetilde{\lambda}^* - \lambda_\mu^*\right)\| = \|\nu^* - \mathcal{P}_{\mathbf{A}}\lambda_\mu^*\| = & \min_{\lambda^* \in \Lambda : \mathcal{P}_{\mathbf{A}}\lambda^* = \nu^*} \|\mathcal{P}_{\mathbf{A}}\left(\lambda^* - \lambda_\mu^*\right)\| \\ & \leq \min_{\lambda^* \in \mathbb{R}^N : \mathcal{P}_{\mathbf{A}}\lambda^* = \nu^*} \|\lambda^* - \lambda_\mu^*\| = \text{dist}(\lambda_\mu^*, \Lambda^*).\end{aligned}$$

Thus, we get the constraint violation bound

$$\|\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b}\| \leq \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu S_T}\left(\|\widetilde{\lambda} - \widetilde{\lambda}^*\| + \text{dist}(\lambda_\mu^*, \Lambda^*)\right).$$

To get the objective suboptimality bound, we start from (23) again. Substituting $\lambda = \widetilde{\lambda}^* = \text{argmin}_{\lambda^* \in \Lambda^*} \|\lambda^* - \widetilde{\lambda}\|$ into (23) gives

$$f_\mu(\overline{\mathbf{x}}_T) - \left\langle \mathbf{b} - \mathbf{A}\overline{\mathbf{x}}_T, \widetilde{\lambda}^* \right\rangle - f_\mu^* \leq \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\left(\|\widetilde{\lambda}^* - \widetilde{\lambda}\|^2 - \|\widetilde{\lambda}^* - \mathbf{z}_T\|^2\right) \leq \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\|\widetilde{\lambda}^* - \widetilde{\lambda}\|^2.$$

By Cauchy-Schwarz inequality and the fact that $\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b} = \mathcal{P}_{\mathbf{A}}(\mathbf{A}\overline{\mathbf{x}}_T - \mathbf{b})$, we have

$$f_\mu(\overline{\mathbf{x}}_T) - f_\mu^* \leq \|\mathbf{b} - \mathbf{A}\overline{\mathbf{x}}_T\|\|\mathcal{P}_{\mathbf{A}}\widetilde{\lambda}^*\| + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\|\widetilde{\lambda}^* - \widetilde{\lambda}\|^2.$$

By the fact that $f(\overline{\mathbf{x}}_T) \leq f_\mu(\overline{\mathbf{x}}_T) \leq f(\overline{\mathbf{x}}_T) + \frac{\mu}{2}D^2$, and the fact that $-f_\mu^* = F_\mu(\lambda_\mu^*) \geq F(\lambda^*) - \frac{\mu}{2}D^2 = -f^* - \frac{\mu}{2}D^2$ (from Lemma 6.2), we obtain

$$f(\overline{\mathbf{x}}_T) - f^* \leq \|\mathbf{b} - \mathbf{A}\overline{\mathbf{x}}_T\|\|\mathcal{P}_{\mathbf{A}}\widetilde{\lambda}^*\| + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu S_T}\|\widetilde{\lambda}^* - \widetilde{\lambda}\|^2 + \frac{\mu}{2}D^2,$$

finishing the proof. □

6.3 Proof of Theorem 3.2

In this section, we give an analysis of the proposed homotopy method building upon the previous results on the primal-dual smoothing. Our improved convergence time analysis under such a homotopy method is built upon previous results, notably the following lemma:

Lemma 6.4 (Yang and Lin (2015))**.** *Consider any convex function $F : \mathbb{R}^N \to \mathbb{R}$ such that the set of optimal points Λ^* defined in* (8) *is non-empty. Then, for any $\lambda \in \mathbb{R}^N$ and any $\varepsilon > 0$,*

$$\|\lambda - \lambda_\varepsilon^\dagger\| \leq \frac{dist(\lambda_\varepsilon^\dagger, \Lambda^*)}{\varepsilon}\left(F(\lambda) - F(\lambda_\varepsilon^\dagger)\right),$$

where $\lambda_\varepsilon^\dagger := argmin_{\lambda_\varepsilon \in \mathcal{S}_\varepsilon}\|\lambda - \lambda_\varepsilon\|$, and $\mathcal{S}_\varepsilon$ is the ε-sublevel set defined in (7)*.*

We start with the following easy corollary of Theorem 6.1.

Corollary 6.1. *Suppose $\{\lambda_t\}_{t=0}^T$ is the sequence produced by Algorithm 1 with the initial condition $\lambda_0 = \lambda_{-1} = \widetilde{\lambda}$, then, for any $\lambda \in \mathbb{R}^N$, we have*

$$F(\lambda_t) \leq F(\lambda) + \theta_{t-1}^2\frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})\|\lambda - \widetilde{\lambda}\|^2}{\mu} + \frac{D^2}{2}\mu, \tag{25}$$

The proof of this corollary is obvious combining (21) of Theorem 6.1 with Lemma 6.2.

The following result, which bounds the convergence time of the dual function, is proved via induction.

Lemma 6.5. *Suppose the assumptions in Theorem 3.2 hold. Let* $\left\{\lambda^{(k)}\right\}_{k=0}^{K}$ *be generated from Algorithm 2. For any* $k = 0, 1, 2, \cdots, K$*, we have*

$$F(\lambda^{(k)}) - F^* \leq \varepsilon_k + \varepsilon,$$

where $\varepsilon_k = \varepsilon_0/2^k$.

Proof of Lemma 6.5. First of all, for $k = 0$, we have $\lambda^{(0)} = 0$ and

$$F(\lambda^{(0)}) = -\max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \leq M,$$

thus, $F(\lambda^{(0)}) - F^* \leq 2M \leq \varepsilon_0 + \varepsilon$, by the assumption that $2M \leq \varepsilon_0$ in Theorem 3.2. Now for any $k > 0$, let $\lambda_\varepsilon^{(k-1)} \in \mathcal{S}_\varepsilon$ be the closest point to $\lambda^{(k-1)}$ specified in Algorithm 2, i.e. $\lambda_\varepsilon^{(k-1)} = \operatorname{argmin}_{\lambda_\varepsilon\in\mathcal{S}_\varepsilon}\|\lambda_\varepsilon - \lambda^{(k-1)}\|$. Suppose the claim holds for $(k-1)$-th stage, where $k > 0$, then, consider the k-th stage.

1. If $F(\lambda^{(k-1)}) - F^* \leq \varepsilon$, then, $\lambda^{(k-1)} \in \mathcal{S}_\varepsilon$, thus, $\|\lambda_\varepsilon^{(k-1)} - \lambda^{(k-1)}\| = 0$. By (25) with $\widetilde{\lambda} = \lambda^{(k-1)}$ from Algorithm 2 and λ chosen to be $\lambda_\varepsilon^{(k-1)}$, we have

   $$F(\lambda^{(k)}) - F(\lambda_\varepsilon^{(k-1)}) \leq \frac{D^2}{2}\mu_k \leq \frac{\varepsilon_k}{2},$$

   Thus, it follows, $F(\lambda^{(k)}) - F^* = F(\lambda^{(k)}) - F(\lambda_\varepsilon^{(k-1)}) + F(\lambda_\varepsilon^{(k-1)}) - F^* \leq \varepsilon_k + \varepsilon$.

2. If $F(\lambda^{(k-1)}) - F^* > \varepsilon$, then, $\lambda^{(k-1)} \notin \mathcal{S}_\varepsilon$ and we claim that

   $$F(\lambda_\varepsilon^{(k-1)}) - F^* = \varepsilon. \tag{26}$$

   Indeed, suppose on the contrary, $F(\lambda_\varepsilon^{(k-1)}) - F^* < \varepsilon$, then, by the continuity of the function F, there exists $\alpha \in (0,1)$ and $\lambda' = \alpha\lambda_\varepsilon^{(k-1)} + (1-\alpha)\lambda^{(k-1)}$ such that $F(\lambda') - F^* = \varepsilon$, i.e. $\lambda' \in \mathcal{S}_\varepsilon$, and $\|\lambda^{(k-1)} - \lambda'\| = \alpha\|\lambda^{(k-1)} - \lambda_\varepsilon^{(k-1)}\| < \|\lambda^{(k-1)} - \lambda_\varepsilon^{(k-1)}\|$, contradicting the fact that $\lambda_\varepsilon^{(k-1)} = \operatorname{argmin}_{\lambda_\varepsilon\in\mathcal{S}_\varepsilon}\|\lambda_\varepsilon - \lambda^{(k-1)}\|$.

