A Variational Approach to Accelerated Dual Methods for Constrained Convex Optimization

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

***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 a discrete-time convergence rate of $\mathcal{O}(1/k^p)$, 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 arises from the discretization of the Euler-Lagrange dynamics associated with the constrained convex program. We demonstrate the performance of the aforementioned continuous-time framework via numerical simulations and evaluate the proposed discrete-time methods on a linear model predictive control problem.**

I. Introduction

Underlying many recent technological advances in artificial intelligence [1], smart devices [2], robotics [3], and wireless communications [4], is the mathematical theory of optimization. Our particular focus is on convex optimization problems with affine constraints, which apply to these respective contexts in the form of large-scale supervised learning [5], cooperative control [6], and wireless routing [7]. 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, i.e., problems of the form

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

where f is convex, n is possibly large, and the linear constraints may represent, for instance, consensus [8] or network flow constraints [7], [9].

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, which uses the second-derivative information and exhibits quadratic and/or super-linear convergence under proper initialization [10]. Newton's method requires evaluation of the Hessian inverse of the objective at each step, whose complexity is cubic in the decision variable dimension. In the case of large-scale supervised learning [5], 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 [11], [12].

On the other hand, first-order methods are popular due to their ease of implementation, low complexity, and robust albeit sublinear[1] $O(1/k)$ convergence [13] (linear convergence $\mathcal{O}(\rho^k),\ 0\leq\rho<1$ 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 [14].

Adaptations of acceleration have been proposed to constrained problems in the primal domain for special cases [15], as well as primal-dual schemes [16]–[18], and proximal dual approaches [7], [19], [20], which reach at least Nesterov's $\mathcal{O}(1/k^2)$ rate, but all require strong convexity. It is worth remarking 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.

Given that linear convergence remains elusive for first-order dual methods for constrained problems, even with strong convexity, we ask the following question: *is Nesterov's optimal rate $\mathcal{O}(1/k^2)$ realizable by first-order accelerated dual methods for constrained problems 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 [21] to dual approaches for constrained optimization (Section III). [21] 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 and Lagrangian mechanics, put forth in [21], 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 objectives that are neither strongly convex nor differentiable. Moreover, we

Work in this paper is supported by the NSF under grants CAREER-ECCS-1651433, IIS- 1447470, the ONR under grant N00014-12-1-0997, ARL MAST CTA, and ASEE SMART.

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

†Computational and Information Sciences, U.S. Army Research Laboratory, Adelphi, MD, 20783. Email: alec.e.koppel.civ@mail.mil.

[1]Throughout the paper, k counts the number of iterations.

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 via applications in continuous time to a synthetic example (Section V-A), as well as in discrete-time for a linear model predictive control problem (Section V-B).

Part of the results in this paper appeared as [22]. However, in this version, we include the proofs and provide additional numerical evaluation on a model predictive control task, which validates our theoretical results on practical problems.

II. Linearly Constrained Convex Optimization

A. Notation and preliminaries

Let $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ be the set of real, nonnegative, and positive numbers. We denote by $\mathbb{R}^{n\times m}$ and $\mathbb{S}^n$ the space of real $n \times m$ and symmetric $n \times n$ matrices, respectively. We denote the minimum and maximum singular values of $A \in \mathbb{R}^{n\times m}$ as $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$, respectively. The indicator function $\mathbb{I}_{\mathcal{X}}\colon \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ of a closed nonempty convex set $\mathcal{X} \subset \mathbb{R}^n$ is defined as $\mathbb{I}_{\mathcal{X}}(x) = 0$ if $x \in \mathcal{X}$, and $\mathbb{I}_{\mathcal{X}}(x) = +\infty$ otherwise. Let $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ be a closed proper function. The effective domain of f is denoted by $\mathrm{dom}(f) = \{x \in \mathbb{R}^n\colon f(x) < \infty\}$. A differentiable function f is (strongly) convex if and only if it satisfies

$$\nabla f(x)^\top (y - x) + \frac{m_f}{2}\|y - x\|_2^2 + f(x) \le f(y), \quad (2)$$

for all $x, y \in \mathrm{dom}(f)$. The parameter $m_f \ge 0$ quantifies the minimum curvature of f. A differentiable function f whose gradient is Lipschitz continuous with parameter $0 \le L_f < \infty$ satisfies

$$f(y) \le f(x) + \nabla f(x)^\top (y - x) + \frac{L_f}{2}\|y - x\|_2^2, \quad (3)$$

for all $x, y \in \mathrm{dom}(f)$. The parameter L_f quantifies the maximum curvature of f. We denote by $\mathcal{F}(m_f, L_f)$ the class of continuously differentiable functions satisfying both (2) and (3). Note that in this class, it must hold that $m_f \le L_f$. We define the condition number of $f \in \mathcal{F}(m_f, L_f)$ as $\kappa_f = L_f/m_f$. For the class $\mathcal{F}(m_f, L_f)$, the case $m_f = 0$ corresponds to weakly convex functions, and the case $L_f = \infty$ corresponds to nondifferentiable convex functions. For $g \in \mathcal{F}(0, \infty)$, we denote ∂g as the subdifferential of g, which is defined as

$$\partial g(x) = \{\gamma \in \mathrm{dom}(g)\colon g(x) + \gamma^\top (y - x) \le g(y)\}. \quad (4)$$

