On Centralized and Distributed Mirror Descent: Convergence Analysis Using Quadratic Constraints

Youbang Sun, Mahyar Fazlyab, *Member, IEEE*, and Shahin Shahrampour, *Senior Member, IEEE*

***Abstract*—Mirror descent (MD) is a powerful first-order optimization technique that subsumes several optimization algorithms including gradient descent (GD). In this work, we leverage quadratic constraints and Lyapunov functions to analyze the stability and characterize the convergence rate of the MD algorithm as well as its distributed variant using semidefinite programming (SDP). For both algorithms, we consider both strongly convex and nonstrongly convex assumptions. For centralized MD and strongly convex problems, we construct an SDP that certifies exponential convergence rates and derive a closed-form feasible solution to the SDP that recovers the optimal rate of GD as a special case. We complement our analysis by providing an explicit $O(1/k)$ convergence rate for convex problems. Next, we analyze the convergence of distributed MD and characterize the rate numerically using an SDP whose dimensions are independent of the network size. To the best of our knowledge, the numerical rate of distributed MD has not been previously reported in the literature. We further prove an $O(1/k)$ convergence rate for distributed MD in the convex setting. Our numerical experiments on strongly convex problems indicate that our framework certifies superior convergence rates compared to the existing rates for distributed GD.**

***Index Terms*—Convex optimization , decentralized optimization, mirror descent, optimization algorithms, quadratic constraints, semidefinite programming.**

I. Introduction

Over the last two decades, distributed optimization over multiagent networks has received a lot of attention in control, optimization, machine learning, and signal processing. In distributed optimization, a group of n agents is connected via a graph and can communicate locally with their neighbors. Each agent is assigned a local objective function $f_i : \mathbb{R}^d \to \mathbb{R}$, and the agents aim to collectively minimize the global objective function

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) \triangleq \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}. \tag{1}$$

The most intuitive gradient-based algorithm to tackle the problem above is *distributed gradient descent* (GD) [1], where at each iteration k, each agent i updates its decision variables by a (private) local GD combined with an averaging of its neighbors' variables. In the unconstrained case, this update is given by

$$x_i^{(k+1)} = x_i^{(k)} - \eta^{(k)}\nabla f_i(x_i^{(k)}) + \beta\sum_{j\in\mathcal{N}_i}(x_j^{(k)} - x_i^{(k)})$$

where $\eta^{(k)} > 0$ is the step size and $\beta > 0$ is the consensus parameter. In the form given above, this update is able to achieve optimal rates for convex problems using a diminishing step-size sequence. Optimality here refers to matching the centralized convergence rate (iteration complexity) up to some errors related to the network structure. However, when the local functions are smooth, the centralized GD algorithm employs a *constant* step-size sequence for which the above-distributed counterpart fails to converge.

The mirror descent (MD) algorithm [2] is a primal–dual method that has been actively studied in recent years. MD can be seen as a generalization of GD, in which the squared Euclidean distance is replaced by Bregman divergence as the regularizer. The freedom in the choice of Bregman divergence makes MD suitable for various problem geometries. MD has been proven to have the same iteration complexity as GD for nonstrongly convex problems [3], and it may even scale better with respect to the dimension of the decision variables [4]. In the strongly convex scenario, MD is less studied, and very recently, its exponential convergence was established under the Polyak–Łojasiewicz (PL) condition [5]. Inspired by the success of MD in centralized optimization, MD has also been studied in the distributed setting. To the best of authors' knowledge, the convergence rate of distributed MD is not established for strongly convex and smooth problems, and only recently, Sun and Shahrampour [6] provided a continuous-time analysis suggesting local exponential rate (without explicitly characterizing the rate).

In this article, we leverage the framework of quadratic constraints (QCs) to certify numerical exponential convergence rates for centralized as well as distributed MD for strongly convex and smooth problems using SDP. For merely convex and smooth problems, we also establish an ergodic $O(1/k)$ convergence rate. We first analyze centralized MD, for which we derive linear matrix inequalities (LMIs) as sufficient conditions for convergence of the algorithm at a specified rate (see Theorem 2, Theorem 6, and Proposition 3). For the strongly convex case, we prove that these LMIs always have a feasible solution that matches the optimal convergence rate of GD when the Bregman divergence is chosen as the squared Euclidean distance (Proposition 4 and Corollary 5). Next, we analyze the convergence of distributed MD and characterize the rate using LMIs (see Theorems 8 and 9). To the best of our knowledge, the exponential rate of distributed MD has not been previously established in the literature. Our numerical experiments on strongly convex problems indicate that our framework certifies superior convergence rates compared to the existing rates for distributed GD.

Manuscript received 17 January 2022; revised 28 June 2022; accepted 14 December 2022. Date of publication 20 December 2022; date of current version 26 April 2023. This work was supported in part by NSF under Grant ECCS-2136206. Recommended by Senior Editor Tetsuya Iwasaki and Guest Editors George J. Pappas, Anuradha M. Annaswamy, Manfred Morari, Claire J. Tomlin, Rene Vidal, and Melanie N. Zeilinger. *(Corresponding author: Shahin Shahrampour.)*

Youbang Sun and Shahin Shahrampour are with the Department of Mechanical and Industrial Engineering, Northeastern University, Boston, MA 02115 USA (e-mail: sun.youb@northeastern.edu; s.shahrampour@northeastern.edu).

Mahyar Fazlyab is with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: mahyarfazlyab@jhu.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAC.2022.3230767.

Digital Object Identifier 10.1109/TAC.2022.3230767

A. Related Literature

1) Distributed Optimization: To ensure that distributed GD (or sub-GD) reach consensus, many methods [1], [7], [8] use diminishing step-size (commonly $1/k$). For distributed MD, similar studies have been conducted for stochastic optimization [9], [10] and online optimization [11], [12]. Doan et al. [13] provide convergence results for both centralized and decentralized MD algorithms. However, convergence rates obtained using diminishing step-size are subexponential and suboptimal under assumptions of strong convexity and smoothness.

To address this issue, a number of recent works introduce an additional variable in the state vectors to track past gradients (see, e.g., [14], [15], [16], [17]). One of the earlier works in this direction is the EXTRA algorithm proposed by Shi et al. [14], which uses the information from the past two iterations to perform each update. For smooth problems, EXTRA provably achieves $O(1/k)$ convergence rate under the convexity assumption and exponential convergence rate under the strong convexity assumption, respectively.

