A Variational Approach to Dual Methods for Constrained Convex Optimization

Mahyar Fazlyab, Alec Koppel, Victor M. Preciado, and Alejandro Ribeiro

***Abstract*— We approach linearly constrained convex optimization problems through their dual reformulation. Specifically, we derive a family of accelerated dual algorithms by adopting a variational perspective in which the dual function of the problem represents the scaled potential energy of a synthetic mechanical system, and the kinetic energy is defined by the Bregman divergence induced by the dual velocity flow. Through application of Hamilton's principle, we derive a continuous-time dynamical system which exponentially converges to the saddle point of the Lagrangian. Moreover, this dynamical system only admits a stable discretization through accelerated higher-order gradient methods, which precisely corresponds to accelerated dual mirror ascent. In particular, we obtain discrete-time convergence rate $\mathcal{O}(1/k^p)$, where $p-1$ is the truncation index of the Taylor expansion of the dual function. For practicality sake, we consider $p = 2$ and $p = 3$ only, respectively corresponding to dual Nesterov acceleration and a dual variant of Nesterov's cubic regularized Newton method. This analysis provides an explanation from whence dual acceleration comes as the discretization of the Euler-Lagrange dynamics associated with the constrained convex program. We demonstrate the performance of the aforementioned continuous-time framework with numerical simulations.**

I. Introduction

Underlying many recent technological advances such as artificial intelligence [1], smart devices [2], robotics [3], [4], and wireless communications [5], is the mathematical theory of optimization. Our particular focus is on convex optimization problems with linear constraints, which apply to these respective contexts in the form of large-scale supervised learning [6], cooperative control [7], and wireless routing [8]. In such settings, obtaining a closed-form solution of the problem is not possible, and hence numerical optimization schemes must be used. Our goal is to derive fast yet efficient iterative methods for linearly constrained convex programming [9], i.e., problems of the form

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),\ \text{s.t. } A\mathbf{x}=\mathbf{b}. \tag{1}$$

where f is convex, n is possibly large, and the linear constraints may represent, for instance, consensus [10] or network flow constraints [8], [11].

In the development of iterative numerical methods for constrained convex problems, there is a fundamental trade-off between computational efficiency and the rate at which we attain optimality. On the one hand, with no regard for computation cost, when the objective is strongly convex, one may apply Newton's method that uses the second-derivative information and exhibits quadratic convergence $\mathcal{O}(\rho^{2^k})$ under proper initialization [12], without which (super)linear $\mathcal{O}(\rho^{k^p})$ rates are possible ($p \geq 1$, $0 \leq \rho \leq 1$) for weakly convex problems. Newton's method requires evaluation of the Hessian inverse of the objective at each step, which is cubic in the decision variable dimension. In the case of large-scale supervised learning [6], for instance, this complexity is prohibitively costly. Quasi-Newton schemes approximate this Hessian inverse computation and in some cases achieve comparable behavior to their exact second-order counterparts [13], [14].

On the other hand, first-order methods are popular due to their ease of implementation, low complexity, and robust, albeit sublinear $\mathcal{O}(1/k)$ convergence [15] (linear $\mathcal{O}(\rho^k)$ when the objective is strongly convex). Accelerated variants of gradient methods, which introduce auxiliary sequences based on recursive averages, have gained popularity due to their ability to improve convergence in the weakly convex case to $\mathcal{O}(1/k^2)$ with comparable complexity [16].

Adaptations of acceleration have been proposed to constrained problems in the primal domain for special cases [17], primal-dual schemes [18]–[20] as well as proximal dual approaches [8], [21], [22], which reach at least Nesterov's $\mathcal{O}(1/k^2)$ rate, but all require strong convexity. The curiosity of these dual approaches is in the fact that when the objective is strongly convex, *linear rates* are achievable by first-order primal methods in the unconstrained setting. Unfortunately, such favorable behavior does not easily carry over to the constrained case, except for special cases [13], [23].

Given that linear convergence remains elusive for first-order dual methods for constrained problems, even with strong convexity, we ask a related question: is Nesterov's optimal rate $\mathcal{O}(1/k^2)$ realizable by first-order accelerated dual methods for constrained problems (Section II) *without strong convexity*? The contribution of this work is an affirmative answer, based on extending a recently discovered connection between Lagrangian mechanics and accelerated mirror descent methods [24] to dual approaches for constrained optimization (Section III). [24] considers primal methods for unconstrained convex programming, whereas our focus is on dual reformulations of linearly constrained convex problems.

In extending the connection between accelerated methods in optimization and Lagrangian mechanics put forth in [24] to dual methods for constrained problems, we attain exponential convergence in continuous time for strongly convex smooth objectives, and $\mathcal{O}(1/k^2)$ rate for discrete time algorithms for

Work in this paper is supported by the NSF under grants CNS-1302222, IIS- 1447470, the ONR under grant N00014-12-1-0997, ARL MAST CTA, and ASEE SMART.

The authors are with Department of Electrical and Systems Engineering, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104. Email: {mahyarfa, akoppel, preciado, aribeiro}@seas.upenn.edu.

objectives that are neither strongly convex nor differentiable. Moreover, we provide a rigorous explanation from whence dual acceleration comes as the stable discretization of a certain continuous-time Euler-Lagrange equation (Section IV), rather than a heuristic reference to adding "momentum" into an optimization scheme. In Section V, we illustrate that favorable convergence behavior translates well into practice.

A longer version of this paper containing the proofs will be available online [25].

II. Linearly Constrained Convex Optimization

Consider the following convex optimization problem,

$$p^\star = \min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),\ \text{s.t. } A\mathbf{x} = \mathbf{b}. \tag{2}$$

where $f(\mathbf{x})\colon \mathbb{R}^n \to \mathbb{R}$ is convex, $A \in \mathbb{R}^{m\times n}$ with $\mathbf{rank}(A) = m < n$, and $\mathbf{b} \in \mathbb{R}^m$. The latter conditions imply that the system of equations $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions, and hence, the problem (2) is feasible and nontrivial. We further assume that the optimal $p^\star$ is finite.

We turn to reformulating (2) in terms of its dual problem. To do so, define the Lagrangian function $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda})\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated with the problem (2) as

$$\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^\top(A\mathbf{x} - \mathbf{b}), \tag{3}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the vector of Lagrange multipliers. Notice that the Lagrangian is convex in $\mathbf{x}$ and concave (affine) in $\boldsymbol{\lambda}$. Further define the dual function $G(\boldsymbol{\lambda})\colon \mathbb{R}^m \to \mathbb{R}$ as

$$G(\boldsymbol{\lambda}) = \min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}). \tag{4}$$

The dual problem is then to maximize the dual function (4) with respect to $\boldsymbol{\lambda}$,

$$d^\star = \max_{\boldsymbol{\lambda}\in\mathbb{R}^m} G(\boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}\in\mathbb{R}^m}\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}). \tag{5}$$

Since the primal problem (2) is convex and feasible, i.e., Slater's condition holds [9], the duality gap is zero, i.e., $d^\star = p^\star$. Thus there is no loss of optimality by approaching the problem by its dual reformulation. For future reference, we define $\mathbf{X}^\star \times \boldsymbol{\Lambda}^\star \subset \mathbb{R}^n \times \mathbb{R}^m$ as the set of primal-dual optimal pairs $(\mathbf{x}^\star,\boldsymbol{\lambda}^\star)$ that satisfy $\mathcal{L}(\mathbf{x}^\star,\boldsymbol{\lambda}^\star) = p^\star = d^\star$. Next we adopt a mechanics perspective in order to derive an exponentially convergent solution to (5) in continuous time.