The indicator function $\mathbb{I}_{\mathcal{X}}(x)$ of a closed, nonempty, and convex set $\mathcal{X} \subset \mathbb{R}^n$ belongs to the class $\mathcal{F}(0, \infty)$. Finally, if $f \in \mathcal{F}(m_f, L_f)$ is *twice* continuously differentiable, its Hessian satisfies $m_f I_n \preceq \nabla^2 f(x) \preceq L_f I_n$.

B. Problem Statement

Consider the following convex optimization problem,

$$p^\star = \mathrm{minimize}_{x\in\mathbb{R}^n} f(x), \text{ s.t. } Ax = b, \quad (5)$$

where $f(x)\colon \mathbb{R}^n \cup \{+\infty\} \to \mathbb{R}$ is convex and possibly real extended-valued. We assume that the system of equations $Ax = b$ has infinitely many solutions; hence, the problem (5) is feasible and nontrivial. Note that this assumption includes the possibility of A not being full row rank. We further assume that the optimal $p^\star$ is finite. Finally, we may allow the decision variable x to be constrained to a nonempty "simple" convex set $\mathcal{X} \subset \mathbb{R}^n$, onto which projection can be easily performed. In this case, we may replace the objective function f in (5) with the non-smooth function $f(x) + \mathbb{I}_{\mathcal{X}}(x)$.

We turn to reformulating (5) in terms of its dual problem. To do so, define the Lagrangian function $\mathcal{L}(x, \lambda)\colon \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ associated with the problem (5) as

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^\top (Ax - b), \quad (6)$$

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

$$d(\lambda) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda). \quad (7)$$

The dual problem is then to maximize the dual function (7) with respect to λ,

$$d^\star = \sup_{\lambda\in\mathbb{R}^m} d(\lambda) = \sup_{\lambda\in\mathbb{R}^m} \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda). \quad (8)$$

Since the primal problem (5) is convex and feasible, the Slater's condition holds [23], resulting in zero duality gap, i.e., $d^\star = p^\star$. Thus, there is no loss of optimality by approaching the problem by its dual reformulation. Next, we adopt a mechanics perspective in order to derive an exponentially convergent solution to (8) in continuous time.

III. Dual Optimization as Lagrangian Mechanics

We now shift our focus to derive a solution to (8) in continuous time, using a variational approach.

A. Bregman Lagrangian and Hamilton's Principle

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

$$D_\psi(\nu, \lambda) = \psi(\nu) - \psi(\lambda) - \nabla\psi(\lambda)^\top (\nu - \lambda)\ . \quad (9)$$

Observe that a special case of ψ is the Euclidean distance, $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$. We denote $x_t \in \mathbb{R}^n$ and $\lambda_t \in \mathbb{R}^m$ as curves parameterized by a continuous time index $t \in \mathbb{T} = [t_0, t_f) \subseteq \mathbb{R}_+$.

As in [21], we identify the *kinetic* energy of a synthetic mechanical system associated with the optimization problem (8) as the Bregman divergence between the position vector and its perturbation by the velocity vector with an appropriate scaling, i.e., $D_\psi(\lambda + e^{-\alpha_t}\dot\lambda, \lambda)$. Further, the *potential* energy is the objective to be minimized, which for dual algorithms is the negative of the dual function $-d(\lambda)$ in (7). Thus, we may

consider the *Bregman Lagrangian* as the following weighted Lagrangian of the synthetic mechanical system,

$$\mathbb{L}(\lambda, \dot{\lambda}, t) = e^{\alpha_t + \gamma_t}\big(D_\psi(\lambda + e^{-\alpha_t}\dot{\lambda}, \lambda) + e^{\beta_t} d(\lambda)\big), \quad (10)$$

where $\alpha_t, \gamma_t, \beta_t \colon \mathbb{T} \to \mathbb{R}$ are smooth functions of time $t \in \mathbb{T}$. By applying Hamilton's Principle (see [24] for details) to (10), it turns out that the ideal scaling conditions

$$\dot{\beta}_t \leq \dot{\gamma}_t = e^{\alpha_t} \quad (11)$$

are required for stability, and simplify the analysis greatly [21]. Hamilton's principle states that minimizing the action functional $J[\lambda] = \int_{\mathbb{T}} \mathbb{L}(\lambda_t, \dot{\lambda}_t, t)dt$ amounts to finding trajectories λ_t that satisfy the Euler-Lagrange equations,

$$\frac{\partial \mathbb{L}}{\partial \lambda_t}(\lambda_t, \dot{\lambda}_t, t) - \frac{d}{dt}\frac{\partial \mathbb{L}}{\partial \dot{\lambda}_t}(\lambda_t, \dot{\lambda}_t, t) = 0. \quad (12)$$