   On the other hand, by induction hypothesis, we have

   $$F(\lambda^{(k-1)}) - F^* \leq \varepsilon_{k-1} + \varepsilon,$$

   which, combining with (26), implies $F(\lambda^{(k-1)}) - F(\lambda_\varepsilon^{(k-1)}) \leq \varepsilon_{k-1}$, and by Lemma 6.4,

   $$\begin{aligned}\|\lambda^{(k-1)} - \lambda_\varepsilon^{(k-1)}\| &\leq \frac{\operatorname{dist}(\lambda_\varepsilon^{(k-1)}, \Lambda^*)}{\varepsilon}\left(F(\lambda^{(k-1)}) - F(\lambda_\varepsilon^{(k-1)})\right)\\ &\leq \frac{C_\delta\left(F(\lambda_\varepsilon^{(k-1)}) - F^*\right)^\beta\left(F(\lambda^{(k-1)}) - F(\lambda_\varepsilon^{(k-1)})\right)}{\varepsilon} \leq \frac{C_\delta\varepsilon_{k-1}}{\varepsilon^{1-\beta}},\end{aligned}$$

   where the second inequality follows from $\varepsilon \leq \delta$ assumed in Theorem 3.2 and the local error bound condition (9). Note that by definition of θ_t in Algorithm 1, $\frac{1}{\theta_{T-1}^2} \geq T^2 \geq \frac{4D^2C_\delta^2\sigma_{\max}(\mathbf{A}^T\mathbf{A})(2M)^\beta}{\varepsilon^{4/(2+\beta)}}$, and $\mu_k = \varepsilon_k/D^2$. Substituting these quantities into (25) with $\widetilde{\lambda} = \lambda^{(k-1)}$ and λ chosen to be $\lambda_\varepsilon^{(k-1)}$, we have

   $$\begin{aligned}F(\lambda^{(k)}) - F(\lambda_\varepsilon^{(k-1)}) &\leq \frac{D^2}{2}\mu_k + \theta_{T-1}^2\frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})\|\lambda_\varepsilon^{(k-1)} - \lambda^{(k-1)}\|^2}{\mu_k}\\ &\leq \frac{\varepsilon_k}{2} + \frac{\varepsilon^{4/(2+\beta)}}{2(2M)^\beta\varepsilon^{2(1-\beta)}}\varepsilon_k = \frac{\varepsilon_k}{2} + \frac{\varepsilon^{\frac{2\beta(1+\beta)}{2+\beta}}}{2(2M)^\beta}\varepsilon_k\\ &\leq \frac{\varepsilon_k}{2}\left(1 + \left(\frac{\varepsilon}{2M}\right)^\beta\right) \leq \varepsilon_k,\end{aligned}$$

   where the second from the last inequality follows from $\varepsilon \leq 1$ and the last inequality follows from $\varepsilon \leq 2M$ assumed in Theorem 3.2. Thus, it follows $F(\lambda^{(k)}) - F^* \leq \varepsilon_k + \varepsilon$.

Overall, we finish the proof. □

Proof of Theorem 3.2. Since the desired accuracy is chosen small enough so that $\varepsilon \leq \frac{\delta}{2}$, and the number of stages $K \geq \lceil \log_2(\varepsilon_0/\varepsilon) \rceil + 1$, it follows $\varepsilon_{K-1} \leq \varepsilon \leq \frac{\delta}{2}$, and thus there exists some threshold $k' \in \{0, 1, 2, \cdots, K-1\}$ such that for any $k \geq k'$, $\varepsilon_k + \varepsilon \leq \delta$. As a consequence, by Lemma 6.5, we have for any $k \geq k'$,

$$F(\lambda^{(k)}) - F^* \leq \varepsilon_k + \varepsilon \leq \delta,$$

i.e. $\lambda^{(k)} \in \mathcal{S}_\delta$, the δ-sublevel set of the function $F(\lambda)$. By the local error bound condition (9), we have

$$\text{dist}(\lambda^{(k)}, \Lambda^*) \leq \left(F(\lambda^{(k)}) - F^*\right)^\beta \leq (\varepsilon_k + \varepsilon)^\beta.$$

Now, consider the $(k+1)$-th stage in the homotopy method. By (14) in Theorem 3.1,

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}}^{(k+1)} - \mathbf{b}\| &\leq \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu_{k+1}S_T}\left(\|\lambda_*^{(k)} - \lambda^{(k)}\| + \text{dist}(\lambda^*_{\mu_{k+1}}, \Lambda^*)\right) \\ &\leq \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu_{k+1}S_T}\left((\varepsilon_k + \varepsilon)^\beta + \text{dist}(\lambda^*_{\mu_{k+1}}, \Lambda^*)\right), \quad (27)\end{aligned}$$

where $\lambda_*^{(k)} = \text{argmin}_{\lambda^* \in \Lambda^*}\|\lambda^* - \lambda^{(k)}\|$, and the second inequality follows from

$$\|\lambda_*^{(k)} - \lambda^{(k)}\| = \text{dist}(\lambda^{(k)}, \Lambda^*) \leq (\varepsilon_k + \varepsilon)^\beta. \quad (28)$$

To bound the second term on the right hand side of (27), note that $\mu_{k+1} = \varepsilon_{k+1}/D^2 = \varepsilon_k/(2D^2) \leq \delta/(2D^2)$. Thus, by Lemma 6.2,

$$\begin{aligned}F(\lambda^*_{\mu_{k+1}}) - F(\lambda^*) &= F(\lambda^*_{\mu_{k+1}}) - F_{\mu_{k+1}}(\lambda^*_{\mu_{k+1}}) + F_{\mu_{k+1}}(\lambda^*_{\mu_{k+1}}) - F(\lambda^*) \\ &\leq \frac{\mu_{k+1}}{2}D^2 + 0 = \mu_{k+1}D^2/2 \leq \delta/2,\end{aligned}$$

thus, it follows $\lambda^*_{\mu_{k+1}} \in \mathcal{S}_\delta$ and by local error bound condition

$$\text{dist}(\lambda^*_{\mu_{k+1}}, \Lambda^*) \leq C_\delta\left(F(\lambda^*_{\mu_{k+1}}) - F(\lambda^*)\right)^\beta \leq C_\delta\left(\varepsilon_k + \varepsilon\right)^\beta.$$

Overall, substituting this bound into (27) ,we get

$$\|\mathbf{A}\overline{\mathbf{x}}^{(k+1)} - \mathbf{b}\| \leq \frac{2\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{\mu_{k+1}S_T}\left(1 + C_\delta\right)\left(\varepsilon_k + \varepsilon\right)^\beta \leq \frac{4\sigma_{\max}(\mathbf{A}^T\mathbf{A})D^2}{\varepsilon_{k+1}T^2}\left(1 + C_\delta\right)\left(\varepsilon_k + \varepsilon\right)^\beta,$$

where we use the fact that $\mu_{k+1} = \varepsilon_{k+1}/D^2$ and $S_T = \sum_{t=0}^{T-1}\frac{1}{\theta_t} \geq \sum_{t=1}^{T} t \geq \frac{T^2}{2}$. Substituting the bound $T^2 \geq \frac{4D^2C_\delta^2\sigma_{\max}(\mathbf{A}^T\mathbf{A})(2M)^\beta}{\varepsilon^{4/(2+\beta)}}$ gives for any $k \geq k'$,

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}}^{(k+1)} - \mathbf{b}\| &\leq \frac{1 + C_\delta}{C_\delta^2(2M)^\beta}\frac{(\varepsilon_k + \varepsilon)^\beta\varepsilon^{4/(2+\beta)}}{\varepsilon_{k+1}} = \frac{2(1 + C_\delta)}{C_\delta^2(2M)^\beta}\frac{(\varepsilon_k + \varepsilon)^\beta\varepsilon^{4/(2+\beta)}}{\varepsilon_k} \\ &\leq \frac{2(1 + C_\delta)}{C_\delta^2(2M)^\beta}\frac{(3\varepsilon_k)^\beta(4\varepsilon_k)^{4/(2+\beta)}}{\varepsilon_k} \leq \frac{24(1 + C_\delta)}{C_\delta^2(2M)^\beta}\varepsilon_k^{1+\frac{\beta^2}{2+\beta}}, \quad (29)\end{aligned}$$

where the equality follows from $\varepsilon_{k+1} = \varepsilon_k/2$, and the second inequality follows from $\varepsilon \leq 2\varepsilon_k$, $\forall k \in \{0, 1, 2, \cdots, K-1\}$. For the objective bound, we have by (13), for any $k \geq k'$,

$$\begin{aligned}f(\overline{\mathbf{x}}^{(k+1)}) - f^* \leq& \|\mathcal{P}_{\mathbf{A}}\lambda_0^*\| \cdot \|\mathbf{A}\overline{\mathbf{x}}^{(k+1)} - \mathbf{b}\| + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu_{k+1}S_T}\|\lambda_*^{(k)} - \lambda^{(k)}\|^2 + \frac{\mu_{k+1}D^2}{2} \\ \leq& \|\mathcal{P}_{\mathbf{A}}\lambda_0^*\|\frac{24(1 + C_\delta)}{C_\delta^2(2M)^\beta}\varepsilon_k^{1+\frac{\beta^2}{2+\beta}} + \frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu_{k+1}S_T}\|\lambda_*^{(k)} - \lambda^{(k)}\|^2 + \frac{\mu_{k+1}D^2}{2}, \quad (30)\end{aligned}$$

where the second inequality follows from (29). Now, for the second term on the right hand side, we have

$$\frac{\sigma_{\max}(\mathbf{A}^T\mathbf{A})}{2\mu_{k+1}S_T}\|\lambda_*^{(k)}-\lambda^{(k)}\|^2 \le \frac{\varepsilon^{4/(2+\beta)}(\varepsilon_k+\varepsilon)^{2\beta}}{4\varepsilon_{k+1}C_\delta^2(2M)^\beta} = \frac{\varepsilon^{4/(2+\beta)}(\varepsilon_k+\varepsilon)^{2\beta}}{2\varepsilon_k C_\delta^2(2M)^\beta}$$
$$\le \frac{(4\varepsilon_k)^{4/(2+\beta)}(3\varepsilon_k)^{2\beta}}{2\varepsilon_k C_\delta^2(2M)^\beta} \le \frac{6\varepsilon_k^{1+\frac{2\beta(1+\beta)}{2+\beta}}}{C_\delta^2(2M)^\beta},$$

where first inequality follows from (28), the equality follows from $\varepsilon_{k+1}=\varepsilon_k/2$, and the second inequality follows from $\varepsilon \le 2\varepsilon_k,\ \forall k \in \{0,1,2,\cdots,K-1\}$. Substituting this bound and $\mu_{k+1} = \varepsilon_{k+1}/D^2 = \varepsilon_k/2D^2$ into (30) gives for any $k \ge k'$,

$$f(\overline{\mathbf{x}}^{(k+1)}) - f^* \le \frac{24\|\mathcal{P}_{\mathbf{A}}\lambda_0^*\|(1+C_\delta)}{C_\delta^2(2M)^\beta}\varepsilon_k^{1+\frac{\beta^2}{2+\beta}} + \frac{6}{C_\delta^2(2M)^\beta}\varepsilon_k^{1+\frac{2\beta(1+\beta)}{2+\beta}} + \frac{1}{4}\varepsilon_k. \tag{31}$$