III. Dual Optimization as Lagrangian Mechanics

We now shift focus to conducting variational analysis in order to derive a solution to (5) in continuous time, the source of our improved convergence results.

A. Bregman Lagrangian and Hamilton's Principle

Begin by equipping the dual domain $\mathbb{R}^m$ with a continuously differentiable convex function $\psi : \mathbb{R}^m \to \mathbb{R}$ that satisfies $\|\nabla\psi(\boldsymbol{\lambda})\| \to \infty$ as $\|\boldsymbol{\lambda}\| \to \infty$ to define a measure of distance in $\mathbb{R}^m$ in terms of the Bregman divergence, i.e., for $\boldsymbol{\lambda},\boldsymbol{\nu} \in \mathbb{R}^m$,

$$D_\psi(\boldsymbol{\nu},\boldsymbol{\lambda}) = \psi(\boldsymbol{\nu}) - \psi(\boldsymbol{\lambda}) - \langle\nabla\psi(\boldsymbol{\lambda}), \boldsymbol{\nu} - \boldsymbol{\lambda}\rangle\ , \tag{6}$$

Observe that a special case of ψ is the Euclidean distance, $\psi(\cdot) = \|\cdot\|_2^2$. We denote $\mathbf{x}_t \in \mathbb{R}^n$ and $\boldsymbol{\lambda}_t \in \mathbb{R}^m$ as curves parameterized by continuous time index $t \in \mathbb{T} \subseteq \mathbb{R}_+$.

As in [24], we identify the *kinetic* energy of a synthetic mechanical system associated with the optimization problem (4) as the Bregman divergence between its state and its state perturbed by its velocity vector with an appropriate scaling, $D_\psi(\boldsymbol{\lambda} + e^{-\alpha_t}\dot{\boldsymbol{\lambda}},\boldsymbol{\lambda})$. Further, the *potential* energy is the objective to be minimized, which for dual algorithms is the negative of the dual function $-G(\boldsymbol{\lambda})$ in (4). Thus, we may consider the *Bregman Lagrangian* as an appropriately weighted Lagrangian of the mechanical system with these identifications, stated as

$$\mathbb{L}(\boldsymbol{\lambda},\dot{\boldsymbol{\lambda}},t) = e^{\alpha_t+\gamma_t}\big(D_\psi(\boldsymbol{\lambda} + e^{-\alpha_t}\dot{\boldsymbol{\lambda}},\boldsymbol{\lambda}) + e^{\beta_t}G(\boldsymbol{\lambda})\big), \tag{7}$$

where $\alpha_t,\gamma_t,\beta_t\colon \mathbb{T}\to\mathbb{R}$ are smooth functions of time $t \in \mathbb{T}$. We next consider applying Hamilton's Principle (see [26] for details) to (7). We find that the ideal scaling conditions

$$\dot\beta_t \le e^{\alpha_t}\ ,\quad \dot\gamma_t = e^{\alpha_t}, \tag{8}$$

are required for stability, and simplify the analysis greatly. Hamilton's principle states that minimizing the action functional $J[\boldsymbol{\lambda}] = \int_{\mathbb{T}} \mathbb{L}(\boldsymbol{\lambda}_t,\dot{\boldsymbol{\lambda}}_t,t)dt$ amounts to finding trajectories $\boldsymbol{\lambda}_t$ that satisfy the Euler-Lagrange equations

$$\frac{\partial\mathbb{L}}{\partial\boldsymbol{\lambda}_t}(\boldsymbol{\lambda}_t,\dot{\boldsymbol{\lambda}}_t,t) - \frac{d}{dt}\frac{\partial\mathbb{L}}{\partial\dot{\boldsymbol{\lambda}}_t}(\boldsymbol{\lambda}_t,\dot{\boldsymbol{\lambda}}_t,t) = 0. \tag{9}$$

Euler-Lagrange equations associated with the Bregman Lagrangian (7) involve the gradient of the dual function, which requires continuous differentiability of $G(\boldsymbol{\lambda})$ in order to exist. Note that by Danskin's theorem [27], the sub-differential of the dual function is defined as

$$\partial G(\boldsymbol{\lambda}) := \{A\overline{\mathbf{x}}(\boldsymbol{\lambda}) - \mathbf{b}\colon \overline{\mathbf{x}}(\boldsymbol{\lambda}) \in \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^n} \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})\}\ , \tag{10}$$

which may not be a singleton since $\overline{\mathbf{x}}(\boldsymbol{\lambda})$ is not necessarily unique. However, we establish that the following smoothness condition on the primal objective $f(\mathbf{x})$ is *sufficient* for the dual function to be continuously differentiable, and will make the evaluation of the Euler-Lagrange system (9) possible for our setting (7).

Assumption 1 (Strong Convexity) *The objective function $f(\mathbf{x})$ is twice continuously differentiable with $\nabla^2 f(\mathbf{x}) \succeq m_f\mathbf{I}$ for some $0 < m_f < \infty$.*

The strong convexity of $\mathbf{x} \mapsto f(\mathbf{x})$ implies the strong convexity of $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x},\boldsymbol{\lambda})$ which further implies that the Lagrangian minimizer $\overline{\mathbf{x}}(\boldsymbol{\lambda})$ [cf. (10)] is unique for each $\boldsymbol{\lambda}$. Uniqueness of $\overline{\mathbf{x}}(\boldsymbol{\lambda})$ implies that the sub-differential of the dual function is a singleton, i.e., the dual function is differentiable. More formally, we have the following smoothness properties for the dual function under Assumption 1.

Lemma 1 *Under Assumption 1, the dual function $G(\boldsymbol{\lambda})$ is twice-continuously differentiable with Lipschitz continuous*

gradient, i.e.,

$$\|\nabla G(\boldsymbol{\lambda}) - \nabla G(\boldsymbol{\nu})\|_2 \leq \frac{\|A\|_2^2}{m_f}\|\boldsymbol{\lambda} - \boldsymbol{\nu}\|_2. \tag{11}$$

for all $\boldsymbol{\lambda}, \boldsymbol{\nu} \in \mathbb{R}^m$, $\|A\|_2 = \lambda_{\max}(A^\top A)^{1/2}$. *Furthermore, the Hessian of the dual function is given by*

$$\nabla^2 G(\boldsymbol{\lambda}) = -A[\nabla^2 f(\overline{\mathbf{x}}(\boldsymbol{\lambda}))]^{-1}A^\top. \tag{12}$$

Under these conditions, we may simplify the partial differential equation (PDE) in (9) to a dynamical system that involves the gradient of the dual function, as we state next.