By using (10) in (12), we find that the Euler-Lagrange equations associated with the Bregman Lagrangian (10) involve the gradient of the dual function, which requires continuous differentiability of $d(\lambda)$ in order to exist. Note that, by Danskin's theorem [25], the sub-differential of the dual function is given by

$$\partial d(\lambda) := \{A\overline{x}(\lambda) - b \colon \overline{x}(\lambda) \in \operatorname*{argmin}_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda)\}\ , \quad (13)$$

which may not be a singleton, since $\overline{x}(\lambda)$ is not necessarily unique. However, we will establish in Lemma 1 that, the following smoothness condition on the primal objective $f(x)$ is *sufficient* for the dual function to be continuously differentiable, and will make the evaluation of the Euler-Lagrange equation in (12) possible.

Assumption 1 (Strong Convexity) *The objective function $f(x)$ is twice continuously differentiable and belongs to the class $\mathcal{F}(m_f, L_f)$ for some $0 < m_f \leq L_f < \infty$.*

The strong convexity of $x \mapsto f(x)$ implies the strong convexity of $x \mapsto \mathcal{L}(x, \lambda)$, which further implies that the Lagrangian minimizer $\overline{x}(\lambda)$ [cf. (13)] is unique for each λ. Uniqueness of $\overline{x}(\lambda)$ implies that the sub-differential of the dual function is a singleton, i.e., the dual function is differentiable. More formally, we prove the following smoothness properties for the dual function.

Lemma 1 *Under Assumption 1, the dual function $d(\lambda)$, defined in (7), satisfies $-d(\lambda) \in \mathcal{F}(m_d, L_d)$ with $m_d = \sigma^2_{\min}(A)/L_f$ and $L_d = \sigma^2_{\max}(A)/m_f$. Further, the gradient and Hessian of the dual function are given by*

$$\nabla d(\lambda) = A\overline{x}(\lambda) - b, \quad (14a)$$

$$\nabla^2 d(\lambda) = -A[\nabla^2 f(\overline{x}(\lambda))]^{-1}A^\top. \quad (14b)$$

Proof: See Appendix VII-A. ■

Under Assumption 1, we can simplify the partial differential equation (PDE) in (12) to a dynamical system involving the gradient of the dual function, as we state next.

Proposition 1 Consider the Lagrangian mechanical system defined by the Bregman Lagrangian in (10) under the ideal scaling conditions (11). Then, under Assumption 1, the Euler-Lagrange equation in (12) for trajectories λ_t that minimizes the action functional $J[\lambda] = \int_{\mathbb{T}} \mathbb{L}(\lambda_t, \dot{\lambda}_t, t)dt$ is equivalent to the following ordinary differential equation (ODE),

$$\ddot{\lambda}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{\lambda}_t \quad (15)$$
$$- e^{2\alpha_t + \beta_t}[\nabla^2 \psi(\lambda_t + e^{-\alpha_t}\dot{\lambda}_t)]^{-1}\nabla d(\lambda_t) = 0\ ,$$

which can be equivalently stated, without inverting the Hessian in (15), as

$$\frac{d}{dt}\nabla\psi(\lambda_t + e^{-\alpha_t}\dot{\lambda}_t) = e^{\alpha_t + \beta_t}\nabla d(\lambda_t). \quad (16)$$

The Euler-Lagrange equations in Proposition 1 are derived by applying Hamilton's Principle for trajectories λ_t in the dual domain $\mathbb{R}^m$. It is unclear, however, what role such trajectories play in solving the optimization problem in (5). In Subsection III-C, we develop another dynamical system in the primal domain, coupled to the Euler-Lagrange equations (16), which yield solutions that converge at a specified rate to a saddle point of the Lagrangian in (6). But first, we study the Lyapunov stability of (16).

B. Lyapunov Stability Analysis

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

Lemma 2 *Consider the Euler-Lagrange dynamics in (16) under the ideal scaling conditions in (11) with initialization $\lambda_0,\ \dot{\lambda}_0 \in \mathbb{R}^m$. Then, under Assumption 1, the dual sub-optimality satisfies*

$$\frac{m_f}{2\sigma^2_{\max}(A)}\|\nabla d(\lambda_t)\|_2^2 \leq d(\lambda^*) - d(\lambda_t) \leq \mathcal{O}(e^{-\beta_t})\ , \quad (17)$$

where the proper selection of the scalar real-valued function β_t determines the rate of convergence. Further, if $L_f/\sigma_{\min}(A) < \infty$, the optimal $\lambda_\star$ is unique and we have that

$$\|\lambda_t - \lambda_\star\|_2^2 \leq \frac{2L_f}{\sigma^2_{\min}(A)}\mathcal{O}(e^{-\beta_t}). \quad (18)$$

Proof: See Appendix VII-B. ■

Lemma 2, which builds upon the results of Theorem 2.1 in [21], establishes a bound on the suboptimality gap evaluated along trajectories satisfying the Euler-Lagrange equations given in (16). In particular, for $\beta_t = ct, c > 0$, this gap vanishes exponentially fast.