Taking $k = K-1$ in (29) and (31) with the fact that $\varepsilon_{K-1} \le \varepsilon \le 1$ gives the desired result. □

6.4 Proof of Lemma 4.1

Proof. For simplicity of notations, we let $\mathbf{x}_i^* = \mathbf{x}_i(\widehat{\lambda}_t)$. First of all, let $H_C(\mathbf{x}_i)$ be the indicator function for the set $C := \{\mathbf{x}_i :\ \|\mathbf{x}_i - \mathbf{b}_i\| \le D\}$, which takes 0 if $\mathbf{x}_i \in C$ and $+\infty$ otherwise. Then, the optimization problem (18) can be equivalently written as an unconstrained problem:

$$\mathbf{x}_i^* = \text{argmax}_{\mathbf{x}_i\in\mathbb{R}^d} -\frac{\mu}{2}\|\mathbf{x}_i - \mathbf{a}_i\|^2 - \|\mathbf{x}_i - \mathbf{b}_i\| - H_C(\mathbf{x}_i) =: g(\mathbf{x}_i), \tag{32}$$

where $\mathbf{a}_i = \widetilde{\mathbf{x}}_i - \frac{1}{\mu}\sum_{j\in\mathcal{N}_i}\mathbf{W}_{ji}\lambda_{t,j}$. Since $\mathbf{x}_i^*$ is the solution, by the optimality condition, $0 \in \partial g(\mathbf{x}_i^*)$, where $\partial g(\mathbf{x}_i^*)$ denotes the set of subdifferentials of g at point $\mathbf{x}_i^*$, i.e.

$$0 \in \mu(\mathbf{x}_i^* - \mathbf{a}_i) + \partial\|\mathbf{x}_i^* - b_i\| + \mathcal{N}_C(\mathbf{x}_i^*),$$

where for any $\mathbf{x} \in \mathbb{R}^d$,

$$\partial\|\mathbf{x}-\mathbf{b}_i\| = \begin{cases} \left\{\frac{\mathbf{x}-\mathbf{b}_i}{\|\mathbf{x}-\mathbf{b}_i\|}\right\}, & \text{if } \mathbf{x} \ne \mathbf{b}_i, \\ \left\{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\| \le 1\right\}, & \text{otherwise,} \end{cases}$$

and $\mathcal{N}_C(\mathbf{x})$ is the normal cone of the set $C = \{\mathbf{x}_i :\ \|\mathbf{x}_i - \mathbf{b}_i\| \le D\}$ at the point $\mathbf{x}$, i.e.

$$\mathcal{N}_C(\mathbf{x}) := \left\{\mathbf{v}\in\mathbb{R}^d :\ \mathbf{v}^T\mathbf{x} \ge \mathbf{v}^T\mathbf{y},\ \forall \mathbf{y}\in C\right\}.$$

This is equivalent to

$$-\mu(\mathbf{x}_i^* - \mathbf{a}_i) - \mathbf{h} \in \mathcal{N}_C(\mathbf{x}_i^*), \tag{33}$$

for some $\mathbf{h} \in \partial\|\mathbf{x}_i^* - \mathbf{b}_i\|$. Note that the function $g(\cdot)$ is a strongly concave function, thus, the solution to the maximization problem (32) is unique, which implies as long as one can find one x_i^* and $\mathbf{h}$ satisfying (33), such a x_i^* must be the only solution. To this point, we consider the following three cases:

1. If $\|\mathbf{b}_i - \mathbf{a}_i\| \le 1/\mu$. Let $\mathbf{x}_i^* = \mathbf{b}_i$ and $\mathbf{h} = \mu(\mathbf{a}_i - \mathbf{b}_i)$, then, $\mathcal{N}_C(\mathbf{x}_i^*) = \{0\}$ and $\|\mathbf{h}\| \le 1$ and $-\mu(\mathbf{x}_i^* - \mathbf{a}_i) - \mathbf{h} = 0 \in \mathcal{N}_C(\mathbf{x}_i^*)$.

2. If $1/\mu < \|\mathbf{b}_i - \mathbf{a}_i\| \le 1/\mu + D$, then, one can take

$$\mathbf{x}_i^* = \mathbf{b}_i - \frac{\mathbf{b}_i - \mathbf{a}_i}{\|\mathbf{b}_i - \mathbf{a}_i\|}\left(\|\mathbf{b}_i - \mathbf{a}_i\| - \frac{1}{\mu}\right) = \mathbf{a}_i + \frac{\mathbf{b}_i - \mathbf{a}_i}{\|\mathbf{b}_i - \mathbf{a}_i\|}\frac{1}{\mu}$$

and $\mathbf{h} = \frac{\mathbf{a}_i - \mathbf{b}_i}{\|\mathbf{a}_i - \mathbf{b}_i\|}$. Note that $\|\mathbf{x}_i^* - \mathbf{b}_i\| = \|\mathbf{a}_i - \mathbf{b}_i\| - 1/\mu \le D$, which again gives $\mathcal{N}_C(\mathbf{x}_i^*) = \{0\}$ and $-\mu(\mathbf{x}_i^* - \mathbf{a}_i) - \mathbf{h} = 0 \in \mathcal{N}_C(\mathbf{x}_i^*)$.

3. If $\|\mathbf{b}_i - \mathbf{a}_i\| > 1/\mu + D$. Then, let $\mathbf{x}_i^* = \mathbf{b}_i - \frac{\mathbf{b}_i - \mathbf{a}_i}{\|\mathbf{b}_i - \mathbf{a}_i\|} D$ and $\mathbf{h} = \frac{\mathbf{a}_i - \mathbf{b}_i}{\|\mathbf{a}_i - \mathbf{b}_i\|}$, which gives

$$\begin{aligned}
-\mu\left(\mathbf{x}_i^* - \mathbf{a}_i\right) - \mathbf{h} =& -\mu\left(\mathbf{b}_i - \mathbf{a}_i - \frac{\mathbf{b}_i - \mathbf{a}_i}{\|\mathbf{b}_i - \mathbf{a}_i\|} D\right) - \frac{\mathbf{a}_i - \mathbf{b}_i}{\|\mathbf{a}_i - \mathbf{b}_i\|} \\
=& -\mu(\mathbf{b}_i - \mathbf{a}_i)\left(1 - \frac{D}{\|\mathbf{b}_i - \mathbf{a}_i\|}\right) - \frac{\mathbf{a}_i - \mathbf{b}_i}{\|\mathbf{a}_i - \mathbf{b}_i\|} \\
=& -\mu(\mathbf{b}_i - \mathbf{a}_i)\left(1 - \frac{D + 1/\mu}{\|\mathbf{b}_i - \mathbf{a}_i\|}\right) = \mu\left(1 - \frac{D + 1/\mu}{\|\mathbf{b}_i - \mathbf{a}_i\|}\right)(\mathbf{a}_i - \mathbf{b}_i).
\end{aligned}$$

Note that the normal $\mathcal{N}_C(\mathbf{x}_i^*) = \{c(\mathbf{a}_i - \mathbf{b}_i), c \geq 0\}$, it follows $-\mu\left(\mathbf{x}_i^* - \mathbf{a}_i\right) - \mathbf{h} \in \mathcal{N}_C(\mathbf{x}_i^*)$.

Overall, we finish the proof. □

6.5 Proof of Theorem 4.1

Since the null space of $\mathbf{A}$ is non-empty and the set

$$\mathcal{X} := \left\{\mathbf{x} \in \mathbb{R}^{nd} : \ \|\mathbf{x}_i - \mathbf{b}_i\| \leq D, i = 1, 2, \cdots, n\right\}$$

is compact, strong duality holds with respect to (16-17). In view of Assumption 2.1(c)(d), we aim to show that the Lagrange dual of (16-17) satisfies the local error bound condition (9) and the set of optimal Lagrange multiplier is unique up to null space of $\mathbf{A}$.