Proposition 1 *Consider the Lagrangian mechanical system defined by the Bregman Lagrangian in* (7) *associated with the optimization problem in* (5) *under the ideal scaling conditions* (8)*. Then under Assumption 1, the Euler-Lagrange equations in* (9) *for trajectories $\boldsymbol{\lambda}_t$ that minimize the action functional $J[\boldsymbol{\lambda}] = \int_{\mathbb{T}} \mathbb{L}(\boldsymbol{\lambda}_t, \dot{\boldsymbol{\lambda}}_t, t)dt$ are given by*

$$\begin{aligned}&\ddot{\boldsymbol{\lambda}}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{\boldsymbol{\lambda}}_t \\ &\quad - e^{2\alpha_t + \beta_t}[\nabla^2\psi(\boldsymbol{\lambda}_t + e^{-\alpha_t}\dot{\boldsymbol{\lambda}}_t)]^{-1}\nabla G(\boldsymbol{\lambda}_t) = 0\ ,\end{aligned} \tag{13}$$

which may be equivalently stated without inverting the Hessian in (13) *as*

$$\frac{d}{dt}\nabla_{\boldsymbol{\lambda}}\psi(\boldsymbol{\lambda}_t + e^{-\alpha_t}\dot{\boldsymbol{\lambda}}_t) = e^{\alpha_t + \beta_t}\nabla G(\boldsymbol{\lambda}_t). \tag{14}$$

The Euler-Lagrange equations presented in Proposition 1 are derived by applying Hamilton's Principle for trajectories $\boldsymbol{\lambda}_t$ in the dual domain $\mathbb{R}^m$. It is unclear, however, what relationship such trajectories play in solving the optimization problem under consideration in (2). In Subsection III-C, we develop another dynamical system in the primal domain, coupled to the Euler-Lagrange equations (14), which yield solutions that converge exponentially to a saddle point of the Lagrangian [cf. (3)]. Before that, we study the stability of (14) next.

B. Stability Analysis

We now turn to studying the convergence properties of the dynamical system given in Proposition 1. First, we present a lemma regarding the evolution of dual sub-optimality to be used to develop continuous-time accelerated dual ascent.

Lemma 2 *Consider the Euler-Lagrange dynamics* (14) *presented in Proposition 1 under the ideal scaling conditions $\dot{\beta}_t \leq \dot{\gamma}_t = e^{\alpha_t}$ in* (8) *with initialization $\boldsymbol{\lambda}_0,\ \dot{\boldsymbol{\lambda}}_0 \in \mathbb{R}^m$. Then under Assumption 1, the dual sub-optimality satisfies*

$$\frac{m_f}{2\|A\|_2^2}\|\nabla G(\boldsymbol{\lambda}_t)\|_2^2 \leq G(\boldsymbol{\lambda}^*) - G(\boldsymbol{\lambda}_t) \leq \mathcal{O}(e^{-\beta_t})\ , \tag{15}$$

where the proper selection of the scalar real-valued function β_t determines the rate of convergence.

Lemma 2, building upon the results of Theorem 2.1 in [24], establishes a bound on the suboptimality gap evaluated along trajectories which satisfy the Euler-Lagrange equations given in (14). In particular, for $\beta_t = ct, c > 0$, this gap vanishes exponentially fast.

C. Evolution of Lagrangian Minimizers

To evaluate the dual gradient $\nabla G(\boldsymbol{\lambda}_t) = A\overline{\mathbf{x}}(\boldsymbol{\lambda}_t) - \mathbf{b}$ in the Euler-Lagrange ordinary differential equations (ODE) (14), we need to continuously evaluate the Lagrangian minimizer $\overline{\mathbf{x}}(\boldsymbol{\lambda}_t) = \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_t)$. As we see below, the smoothness of $f(\mathbf{x})$ allows us to continuously compute this minimizer without performing the minimization all the time, under appropriate initialization. That is, apply the chain rule to the dual feasibility identity $\nabla_{\mathbf{x}}\mathcal{L}(\overline{\mathbf{x}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = 0$ and rearrange terms to obtain

$$\frac{d}{d\boldsymbol{\lambda}^\top}\overline{\mathbf{x}}(\boldsymbol{\lambda}) = -[\nabla^2 f(\overline{\mathbf{x}}(\boldsymbol{\lambda}))]^{-1}A^\top. \tag{16}$$

where $[\frac{d}{d\boldsymbol{\lambda}^\top}\overline{\mathbf{x}}(\boldsymbol{\lambda})]_{ij} = \frac{d}{d\lambda_j}\overline{x}_i(\boldsymbol{\lambda})$. Notice that the last result requires $f(\mathbf{x})$ to be twice continuously differentiable with invertible Hessian (Assumption 1). Given the time evolution of $\boldsymbol{\lambda}_t$, $\overline{\mathbf{x}}_t := \overline{\mathbf{x}}(\boldsymbol{\lambda}_t)$ obeys the ODE

$$\frac{d}{dt}\overline{\mathbf{x}}(\boldsymbol{\lambda}_t) = \Big[\frac{d}{d\boldsymbol{\lambda}_t^\top}\overline{\mathbf{x}}(\boldsymbol{\lambda}_t)\Big]\frac{d}{dt}\boldsymbol{\lambda}_t. \tag{17}$$

with initial condition $\overline{\mathbf{x}}_0 = \overline{\mathbf{x}}(\boldsymbol{\lambda}_0) = \arg\min_{\mathbf{x}\in\mathbb{R}^n}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_0)$. Combining the expressions in (17) and (16) along with (14) yields the continuous-time dynamical system

$$\dot{\nabla}\psi(\boldsymbol{\lambda}_t + e^{-\alpha_t}\dot{\boldsymbol{\lambda}}_t) = e^{\alpha_t+\beta_t}(A\overline{\mathbf{x}}_t - \mathbf{b}), \tag{18a}$$

$$\dot{\overline{\mathbf{x}}}_t = -[\nabla^2 f(\overline{\mathbf{x}}_t)]^{-1}A^\top\dot{\boldsymbol{\lambda}}_t. \tag{18b}$$

The states of (18) are $\boldsymbol{\lambda}_t, \dot{\boldsymbol{\lambda}}_t$ and $\overline{\mathbf{x}}_t$. Intuitively, the first ODE (18a) executes an accelerated gradient flow on the dual function while the second ODE (18b) maintains the dual feasibility $\nabla_{\mathbf{x}}\mathcal{L}(\overline{\mathbf{x}}_t, \boldsymbol{\lambda}_t) = \mathbf{0}$ all the time.