C. Evolution of Lagrangian Minimizers

To evaluate the dual gradient $\nabla d(\lambda_t) = A\overline{x}(\lambda_t) - b$ in the Euler-Lagrange ODE in (16), we need to continuously evaluate the Lagrangian minimizer $\overline{x}(\lambda_t) = \arg\min_x \mathcal{L}(x, \lambda_t)$. As we show below, the smoothness of $f(x)$ allows us to continuously compute this minimizer without performing the minimization all the time, under appropriate initialization. That

is, by applying the chain rule to the dual feasibility identity $\nabla_x \mathcal{L}(\overline{x}(\lambda), \lambda) = 0$ and rearranging terms, we obtain

$$\frac{d}{d\lambda^\top}\overline{x}(\lambda) = -[\nabla^2 f(\overline{x}(\lambda))]^{-1} A^\top, \tag{19}$$

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

$$\frac{d}{dt}\overline{x}(\lambda_t) = \Big[\frac{d}{d\lambda_t^\top}\overline{x}(\lambda_t)\Big]\frac{d}{dt}\lambda_t, \tag{20}$$

with initial condition $\overline{x}_0 = \overline{x}(\lambda_0) = \arg\min_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda_0)$. Combining the expressions in (20) and (19) along with (16) yields the following continuous-time dynamical system,

$$\dot{\nabla}\psi(\lambda_t + e^{-\alpha_t}\dot{\lambda}_t) = e^{\alpha_t+\beta_t}(A\overline{x}_t - b), \tag{21a}$$

$$\dot{\overline{x}}_t = -[\nabla^2 f(\overline{x}_t)]^{-1} A^\top \dot{\lambda}_t. \tag{21b}$$

The states of (21) are $\lambda_t, \dot{\lambda}_t$ and $\overline{x}_t$. Intuitively, the first ODE (21a) executes an accelerated gradient flow on the dual function, while the second ODE (21b) maintains the dual feasibility $\nabla_x \mathcal{L}(\overline{x}_t, \lambda_t) = 0$ all the time.

Next, we prove that the dynamical system defined in (21) converges to the saddle point of the Lagrangian and, hence, solves (5) at an exponential rate (depending on the choice of β_t–see Section IV).

Theorem 1 *Under Assumption 1 and the ideal scaling conditions in* (11) *, the primal-dual flow* $(\overline{x}_t, \lambda_t)$ *defined by the dynamical system* (21) *with initialization* $\lambda_0,\ \dot{\lambda}_0 \in \mathbb{R}^m, \overline{x}_0 = \arg\min_x \mathcal{L}(x, \lambda_0)$ *satisfies*

$$\frac{m_f}{2\sigma^2_{\max}(A)}\|A\overline{x}_t - b\|_2^2 \leq \mathcal{L}(x^*, \lambda^*) - \mathcal{L}(\overline{x}_t, \lambda_t) \leq \mathcal{O}\big(e^{-\beta_t}\big). \tag{22}$$

Proof: See Appendix VII-C. ■

Theorem 1 establishes that the dynamical system (21) yields solutions that converge to the saddle point of the Lagrangian, possibly exponentially fast, depending on the selection of the scalar-valued function β_t. The solutions are dual feasible all the time. However, primal feasibility is achieved exponentially fast, and requires strong convexity of the primal objective function–see the left inequality in (22).

D. Parameter Selection

To achieve exponential convergence, one may set $\beta_t = ct,\ c > 0$ in (21), and set the other parameters accordingly to satisfy the ideal scaling in (11), which were used to derive the Euler-Lagrange PDE. This rate is valid for the case that the primal objective is twice-continuously differentiable and strongly convex (Assumption 1).

Alternatively, one can consider the choice of a logarithmic (polynomial) scaling, such as, $e^{\alpha_t} = p/t$, $e^{\beta_t} = Ct^p$, $e^{\gamma_t} = t^p,\ p > 0$, with $C > 0$ being an arbitrary positive scalar. In this case, the Euler-Lagrange ODE in (15) simplifies to

$$\ddot{\lambda}_t + \frac{p+1}{t}\dot{\lambda}_t - Cp^2 t^{(p-2)}\left(\nabla^2\psi(\lambda_t + \frac{t}{p}\dot{\lambda}_t)\right)^{-1}\nabla d(\lambda_t) = 0. \tag{23}$$

Using Lemma 2, one can easily show that (23) attains a polynomial convergence rate $\mathcal{O}(1/t^p)$. The question then becomes: which discretization of (15) remains stable while converging as fast as possible? We now shift our focus to this question.