We start by rewriting (16-17) as follows: Let $\mathbf{y}_i = \mathbf{x}_i - \mathbf{b}_i$, and $\mathbf{y} = [\mathbf{y}_1^T,\ \mathbf{y}_2^T, \cdots,\ \mathbf{y}_n^T]^T$, then, (16-17) is equivalent to

$$\begin{aligned}
\min\ & \sum_{i=1}^n \|\mathbf{y}_i\| \\
s.t.\ & \mathbf{A}\mathbf{y} + \mathbf{A}\mathbf{b} = 0, \|\mathbf{y}_i\| \leq D,\ i = 1, 2, \cdots, n.
\end{aligned}$$

Then, for any $\lambda \in \mathbb{R}^{nd}$, the Lagrange dual function

$$\begin{aligned}
F(\lambda) =& \max_{\|\mathbf{y}_i\| \leq D,\ i=1,2,\cdots,n} -\sum_{i=1}^n \|\mathbf{y}_i\| - \langle \lambda, \mathbf{A}\mathbf{y} + \mathbf{A}\mathbf{b}\rangle \\
=& \underbrace{\max_{\|\mathbf{y}_i\| \leq D,\ i=1,2,\cdots,n} -\sum_{i=1}^n \left(\|\mathbf{y}_i\| + \left\langle \lambda, \mathbf{A}_{[i]}\mathbf{y}_i\right\rangle\right) - \langle \lambda, \mathbf{A}\mathbf{b}\rangle}_{\text{(I)}},
\end{aligned}$$

where

$$\mathbf{A}_{[i]} = [\mathbf{W}_{1i}\ \mathbf{W}_{2i}\ \cdots\ \mathbf{W}_{ni}]^T$$

i-th column block of the matrix $\mathbf{A}$ corresponding to $\mathbf{y}_i$. Note that maximization of (I) is separable with respect to the index i, we have for any $i \in \{1, 2, \cdots, n\}$,

$$\begin{aligned}
\max_{\|\mathbf{y}_i\| \leq D} -\|\mathbf{y}_i\| - \left\langle \lambda, \mathbf{A}_{[i]}\mathbf{y}_i\right\rangle =& \max_{\|\mathbf{y}_i\| \leq D} -\|\mathbf{y}_i\| - \left\langle \mathbf{A}_{[i]}^T\lambda, \mathbf{y}_i\right\rangle \\
=& \begin{cases} 0, & \text{if } \|\mathbf{A}_{[i]}^T\lambda\| \leq 1 \\ (\|\mathbf{A}_{[i]}^T\lambda\| - 1) \cdot D, & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus, one can write $F(\lambda)$ as follows

$$F(\lambda) = -\left\langle \mathbf{A}^T\lambda, \mathbf{b}\right\rangle + D\sum_{i=1}^n (\|\mathbf{A}_{[i]}^T\lambda\| - 1) \cdot I\left(\|\mathbf{A}_{[i]}^T\lambda\| > 1\right), \tag{34}$$

where $I\left(\|\mathbf{A}_{[i]}^T\lambda\| > 1\right)$ is the indicator function which takes 1 if $\|\mathbf{A}_{[i]}^T\lambda\| > 1$ and 0 otherwise. To this point, we make another change of variables by setting $\nu_i = \mathbf{A}_{[i]}^T\lambda,\ i = 1, 2, \cdots, n$ and

$\nu = [\nu_1^T\ \nu_2^T\ \cdots\ \nu_n^T]^T$. Note that $\{\mathbf{A}^T\lambda : \lambda \in \mathbb{R}^{nd}\} = \mathcal{R}(\mathbf{A}^T)$. By the null space property (15), the range space of $\mathbf{A}^T$ has the following explicit representation:

$$\mathcal{R}(\mathbf{A}^T) = \left\{ \nu \in \mathbb{R}^{nd} :\ \mathbf{u} = [\nu_1^T, \cdots, \nu_n^T]^T,\ \sum_{i=1}^{n} \nu_i = 0 \right\}. \tag{35}$$

Thus, minimizing (34) is equivalent to solving the following constrained optimization problem:

$$\min_{\nu \in \mathbb{R}^{nd}} \ -\langle \nu, \mathbf{b} \rangle + D \sum_{i=1}^{n} (\|\nu_i\| - 1) \cdot I\left(\|\nu_i\| > 1\right), \tag{36}$$

$$s.t.\ \sum_{i=1}^{n} \nu_i = 0, \tag{37}$$

Denote

$$G(\nu) = -\langle \nu, \mathbf{b} \rangle + D \sum_{i=1}^{n} (\|\nu_i\| - 1) \cdot I\left(\|\nu_i\| > 1\right). \tag{38}$$

The following lemma, which characterizes the set of solutions to (36-37), paves the way of our analysis.

Lemma 6.6. *The solution to (36-37) is attained within the region:* $\mathcal{B} = \{\nu \in \mathbb{R}^{nd}, \|\nu_i\| \leq 1,\ \forall i\}$. *Furthermore, for any* $\nu' \in \mathbb{R}^{nd}$ *satisfying* (37) *but not in* $\mathcal{B}$*, there exists a point* $\overline{\nu}' \in \mathcal{B}$ *such that* (37) *is satisfied and*

$$G(\nu') - G(\overline{\nu}') \geq \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \|\nu' - \overline{\nu}'\|.$$

Proof of Lemma 6.6. Consider any $\nu' \in \mathbb{R}^{nd}$ not in the set $\mathcal{B}$, then, define the set $\mathcal{J}$ as the set of coordinates j in $\{1, ..., n\}$ such that $\|\nu_j'\| > 1$. Since ν' is not in the set $\mathcal{B}$, we know $\mathcal{J}$ is nonempty. Then, let $L := \max_{j \in \mathcal{J}} \|\nu_j'\| > 1$. Consider the vector $\overline{\nu}' := \nu'/L$, then, since ν' is a solution to (36-37), $\sum_{i=1}^{n} \nu_i' = 0$, which implies $\sum_{i=1}^{n} \overline{\nu}_i' = 0$. Furthermore, we obviously have $\|\overline{\nu}_i'\| \leq 1,\ \forall i$. Now, we are going to show that $G(\nu') > G(\overline{\nu}')$, thereby reaching a contradiction. Consider the

difference

$$
\begin{aligned}
&G(\nu') - G(\overline{\nu}') \\
&= \langle \overline{\nu}' - \nu', \mathbf{b} \rangle + D \sum_{i=1}^{n} (\|\nu_i'\| - 1) \cdot I\left(\|\nu_i'\| > 1\right) \\
&= \sum_{i=1}^{n-1} \langle \overline{\nu}_i' - \nu_i', \mathbf{b}_i - \mathbf{b}_n \rangle + D \sum_{i=1}^{n} (\|\nu_i'\| - 1) \cdot I\left(\|\nu_i'\| > 1\right) \quad \left(\text{by the fact } \sum_{i=1}^{n} \overline{\nu}_i' = \sum_{i=1}^{n} \nu_i' = 0 \right) \\
&\geq \sum_{i=1}^{n-1} \langle \overline{\nu}_i' - \nu_i', \mathbf{b}_i - \mathbf{b}_n \rangle + (L-1)D \\
&\geq - \sum_{i=1}^{n-1} \|\overline{\nu}_i' - \nu_i'\| \cdot \|\mathbf{b}_i - \mathbf{b}_n\| + (L-1)D \ \text{ (by Cauchy-Schwarz)} \\
&\geq - \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \sum_{i=1}^{n-1} \|\overline{\nu}_i' - \nu_i'\| + (L-1)D \\
&= - \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \sum_{i=1}^{n-1} \|\overline{\nu}_i'\|(L-1) + (L-1)D \ \text{ (By definition } \overline{\nu}' := \nu'/L) \\
&\geq - \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \sum_{i=1}^{n-1} \|\overline{\nu}_i'\|(L-1) + 2(L-1) \cdot n \cdot \max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \quad \left(D \geq 2n \cdot \max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \\
&\geq \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \sum_{i=1}^{n} \|\overline{\nu}_i'\|(L-1) \ \text{ (by the fact } \|\overline{\nu}_i'\| \leq 1) \\
&\geq \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \sum_{i=1}^{n} \|\nu_i' - \overline{\nu}_i'\| \geq \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\| \right) \|\nu' - \overline{\nu}'\|, \quad \text{(By definition } \overline{\nu}' := \nu'/L)
\end{aligned}
$$

and the lemma follows. □

By the previous lemma, in order to characterize the set of solutions to (36-37), it is enough to look at the following more restricted problem:

$$\min_{\nu \in \mathbb{R}^{nd}} \ -\langle \nu, \mathbf{b} \rangle, \tag{39}$$

$$s.t. \ \sum_{i=1}^{n} \nu_i = 0, \tag{40}$$

$$\|\nu_i\|^2 \leq 1, \ i = 1, 2, \cdots, n, \tag{41}$$

where we used the fact that $G(\nu) = -\langle \nu, \mathbf{b} \rangle$ when $\|\nu_i\|^2 \leq 1, \ \forall i$. This is a quadratic constrained problem. Now, we show the key lemma that $G(\nu)$ satisfies the local error bound with parameter $\beta = 1/2$ over the restricted set (40) and (41).

Lemma 6.7. *The solution to (39-41) is unique. Furthermore, let $\nu^* \in \mathbb{R}^{nd}$ be the solution to (39-41). There exists a constant $C_0 > 0$ such that for any $\nu \in \mathbb{R}^{nd}$ satisfying (40-41),*

$$\|\nu - \nu^*\| \leq C_0 \left(G(\nu) - G(\nu^*) \right)^{1/2}.$$

The proof of Lemma 6.7 is somewhat lengthy, but it follows a simple intuition that if the solution point lies on the boundary of a ball, then, sliding a point away from the solution results in a locally quadratic growth of the objective when it is linear. We split the proof into two cases below.

6.5.1 Proof of Lemma 6.7: Case 1

Case 1: The solution of the original geometric median (16-17) is achieved at one of the vectors $\{\mathbf{b}_1, \ \mathbf{b}_2, \ \cdots, \ \mathbf{b}_n\}$.