Next, we establish that the dynamical system defined by (18) converges to the saddle point of the Lagrangian, and hence, solves (2) at an exponential rate (depending on the choice of β_t – see Section IV).

Theorem 1 *Under Assumption 1, the primal-dual flow $(\overline{\mathbf{x}}_t, \boldsymbol{\lambda}_t)$ defined by the dynamical system* (18) *with initialization $\boldsymbol{\lambda}_0,\ \dot{\boldsymbol{\lambda}}_0 \in \mathbb{R}^m, \overline{\mathbf{x}}_0 = \arg\min_{\mathbf{x}}\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_0)$ and under the ideal scaling conditions* (8) *obeys the following bounds,*

$$\frac{m_f}{2\|A\|_2^2}\|A\overline{\mathbf{x}}_t - \mathbf{b}\|_2^2 \leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) - \mathcal{L}(\overline{\mathbf{x}}_t, \boldsymbol{\lambda}_t) \leq \mathcal{O}(e^{-\beta_t}). \tag{19}$$

Theorem 1 establishes that the dynamical system (18) yields solutions that converge to the saddle point of the Lagrangian, possibly exponentially fast, depending on the selection of β_t. The solutions are dual feasible all the time. However, primal feasibility is achieved asymptotically, and requires strong convexity of the primal objective function [cf. the left inequality in (19)].

D. Parameter Selection

To achieve exponential convergence, one may set $\beta_t = ct,\ c > 0$ in (18), and set the other parameters accordingly to satisfy the ideal scaling (8) used to derive the Euler-Lagrange PDE. For this choice we have a *first-order* dual method for constrained optimization in continuous time with

exponential convergence. These results are for the case that the primal objective is twice-continuously differentiable and strongly convex (Assumption 1).

Alternatively, consider the choice of a logarithmic (polynomial) regime as $e^{\alpha_t} = p/t$, $e^{\beta_t} = Ct^p$, $e^{\gamma_t} = t^p, p > 0$, with $C > 0$ an arbitrary positive scalar. For this case, the Euler-Lagrange ODE (13) simplifies to

$$\ddot{\boldsymbol{\lambda}}_t + \frac{p+1}{t}\dot{\boldsymbol{\lambda}}_t - Cp^2 t^{(p-2)}\big(\nabla^2\psi(\boldsymbol{\lambda}_t + \frac{t}{p}\dot{\boldsymbol{\lambda}}_t)\big)^{-1}\nabla G(\boldsymbol{\lambda}_t) = 0. \tag{20}$$

Lemma 2 implies that (20) attains a polynomial convergence rate $\mathcal{O}(1/t^p)$.

In the next section, we shift focus to the discretization of the ODE (18). However, this discretization is subtle, and not any parameter selection of α_t, β_t, and γ_t which satisfies the ideal scaling (8) stably translates into discrete-time algorithms. Therefore, we consider algorithmic schemes for the polynomial scaling case, i.e. discretizing the polynomial Euler-Lagrange ODE (20). Since the discretized scheme no longer requires continuous evaluation of the dual gradient, we may relax the smoothness conditions of Assumption 1 so as to encompass a broader category of linearly constrained problems – see Section IV.

Remark 1 (Euler-Lagrange vs. Dual Ascent) Under Assumption 1, the dual function is twice differentiable with Lipschitz gradient. However, it is not necessarily strongly concave. [1] Therefore, under the current technical setting, standard dual gradient update law $\dot{\boldsymbol{\lambda}}_t = \nabla G(\boldsymbol{\lambda}_t)$ together with (18b) yield $\mathcal{O}(1/t)$ convergence rate [28, § 2], [29], as opposed to the $\mathcal{O}(1/t^p)$ rate of polynomial Euler-Lagrange ODE (20).

IV. Discrete-Time Dual Acceleration

In this section, we consider the discretization of Euler-Lagrange (18) with polynomial parameter selection [cf. (20)]. In [24], a rate-matching discretization scheme is proposed which relies on higher-order gradient methods to stably discretize the polynomial Eluer-Lagrange ODE (20). This scheme requires evaluation of the first $p-1$ derivatives of the dual function $G(\boldsymbol{\lambda})$ along with Lipschitz continuity of the $p-1$-th derivative to preserve the continuous-time convergence rate to $\mathcal{O}(1/k^p)$, where k is the discrete iteration index counting the number of dual updates. Explicity, the *accelerated higher-order gradient method* performs the following updates for minimizing the convex function $\boldsymbol{\lambda} \mapsto -G(\boldsymbol{\lambda})$ over $\mathbb{R}^m$: start with $\boldsymbol{\lambda}_0 = \boldsymbol{\mu}_{-1} \in \mathbb{R}^m$ and inductively define

$$\begin{aligned}
\boldsymbol{\nu}_k &= \arg\min_{\boldsymbol{\nu}\in\mathbb{R}^m} \{-G_{p-1}(\boldsymbol{\nu};\boldsymbol{\lambda}_k) + \frac{NL_{p-1}}{p!}\|\boldsymbol{\nu}-\boldsymbol{\lambda}_k\|^p\},\\
\boldsymbol{\mu}_k &= \arg\min_{\boldsymbol{\mu}\in\mathbb{R}^m}\{-Cpk^{(p-1)}\nabla G(\boldsymbol{\nu}_k)^\top\boldsymbol{\mu} + \frac{L_{p-1}}{(p-1)!}D_\psi(\boldsymbol{\mu},\boldsymbol{\mu}_{k-1})\},\\
\boldsymbol{\lambda}_{k+1} &= \frac{p}{k+p}\boldsymbol{\mu}_k + \frac{k}{k+p}\boldsymbol{\nu}_k.
\end{aligned} \tag{21}$$

where $G_{p-1}(\boldsymbol{\nu};\boldsymbol{\lambda}_k)$ is $p-1$-th order Taylor expansion of $G(\boldsymbol{\nu})$ around $\boldsymbol{\lambda}_k$, $k^{(p-1)} := k(k+1)\cdots(k+p-2)$ is the rising factorial, $C \le (N^2-1)^{\frac{p-2}{2}}((2N)^{p-1}p^p)$ is a positive scalar which arises in the polynomial scaling (20), $D_\psi(\cdot,\cdot)$ is the Bregman divergence (6), $N > 1$ is arbitrary, and $L_{p-1} > 0$ is the Lipschitz constant of the $p-1$-th derivative of $G(\boldsymbol{\lambda})$.

The first update is a *higher-order* dual gradient step at $\boldsymbol{\lambda}_k$ to generate the auxiliary dual variable $\boldsymbol{\nu}_k \in \mathbb{R}^m$, whose update depends on the truncated Taylor expansion $G_{p-1}(\boldsymbol{\nu};\boldsymbol{\lambda}_k)$ of the dual function – special instances are discussed in the following subsections. The second step resembles a dual mirror ascent step, generating an auxiliary sequence $\boldsymbol{\mu}_k \in \mathbb{R}^m$ in which the dual gradient is computed at $\boldsymbol{\nu}_k$ (rather than $\boldsymbol{\mu}_{k-1}$ in standard dual ascent). Finally, we consider the update of the actual Lagrange multiplier $\boldsymbol{\lambda}_{k+1}$ as a weighted combination of dual auxiliary sequences $\boldsymbol{\nu}_k$ and $\boldsymbol{\mu}_k$.