We remark that the discretization process of the ODE in (21) is not trivial, and not any selection of α_t, β_t, and γ_t satisfying the ideal scaling (11) stably preserves the convergence rate. Since the discretization is simpler for the polynomial scaling case, i.e., discretizing the polynomial Euler-Lagrange ODE in (23), we henceforth focus on this parameter selection. Further, the discretized scheme no longer requires a continuous evaluation of the dual gradient; therefore, we can relax the smoothness conditions in Assumption 1, so as to encompass a broader category of linearly constrained problems.

IV. Discrete-Time Dual Acceleration

In this section, we consider the discretization of Euler-Lagrange in (21) with polynomial parameter selection–see (23). In [21], a rate-matching discretization scheme relying on higher-order gradient methods to stably discretize the polynomial Eluer-Lagrange ODE in (23) is proposed. This scheme requires the evaluation of the first $p-1$ derivatives of the dual function $d(\lambda)$, along with Lipschitz continuity of the $(p-1)$-th derivative to obtain an $\mathcal{O}(1/k^p)$ convergence rate, 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 $-d(\lambda)$ over $\mathbb{R}^m$: start with $\lambda_0 = \mu_{-1} \in \mathbb{R}^m$ and define the following iterative sequences,

$$\begin{aligned}
&\nu_k = \operatorname{argmin}_{\nu\in\mathbb{R}^m}\ \{-d_{p-1}(\nu; \lambda_k) + \frac{NL_{p-1}}{p!}\|\nu - \lambda_k\|^p\}\ ,\\
&\mu_k = \operatorname{argmin}_{\mu\in\mathbb{R}^m}\{-Cpk^{(p-1)}\nabla d(\nu_k)^\top\mu + \frac{L_{p-1}}{(p-1)!}D_\psi(\mu, \mu_{k-1})\}\ ,\\
&\lambda_{k+1} = \frac{p}{k+p}\mu_k + \frac{k}{k+p}\nu_k\ ,
\end{aligned} \tag{24}$$

where $d_{p-1}(\nu; \lambda_k)$ is the $(p-1)$-th order Taylor expansion of $d(\nu)$ around λ_k. Further, $k^{(p-1)} = k(k+1)\cdots(k+p-2)$ is the rising factorial; $C \leq (N^2-1)^{\frac{p-2}{2}}((2N)^{p-1}p^p)$ is a positive scalar which arises in the polynomial scaling (23); $D_\psi(\cdot,\cdot)$ is the Bregman divergence (9); $N > 1$ is arbitrary; and $L_{p-1} > 0$ is the Lipschitz constant of the $p-1$-th derivative of $d(\lambda)$.

The first update is a *higher-order* dual gradient step at λ_k to generate the auxiliary dual variable $\nu_k \in \mathbb{R}^m$, whose update depends on the truncated Taylor expansion $d_{p-1}(\nu; \lambda_k)$ 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 $\mu_k \in \mathbb{R}^m$ in which the dual gradient is computed at ν_k (rather than μ_{k-1} in standard dual ascent). Finally, we consider the update of the actual Lagrange multiplier λ_{k+1} as a convex combination of the dual auxiliary sequences ν_k and μ_k.

The iteration in (24) represents the discrete-time counter part of the ODE (23), and exhibits an $\mathcal{O}(1/k^p)$ converges rate [21, §3.4]. Hypothetically, one may achieve faster convergence by increasing p in (24), 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 [14] and cubic regularized Newton's method [26], [27].

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

Specialized to the case $p = 2$, the implementation of (24) requires the dual function to be continuously differentiable with a *Lipschitz gradient*. Assumption 1 is too strong for these regularity requirements, according to Lemma 1. We could still satisfy these conditions under a less restrictive assumption; that is, we may relax the requirement of strong convexity and smoothness of $f(x)$ by adopting an augmented Lagrangian approach. To do so, we define the augmented Lagrangian as

$$\mathcal{L}_\rho(x, \lambda) = f(x) + \lambda^\top (Ax - b) + \frac{\rho}{2}\|Ax - b\|^2 , \qquad (25)$$

where $\rho > 0$ is arbitrary. Notice that the quadratic term is zero on the feasible set and, hence, does not alter the saddle points. For a fixed λ, the Lagrangian minimizer $\overline{x}(\lambda) = \arg\min \mathcal{L}_\rho(x, \lambda)$ is no longer unique, opening the possibility for the dual function to be non-differentiable – see (13). However, as we show next, the quadratic penalty term in (25) renders a continuously differentiable dual function $d(\lambda) = \min_x \mathcal{L}_\rho(x, \lambda)$ with a Lipschitz gradient, under less restrictive Assumptions on $f(x)$.