A closely relevant literature is the continuous-time distributed GD, where the algorithms are constructed by a set of ordinary differential equations (ODEs). These works are mostly based on the idea of *integral feedback*, which can be thought as the continuous-time analog of gradient tracking. In this case, each agent uses an integration term as a part of the ODE (see, e.g., [18], [19], [20], [21]). In these works, the analysis is carried out by proving the Lyapunov stability for the corresponding continuous-time dynamics, and exponential stability can be obtained in certain cases [20]. For MD, the continuous-time algorithm in [6] and [22] and the discrete-time algorithm in [23] both adapt the integral feedback (or gradient tracking) method and propose algorithms that do not suffer from suboptimal convergence rates. Specifically, Sun et al. [6] propose a continuous-time distributed MD that achieves a "local" exponential rate for strongly convex problems, and Yu et al. [23] provide an $O(1/k)$ convergence rate under the convexity assumption in discrete time. Nevertheless, the exponential rate of (discrete-time) distributed MD for strongly convex and smooth problems remains an open problem, which we target in the current work.

2) Integral QCs: Deriving convergence rates for iterative optimization algorithms in the worst case is an integral part of algorithm design. However, this procedure is not principled, requires a case-by-case analysis, and might lead to conservative rates. To automate convergence analysis and derive sharp convergence rates, several past works have used integral QCs (IQCs) and semidefinite programming in various settings [24], [25], [26], [27], [28], [29], [30], pioneered by the work in [24]. IQCs are a tool from robust control to analyze dynamical systems that contain components that are nonlinear, uncertain, or difficult to model [31]. The basic idea is to abstract these troublesome components by constraints on their input and output signals. This approach to algorithm analysis can also guide the search for parameter selection in algorithm design. The works in [32] and [33] are of particular relevance to our work. They both provide the IQC-based analysis of distributed gradient-based algorithms in strongly convex settings. Compared to these works, our framework focuses on distributed MD in both strongly convex and convex settings.

II. Preliminaries

A. Notations

The identity matrix of dimension n is denoted by I_n and the n-dimensional vector with all entries 1 is represented by $\mathbf{1}_n$. We denote the set of n-dimensional symmetric matrices by $\mathbb{S}^n$. The positive (negative) semidefiniteness of matrix M is denoted as $M \succeq 0$ ($M \preceq 0$). We use $\otimes$ and $\|\cdot\|$ to denote the Kronecker product and spectral norm, respectively. We define the norm of vector v with respect to a positive semidefinite matrix M as $\|v\|_M$. The indicator function of a set $\mathcal{X} \subseteq \mathbb{R}^d$ is defined as $\mathbb{I}_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$ and $\mathbb{I}_{\mathcal{X}}(x) = +\infty$ otherwise.

Definition 1 (Strong convexity): A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ_f-strongly convex on $\mathbb{R}^d$ if the following inequality is true for all $x, y \in \mathbb{R}^d$:

$$f(x) + \nabla f(x)^\top (y - x) + \frac{\mu_f}{2} \|y - x\|^2 \leq f(y).$$

Definition 2 (Lipschitz smoothness): A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is L_f-smooth on $\mathbb{R}^d$ if $\frac{L_f}{2}\|x\|^2 - f(x)$ is convex, which implies that for all $x, y \in \mathbb{R}^d$

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L_f}{2} \|y - x\|^2 .$$

We further denote the condition number of function f by $\kappa_f \triangleq \frac{L_f}{\mu_f} \geq 1$. When $\mu_f = 0$, the function is only convex.

Proposition 1: Suppose f is μ_f-strongly convex and L_f-smooth on $\mathbb{R}^d$. Then, the following inequality holds for all $x, y \in \mathbb{R}^d$, and $u = \nabla f(x),\ v = \nabla f(y)$:

$$\begin{bmatrix} x - y \\ u - v \end{bmatrix}^\top \begin{bmatrix} \frac{-\mu_f L_f}{\mu_f + L_f} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & \frac{-1}{\mu_f + L_f} I_d \end{bmatrix} \begin{bmatrix} x - y \\ u - v \end{bmatrix} \geq 0. \tag{2}$$

The above QC follows from the combination of strong convexity and Lipschitz smoothness [24], [34].

B. Centralized MD Algorithm

We start by providing some background on the centralized MD algorithm. For simplicity in the exposition, we study the unconstrained case, but our analysis can also be extended to the constrained case. Let us start with the GD algorithm, whose update is equivalent to the following minimization:

$$x^{(k+1)} = \operatorname*{argmin}_x \Big\{ f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{2\eta} \|x - x^{(k)}\|^2 \Big\}$$

where $\eta > 0$ is the step size. In each iteration, the algorithm seeks to minimize a first-order approximation of the function with a Euclidean regularizer. As a generalization of GD, MD replaces the squared Euclidean distance with Bregman divergence, which is defined with respect to a distance generating function (DGF) $\phi : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$\mathcal{D}_\phi(x, x') \triangleq \phi(x) - \phi(x') - \langle \nabla \phi(x'), x - x' \rangle. \tag{3}$$

Assumption 1: The DGF $\phi : \mathbb{R}^d \to \mathbb{R}$ is μ_ϕ-strongly convex and L_ϕ-smooth.

The centralized (unconstrained) MD algorithm with step-size η is written as

$$x^{(k+1)} = \operatorname*{argmin}_x \Big\{ f(x^{(k)}) + \nabla f(x^{(k)})^\top (x - x^{(k)}) + \frac{1}{\eta} \mathcal{D}_\phi(x, x^{(k)}) \Big\} \tag{4}$$

where if we choose the Bregman divergence to be the squared Euclidean distance, the update above reduces to GD.

We can also view the MD update through a different lens using the *convex conjugate* of function ϕ. The convex conjugate of function ϕ, denoted by $\phi^\star$, is defined as $\phi^\star(z) \triangleq \sup_{x \in \mathbb{R}^d} \{\langle x, z \rangle - \phi(x)\}$. Assumption 1 guarantees that $\phi^\star$ is L_ϕ^{-1}-strongly convex and μ_ϕ^{-1}-smooth.

We refer the reader to [35] for further details. Correspondingly, the following equivalence can be established:

$$z' = \nabla\phi(x') \iff x' = \nabla\phi^\star(z').$$

Then, the centralized MD update can be rewritten in the following form:

$$\begin{aligned} z^{(k+1)} &= z^{(k)} - \eta\nabla f(x^{(k)}) \\ x^{(k+1)} &= \nabla\phi^\star(z^{(k+1)}) \end{aligned} \tag{5}$$

or, equivalently, $z^{(k+1)} = z^{(k)} - \eta(\nabla f \circ \nabla\phi^\star)(z^{(k)})$, which is reminiscent of GD. We can see that MD is more general than GD in that we can exploit the geometry of the problem using an appropriate choice of ϕ, which makes MD more suitable for problems such as convex clustering, matrix optimization with regularization, etc. [36], [37].