The algorithm (21), the discrete-time counter part of the ODE (20), exhibits the converges rate $\mathcal{O}(1/k^p)$ [24, §3.4]. Hypothetically, one may achieve faster convergence by increasing p in (21) at the expense of requiring higher-order gradient information about the dual function. Since this information is precluded from use in practical settings, we restrict our focus to $p = 2$ and $p = 3$, corresponding to dual variants of Nesterov acceleration [16] and cubic regularized Newton's method [30], [31] when discretizing (20).

A. Accelerated Method of Multipliers: $p = 2$

Specialized to the case that $p = 2$, implementation and convergence of (21) requires the dual function to be continuously differentiable with *Lipschitz gradient*. Assumption 1 is *sufficient* for these regularity conditions to hold, according to Lemma 1. We could still satisfy these conditions under a less restrictive assumption; that is, we may relax requirement of strong convexity and smoothness of $f(\mathbf{x})$ by adopting an augmented Lagrangian approach. To do so, define the augmented Lagrangian as

$$\mathcal{L}_\rho(\mathbf{x},\boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^\top(A\mathbf{x}-\mathbf{b}) + \frac{\rho}{2}\|A\mathbf{x}-\mathbf{b}\|^2\,, \tag{22}$$

where $\rho > 0$ is arbitrary. Notice that the quadratic term does not alter the saddle points $(\mathbf{x}^\star,\boldsymbol{\lambda}^\star) \in \mathbf{X}^\star \times \boldsymbol{\Lambda}^\star$ as it is null on the feasible set. For a fixed $\boldsymbol{\lambda}$, the Lagrangian minimizer $\overline{\mathbf{x}}(\boldsymbol{\lambda}) = \arg\min \mathcal{L}_\rho(\mathbf{x},\boldsymbol{\lambda})$ is no longer unique, opening the possibility for the dual function to be non-differentiable [cf. (10)]. However, as we show next, the quadratic penalty term in (22) renders a continuously differentiable dual function $G(\boldsymbol{\lambda}) = \min_{\mathbf{x}} \mathcal{L}_\rho(\mathbf{x},\boldsymbol{\lambda})$ with Lipschitz gradient, under less restrictive Assumptions on $f(\mathbf{x})$.

[1] In fact, according to (12), strong concavity of the dual function requires $f(\mathbf{x})$ to have uniformly bounded Hessian, i.e., $\nabla^2 f(\mathbf{x}) \preceq M_f\mathbf{I}_n$ for some $0 < M_f < \infty$ and all $\mathbf{x} \in \mathbb{R}^n$.

Algorithm 1 $p = 2$: Accelerated Method of Multipliers

Require: Augmented Lagrangian $\mathcal{L}_\rho(\mathbf{x}, \boldsymbol{\lambda})$ [cf. (22)], scaling parameters C and N such that $C \leq 1/(8N)$, $N > 1$, augmentation constant $\rho > 0$ (determines algorithm step-size).
initialize primal $\mathbf{x}_0 \in \mathbb{R}^n$, dual variables $\boldsymbol{\lambda}_0 = \boldsymbol{\mu}_{-1} \in \mathbb{R}^m$
for $k = 0, 1, 2, \ldots$ **do**
Compute primal minimizer $\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{L}_\rho(\mathbf{x}, \boldsymbol{\lambda}_k)$.
Evaluate dual gradient $\nabla G(\boldsymbol{\lambda}_k) = A\mathbf{x}_{k+1} - \mathbf{b}$.
Compute dual ascent step

$$\boldsymbol{\nu}_k = \arg\min_{\boldsymbol{\nu}\in\mathbb{R}^m} \{-\nabla G(\boldsymbol{\lambda}_k)^\top(\boldsymbol{\nu} - \boldsymbol{\lambda}_k) + \tfrac{N}{2\rho}\|\boldsymbol{\nu} - \boldsymbol{\lambda}_k\|^2\}.$$

Compute auxiliary minimizer $\mathbf{y}_{k+1} \in \arg\min_{\mathbf{y}\in\mathbb{R}^n} \mathcal{L}_\rho(\mathbf{y}, \boldsymbol{\nu}_k)$.
Evaluate dual gradient $\nabla G(\boldsymbol{\nu}_k) = A\mathbf{y}_{k+1} - \mathbf{b}$.
Execute mirror ascent w.r.t. Bregman divergence D_ψ

$$\boldsymbol{\mu}_k = \operatorname{argmin}_{\boldsymbol{\mu}\in\mathbb{R}^m}\{-2Ck\nabla G(\boldsymbol{\nu}_k)^\top\boldsymbol{\mu} + \frac{1}{\rho}D_\psi(\boldsymbol{\mu},\boldsymbol{\mu}_{k-1})\}$$

Update Lagrange multiplier $\boldsymbol{\lambda}_{k+1}$ as weighted average of auxiliary dual variables $\boldsymbol{\nu}_k$ and $\boldsymbol{\mu}_k$

$$\boldsymbol{\lambda}_{k+1} = \frac{2}{k+2}\boldsymbol{\mu}_k + \frac{k}{k+2}\boldsymbol{\nu}_k.$$

end for

Lemma 3 *When the primal objective $f(\mathbf{x})$ is convex (and possibly extended-real-valued), the dual function $G(\boldsymbol{\lambda})$ associated with the augmented Lagrangian* (22) *is continuously differentiable with $1/\rho$-Lipschitz gradient given by $\nabla G(\boldsymbol{\lambda}) = A\overline{\mathbf{x}}(\boldsymbol{\lambda}) - \mathbf{b}$.*

Intuitively, the penalty term induces favorable curvature profiles of strong convexity onto weakly convex functions [2]. Now, we substitute use of the dual function linearization,

$$G_1(\boldsymbol{\lambda}; \boldsymbol{\lambda}_k) = G(\boldsymbol{\lambda}_k) + \nabla G(\boldsymbol{\lambda}_k)^\top(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k), \tag{23}$$

into the algorithm (21). The resulting accelerated variation of the method of multipliers (AMM) is summarized in Algorithm 1. Observe that we have $G(\boldsymbol{\lambda}_k) = \mathcal{L}(\overline{\mathbf{x}}(\boldsymbol{\lambda}_k), \boldsymbol{\lambda}_k)$ and $\nabla G(\boldsymbol{\lambda}_k) = A\overline{\mathbf{x}}(\boldsymbol{\lambda}_k) - \mathbf{b}$. Therefore, for evaluating the dual function and its gradient at $\boldsymbol{\lambda}_k$ in (23), an *exact* minimization $\overline{\mathbf{x}}(\boldsymbol{\lambda}_k) \in \arg\min_{\mathbf{x}} \mathcal{L}_\rho(\mathbf{x}, \boldsymbol{\lambda}_k)$ is required. We now establish the convergence of dual accelerated methods at Nesterov's optimal $\mathcal{O}(1/k^2)$ *without* strong convexity.