Lemma 3 Assume that $f(x)$ belongs to the class $\mathcal{F}(0, L_f)$ with $0 < L_f \leq \infty$. Then, the dual function $d(\lambda)$ associated with the augmented Lagrangian (25) satisfies $-d(\lambda) \in \mathcal{F}(m_d, L_d)$, where the parameters of strong concavity and Lipschitz continuity are given by

$$m_d = \frac{\sigma^2_{\min}(A)}{2\rho\sigma^2_{\min}(A) + L_f}, \; L_d = \frac{2\rho\sigma^2_{\min}(A) + L_f}{\rho^2\sigma^2_{\min}(A) + \rho L_f}. \qquad (26)$$

Proof: See Appendix VII-D. ■

We can make several observations from Lemma 3. First, Lagrangian augmentation admits a Lipschitz dual gradient, even if the primal objective function is not strongly convex [cf. Lemma 1]. Intuitively, the penalty term induces favorable curvature profiles of strong convexity onto weakly convex functions[2]. Second, when the primal objective function is continuously differentiable and the matrix A is full row rank (so that $L_f < \infty$ and $\sigma^2_{\min}(A) > 0$), the dual function becomes strongly concave. Finally, when the primal objective is nonsmooth (i.e., when $L_f = \infty$), the dual function becomes continuously differentiable with $1/\rho$-Lipschitz gradient, which is relatively well-known (see, e.g. [28]).

In the following, we focus on the case in which f is non-differentiable ($L_f = \infty$), which admits a weakly concave dual function ($m_d = 0$). We substitute the linearization of the dual function,

$$d_1(\lambda; \lambda_k) = d(\lambda_k) + \nabla d(\lambda_k)^\top (\lambda - \lambda_k), \qquad (27)$$

into (24). The resulting accelerated variation of the method of multipliers (AMM) is summarized in Algorithm 1. Observe that we have $d(\lambda_k) = \mathcal{L}(\overline{x}(\lambda_k), \lambda_k)$ and $\nabla d(\lambda_k) = A\overline{x}(\lambda_k) - b$. Therefore, for evaluating the dual function and its gradient at λ_k in (27), an *exact* minimization $\overline{x}(\lambda_k) \in \arg\min_x \mathcal{L}_\rho(x, \lambda_k)$ is required. We now establish the convergence of dual accelerated methods at Nesterov's optimal $\mathcal{O}(1/k^2)$ *without* strong convexity.

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

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

Require: Augmented Lagrangian $\mathcal{L}_\rho(x, \lambda)$ (see (25)), scaling parameters C and N such that $C \leq 1/(8N)$, $N > 1$, augmentation constant $\rho > 0$ (determines algorithm step-size).
initialize primal $x_0 \in \mathbb{R}^n$, dual variables $\lambda_0 = \mu_{-1} \in \mathbb{R}^m$
for $k = 0, 1, 2, \ldots$ **do**
Compute primal minimizer,

$$x_{k+1} \in \arg\min_{x \in \mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k). \qquad (28a)$$

Evaluate dual gradient,

$$\nabla d(\lambda_k) = Ax_{k+1} - b. \qquad (28b)$$

Compute dual ascent step,

$$\nu_k = \arg\min_{\nu \in \mathbb{R}^m} \{-\nabla d(\lambda_k)^\top (\nu - \lambda_k) + \tfrac{N}{2\rho}\|\nu - \lambda_k\|^2\}. \qquad (28c)$$

Compute auxiliary minimizer,

$$y_{k+1} \in \arg\min_{y \in \mathbb{R}^n} \mathcal{L}_\rho(y, \nu_k). \qquad (28d)$$

Evaluate dual gradient,

$$\nabla d(\nu_k) = Ay_{k+1} - b. \qquad (28e)$$

Execute mirror ascent w.r.t. Bregman divergence D_ψ,

$$\mu_k = \operatorname*{argmin}_{\mu \in \mathbb{R}^m} \{-2Ck\nabla d(\nu_k)^\top \mu + \frac{1}{\rho} D_\psi(\mu, \mu_{k-1})\}. \qquad (28f)$$

Update multiplier λ_{k+1} as weighted avg. of aux. vars. ν_k, μ_k,

$$\lambda_{k+1} = \frac{2}{k+2}\mu_k + \frac{k}{k+2}\nu_k. \qquad (28g)$$

end for

Theorem 2 (Accelerated Method of Multipliers) Consider the optimization problem in (5) with $f \in \mathcal{F}(0, \infty)$ and the corresponding augmented Lagrangian (25). Then, the primal sequence $\{y_k\}_{k \geq 0}$ and the dual sequence $\{\nu_k\}_{k \geq 0}$ obtained from Algorithm 1 satisfy

$$\frac{\rho}{2}\|Ay_{k+1} - b\|_2^2 \leq p^\star - d(\nu_k) \leq \mathcal{O}(\frac{1}{\rho k^2}). \qquad (29)$$