Denoting $x^\star$ and $z^\star$ as the fixed points of (5), we have

$$z^\star = z^\star - \eta\nabla f(x^\star) \quad x^\star = \nabla\phi^\star(z^\star)$$

which implies that $x^\star$ is a minimizer of f.

III. Convergence Analysis of Centralized MD

In this section, we provide a convergence analysis of the centralized MD using semidefinite programming. Our starting point is to describe all the nonlinear functions in the algorithm, namely ∇f and $\nabla\phi^\star$ by QCs on their input–output pairs, resulting in a *quadratically-constrained linear system*. We then find a suitable "rate-generating" Lyapunov function for this constrained system using semidefinite programming. We derive exponential (respectively, subexponential) convergence rates for strongly convex (respectively, convex) functions.

A. Exponential Convergence for Strongly Convex f

In the following theorem, we characterize an LMI that depends on parameters of f (μ_f and L_f), parameters of ϕ (μ_ϕ and L_ϕ), and several decision variables (including the step-size η and the convergence rate $\rho \in (0,1)$). We prove that if the LMI is satisfied, the iterates converge exponentially fast to the unique fixed point $(x^\star, z^\star)$ with the rate ρ.

Theorem 2: Let Assumption 1 hold and assume that f is μ_f-strongly convex and L_f-smooth. Define matrices M_{sc}, M_f, M_ϕ as follows:

$$M_{sc} = \begin{bmatrix} \frac{1-\rho}{2\mu_\phi} I_d & 0 & 0 \\ 0 & 0 & \frac{-\eta}{2} I_d \\ 0 & \frac{-\eta}{2} I_d & \frac{\eta^2}{2\mu_\phi} I_d \end{bmatrix}$$

$$M_f = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \frac{-\mu_f L_f}{\mu_f + L_f} I_d & \frac{1}{2} I_d \\ 0 & \frac{1}{2} I_d & \frac{-1}{\mu_f + L_f} I_d \end{bmatrix}$$

$$M_\phi = \begin{bmatrix} \frac{-1}{\mu_\phi + L_\phi} I_d & \frac{1}{2} I_d & 0 \\ \frac{1}{2} I_d & \frac{-\mu_\phi L_\phi}{\mu_\phi + L_\phi} I_d & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{6}$$

If there exist some $\rho \in (0,1), \eta > 0, \sigma_f \geq 0, \sigma_\phi \geq 0$, such that the following matrix inequality holds:

$$M_{sc} + \sigma_f M_f + \sigma_\phi M_\phi \preceq 0 \tag{7}$$

then the MD algorithm in (5) converges exponentially fast with the rate of ρ. In particular

$$\|x^{(k)} - x^\star\|^2 \leq \frac{2\mathcal{D}_{\phi^\star}(z^{(0)}, z^\star)}{\mu_\phi} \rho^k.$$

Proof: Denote $u^{(k)} \triangleq \nabla f(x^{(k)})$ and define the stacked vector

$$e^{(k)^\top} = \begin{bmatrix} (z^{(k)} - z^\star)^\top & (x^{(k)} - x^\star)^\top & (u^{(k)} - u^\star)^\top \end{bmatrix}. \tag{8}$$

Then, from Proposition 1, we obtain the following quadratic inequalities:

$$e^{(k)\top} M_f e^{(k)} \geq 0, \quad e^{(k)\top} M_\phi e^{(k)} \geq 0 \quad \forall k$$

which are imposed by ∇f and $\nabla\phi$, respectively. Consider the Lyapunov candidate $V^{(k)} = \rho^{-k}\mathcal{D}_{\phi^\star}(z^{(k)}, z^\star)$. Recall that $\phi^\star$ is L_ϕ^{-1}-strongly convex and μ_ϕ^{-1}-smooth, so the Lyapunov function is, indeed, nonnegative and continuously differentiable. Using Lemma 10 (provided in the appendix of [38]), we can calculate the Lyapunov function difference between two consecutive iterations as

$$V^{(k+1)} - V^{(k)} \leq \rho^{-k-1} e^{(k)\top} M_{sc} e^{(k)}. \tag{9}$$

Utilizing the two quadratic inequalities imposed by the nonlinearities, we can write

$$\begin{aligned} V^{(k+1)} - V^{(k)} &\leq \rho^{-k-1} e^{(k)\top} M_{sc} e^{(k)} \\ &\leq \rho^{-k-1} e^{(k)\top} (M_{sc} + \sigma_f M_f + \sigma_\phi M_\phi) e^{(k)}. \end{aligned}$$

Now if the LMI in (7) holds, the Lyapunov function is nonincreasing, which yields

$$\mathcal{D}_{\phi^\star}(z^{(k)}, z^\star) = \rho^k V^{(k)} \leq \rho^k V^{(0)} = \rho^k \mathcal{D}_{\phi^\star}(z^{(0)}, z^\star). \tag{10}$$

Observing $\mathcal{D}_{\phi^\star}(z^{(k)}, z^\star) = \mathcal{D}_\phi(x^\star, x^{(k)})$ and

$$\frac{\mu_\phi}{2}\|x^{(k)} - x^\star\|^2 \leq \mathcal{D}_\phi(x^\star, x^{(k)}).$$

Theorem 2 provides a matrix inequality feasibility problem that establishes the exponential convergence rate of MD for a given ρ. This matrix inequality is linear in $(\rho, \sigma_f, \sigma_\phi)$ (but not in η), allowing us to find the smallest ρ by the semidefinite program

$$\begin{aligned} \underset{\rho, \sigma_\phi, \sigma_f}{\text{minimize}} \quad & \rho \\ \text{subject to} \quad & 0 < \rho \leq 1 \\ & \eta, \sigma_\phi, \sigma_f \geq 0 \\ & M_{sc} + \sigma_f M_f + \sigma_\phi M_\phi \preceq 0. \end{aligned} \tag{11}$$

If, in addition, we want to optimize ρ over the step-size η, we can use Schur Complements to "convexify" the matrix inequality with respect to η. We state this result formally in the next proposition.