Theorem 2 *For the optimization problem* (2) *and the corresponding augmented Lagrangian* (22)*, consider the algorithm* (21) *with $p = 2$, the linearized dual function $G_1(\boldsymbol{\lambda}; \boldsymbol{\lambda}_k)$ defined in* (23)*, and constants $L_1 = \rho^{-1}$, $N > 1$, and $C \leq 1/(8N)$, as presented in Algorithm 1. Then, the primal sequence $\{\mathbf{y}_k\}$ and the dual sequence $\{\boldsymbol{\nu}_k\}$ satisfy*

$$\frac{\rho}{2}\|A\mathbf{y}_{k+1} - \mathbf{b}\|_2^2 \leq p^\star - G(\boldsymbol{\nu}_k) \leq \mathcal{O}(1/(\rho k^2)). \tag{24}$$

The result of Theorem 2 establishes a $\mathcal{O}(1/k^2)$ rate for accelerated dual ascent when the primal objective is weakly convex and the constraints are linear. In the next subsection, we consider further accelerating this scheme through the use of second-order information of the dual function, which corresponds to the case $p = 3$.

[2]Notice, however, that the augmented Lagrangian is not strongly convex despite adding the quadratic term.

Algorithm 2 $p = 3$: Accelerated Dual Newton Method

Require: (Non-augmented) Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ [cf. (3)], scaling parameters C and N such that $C \leq \sqrt{N^2-1}/(108N^2)$, $N > 1$, step-size parameter $L_2 = C_f\,\|A\|_2^3/m_f^3$ [cf. (26)].
initialize primal $\mathbf{x}_0 \in \mathbb{R}^n$, dual variables $\boldsymbol{\lambda}_0 = \boldsymbol{\mu}_{-1} \in \mathbb{R}^m$
for $k = 0, 1, 2, \ldots$ **do**
Compute primal minimizer $\mathbf{x}_{k+1} \in \arg\min_{\mathbf{x}\in\mathbb{R}^n} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_k)$
Evaluate dual gradient $\nabla G(\boldsymbol{\lambda}_k) = A\mathbf{x}_{k+1} - \mathbf{b}$.
Evaluate dual Hessian $\nabla^2 G(\boldsymbol{\lambda}_k) = -A[\nabla^2 f(\mathbf{x}_{k+1})]^{-1}A^\top$.
Compute cubic dual Newton step

$$\boldsymbol{\nu}_k = \operatorname*{argmin}_{\boldsymbol{\nu}\in\mathbb{R}^m}\{-\nabla G(\boldsymbol{\lambda}_k)^\top(\boldsymbol{\nu} - \boldsymbol{\lambda}_k) + \tfrac{1}{2}(\boldsymbol{\nu} - \boldsymbol{\lambda}_k)^\top\nabla^2 G(\boldsymbol{\lambda}_k)(\boldsymbol{\nu} - \boldsymbol{\lambda}_k) + \tfrac{NL_2}{3!}\|\boldsymbol{\nu} - \boldsymbol{\lambda}_k\|^3\},$$

Compute auxiliary minimizer $\mathbf{y}_{k+1} \in \arg\min_{\mathbf{y}\in\mathbb{R}^n} \mathcal{L}(\mathbf{y}, \boldsymbol{\nu}_k)$.
Evaluate dual gradient $\nabla G(\boldsymbol{\nu}_k) = A\mathbf{y}_{k+1} - \mathbf{b}$.
Execute mirror ascent w.r.t. Bregman divergence D_ψ

$$\boldsymbol{\mu}_k = \operatorname*{argmin}_{\boldsymbol{\mu}\in\mathbb{R}^m}\{-3C(k^2+k)\nabla G(\boldsymbol{\nu}_k)^\top\boldsymbol{\mu} + \tfrac{L_2}{2}D_\psi(\boldsymbol{\mu}, \boldsymbol{\mu}_{k-1})\},$$

Update Lagrange multiplier $\boldsymbol{\lambda}_{k+1}$ as weighted average of auxiliary dual variables $\boldsymbol{\nu}_k$ and $\boldsymbol{\mu}_k$

$$\boldsymbol{\lambda}_{k+1} = \frac{3}{k+3}\boldsymbol{\mu}_k + \frac{k}{k+3}\boldsymbol{\nu}_k.$$

end for

B. $p = 3$: Accelerated Dual Newton Method

Next we turn to developing a dual variant of the cubic regularized Newton method [30]. This development corresponds to specializing the dual higher-order gradient scheme (21) to the case that $p = 3$. To do so, we require the dual function to be twice continuously differentiable with *Lipschitz Hessian*, which may be guaranteed by Assumption 1 as well as the following condition.

Assumption 2 (Lipschitz Hessian) *The objective function $f(\mathbf{x})$ is twice continuously differentiable with Lipschitz continuous Hessian $\nabla^2 f(\mathbf{x})$, i.e.,*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq C_f\|\mathbf{x} - \mathbf{y}\|. \tag{25}$$

For some $0 \leq C_f < \infty$ and all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Assumption 1 together with Assumption 2 allow us to establish that the dual function satisfies the requisite smoothness properties necessary to derive an accelerated dual Newton method, as we state next.

Lemma 4 *Under Assumptions 1 and 2, the dual function $G(\boldsymbol{\lambda})$ is twice-continuously differentiable with its Hessian $\nabla^2 G(\boldsymbol{\lambda}) = -A\nabla^2 f(\overline{\mathbf{x}}(\boldsymbol{\lambda}))^{-1}A^\top$ satisfying*

$$\|\nabla^2 G(\boldsymbol{\lambda}) - \nabla^2 G(\boldsymbol{\nu})\|_2 \leq \frac{C_f}{m_f^3}\|A\|_2^3\|\boldsymbol{\lambda} - \boldsymbol{\nu}\|_2. \tag{26}$$

for all $\boldsymbol{\lambda}, \boldsymbol{\nu} \in \mathbb{R}^m$.