Proof: The proof of the right inequality follows from Corollary 3.5 in [21] with the negative of the dual function, $-d(\lambda) \in \mathcal{F}(0, 1/\rho)$, used as the convex objective function to be minimized. The left inequality follows from Lipschitz continuity of the dual gradient with parameter ρ^{-1}. More precisely, since the dual function satisfies $-d(\lambda) \in \mathcal{F}(0, \rho^{-1})$, we can write (see [29, §9])

$$\frac{\rho}{2}\|\nabla d(\nu)\|_2 \leq d^\star - d(\nu). \qquad (30)$$

By substituting $\nu = \nu_k$ in (30) and recalling that $y_{k+1} \in \operatorname{argmin}_{y \in \mathbb{R}^n} \mathcal{L}_\rho(y, \nu_k)$ so that $\nabla d(\nu_k) = Ay_{k+1} - b$ (see (28d) and (28e) in Algorithm 1), the left inequality in (29) is obtained. The proof is complete. ■

The result of Theorem 2 establishes an $\mathcal{O}(1/k^2)$ rate for accelerated dual ascent when the primal objective is non-differentiable and weakly convex. In the next subsection, we show how to further accelerate this scheme through the use of second-order information of the dual function, which corresponds to the case $p = 3$ of (24). Before doing so, we present a remark for the case in which the primal objective function $f(x)$ is weakly convex and differentiable.

Remark 1 (Exponential Convergence) By Lemma 3, when the primal objective function is continuously differentiable (i.e., $f \in \mathcal{F}(0, L_f)$ for some $L_f < \infty$), the dual function becomes strongly concave. In this case, it is possible to achieve exponential convergence. More specifically, we can make use of Nesterov's accelerated method [13] in the dual domain to obtain exponential convergence. More specifically, the algorithm reads as follows

$$x_{k+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^n} \mathcal{L}_\rho(x, \lambda_k), \tag{31a}$$

$$\nu_k = \lambda_k + s(Ax_{k+1} - b), \tag{31b}$$

$$\lambda_{k+1} = \nu_k + \beta(\nu_k - \nu_{k-1}) \tag{31c}$$

where $s = L_d^{-1}$ is the step size, and $\beta > 0$ is the momentum parameter. The inner minimization in (31a) yields the dual function oracle at λ_k, so that $d(\lambda_k) = \mathcal{L}_\rho(x_{k+1}, \lambda_k)$ and $\nabla d(\lambda_k) = Ax_{k+1} - b$. Therefore, the updates in (31b) and (31c) correspond to an accelerated dual ascent. It can be shown that, for the parameter selection $\beta = \frac{\sqrt{L_d}-\sqrt{m_d}}{\sqrt{L_d}+\sqrt{m_d}}$ with m_d and L_d given as in Lemma 3, exponential convergence is achieved with an an optimal decay factor equal to $1 - \sqrt{m_d/L_d}$ [13]. However, this selection requires the knowledge of L_f, the Lipschitz constant of the primal gradient function–see (26).

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

Next, we turn to developing a dual variant of the cubic regularized Newton method [26]. This development corresponds to specializing the dual higher-order gradient scheme (24) to $p = 3$. To do so, we require the dual function to be twice continuously differentiable with *Lipschitz Hessian*, which can be guaranteed by Assumption 1, as well as the following condition.

Assumption 2 (Lipschitz Primal Hessian) The objective function $f(x)$ is twice continuously differentiable with a Lipschitz continuous Hessian, i.e.,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \le C_f \|x - y\|, \tag{33}$$

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

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

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

Require: (Non-augmented) Lagrangian $\mathcal{L}(x, \lambda)$ [cf. (6)], scaling parameters C and N such that $C \le \sqrt{N^2-1}/(108N^2)$, $N > 1$, step-size parameter $L_2 = C_f \|A\|_2^3/m_f^3$ [cf. (34)].
initialize primal $x_0 \in \mathbb{R}^n$, dual variables $\lambda_0 = \mu_{-1} \in \mathbb{R}^m$
for $k = 0, 1, 2, \ldots$ **do**
Compute primal minimizer,

$$x_{k+1} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda_k). \tag{32a}$$

Evaluate dual gradient and dual Hessian,

$$\nabla d(\lambda_k) = Ax_{k+1} - b,$$
$$\nabla^2 d(\lambda_k) = -A[\nabla^2 f(x_{k+1})]^{-1}A^\top. \tag{32b}$$

Compute cubic dual Newton step,

$$\nu_k = \operatorname*{argmin}_{\nu\in\mathbb{R}^m}\{-\nabla d(\lambda_k)^\top(\nu - \lambda_k) + \tfrac{1}{2}(\nu - \lambda_k)^\top\nabla^2 d(\lambda_k)(\nu - \lambda_k) + \tfrac{NL_2}{3!}\|\nu - \lambda_k\|^3\}, \tag{32c}$$

Compute auxiliary minimizer,

$$y_{k+1} \in \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \mathcal{L}(y, \nu_k). \tag{32d}$$