Proposition 3: The optimization problem in (11) is equivalent to the following SDP:

$$\begin{aligned} \underset{\eta, \rho, \sigma_\phi, \sigma_f}{\text{minimize}} \quad & \rho \\ \text{subject to} \quad & 0 < \rho \leq 1 \\ & \eta, \sigma_\phi, \sigma_f \geq 0 \\ & \begin{bmatrix} \frac{\sigma_\phi}{\mu_\phi + L_\phi} + \frac{\rho - 1}{2\mu_\phi} & \frac{-\sigma_\phi}{2} & 0 & 0 \\ \frac{-\sigma_\phi}{2} & \frac{\mu_\phi L_\phi \sigma_\phi}{\mu_\phi + L_\phi} + \frac{\mu_\phi}{2} + \frac{\mu_f L_f \sigma_f}{\mu_f + L_f} & \frac{-\sigma_f}{2} & \frac{-\sqrt{\mu_\phi}}{\sqrt{2}} \\ 0 & \frac{-\sigma_f}{2} & \frac{\sigma_f}{\mu_f + L_f} & \frac{\eta}{\sqrt{2\mu_\phi}} \\ 0 & \frac{-\sqrt{\mu_\phi}}{\sqrt{2}} & \frac{\eta}{\sqrt{2\mu_\phi}} & 1 \end{bmatrix} \\ & \times \succeq 0. \end{aligned} \tag{12}$$

We refer to the appendix of [38] for the proof of this proposition. We now show that the SDP in (12) has a feasible solution for which we can analytically calculate the convergence rate.

Proposition 4: The following selection

$$\eta = \sigma_f = \frac{2\mu_\phi}{\mu_f + L_f}$$
$$\sigma_\phi = \frac{4\mu_f L_f}{(\mu_f + L_f)^2}\frac{(1+\kappa_\phi)}{\kappa_\phi(\kappa_\phi - 1)}$$
$$\rho_{opt} = 1 - \frac{4\mu_f L_f}{(\mu_f + L_f)^2 \kappa_\phi^2} \tag{13}$$

is a feasible solution to the SDP in (12).

The proof of the proposition can be found in the appendix of [38]. Note that ρ_{opt} is an upper bound on the optimal value of (12).

The recent work in [5] also proposed an explicit rate of $1 - \frac{1}{5\kappa_\phi^2 \kappa_f^2}$ for MD under the PL condition. Although the PL condition is weaker than strong convexity, ρ_{opt} is strictly smaller than the rate in [5]. Furthermore, in our result, we do not make full use of strong convexity: We only require the quadratic inequality (2) to hold for the pair $(x, x^\star)$ (x arbitrary and $x^\star$ the fixed point of the algorithm), whereas for strongly convex f this inequality holds for *all* (x, y). Our rate also recovers the optimal rate of GD as a special case.

Corollary 5: For $\phi(x) = \frac{1}{2}\|x\|^2$ the optimal rate ρ_{opt} in (13) coincides with the optimal convergence rate of GD.

Proof: If $\phi(x) = \frac{1}{2}\|x\|^2$, we have that $\phi^\star(z) = \frac{1}{2}\|z\|^2$ and (5) is equivalent to GD. In this case, the condition number $\kappa_\phi = \frac{L_\phi}{\mu_\phi} = 1$, and ρ_{opt} reduces to the optimal convergence rate for GD (see [34, Th. 2.1.15]).

B. $O(1/k)$ Convergence for Convex f

We now propose an LMI, which establishes subexponential convergence rate for the MD algorithm when the objective function is convex ($\mu_f = 0$).

Theorem 6: Let Assumption 1 hold and assume that f is convex ($\mu_f = 0$) and L_f-smooth ($0 < L_f < \infty$), and define the matrix M_c as follows:

$$M_c = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{\epsilon-\eta}{2} I \\ 0 & \frac{\epsilon-\eta}{2} I & \frac{\eta^2}{2\mu_\phi} I \end{bmatrix}. \tag{14}$$

If there exist some $\eta > 0, \sigma_f \geq 0, \sigma_\phi \geq 0, \epsilon \geq 0$, such that the following matrix inequality holds:

$$M_c + \sigma_f M_f + \sigma_\phi M_\phi \preceq 0 \tag{15}$$

then the ergodic mean of function value at iteration K satisfies

$$f(\bar{x}^{(K)}) - f(x^\star) \leq \frac{\mathcal{D}_{\phi^\star}(z^{(0)}, z^\star)}{\epsilon K}$$

where $\bar{x}^{(K)} = \frac{1}{K}\sum_{i=1}^K x^{(i)}$.

We remark that a similar analysis can be applied to Theorem 6 to find the best step-size that maximizes ϵ. The details are omitted due to space limitation.

Remark 1 (Constrained MD): Consider the constrained version of centralized (lazy) MD [39]

$$z^{(k+1)} = z^{(k)} - \eta \nabla f(x^{(k)})$$
$$s^{(k)} = \nabla \phi^\star(z^{(k)})$$
$$x^{(k)} = \arg\min_{x \in \mathcal{X}} \mathcal{D}_\phi(x, s^{(k)}) \tag{16}$$

where $\mathcal{X}$ is a convex subset of $\mathbb{R}^d$. By defining $g(x) = \mathbb{I}_{\mathcal{X}}(x)$ as the indicator function of the set $\mathcal{X}$ and denoting its subdifferential by ∂g, the optimality condition that characterizes $x^{(k)}$ is

$$\nabla\phi(x^{(k)}) - z^k \in \partial g(x^{(k)}).$$

Using the fact that the subdifferential ∂g is monotone (since $\mathcal{X}$ is convex), we can rewrite (16) as

$$z^{(k+1)} = z^{(k)} - \eta u^{(k)}$$
$$u^{(k)} \triangleq \nabla f(x^{(k)})$$
$$v^{(k)} \triangleq \nabla \phi(x^k) \tag{17}$$

subject to the QC

$$(v^{(k)} - v^\star - (z^{(k)} - z^\star))^\top (x^{(k)} - x^\star) \geq 0 \quad \forall k.$$

Furthermore, we can write two separate QCs for the relationships $u^{(k)} = \nabla f(x^{(k)})$ and $v^{(k)} = \nabla\phi(x^{(k)})$. We can, therefore, employ the same approach and derive an LMI as a sufficient condition to establish exponential and $O(1/k)$ convergence rates for strongly convex and convex problems, respectively.

IV. Convergence Analysis of Distributed MD

In the distributed setup, we have a network of agents, characterized by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node in $\mathcal{V} = \{1, \ldots, n\}$ represents an agent, and the connection between two agents i and j is captured by the edge $\{i, j\} \in \mathcal{E}$. We use $\mathcal{N}_i \triangleq \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}$ to denote the neighborhood of agent i. The graph Laplacian is represented by $\mathcal{L} \in \mathbb{R}^{n \times n}$.

Assumption 2: The graph $\mathcal{G}$ is undirected and connected, i.e., there exists a path between any two distinct agents $i, j \in \mathcal{V}$.

The connectivity assumption implies that $\mathcal{L}$ has a unique null eigenvalue; that is, $\mathcal{L}\mathbf{1}_n = 0$.