We now turn to the algorithm (21) where the first and second-order derivatives of the dual function are used. Define the second-order Taylor expansion of the dual function as

$$G_2(\boldsymbol{\lambda}; \boldsymbol{\lambda}_k) = G(\boldsymbol{\lambda}_k) + \nabla G(\boldsymbol{\lambda}_k)^\top (\boldsymbol{\lambda} - \boldsymbol{\lambda}_k) + \frac{1}{2}(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k)^\top \nabla^2 G(\boldsymbol{\lambda}_k)(\boldsymbol{\lambda} - \boldsymbol{\lambda}_k), \quad (27)$$

Observe that we have $G(\boldsymbol{\lambda}_k) = \mathcal{L}(\overline{\mathbf{x}}(\boldsymbol{\lambda}_k), \boldsymbol{\lambda}_k)$, $\nabla G(\boldsymbol{\lambda}_k) = A\overline{\mathbf{x}}(\boldsymbol{\lambda}_k) - \mathbf{b}$, and $\nabla^2 G(\boldsymbol{\lambda}_k) = -A[\nabla^2 f(\overline{\mathbf{x}}(\boldsymbol{\lambda}_k))]^{-1} A^\top$. Hence, we need to incorporate the exact minimization step $\overline{\mathbf{x}}(\boldsymbol{\lambda}_k) \in \arg\min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_k)$ into the updates. The resulting iterative scheme is summarized in Algorithm 2. With Lemma 4 established, and the technical setting clarified, we present our main result for accelerated second-order dual methods.

Theorem 3 *For the optimization problem* (2) *and the corresponding Lagrangian* (3)*, consider the algorithm* (21) *with* $p = 3$*, the quadratic approximation of the dual function* $G_2(\boldsymbol{\lambda}; \boldsymbol{\lambda}_k)$ *defined in* (27)*, and constants* $L_2 = C_f \|A\|_2^3 / m_f^3$*,* $N > 1$*, and* $C \leq \sqrt{N^2 - 1}/(108N^2)$*. Then, under Assumption 1 and Assumption 2, the primal sequence* $\{\mathbf{y}_k\}$ *and the dual sequence* $\{\boldsymbol{\nu}_k\}$ *generated by Algorithm 2 satisfy*

$$\frac{m_f}{2\|A\|_2^2} \|A\mathbf{y}_{k+1} - \mathbf{b}\|_2^2 \leq p^\star - G(\boldsymbol{\nu}_k) \leq \mathcal{O}(1/k^3). \quad (28)$$

In Theorem 3 we establish that a second-order accelerated dual approach to solving (1) yields the sequence which converges to a dual optimal point (5), and hence, by strong duality we converge to a saddle point of the Lagrangian (3). The rate at which we find the primal-dual optimal pair is $\mathcal{O}(1/k^3)$, which is state of the art relative to existing dual approaches to optimization with linear constraints. This result requires strong convexity of the primal objective (Assumption 1) and Lipschitz continuity of its Hessian (Assumption 2). In the next section we illustrate that favorable convergence properties of (18) in continuous time translate into practice.

V. Numerical Evaluation

We consider the following synthetic optimization problem

$$p^\star = \min_{\mathbf{x} \in \mathbb{R}^n} \ \frac{1}{2}\mathbf{x}^\top P \mathbf{x} + \mu \exp(\mathbf{1}_n^\top \mathbf{x}) \ \text{ s.t. } \ A\mathbf{x} = \mathbf{b}, \quad (29)$$

The above problem is convex with a twice-differentiable strongly convex objective when $\mu \geq 0$ and $P \succ 0$. For the problem data, we have $\mu = 0.01$, $n = 100$, $m = 50$, $P = QQ^\top$ where elements of $Q \in \mathbb{R}^n$ are independently drawn from standard normal distribution. The elements of $A \in \mathbb{R}^{m \times n}$ are also drawn from standard normal distribution. Finally, $\mathbf{b} = A\mathbf{z} \in \mathcal{R}(A)$ where $\mathbf{z} \in \mathbb{R}^n$ is an instance of standard normal distribution. For our problem data, the condition number of P is 4.8902×10^4. We then consider different selection of the scaling parameters, namely: exponential regime where $\beta_t = t, e^{\alpha_t} = t$ (which corresponds to $\mathcal{O}(e^{-t})$ convergence rate), and polynomial regime where

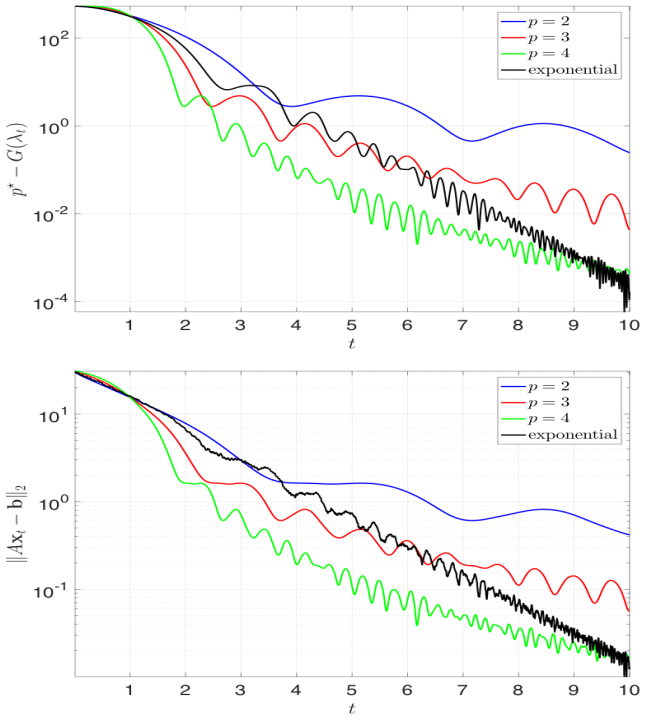

Fig. 1: The continuous-time accelerated dual ascent ODE (18) applied to (29) for various realizations of the scaling parameters [cf. Subsection III-D]. (Top) Evolution of dual sub-optimality gap $p^\star - G(\boldsymbol{\lambda}_t)$ against t in log-linear scale. (Bottom) Evolution of $\|A\overline{\mathbf{x}}_t - \mathbf{b}\|_2$ against t in log-linear scale [cf. (19)]. By increasing p, faster convergence rates are attained at the expense of more frequent sampling by the solver for a stable path generation.

$e^{\beta_t} = t^p, e^{\alpha_t} = p/t, p = 2, 3, 4$ (which corresponds to $\mathcal{O}(1/t^p)$ convergence rate). We then solve a modified version of (18):

$$\dot{\nabla}\psi(\boldsymbol{\lambda}_t + e^{-\alpha_t}\dot{\boldsymbol{\lambda}}_t) = e^{\alpha_t + \beta_t}(A\overline{\mathbf{x}}_t - \mathbf{b}), \quad (30a)$$

$$\dot{\overline{\mathbf{x}}}_t = -[\nabla^2 f(\overline{\mathbf{x}}_t)]^{-1}(\nabla f(\overline{\mathbf{x}}_t) + A^\top \boldsymbol{\lambda}_t + A^\top \dot{\boldsymbol{\lambda}}_t). \quad (30b)$$