Assume without loss of generality that it is achieved at $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n = \mathbf{b}_n$, then, one know that the minimum of (16-17) is $\sum_{i=1}^{n-1} \|\mathbf{b}_i - \mathbf{b}_n\|$. Furthermore, since we assume $\{\mathbf{b}_1,\ \mathbf{b}_2,\ \cdots,\ \mathbf{b}_n\}$ is not co-linear, the solution is unique, and thus, for all feasible $\mathbf{x} \neq [\mathbf{b}_n^T,\ \mathbf{b}_n^T,\ \cdots,\ \mathbf{b}_n^T]^T$, $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{b}_i\| > \sum_{i=1}^{n-1} \|\mathbf{b}_n - \mathbf{b}_i\|$.

First, one can get rid of constraint (40) in (39-41) by substituting $\nu_n = -\sum_{i=1}^{n-1} \nu_i$ and equivalently form the following optimization problem:

$$\min_{\nu \in \mathbb{R}^{nd}} \quad -\sum_{i=1}^{n-1} \langle \nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle , \tag{42}$$

$$s.t.\ \ \|\nu_i\|^2 \leq 1,\ i = 1, 2, \cdots, n-1, \tag{43}$$

$$\left\| \sum_{i=1}^{n-1} \nu_i \right\| \leq 1. \tag{44}$$

Then, to show the uniqueness of the solution to (39-41), it is enough to show the solution to (42-44) is unique. To see the the uniqueness, suppose we temporarily delete constraint (44), then we obtain a relaxed problem:

$$\min_{\nu \in \mathbb{R}^{nd}} \quad -\sum_{i=1}^{n-1} \langle \nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle ,$$

$$s.t.\ \ \|\nu_i\|^2 \leq 1,\ i = 1, 2, \cdots, n-1,$$

which is separable and we know trivially that for each index i, the solution to

$$\min_{\nu_i \in \mathbb{R}^{d}} -\langle \nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle ,\ \ s.t.\ \|\nu_i\|^2 \leq 1,$$

is attained *uniquely* at $\nu_i^* = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|}$. This gives the objective value $-\sum_{i=1}^{n-1} \|\mathbf{b}_n - \mathbf{b}_i\|$ to the relaxed problem. On the other hand, by strong duality, the optimal objective of the original problem (39-41) is also $-\sum_{i=1}^{n-1} \|\mathbf{b}_n - \mathbf{b}_i\|$. The fact that the optimal objective does not change even when adding an extra constraint $\left\|\sum_{i=1}^{n-1} \nu_i\right\| \leq 1$ implies that $\nu_i^* = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|},\ i = 1, 2, \cdots, n-1$ is feasible with respect to (39-41), and the solution to (39-41) cannot be attained at any feasible point other than $\nu_i^* = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|},\ i = 1, 2, \cdots, n-1$. As a consequence, the solution to (39-41) is also unique, which is $\nu_i^* = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|},\ i = 1, 2, \cdots, n-1$ and $\nu_n^* = -\sum_{i=1}^{n-1} \nu_i$.

Next, we are going to show a local error bound condition for (42-44), and then pass the result back to (39-41). To this point, we consider any perturbation $\Delta\nu = [\Delta\nu_1^T,\ \Delta\nu_2^T,\ \cdots,\ \Delta\nu_n^T]^T$ around the solution to (42-44) so that $\nu^* + \Delta\nu$ is within the feasible set $\left\{\nu \in \mathbb{R}^{nd} :\ \|\nu_i\|^2 \leq 1,\ i = 1, 2, \cdots, n-1, \left\|\sum_{i=1}^{n-1} \nu_i\right\| \leq 1.\right\}$. It follows $\sum_{i=1}^{n} (\nu_i^* + \Delta\nu_i) = 0$, which implies $\Delta\nu_n = -\sum_{i=1}^{n-1} \Delta\nu_i$. Furthermore, $\|\nu_i^* + \Delta\nu_i\| \leq 1,\ \forall i = 1, 2, \cdots, n-1$ and $\left\|\sum_{i=1}^{n-1} (\nu_i^* + \Delta\nu_i)\right\| \leq 1$.

Denote $q(\nu) := -\sum_{i=1}^{n-1} \langle \nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle$. Then, we have

$$q(\nu^* + \Delta\nu) - q(\nu^*) = -\sum_{i=1}^{n-1} \langle \Delta\nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle . \tag{45}$$

Recall that $\|\nu_i^* + \Delta\nu_i\| \leq 1$ and $\nu_i^* = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|}$, it follows,

$$\left\| \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|} + \Delta\nu_i \right\|^2 \leq 1.$$

Expanding the squares gives

$$1 + 2 \left\langle \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|}, \Delta\nu_i \right\rangle + \|\Delta\nu_i\|^2 \leq 1.$$

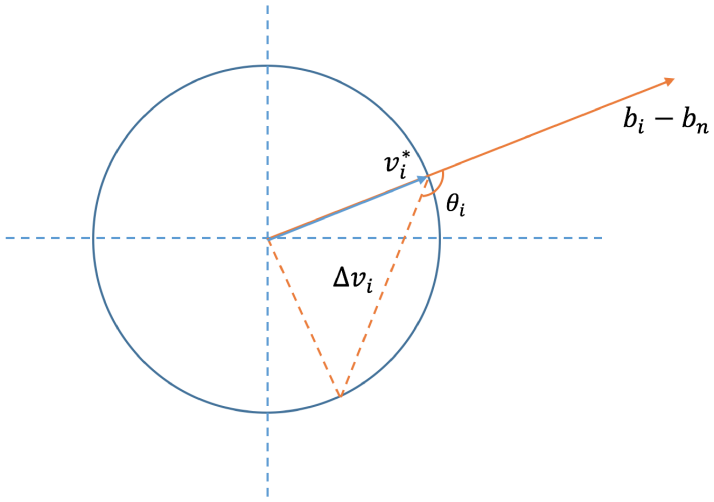


Figure 2: Geometric interpretation of the local perturbation by $\Delta\nu_i$ around the solution ν_i^*. For any perturbation $\Delta\nu_i$ of fixed length, the maximum of $\langle \mathbf{b}_i - \mathbf{b}_n, \Delta\nu_i \rangle$ is achieved when $\|\nu^* + \Delta\nu_i\| = 1$, i.e. $\nu^* + \Delta\nu_i$ is on the boundary of the unit ball, in which case we have $\cos\theta_i = -\|\Delta\nu_i\|/2$ and $\langle \mathbf{b}_i - \mathbf{b}_n, \Delta\nu_i \rangle = \|\mathbf{b}_i - \mathbf{b}_n\| \cdot \|\Delta\nu_i\| \cos\theta_i = -\|\mathbf{b}_i - \mathbf{b}_n\| \cdot \|\Delta\nu_i\|^2/2$.

Rearranging the terms gives

$$\langle \mathbf{b}_i - \mathbf{b}_n, \Delta\nu_i \rangle \leq -\|\mathbf{b}_i - \mathbf{b}_n\| \cdot \|\Delta\nu_i\|^2/2.$$

A geometric interpretation of this bound is given in Fig. 2. Substituting this bound into (45) gives

$$\begin{aligned} q(\nu^* + \Delta\nu) - q(\nu^*) \geq & \sum_{i=1}^{n-1} \|\mathbf{b}_i - \mathbf{b}_n\| \cdot \frac{\|\Delta\nu_i\|^2}{2} \\ \geq & \frac{1}{2} \left(\min_i \|\mathbf{b}_i - \mathbf{b}_n\| \right) \sum_{i=1}^{n-1} \|\Delta\nu_i\|^2. \end{aligned}$$

Note that since $\{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_n\}$ are distinct, $\min_i \|\mathbf{b}_i - \mathbf{b}_n\| > 0$ and this gives a local error bound condition for (42-44) with parameter $\beta = \frac{1}{2}$. Finally, since $\Delta\nu_n = -\sum_{i=1}^{n-1} \Delta\nu_i$, it follows,

$$\begin{aligned} q(\nu^* + \Delta\nu) - q(\nu^*) \geq & \frac{1}{2} \left(\min_i \|\mathbf{b}_i - \mathbf{b}_n\| \right) \sum_{i=1}^{n-1} \|\Delta\nu_i\|^2 \\ \geq & \frac{1}{2(n-1)} \left(\min_i \|\mathbf{b}_i - \mathbf{b}_n\| \right) \left\| \sum_{i=1}^{n-1} \Delta\nu_i \right\|^2 = \frac{1}{2(n-1)} \left(\min_i \|\mathbf{b}_i - \mathbf{b}_n\| \right) \|\Delta\nu_n\|^2, \end{aligned}$$

where the second inequality follows from Cauchy-Schwarz inequality that

$$\sqrt{\sum_{i=1}^{n-1} \|\Delta\nu_i\|^2} \sqrt{n-1} \geq \sum_{i=1}^{n-1} \|\Delta\nu_i\| \geq \left\| \sum_{i=1}^{n-1} \Delta\nu_i \right\|.$$

Since $G(\nu + \Delta\nu) - G(\nu^*) = q(\nu^* + \Delta\nu) - q(\nu^*)$, it follows

$$G(\nu + \Delta\nu) - G(\nu^*) \geq \frac{1}{4(n-1)} \left(\min_i \|\mathbf{b}_i - \mathbf{b}_n\| \right) \sum_{i=1}^{n} \|\Delta\nu_i\|^2 = \frac{1}{4(n-1)} \left(\min_i \|\mathbf{b}_i - \mathbf{b}_n\| \right) \|\Delta\nu\|^2.$$

Finishing the proof for case 1.