A. Distributed MD Algorithm

We first introduce the distributed MD update, in which each agent i in the network implements the following iterative algorithm:

$$z_i{}^{(k+1)} = z_i{}^{(k)} - \eta_1 \left(\nabla f_i(x_i{}^{(k)}) + y_i{}^{(k)}\right) - \eta_2 \sum_{j \in \mathcal{N}_i} (z_i{}^{(k)} - z_j{}^{(k)})$$
$$y_i{}^{(k+1)} = y_i{}^{(k)} + \eta_2 \sum_{j \in \mathcal{N}_i} (z_i{}^{(k)} - z_j{}^{(k)})$$
$$x_i{}^{(k+1)} = \nabla \phi^\star(z_i{}^{(k+1)}). \tag{18}$$

The first update uses private gradient information as well as the dual variables from the neighbors. It also depends on a variable $y_i{}^{(k)}$ which acts as an integrator. This algorithm is similar to the discretized version of the distributed MD proposed in [22] using the idea of integral feedback. However, the method differs slightly in the local averaging in that the algorithm in [22] performs local averaging with respect to the primal variable, and here, the averaging is done on the dual variable $z_i{}^{(k)}$.

It is evident that the behavior of this system relies on the network structure through the dependence on the Laplacian of the graph capturing the network. Since $\mathcal{L} \in \mathbb{S}^n$, the LMIs will consist of matrices whose dimensions scale with n, which is not suitable when n is large.

Following the idea in [32] and [33], we transform the updates such that the dependence on the *full structure* of the network is avoided. Define

$$W \triangleq I_n - \eta_2 \mathcal{L} = \Delta W + \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$$

and further denote the spectral norm of ΔW by $\lambda \triangleq \|\Delta W\|$. The quantity $1-\lambda$ is also known as the spectral gap.

To represent the updates collectively for all the agents, we define the stacked vectors

$$\begin{aligned} z^{(k)} &= [z_1^{(k)\top}, \dots, z_n^{(k)\top}]^\top \\ y^{(k)} &= [y_1^{(k)\top}, \dots, y_n^{(k)\top}]^\top \\ u^{(k)} &= \nabla \mathbf{f}(x^{(k)}) \triangleq [\nabla f_1(x_1^{(k)})^\top, \dots, \nabla f_n(x_n^{(k)})^\top]^\top \\ x^{(k)} &= [\nabla \phi^\star(z_1^{(k)})^\top, \dots, \nabla \phi^\star(z_n^{(k)})^\top]^\top \\ v^{(k)} &= (\Delta W \otimes I_d) z^{(k)}. \end{aligned} \tag{19}$$

We can now rewrite (18) as

$$\begin{aligned} z^{(k+1)} &= \left(\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes I_d\right) z^{(k)} - \eta_1 (u^{(k)} + y^{(k)}) + v^{(k)} \\ y^{(k+1)} &= y^{(k)} + \left(\left(I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right) \otimes I_d\right) z^{(k)} - v^{(k)} \\ v^{(k)} &= (\Delta W \otimes I_d) z^{(k)} \\ x^{(k)} &= \nabla \phi^\star(z^{(k)}) \\ u^{(k)} &= \nabla \mathbf{f}(x^{(k)}). \end{aligned} \tag{20}$$

To represent (20) in a state-space form, we can write

$$\begin{aligned} \begin{bmatrix} z^{(k+1)} \\ y^{(k+1)} \end{bmatrix} &= \begin{bmatrix} \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes I_d & -\eta_1 I_{nd} \\ (I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes I_d & I_{nd} \end{bmatrix} \begin{bmatrix} z^{(k)} \\ y^{(k)} \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & -\eta_1 I_{nd} & I_{nd} \\ 0 & 0 & -I_{nd} \end{bmatrix} \begin{bmatrix} x^{(k)} \\ u^{(k)} \\ v^{(k)} \end{bmatrix}. \end{aligned} \tag{21}$$

Additionally, we know the following constraints on the updates:

$$\begin{aligned} \begin{bmatrix} 0 \\ 0 \end{bmatrix} &= \begin{bmatrix} 0 & \mathbf{1}_n\mathbf{1}_n^\top \otimes I_d \\ 0 & 0 \end{bmatrix} \begin{bmatrix} z^{(k)} \\ y^{(k)} \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \mathbf{1}_n\mathbf{1}_n^\top \otimes I_d \end{bmatrix} \begin{bmatrix} x^{(k)} \\ u^{(k)} \\ v^{(k)} \end{bmatrix}. \end{aligned} \tag{22}$$

We define the state vector $\xi^{(k)\top} \triangleq \begin{bmatrix} z^{(k)\top} & y^{(k)\top} \end{bmatrix}$ as well as the input vector $\zeta^{(k)\top} \triangleq \begin{bmatrix} x^{(k)\top} & u^{(k)\top} & v^{(k)\top} \end{bmatrix}$. We can rewrite (21) and (22) as

$$\xi^{(k+1)} = A\xi^{(k)} + B\zeta^{(k)} \quad 0 = F\xi^{(k)} + G\zeta^{(k)} \tag{23}$$

where A, B, F, G are of appropriate dimensions. For ease of notation, we denote $H \triangleq \begin{bmatrix} F & G \end{bmatrix}$.

For the purpose of convergence analysis, we characterize the fixed point of (21). Define $x^\star \triangleq \mathbf{1}_n \otimes x_\star$, where $x_\star \in \mathbb{R}^d$ is a minimizer of (1), and let $z^\star \triangleq \nabla\phi(x^\star)$, $u^\star \triangleq \nabla \mathbf{f}(x^\star)$, $y^\star \triangleq -\nabla \mathbf{f}(x^\star)$, and $v^\star = 0$. By letting $z^{(k)}, y^{(k)}, v^{(k)}, x^{(k)}$, and $u^{(k)}$ in (21) take the values of $z^\star, y^\star, v^\star, x^\star$, and $u^\star$, it is easy to show that $z^{(k+1)} = z^{(k)}, y^{(k+1)} = y^{(k)}$ using Assumption 2.

B. Exponential Convergence of Distributed MD

In the following theorem, we present the main result of this section. We provide two LMIs to characterize the convergence rate of distributed MD. The LMIs are written in terms of several decision variables, including the step-size η_1 and the convergence rate ρ. If we can find a feasible solution for these LMIs, the distributed MD is guaranteed to converge exponentially fast.

Before stating the theorem, we state the following lemma, which will allow us to simplify the resulting SDP.