The initial conditions are $\boldsymbol{\lambda}_0 = \dot{\boldsymbol{\lambda}}_0 = 0$, and $\overline{\mathbf{x}}_0 = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}_0)$, and the Bregman distance function ψ is chosen to be the Euclidean norm. Comparing to (18), the above ODE includes an additional correction term $\nabla_{\mathbf{x}} \mathcal{L}(\overline{\mathbf{x}}_t, \boldsymbol{\lambda}_t) = \nabla f(\overline{\mathbf{x}}_t) + A^\top \boldsymbol{\lambda}_t$ which corrects the trajectory and maintains the dual feasibility condition $\nabla_{\mathbf{x}} \mathcal{L}(\overline{\mathbf{x}}_t, \boldsymbol{\lambda}_t) = \mathbf{0}$ in presence of numerical errors [32]. We simulate the corresponding ODEs over the time interval $t \in [10^{-1}, 10]$. Note that the method must start at $t > 0$ due to the presence of the asymptote at the origin $t = 0$ in (20). We numerically integrate these dynamics using MATLAB ODE23 solver. The results are depicted in Figure 1. We plot the dual sub-optimality $p^\star - G(\boldsymbol{\lambda}_t)$ over t for various selection of the scaling parameters. We also plot $\|A\overline{\mathbf{x}}_t - \mathbf{b}\|_2$ as a barometer for attaining primal feasibility, which we clearly observe. Notice that convergence improves with increased p. Also, exponential convergence is attained by selecting the scaling parameters accordingly.

VI. Conclusion

We develop a family of accelerated methods, inspired by [24], that solves linearly-constrained convex optimization problems through their dual reformulation. Specifically, we derive a family of continuous-time dynamical systems that yields exponentially convergent trajectories to the saddle point of the problem when the objective function is twice continuously differentiable and strongly convex. Next we use an accelerated higher-order gradient method proposed in [24] to develop the discrete-time counterparts of this family. We specialize the algorithm to accelerated method of multipliers and accelerated dual Newton method with convergence rates of $\mathcal{O}(1/k^2)$ and $\mathcal{O}(1/k^3)$, respectively. The latter algorithm requires twice differentiability and strong convexity but the former is realizable under the less restrictive assumption of weak convexity and non-differentiability. The discrete-time algorithm assumes the Lagrangian (inner) minimization is performed exactly at each iteration, which translates into the exact dual gradient (or dual Hessian) information to be available. However, accelerated dual methods have been shown to necessarily suffer from inexact information [33]. Studying the convergence of the developed schemes with Lagrangian inexact minimization is the subject of future research.

References

[1] K. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[2] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc wsns with noisy links - part i: distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.

[3] F. Bullo, J. Cortés, and S. Martínez, *Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms:*, ser. Princeton Series in Applied Mathematics.

[4] A. Koppel, J. Fink, G. Warnell, E. Stump, and A. Ribeiro, "Online learning for characterizing," in *2016 IEEE International Conference in Intelligent Robots and Systems (IROS) (to appear)*. IEEE, 2016.

[5] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *Signal Processing, IEEE Transactions on*, vol. 58, no. 12, pp. 6369–6386, Dec 2010.

[6] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," *Signal Processing Magazine, IEEE*, vol. 31, no. 5, pp. 32–43, Sept 2014.

[7] Y. Cao, W. Yu, W. Ren, and G. Chen, "An overview of recent progress in the study of distributed multi-agent coordination," *Industrial Informatics, IEEE Transactions on*, vol. 9, no. 1, pp. 427–438, 2013.

[8] A. Ribeiro, "Optimal resource allocation in wireless communication and networking," *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, pp. 1–19, 2012.

[9] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar *et al.*, "Convex analysis and optimization," 2003.

[10] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[11] M. Zargham, A. Ribeiro, A. Ozdaglar, and A. Jadbabaie, "Accelerated dual descent for network flow optimization," *IEEE Transactions on Automatic Control*, vol. 59, no. 4, pp. 905–920, 2014.

[12] J. Goodman, "Newton's method for constrained optimization," *Mathematical Programming*, vol. 33, no. 2, pp. 162–171, 1985.

[13] A. Mokhtari, W. Shi, Q. Ling, and A. Ribeiro, "A decentralized second-order method with exact linear convergence rate for consensus optimization," *arXiv preprint arXiv:1602.00596*, 2016.

[14] R. Tutunov, H. B. Ammar, and A. Jadbabaie, "A distributed newton method for large scale consensus optimization," *arXiv preprint arXiv:1606.06593*, 2016.

[15] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[16] ——, "A method of solving a convex programming problem with convergence rate o (1/k2)."

[17] I. Daubechies, M. Fornasier, and I. Loris, "Accelerated projected gradient method for linear inverse problems with sparsity constraints," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5-6, pp. 764–792, 2008.

[18] A. Koppel, F. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization," *IEEE Trans. Signal Process.*, p. 15, Oct 2015.

[19] Y. Xu, "Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming," *arXiv preprint arXiv:1606.09155*, 2016.

[20] Y. Chen, G. Lan, and Y. Ouyang, "Optimal primal-dual methods for a class of saddle point problems," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 1779–1814, 2014.

[21] T. Goldstein, B. O'Donoghue, S. Setzer, and R. Baraniuk, "Fast alternating direction optimization methods," *SIAM Journal on Imaging Sciences*, vol. 7, no. 3, pp. 1588–1623, 2014.

[22] M. Kadkhodaie, K. Christakopoulou, M. Sanjabi, and A. Banerjee, "Accelerated alternating direction method of multipliers," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 497–506.

[23] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[24] A. Wibisono, A. C. Wilson, and M. I. Jordan, "A variational perspective on accelerated methods in optimization," *Proceedings of the National Academy of Sciences*, p. 201614734, 2016.

[25] M. Fazlyab, A. Koppel, A. Ribeiro, and V. M. Preciado, "Accelerated dual methods for constrained convex optimization," *IEEE Transactions on Automatic Control (under preparation)*, 2017.

[26] C. Bailey, "Hamilton's principle and the calculus of variations," *Acta Mechanica*, vol. 44, no. 1-2, pp. 49–57, 1982.

[27] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.

[28] W. Krichene, A. Bayen, and P. L. Bartlett, "Accelerated mirror descent in continuous and discrete time," in *Advances in Neural Information Processing Systems*, 2015, pp. 2845–2853.

[29] W. Su, S. Boyd, and E. J. Candes, "A differential equation for modeling nesterov?s accelerated gradient method: theory and insights," *Journal of Machine Learning Research*, vol. 17, no. 153, pp. 1–43, 2016.

[30] Y. Nesterov, "Accelerating the cubic regularization of newton?s method on convex problems," *Mathematical Programming*, vol. 112, no. 1, pp. 159–181, 2008.

[31] M. Baes, "Estimate sequence methods: extensions and approximations," 2009.

[32] M. Fazlyab, S. Paternain, V. M. Preciado, and A. Ribeiro, "Prediction-correction interior-point method for time-varying convex optimization," *arXiv preprint arXiv:1608.07544*, 2016.

[33] O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, vol. 146, no. 1-2, pp. 37–75, 2014.