6.5.2 Proof of Lemma 6.7: Case 2

Case 2: The solution of the original geometric median (16-17) is NOT achieved at any of the vectors $\{\mathbf{b}_1,\ \mathbf{b}_2,\ \cdots,\ \mathbf{b}_n\}$.

We start by rewriting problem (39-41) as an equivalent feasibility problem:

$$\begin{cases} -\langle \nu, \mathbf{b}\rangle - G(\nu^*) \le 0, \\ \|\nu_i\|^2 \le 1,\ \ i = 1, 2, \cdots, n, \\ \sum_{i=1}^n \nu_i = 0. \end{cases} \tag{46}$$

The uniqueness in this case comes from the following lemma.

Lemma 6.8. *The solution $\nu^* \in \mathbb{R}^{nd}$ to* (46) *is unique and satisfies $\|\nu_i^*\| = 1,\ \forall i = 1, 2, \cdots, n$.*

To understand the feasibility problem (46) and prove Lemma 6.8, we start with the following definition:

Definition 6.1 (Wang and Pang (1994))**.** *Consider any inequality system $f_i(\mathbf{x}) \le 0,\ i = 1, 2, \cdots, m$. An inequality $f_i(\mathbf{x}) \le 0$ in the system is said to be singular if $f_i(\mathbf{x}) = 0$ for any solution to the system. If every inequality in the system is singular, we say the inequality system is singular.*

The following basic lemma regarding general feasibility problems is also proved in (Wang and Pang (1994)).

Lemma 6.9 (Lemma 2.1 of Wang and Pang (1994))**.** *Consider any inequality system $f_i(\mathbf{x}) \le 0,\ i = 1, 2, \cdots, m$ with non-empty solution set S. Suppose each of f_i is convex. Denote*

$$\begin{aligned} K &:= \{k \in \{1, 2, \cdots, m\} : f_k(\mathbf{x}) \le 0 \text{ is nonsingular}\}, \\ J &:= \{j \in \{1, 2, \cdots, m\} : f_j(\mathbf{x}) \le 0 \text{ is singular}\}. \end{aligned}$$

Then, the sub-system $f_j(\mathbf{x}) \le 0, j \in J$ alone is singular.

Proof of Lemma 6.8. Suppose ν^* is one of the solutions to (46). Suppose without loss of generality, the ball constraint $\|\nu_n\|^2 \le 1$ in (46) is nonsingular. Then, by Lemma 6.9, the subsystem

$$\begin{cases} -\langle \nu, \mathbf{b}\rangle - G(\nu^*) \le 0, \\ \|\nu_i\|^2 \le 1,\ \ i = 1, 2, \cdots, n-1, \\ \sum_{i=1}^n \nu_i = 0. \end{cases} \tag{47}$$

is still singular. This implies the optimal objective value of the following problem

$$\begin{aligned} \min_{\nu \in \mathbb{R}^{nd}} \ & -\langle \nu, \mathbf{b}\rangle, \\ s.t.\ & \sum_{i=1}^n \nu_i = 0, \\ & \|\nu_i\|^2 \le 1,\ i = 1, 2, \cdots, n-1, \end{aligned}$$

is still $G(\nu^*)$. Similar as before, one can get rid of the equality using $\nu_n = -\sum_{i=1}^{n-1} \nu_i$ and form an equivalent problem:

$$\begin{aligned} \min_{\nu \in \mathbb{R}^{nd}} \ & -\sum_{i=1}^{n-1} \langle \nu_i, \mathbf{b}_i - \mathbf{b}_n\rangle, \\ s.t.\ & \|\nu_i\|^2 \le 1,\ i = 1, 2, \cdots, n-1. \end{aligned}$$

This is a separable problem and obviously the optimal objective of this problem is $-\sum_{i=1}^{n-1} \|\mathbf{b}_i - \mathbf{b}_n\|$, which implies $G(\nu^*) = -\sum_{i=1}^{n-1} \|\mathbf{b}_i - \mathbf{b}_n\|$. However, by strong duality and the uniqueness of the geometric median problem (16-17), this further implies the solution to (16-17) is attained uniquely at $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n = \mathbf{b}_n$, contradicting the assumption that the solution to (16-17) is NOT achieved at any of the vectors $\{\mathbf{b}_1,\ \mathbf{b}_2,\ \cdots,\ \mathbf{b}_n\}$. Thus, we have shown that it is not possible to have one of the ball constraint being loose. This trivially implies it is not possible to have any two

or more ball constraints being loose and hence we know that any solution ν^* to (46) must satisfy $\|\nu_i^*\| = 1,\ \forall i = 1, 2, \cdots, n$.

Now suppose on the contrary such a solution is not unique. Let $\nu^*,\ \widetilde{\nu}^* \in \mathbb{R}^{nd}$ be two distinct solutions. Then, they must be different at some index j, i.e. $\exists j$ such that $\nu_j^* \neq \widetilde{\nu}_j^*$ and they satisfy $\|\nu_j^*\| = \|\widetilde{\nu}_j^*\| = 1$ by the previous argument. However, since the solution set to (46) must be convex (which follows trivially from the fact that all constraints are convex), any convex combination of $\nu^*,\ \widetilde{\nu}^*$ must be the solution. Specifically, the solution $\frac{\nu^*+\widetilde{\nu}^*}{2}$ has its j-th index $\left\|\frac{\nu_j^*+\widetilde{\nu}_j^*}{2}\right\| < 1$, contradicting the fact that any solution ν^* must satisfy $\|\nu_i^*\| = 1,\ \forall i = 1, 2, \cdots, n$. □

Now, we proceed to prove Lemma 6.7 for this case. The proof is inspired by a crucial "linearization" technique transforming general quadratic systems to linear systems which we are able to understand (e.g. Wang and Pang (1994), Luo and Luo (1994)). Consider any feasible $\nu \in \mathbb{R}^{nd}$ regarding (39)-(41). Then, for any index i, we have

$$\begin{aligned}\|\nu_i - \nu_i^*\|^2 = \|\nu_i\|^2 - 2\langle \nu_i, \nu_i^*\rangle + \|\nu_i^*\|^2 &= \|\nu_i\|^2 - 2\langle \nu_i - \nu_i^*, \nu_i^*\rangle - \|\nu_i^*\|^2 \\ &= \|\nu_i\|^2 - 1 + 2\langle \nu_i^* - \nu_i, \nu_i^*\rangle \leq 2\langle \nu_i^* - \nu_i, \nu_i^*\rangle, \quad (48)\end{aligned}$$

where in the third equality we use Lemma 6.8 that $\|\nu_i^*\| = 1$. We aim to bound the second term $\langle \nu_i^* - \nu_i, \nu_i^*\rangle$.

By Lemma 6.8, we have the following system has NO solution:

$$\begin{cases} -\langle \nu, \mathbf{b}\rangle - G(\nu^*) \leq 0, \\ \|\nu_i\|^2 - 1 < 0, \ \ i = 1, 2, \cdots, n, \\ \sum_{i=1}^n \nu_i = 0. \end{cases} \quad (49)$$

This is equivalent to claiming the following *linear* system has no solution:

$$\begin{cases} -\langle \mathbf{b}, \mathbf{y}\rangle \leq 0, \\ \langle \nu_i^*, \mathbf{y}_i\rangle < 0, \ \ i = 1, 2, \cdots, n, \\ \sum_{i=1}^n \mathbf{y}_i = 0. \end{cases} \quad (50)$$

To see why this is true, suppose on the contrary, (50) indeed has a solution. Let $\mathbf{y}^*$ be its solution, then we have $\alpha\mathbf{y}^*$ is also a solution for any $\alpha > 0$. This in turn implies

$$-\langle \mathbf{b}, \nu^* + \alpha\mathbf{y}^*\rangle - G(\nu^*) \leq -\langle \mathbf{b}, \nu^*\rangle - G(\nu^*) \leq 0,$$

and

$$\sum_{i=1}^n (\nu_i + \alpha\mathbf{y}_i^*) = \alpha \sum_{i=1}^n \mathbf{y}_i = 0.$$

Furthermore, for sufficiently small α, e.g. we can choose any $\alpha \leq \min_i \frac{\langle \nu_i^*, \mathbf{y}_i^*\rangle}{\|\mathbf{y}_i^*\|^2}$, the following holds,

$$\langle \nu_i^*, \alpha\mathbf{y}_i^*\rangle + \alpha^2\|\mathbf{y}_i^*\|^2 \leq 0.$$

This implies

$$\|\nu_i^* + \alpha\mathbf{y}_i^*\| = \|\nu_i^*\|^2 + 2\langle \nu_i^*, \alpha y_i^*\rangle + \|\mathbf{y}_i^*\| - 1 \leq \langle \nu_i, \alpha\mathbf{y}_i^*\rangle < 0,$$

and thus $\nu_i^* + \alpha\mathbf{y}_i^*$ is a solution to (49). On the other hand, suppose (49) has a solution, then, one can show similarly (50) has a solution.

To analyze (50), we employ the classical Motzkin's alternative theorem:

Lemma 6.10 (Motzkin (1952), Theorem D6)**.** *Suppose* $\mathbf{A} \neq 0$*. Either*

$$\mathbf{Ax} > 0, \ \ \mathbf{Bx} \geq 0, \ \ \mathbf{Cx} = 0,$$

has a solution, or there exists $\mathbf{u}, \mathbf{v}, \mathbf{w}$ *such that*

$$\mathbf{A}^T\mathbf{u} + \mathbf{B}^T\mathbf{v} + \mathbf{C}^T\mathbf{w} = 0, \ \ \mathbf{u} \geq 0, \ \ \mathbf{v} \geq 0, \mathbf{u} \neq 0,$$

but not both, where the inequalities are taken to be entrywise.