Lemma 7 (Lemma 6 in [32]): Suppose that square matrices J_1, J_2 satisfy $J_1^2 = J_1, J_2^2 = J_2, J_1J_2 = J_2J_1 = 0$. For square matrices Q_1 and Q_2, define $Q \triangleq Q_1 \otimes J_1 + Q_2 \otimes J_2$. Then, the following are equivalent: 1) $Q \succeq 0$. 2) $Q_1 \succeq 0, Q_2 \succeq 0$.

Theorem 8: Let Assumptions 1 and 2 hold and assume all local functions f_i are μ_f-strongly convex and L_f-smooth. Define the following matrices:

$$\begin{aligned} A_1 &= \begin{bmatrix} 0 & -\eta_1 \\ 1 & 1 \end{bmatrix}, B_1 = \begin{bmatrix} 0 & -\eta_1 & 1 \\ 0 & 0 & -1 \end{bmatrix} \\ A_2 &= \begin{bmatrix} 1 & -\eta_1 \\ 0 & 1 \end{bmatrix}, B_2 = \begin{bmatrix} 0 & -\eta_1 & 1 \\ 0 & 0 & -1 \end{bmatrix} \\ H_1 &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, H_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

and let R_i be a matrix whose columns form a basis for the null space of H_i for $i = 1, 2$. Furthermore, define

$$M_f = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{-\mu_f L_f}{\mu_f + L_f} & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & \frac{-1}{\mu_f + L_f} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M_\lambda = \begin{bmatrix} \lambda^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

$$M_\phi = \begin{bmatrix} \frac{-1}{\mu_\phi + L_\phi} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{-\mu_\phi L_\phi}{\mu_\phi + L_\phi} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

If there exists $\rho \in (0,1), \eta_1 \geq 0, P \in \mathbb{S}^2, \; P \succ 0, \sigma_f \geq 0, \sigma_\phi \geq 0, \sigma_\lambda \geq 0$, such that the following matrix inequality holds for $i = 1, 2$:

$$\begin{aligned} R_i^\top \Bigg(\begin{bmatrix} A_i^\top P A_i - \rho P & A_i^\top P B_i \\ B_i^\top P A_i & B_i^\top P B_i \end{bmatrix} \\ + \sigma_f M_f + \sigma_\lambda M_\lambda + \sigma_\phi M_\phi \Bigg) R_i \preceq 0 \end{aligned} \tag{24}$$

then the distributed MD algorithm (18) initialized at $y^{(0)} = 0$ converges exponentially with a rate of ρ as follows:

$$\|\xi^{(k)} - \xi^\star\|_{P \otimes I_{nd}}^2 \leq \rho^k \|\xi^{(0)} - \xi^\star\|_{P \otimes I_{nd}}^2.$$

Proof: Define the vector $e^{(k)\top} = [\xi^{(k)\top} \quad \zeta^{(k)\top}]$. For any $\Sigma_{eq} \in \mathbb{S}^2$, we can establish the following (in)equalities:

$$e^{(k)\top}(M_f \otimes I_{nd})e^{(k)} \geq 0$$
$$e^{(k)\top}(M_\phi \otimes I_{nd})e^{(k)} \geq 0$$
$$e^{(k)\top}(M_\lambda \otimes I_{nd})e^{(k)} \geq 0$$
$$e^{(k)\top}H^\top(\Sigma_{eq} \otimes I_{nd})He^{(k)} = 0.$$

The first two inequalities are derived from Proposition 1, the third inequality is due to the fact that $\lambda = \|\Delta W\|$, and the equality follows from the affine constraint in (22).

Define the Lyapunov function

$$V^{(k)} = \rho^{-k}(\xi^{(k)} - \xi^\star)^\top P'(\xi^{(k)} - \xi^\star)$$

where $P' \triangleq P \otimes I_{nd}$. Then, using (23), we can write

$$V^{(k+1)} - V^{(k)} = \rho^{-k-1}e^{(k)\top}\begin{bmatrix} A^\top P'A - \rho P' & A^\top P'B \\ B^\top P'A & B^\top P'B \end{bmatrix} e^{(k)}.$$

Now, if the following LMI holds:

$$\begin{bmatrix} A^\top P'A - \rho P' & A^\top P'B \\ B^\top P'A & B^\top P'B \end{bmatrix} + \sigma_f M_f \otimes I_{nd} + \sigma_\lambda M_\lambda \otimes I_{nd} + \sigma_\phi M_\phi \otimes I_{nd} + H^\top(\Sigma_{eq} \otimes I_{nd})H \preceq 0 \quad (25)$$

then for any $e^{(k)}$, we have that

$$\rho^{-k-1}e^{(k)\top}\begin{bmatrix} A^\top P'A - \rho P' & A^\top P'B \\ B^\top P'A & B^\top P'B \end{bmatrix} e^{(k)} \leq 0$$

or, equivalently,

$$(\xi^{(k)} - \xi^\star)^\top P'(\xi^{(k)} - \xi^\star) \leq \rho^k(\xi^{(0)} - \xi^\star)^\top P'(\xi^{(0)} - \xi^\star).$$

In words, the squared norm of system variables decreases exponentially fast to zero. Next, we simplify the LMI such that the dimension is not dependent on the agent number n. Our approach follows that in [32]. Define J_1 and J_2 in Lemma 7 as $J_1 = (I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top) \otimes I_d$, $J_2 = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \otimes I_d$. It is easy to verify that these matrices satisfy the constraints in Lemma 7. We then have that

$$A = A_1 \otimes J_1 + A_2 \otimes J_2$$
$$B = B_1 \otimes J_1 + B_2 \otimes J_2$$
$$H = H_1 \otimes J_1 + H_2 \otimes J_2.$$

Consider the following for $i = 1, 2$:

$$\begin{bmatrix} A_i^\top PA_i - \rho P & A_i^\top PB_i \\ B_i^\top PA_i & B_i^\top PB_i \end{bmatrix} + \sigma_f M_f + \sigma_\lambda M_\lambda + \sigma_\phi M_\phi + H_i^\top \Sigma_{eq} H_i \preceq 0. \quad (26)$$

Since matrices J_1, J_2 satisfy the conditions in Lemma 7, if we consider $Q, Q_1,$ and Q_2 as the negative left-hand side of (25) and (26), respectively, then a feasible set of solutions that satisfy (26) is equivalently a feasible set of solutions for (25). Then, we only need to show the equivalence of (24) and (26), which follows from [40, Lemma 3.1]. This equivalence is also used in the proof of [32, Th. 7].

The theorem provides two LMIs that establish the exponential convergence rate of distributed MD. As we can see the LMIs are more involved compared to the centralized case, and it is challenging to find even a suboptimal analytical rate.

We finally remark that common analysis on distributed MD involves general primal–dual norms [11], whereas QCs are defined with respect to the Euclidean norm. The use of general primal–dual norms in nonstrongly convex problems helps with improving the rate up to a multiplicative factor of $\sqrt{d}$. However, in a strongly convex case, the rate is exponentially fast, and a more general analysis can only change the iteration complexity by at most logarithmic factors of d, which is an interesting avenue to investigate in the future.

C. $O(1/k)$ Convergence for Convex Functions

In the following theorem, we present the counterpart of Theorem 8 for convex problems.

Theorem 9: Let Assumptions 1 and 2 hold and assume all local functions f_i are convex ($\mu_f = 0$) and L_f-smooth. Recall the definitions of matrices $A_1, A_2, B_1, B_2, H_1, H_2, R_1, R_2, M_f, M_\lambda, M_\phi$ in Theorem 8 and define the following additional matrices:

$$M_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & L_f & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad M_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

If there exist $\eta_1 \geq 0, P \in \mathbb{S}^2, P \succ 0, \sigma_f \geq 0, \sigma_\phi \geq 0, \sigma_\lambda \geq 0, \epsilon \geq 0$, such that the following matrix inequality holds for $i = 1, 2$:

$$R_i^\top\left(\begin{bmatrix} A_i^\top PA_i - P & A_i^\top PB_i \\ B_i^\top PA_i & B_i^\top PB_i \end{bmatrix} + \sigma_f M_f + \sigma_\lambda M_\lambda + \sigma_\phi M_\phi + \epsilon M_i\right)R_i \preceq 0 \quad (27)$$

then, the iterates of the distributed MD algorithm (18) initialized at $y^{(0)} = 0$ satisfy the following inequality:

$$\sum_{i=1}^n \left(f(\bar{x}_i^{(K)}) - f^\star\right) \leq \frac{V^{(0)}}{\epsilon K}$$

where $\bar{x}_i^{(K)} \triangleq \frac{1}{K}\sum_{k=0}^{K-1} x_i^{(k)}$.

We refer to the appendix of [38] for the proof of this theorem. Given that $f(\bar{x}_i^{(K)}) - f^\star$ is nonnegative, it is easy to see that the function evaluated at the ergodic average of each agent iterate converges to a minimum with a rate of $O(1/K)$.

D. Evaluating the Tightness of Results

For the distributed MD algorithm, we provide numerical results based on Theorem 8. First, we demonstrate the influence of the network structure, and then, we compare the rate recovered by Theorem 8 to existing theoretical rates on distributed GD when it achieves exponential convergence.

1) Impact of the Network Structure on Convergence Rate: We calculate the worst case convergence rate with several choices of λ and plot it with respect to the step-size η_1. We set the local functions to have condition number $\kappa_f = 2$ and the DGF to have condition number $\kappa_\phi = 2$. Each curve in the plot represents a certain λ and is obtained by scanning feasible values for the decision variables in the LMIs (24). From Fig. 1(a), we can see that there exists an optimal step-size to obtain the best convergence rate and that as λ increases, the best rate becomes worse. Hence, for any given network structure and its corresponding Laplacian matrix, we should select η_2 such that λ is minimized. This is consistent with results on distributed optimization, where having a larger λ deteriorates the performance.

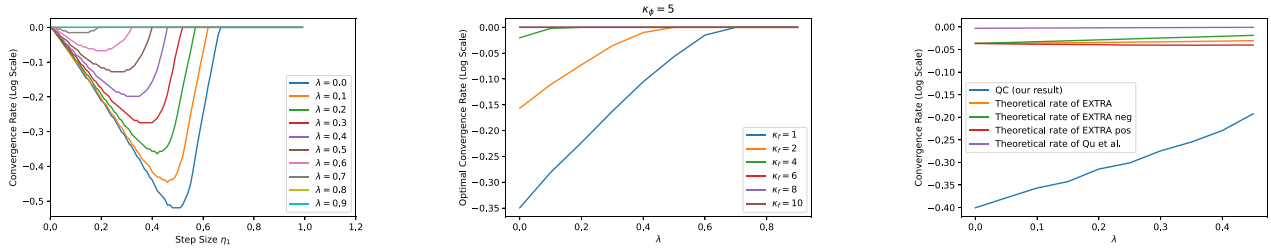

Fig. 1. Optimal convergence rate for distributed MD obtained by solving LMIs under various assumptions. (a) Convergence rates generated from Theorem 8 versus step-size η_1. (b) Fixed $\kappa_\phi = 5$. Varying λ and κ_f, optimal learning rate is chosen. (c) Comparison of the convergence rates for different methods.

In Fig. 1(b), we keep $\kappa_\phi = 5$ constant and study the optimal convergence rate for different λ and κ_f. When the condition number increases, the optimal rate worsens. This behavior aligns with GD, where $\kappa_\phi = 1$.

2) Comparison With Distributed GD: To the best of our knowledge, there is currently no work that provides an exponential convergence rate for distributed MD algorithm. Hence, we select two previous works on distributed GD, namely [14] and [15], and compare our performance with the theoretical rates provided in these works. In order to provide a fair comparison, we must set $\kappa_\phi = 1$ to ensure that MD reduces to GD. We also set the local functions to have condition number $\kappa_f = 3$.

Of the two related works above, EXTRA [14] is of particular relevance to our algorithm. If the matrix $\tilde{W}$ in EXTRA is set to be $(I_n + W)/2$, the EXTRA algorithm coincides with our algorithm with the exception of having a coefficient difference of $\frac{1}{2}$ for the tracking term. Note that the theoretical convergence rate of EXTRA relies on the spectral norm of ΔW as well as the *smallest nonzero eigenvalue* λ_n of W. We plot the convergence rate of EXTRA under the following three different scenarios:

1) $\lambda_n = \lambda$ (EXTRA pos);
2) $\lambda_n = -\lambda$ (EXTRA neg);
3) $\lambda_n \approx 0$ (EXTRA).

From Fig. 1(c), we can see that when λ is small, the rate recovered by Theorem 8 significantly outperforms EXTRA. As λ increases, the convergence rate calculated for our method starts increasing. We also include the theoretical convergence results from Qu et al. [15], which are consistently outperformed by EXTRA.

Note that the point of this plot is not to declare a winner among algorithms. The goal is to show that the richness of the Lyapunov function and QC analysis provides machinery to obtain better convergence rates, especially compared to the rates that are algorithm specific. In this case, our algorithm can coincide with EXTRA but still our analysis provides better rates. Our observation is in line with empirical results of [32].