Now, applying Motzkin's alternative to (50), we have there exists a $\mathbf{u} \in \mathbb{R}^{2n+1}$ such that

$$-u_0\mathbf{b}^T + \sum_{i=1}^{n} u_i\,[\nu_i^*] + \sum_{i=1}^{n} u_{n+i}[\mathbf{e}_i] = 0,\ \ [u_1,\ u_2,\ \cdots,\ u_n] \neq 0,\ \mathbf{u} \geq 0, \tag{51}$$

where we define the block notation "$[\cdot]$" as follows

$$[\nu_i^*] = [\mathbf{0}, \cdots,\ \mathbf{0},\ (\nu_i^*)^T,\ \mathbf{0}, \cdots,\ \mathbf{0}] \in \mathbb{R}^{nd},$$

which takes ν_i^* at the i-th block of dimension d and $\mathbf{0}$ on other blocks. Also,

$$[\mathbf{e}_i] = [\mathbf{e}_i^T,\ \mathbf{e}_i^T,\ \cdots,\ \mathbf{e}_i^T] \in \mathbb{R}^{nd},$$

which takes unit basis vector $\mathbf{e}_i \in \mathbb{R}^d$ on all blocks.

Claim 1: $u_i > 0,\ \forall i = 1, 2, \cdots, n$.

To see why this is true, suppose on the contrary one of the u_i's is 0. Without loss of generality, we can assume $u_n = 0$. Then, by Motzkin's alternative again on (51), the following system has no solution:

$$\begin{cases} -\langle \mathbf{b}, \mathbf{y} \rangle \leq 0, \\ \langle \nu_i^*, \mathbf{y}_i \rangle < 0, \ \ i = 1, 2, \cdots, n-1, \\ \sum_{i=1}^{n} \mathbf{y}_i = 0. \end{cases} \tag{52}$$

By a similar equivalence relation as that of (49) and (50), this implies the following system has no solution,

$$\begin{cases} -\langle \nu, \mathbf{b} \rangle - G(\nu^*) \leq 0, \\ \|\nu_i\|^2 - 1 < 0, \ \ i = 1, 2, \cdots, n-1, \\ \sum_{i=1}^{n} \nu_i = 0, \end{cases}$$

which, by substituting $\nu_n = -\sum_{i=1}^{n-1} \nu_i$, implies the following system has no solution:

$$\begin{cases} -\sum_{i=1}^{n-1} \langle \nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle - G(\nu^*) \leq 0, \\ \|\nu_i\|^2 - 1 < 0, \ \ i = 1, 2, \cdots, n-1. \end{cases} \tag{53}$$

However, we know that the solution to the following minimization problem:

$$\min_{\nu \in \mathbb{R}^{nd}} \ -\sum_{i=1}^{n-1} \langle \nu_i, \mathbf{b}_i - \mathbf{b}_n \rangle\,, \ \ s.t.\ \|\nu_i\|^2 \leq 1,\ i = 1, 2, \cdots, n-1,$$

is attained uniquely at $\nu_i = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|}$ and the optimal objective value is $-\sum_{i=1}^{n-1} \|\mathbf{b}_i - \mathbf{b}_n\|$ which must be *strictly less than* $G(\nu^*)$ by strong duality and the fact that the solution to (16-17) is not attained at $\mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n = \mathbf{b}_n$. As a consequence, if we set

$$\widetilde{\nu}_i = \frac{\mathbf{b}_i - \mathbf{b}_n}{\|\mathbf{b}_i - \mathbf{b}_n\|} \frac{-G(\nu^*)}{\sum_{i=1}^{n-1} \|\mathbf{b}_i - \mathbf{b}_n\|},\ i = 1, 2, \cdots, n-1,$$

then, $\|\widetilde{\nu}_i\| < 1,\ \forall i = 1, 2, \cdots, n-1$ and $-\sum_{i=1}^{n-1} \langle \widetilde{\nu}_i, \mathbf{b}_i - \mathbf{b}_n \rangle - G(\nu^*) = 0$, which implies (53) has a solution and we reach a contradiction.

Now, rewriting (51), we have

$$[u_1(\nu_1^*)^T,\ u_2(\nu_2^*)^T,\ \cdots,\ u_n(\nu_n^*)^T] = u_0\mathbf{b} - \sum_{i=1}^{n} u_{n+i}[\mathbf{e}_i],$$

multiplying both sides by $[\nu_1^* - \nu_1,\ \nu_2^* - \nu_2,\ \cdots,\ \nu_n^* - \nu_n]$, which implies

$$\begin{aligned} \sum_{j=1}^{n} u_j \left\langle \nu_j^* - \nu_j, \nu_j^* \right\rangle = u_0 \sum_{j=1}^{n} \left\langle \mathbf{b}_j, \nu_j^* - \nu_j \right\rangle - \sum_{i=1}^{n} \sum_{j=1}^{n} u_{n+i} \left\langle \mathbf{e}_i, \nu_j^* - \nu_j \right\rangle \\ = u_0 \sum_{j=1}^{n} \left\langle \mathbf{b}_j, \nu_j^* - \nu_j \right\rangle = u_0(G(\nu) - G(\nu^*)), \end{aligned}$$

where the second from the last equality follows from $\sum_{i=1}^n \nu_i = \sum_{i=1}^n \nu_i^* = 0$. Thus, for any index $j \in \{1, 2, \cdots, n\}$,

$$\begin{aligned}\left\langle \nu_j^* - \nu_j, \nu_j^* \right\rangle =& \sum_{i \neq j} \frac{u_i}{u_j} \left\langle \nu_i - \nu_i^*, \nu_i^* \right\rangle + \frac{u_0}{u_j}(G(\nu) - G(\nu^*)) \quad \text{(by the fact } u_j > 0) \\ \leq& \sum_{i \neq j} \frac{u_i}{u_j}(\|\nu_i\|^2 - \|\nu_i^*\|^2) + \frac{u_0}{u_j}(G(\nu) - G(\nu^*)) \quad \text{(by convexity and } u_i > 0) \\ \leq& \frac{u_0}{u_j}(G(\nu) - G(\nu^*)) \quad \text{(by feasibility that} \|\nu_i\|^2 \leq 1 = \|\nu_i^*\|^2).\end{aligned}$$

Substituting this bound into (48) gives

$$\|\nu_j^* - \nu_j\|^2 \leq \frac{2u_0}{u_j}(G(\nu) - G(\nu^*)),\ \forall j \in \{1, 2, \cdots, n\},$$

and thus,

$$\|\nu^* - \nu\|^2 = \sum_{j=1}^n \|\nu_j^* - \nu_j\|^2 \leq \sum_j \frac{2u_0}{u_j}(G(\nu) - G(\nu^*)),$$

finishing the proof.

6.5.3 Putting everything together

Combining Lemma 6.6 and Lemma 6.7 we can easily show the following:

Lemma 6.11. *The solution ν^* to (36-37) is unique and furthermore, for any $\delta > 0$ and any point $\nu = [\nu_1^T,\ \nu_2^T,\ \cdots,\ \nu_n^T]^T \in \mathbb{R}^{nd}$ such that $\sum_{i=1}^n \nu_i = 0$ and $G(\nu) - G(\nu^*) \leq \delta$, we have there exists a constant C_δ depending on δ such that*

$$G(\nu) - G(\nu^*) \geq C_\delta \|\nu - \nu^*\|^2.$$

Proof of Lemma 6.11. Since the solution to (36-37) is attained in the constraint set (40-41) by Lemma 6.6, the uniqueness follows directly from Lemma 6.7.

Now, for any $\nu \in \mathbb{R}^{nd}$, such that $\sum_{i=1}^n \nu_i = 0$, and $\|\nu_i\| > 1$ for some index i,

$$G(\nu) - G(\nu^*) = G(\nu) - G(\overline{\nu}) + G(\overline{\nu}) - G(\nu^*) \geq \left(\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|\right) \|\nu - \overline{\nu}\| + C_0^2 \|\overline{\nu} - \nu^*\|^2.$$

where the vector $\overline{\nu}$ is defined in Lemma 6.6, the second inequality follows from Lemma 6.6 that $G(\nu) - G(\overline{\nu}) \geq (\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|) \|\nu - \overline{\nu}\|$ and Lemma 6.7 that $G(\overline{\nu}) - G(\nu^*) \geq C_0^2 \|\overline{\nu} - \nu^*\|^2$.

Thus, for any ν such that $G(\nu) - G(\nu^*) \leq \delta$, we have

$$\frac{\delta}{\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|} \geq \|\nu - \overline{\nu}\|,$$

which implies

$$\|\nu - \overline{\nu}\| \geq \begin{cases} \frac{\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|}{\delta} \|\nu - \overline{\nu}\|^2, & \text{if } \frac{\delta}{\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|} > 1, \\ \|\nu - \overline{\nu}\|^2, & \text{otherwise.} \end{cases}$$

Thus,

$$\begin{aligned}G(\nu) - G(\nu^*) \geq& C_0^2 \|\overline{\nu} - \nu^*\|^2 + \frac{\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|}{\max\left\{\frac{\delta}{\max_{i,j} \|\mathbf{b}_i - \mathbf{b}_j\|}, 1\right\}} \|\nu - \overline{\nu}\|^2 \\ \geq& C_\delta (\|\overline{\nu} - \nu^*\| + \|\nu - \overline{\nu}\|)^2 \geq C_\delta \|\nu - \nu^*\|^2,\end{aligned}$$

for some $C_\delta > 0$, where the second inequality follows from $\|w + z\|^2 \leq 2\|w\|^2 + 2\|z\|^2,\ \forall w, z$ and the last inequality follows from triangle inequality.

On the other hand, for any $\nu \in \mathbb{R}^{nd}$, such that $\sum_{i=1}^n \nu_i = 0$, and $\|\nu_i\| \leq 1$ for all indices i, by Lemma 6.7

$$G(\nu) - G(\nu^*) \geq C_0^2 \|\nu - \nu^*\|^2.$$

Overall, we finish the proof. □

6.5.4 Finishing the proof of Theorem 4.1

We recall the following well-known Hoffman's error bound:

Lemma 6.12 (Theorem 9 of Pang (1997))**.** *Given a convex polyhedron expressed as the solution set of a system of linear inequalities and equations defined by a pair of matrices* $(\mathbf{A}, \mathbf{B})$*:*

$$S := \left\{ \mathbf{x} \in \mathbb{R}^d : \ \mathbf{A}\mathbf{x} \leq a, \ \mathbf{B}\mathbf{x} = \mathbf{b} \right\}.$$

There exists a scalar $c > 0$ *such that for all* $(\mathbf{a}, \mathbf{b})$ *for which* S *is non-empty,*

$$\textit{dist}(\mathbf{x}, S) \leq c \left(\|(\mathbf{A}\mathbf{x} - \mathbf{a})_+\| + \|\mathbf{B}\mathbf{x} - \mathbf{b}\| \right), \ \forall \mathbf{x} \in \mathbb{R}^d,$$

where for any vector $\mathbf{y} \in \mathbb{R}^n$*,* $\|(\mathbf{y})_+\| := \sqrt{\sum_{i=1}^n \max\{y_i, 0\}^2}$.

The idea is to translate the local error bound on function $G(\nu)$ (i.e. Lemma 6.11) back to the local error bound on the original dual function $F(\lambda)$ using the equivalence relation between minimizing the dual function (34) and problem (36-37). Recall the definition of $F(\lambda)$ in (34) and $G(\nu)$ in (38), we have $F(\lambda) = G(\nu)$ for any $\lambda \in \mathbb{R}^{nd}$ such that $\mathbf{A}^T\lambda = \nu$. Thus, by Lemma 6.11, with ν replaced by $\mathbf{A}^T\lambda$ and $G(\nu)$ replaced by $F(\lambda)$,

$$\|\mathbf{A}^T\lambda - \nu^*\| \leq C_0(F(\lambda) - F^*)^{1/2},$$

where F^* is the optimal dual function value, and we use the fact that F^* equals $G(\nu^*)$, the optimal objective of (36-37). Since the solution ν^* to (36-37) is unique, the set of optimal Lagrange multipliers (i.e. the set of minimizers of (34)) $\Lambda^* = \left\{ \lambda \in \mathbb{R}^{nd} : \ \mathbf{A}^T\lambda = \nu^* \right\}$. By Hoffman's bound with $S = \Lambda^*$, we have

$$\text{dist}(\lambda, \Lambda^*) \leq c\|\mathbf{A}^T\lambda - \nu^*\|$$

for some positive constant c. Thus,

$$\text{dist}(\lambda, \Lambda^*) \leq \frac{C_0}{c}(F(\lambda) - F^*)^{1/2}.$$

Furthermore, since for any $\lambda^* \in \Lambda^*$ there exists a unique ν^* such that $\mathbf{A}^T\lambda^* = \nu^*$, it follows $\mathcal{P}_{\mathbf{A}}\lambda^* = \mathbf{A}(\mathbf{A}^T\mathbf{A})^\dagger\mathbf{A}^T\lambda^* = \mathbf{A}(\mathbf{A}^T\mathbf{A})^\dagger\nu^*$.

6.6 Simulation setups and additional simulation results

In this section, we give more details about our simulation along with more simulation results. First of all, in all three cases of Section 5, the randomly generated graph are connected. The way we ensure its connectivity is to first connect all nodes together by assigning $(n-1)$ edges, and then, randomly pick the remaining edges from the edge set of $n(n+1)/2$ edges according to the connectivity ratio. An example graph containing 20 nodes with connectivity ratio of 0.13 is shown in Fig. 3.

The parameters of algorithms are set as follows: (1) For the DSM algorithm, the learning rate $\alpha = 10$. (2) For the EXTRA algorithm, the learning rate $\alpha = 5$ when $n = 20$ and $\alpha = 20$ when $n = 50, 100$. (3) For the Jacobian ADMM, the proximal weight $\rho = 2\sigma_{\max}(\mathbf{A})$, where $\sigma_{\max}(\mathbf{A})$ is the maximum eigenvalue of $\mathbf{A}$. (4) For the smoothing algorithm, we fix the smoothing parameter $\mu = 10^{-5}$ throughout the experiments. (5) For our proposed algorithm, we set $D = 10\sqrt{d}$, where d is the dimension of the data and the desired accuracy $\varepsilon = 10^{-3}$. During the k-th stage, the time horizon $T^{(k)} = \frac{D}{\varepsilon^{0.8}} \cdot \frac{k}{K}$, where $K = \lceil \log_2(1/\varepsilon) \rceil + 1$ is the total number of rounds. The reason why we consider increasing the time horizon gradually is that we observe in practice the algorithm converges very fast during the first few stages and it is not necessary to run a long time. The aforementioned parameters of all algorithms are chosen in an ad-hoc way to ensure good performances.

Here, we perform additional simulations to show that our algorithm also works well under other scenarios where we change the dimension of the data. In the experiment below, the number of agents is set to be $n = 100$ and all the parameters are as described above. We vary the dimension of the data from 20 to 200, where each entry of the data points is still uniformly distributed over $[0, 10]$. The results are shown in Fig. 4.

Finally we demonstrate the performance of our algorithm under different network connectivity ratios. In the experiment below, the number of agents is set to be $n = 150$, dimension $d = 100$, and all the parameters are as described above. The results are shown in Fig. 5.

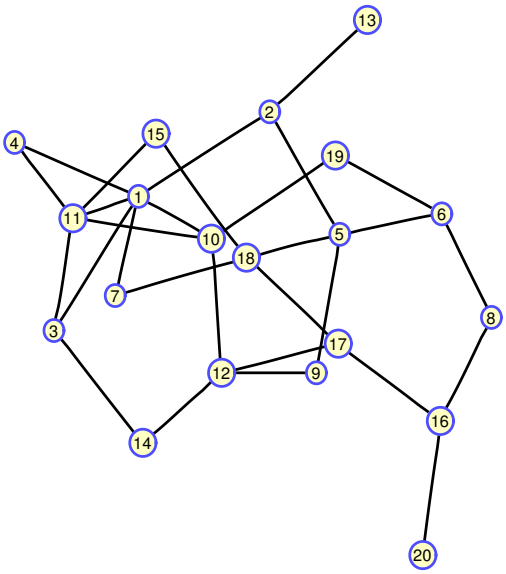
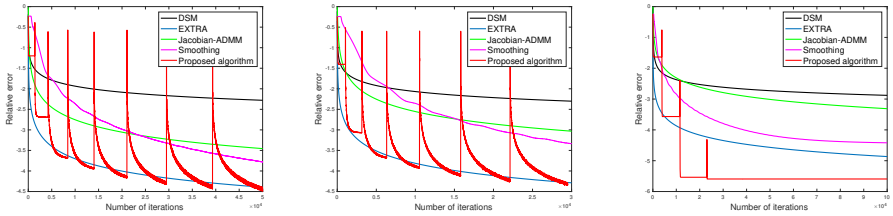
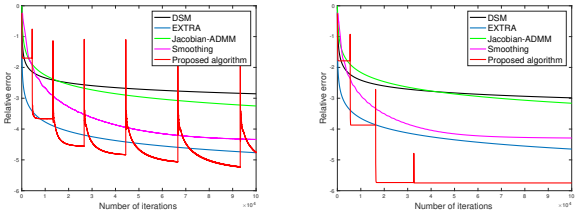


Figure 3: Illustration of a randomly generated connected graph with $n = 20$ and connectivity ratio=0.13.

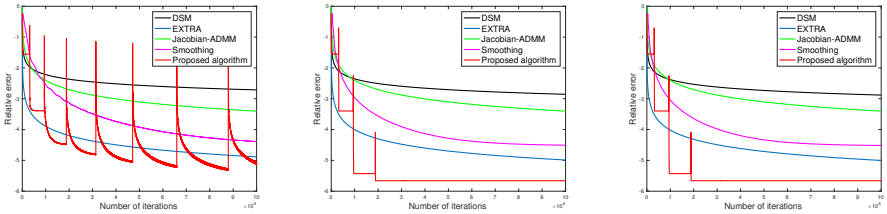


(a) $d = 20$, ratio=0.15. (b) $d = 50$, ratio=0.1. (c) $d = 150$, ratio=0.1.



(d) $d = 200$, ratio=0.1. (e) $d = 300$, ratio=0.1.

Figure 4: Performance of different algorithms under various dimensions of the vectors.



(a) $Ratio = 0.05$ (b) $Ratio = 0.15$ (c) $Ratio = 0.3$

Figure 5: Comparison of different algorithms on networks of different connectivity ratios.