V. Conclusion

In this article, we adapted the IQC approaches in [24], [25], [32], and [33] to pose the convergence analysis of the MD algorithm as a semidefinite program. We characterized the convergence rate for both centralized and distributed settings, and empirical evaluations were performed under the assumption of strongly convex and smooth local objective functions. For the centralized case, we derived a closed-form feasible solution to the SDP for the convergence rate, which depends on the condition number of the DGF. For the decentralized case, we numerically derived the convergence rates using SDP. These SDPs do not scale with the ambient dimension and the network size. Using the QC framework, we further proved the $O(1/k)$ convergence rate for centralized and distributed MD in the convex and smooth setting. It would be interesting to derive analytical rates for the distributed case. Another important direction is the analysis of the MD algorithm with primal–dual norms. This is a challenging problem as current SDP approaches rely on the Euclidean norm and they do not lend themselves to general primal–dual norms.

References

[1] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[2] A. Nemirovsky and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*. New York, NY, USA: Wiley, 1983.
[3] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Res. Lett.*, vol. 31, no. 3, pp. 167–175, 2003.
[4] A. Ben-Tal, T. Margalit, and A. Nemirovski, "The ordered subsets mirror descent optimization method with applications to tomography," *SIAM J. Optim.*, vol. 12, no. 1, pp. 79–108, 2001.
[5] A. Radhakrishnan, M. Belkin, and C. Uhler, "Linear convergence and implicit regularization of generalized mirror descent with time-dependent mirrors," 2020, *arXiv:2009.08574*.
[6] Y. Sun and S. Shahrampour, "Linear convergence of distributed mirror descent with integral feedback for strongly convex problems," in *Proc. IEEE Conf. Decis. Control*, 2021, pp. 1683–1688.
[7] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control*, vol. 55, no. 4, pp. 922–938, Apr. 2010.
[8] P. Lin, W. Ren, and Y. Song, "Distributed multi-agent optimization subject to nonidentical constraints and communication delays," *Automatica*, vol. 65, pp. 120–131, 2016.
[9] D. Yuan, Y. Hong, D. W. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, 2018.
[10] M. Rabbat, "Multi-agent mirror descent for decentralized stochastic optimization," in *Proc. IEEE 6th Int. Workshop Comput. Adv. Multi-Sensor Adaptive Process.*, 2015, pp. 517–520.
[11] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Trans. Autom. Control*, vol. 63, no. 3, pp. 714–725, Mar. 2018.
[12] D. Yuan, Y. Hong, D. W. C. Ho, and S. Xu, "Distributed mirror descent for online composite optimization," *IEEE Trans. Autom. Control*, vol. 66, no. 2, pp. 714–729, Feb. 2021.
[13] T. T. Doan, S. Bose, D. H. Nguyen, and C. L. Beck, "Convergence of the iterates in mirror descent methods," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 114–119, Jan. 2019.
[14] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.
[15] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, Sep. 2018.

[16] Y. Sun, G. Scutari, and A. Daneshmand,, "Distributed optimization based on gradient tracking revisited: Enhancing convergence rate via surrogation," *SIAM J. Optim.*, vol. 32, no. 2, pp. 354–385, 2022.
[17] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Math. Program.*, vol. 187, pp. 409–457, 2020.
[18] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Trans. Autom. Control*, vol. 59, no. 3, pp. 781–786, Mar. 2014.
[19] X. Zeng, P. Yi, and Y. Hong, "Distributed continuous-time algorithm for constrained convex optimizations via nonsmooth analysis approach," *IEEE Trans. Autom. Control*, vol. 62, no. 10, pp. 5227–5233, Oct. 2017.
[20] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2015.
[21] S. Yang, Q. Liu, and J. Wang, "A multi-agent system with a proportional-integral protocol for distributed constrained optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 7, pp. 3461–3467, Jul. 2017.
[22] Y. Sun and S. Shahrampour, "Distributed mirror descent with integral feedback: Asymptotic convergence analysis of continuous-time dynamics," *IEEE Control Syst. Lett.*, vol. 5, no. 5, pp. 1507–1512, Nov. 2021.
[23] Y. Yu and B. Açıkmeşe, "RLC circuits-based distributed mirror descent method," *IEEE Control Syst. Lett.*, vol. 4, no. 3, pp. 548–553, Jul. 2020.
[24] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optim.*, vol. 26, no. 1, pp. 57–95, 2016.
[25] M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado, "Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems," *SIAM J. Optim.*, vol. 28, no. 3, pp. 2654–2689, 2018.
[26] B. Hu and L. Lessard, "Dissipativity theory for Nesterov's accelerated method," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1549–1557.
[27] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Smooth strongly convex interpolation and exact worst-case performance of first-order methods," *Math. Program.*, vol. 161, no. 1/2, pp. 307–345, 2017.
[28] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2861–2868, Jul. 2019.
[29] N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar, "Robust accelerated gradient methods for smooth strongly convex functions," *SIAM J. Optim.*, vol. 30, no. 1, pp. 717–751, 2020.
[30] C. Scherer and C. Ebenbauer, "Convex synthesis of accelerated gradient algorithms," *SIAM J. Control Optimi.*, vol. 59, no. 6, pp. 4615–4645, 2021.
[31] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 42, no. 6, pp. 819–830, Jun. 1997.
[32] A. Sundararajan, B. Hu, and L. Lessard, "Robust convergence analysis of distributed optimization algorithms," in *Proc. 55th Annu. Allerton Conf. Commun., Control, Comput.*, 2017, pp. 1206–1212.
[33] A. Sundararajan, B. Van Scoy, and L. Lessard, "Analysis and design of first-order distributed optimization algorithms over time-varying graphs," *IEEE Trans. Control Netw. Syst.*, vol. 7, no. 4, pp. 1597–1608, Dec. 2020.
[34] Y. Nesterov et al., *Lectures on Convex Optimization*, vol. 137. Berlin, Germany: Springer, 2018.
[35] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*. Berlin, Germany: Springer, 2012.
[36] D. Lashkari and P. Golland, "Convex clustering with exemplar-based models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20, 2007, pp. 825–832,.
[37] A. Benfenati, E. Chouzenoux, and J.-C. Pesquet, "Proximal approaches for matrix optimization problems: Application to robust precision matrix estimation," *Signal Process.*, vol. 169, 2020, Art. no. 107417.
[38] Y. Sun, M. Fazlyab, and S. Shahrampour, "On centralized and distributed mirror descent: Convergence analysis using quadratic constraints," 2021, *arXiv:2105.14385*.
[39] E. Hazan, "Introduction to online convex optimization," *Found. Trends Optim.*, vol. 2, no. 3/4, pp. 157–325, 2016.
[40] P. Gahinet and P. Apkarian, "A linear matrix inequality approach to H_∞ control," *Int. J. Robust Nonlinear Control*, vol. 4, no. 4, pp. 421–448, 1994.