Evaluate dual gradient,

$$\nabla d(\nu_k) = Ay_{k+1} - b. \tag{32e}$$

Execute mirror ascent w.r.t. Bregman divergence D_ψ,

$$\mu_k = \operatorname*{argmin}_{\mu\in\mathbb{R}^m}\{-3C(k^2+k)\nabla d(\nu_k)^\top\mu + \tfrac{L_2}{2}D_\psi(\mu, \mu_{k-1})\}, \tag{32f}$$

Update Lagrange multiplier λ_{k+1} as weighted average of auxiliary dual variables ν_k and μ_k,

$$\lambda_{k+1} = \frac{3}{k+3}\mu_k + \frac{k}{k+3}\nu_k. \tag{32g}$$

end for

Lemma 4 (Lipschitz Dual Hessian) Under Assumptions 1 and 2, the dual function $d(\lambda)$ is twice-continuously differentiable, and its Hessian $\nabla^2 d(\lambda) = -A\nabla^2 f(\overline{x}(\lambda))^{-1}A^\top$ satisfies

$$\|\nabla^2 d(\lambda) - \nabla^2 d(\nu)\|_2 \le \frac{C_f}{m_f^3}\|A\|_2^3\|\lambda - \nu\|_2, \tag{34}$$

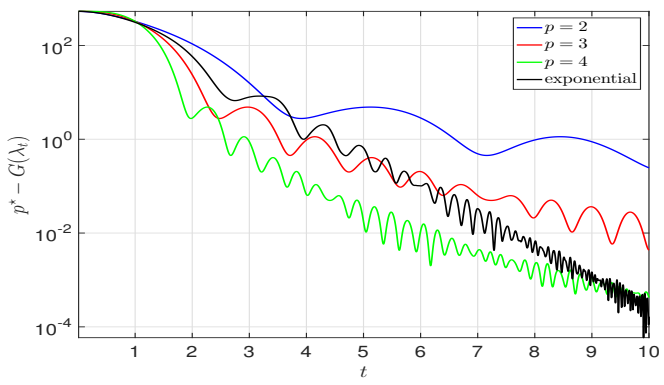
for all $\lambda, \nu \in \mathbb{R}^m$.

Proof: See Appendix (VII-E) ■

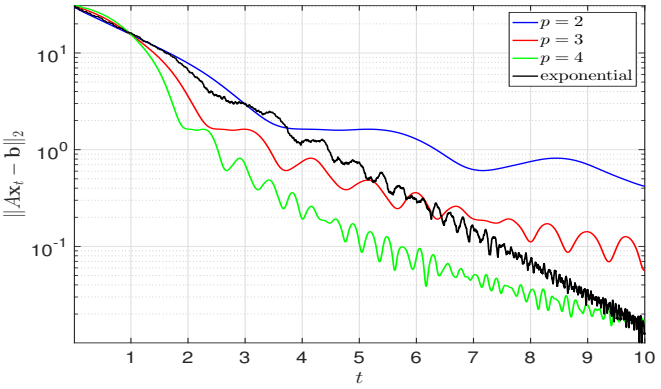
We now turn to the recursions in (24), where the first- and second-order derivatives of the dual function are used. Consider the second-order Taylor approximation of the dual function as

$$d_2(\lambda; \lambda_k) = d(\lambda_k) + \nabla d(\lambda_k)^\top(\lambda - \lambda_k) + \frac{1}{2}(\lambda - \lambda_k)^\top\nabla^2 d(\lambda_k)(\lambda - \lambda_k). \tag{35}$$

Observe that we have $d(\lambda_k) = \mathcal{L}(\overline{x}(\lambda_k), \lambda_k)$, $\nabla d(\lambda_k) = A\overline{x}(\lambda_k) - b$, and $\nabla^2 d(\lambda_k) = -A[\nabla^2 f(\overline{x}(\lambda_k))]^{-1}A^\top$. Hence, we need to incorporate the exact minimization step $\overline{x}(\lambda_k) \in \arg\min_x \mathcal{L}(x, \lambda_k)$ into the updates. The resulting iterative



(a) Dual sub-optimality vs. time t.



(b) Contraint violation vs. time t.

Fig. 1: Continuous-time accelerated dual ascent ODE (21) applied to (37) for various realizations of the scaling parameters (see Subsection III-D). Left: Figure 1a displays the evolution of dual sub-optimality gap $p^\star - d(\lambda_t)$ against t in log-linear scale. In Figure 1b, we plot the evolution of $\|A\overline{x}_t - b\|_2$ against t in log-linear scale–see (22). By increasing p, faster convergence rates are attained at the expense of more frequent sampling by the solver for a stable path generation..

scheme is summarized in Algorithm 2. With Lemma 4 established, and the technical setting clarified, we now present our main result for accelerated second-order dual methods.

Theorem 3 (Accelerated Cubic-Regularized Dual Newton Method) Consider the optimization problem in (5) and the corresponding Lagrangian in (6), where the objective function $f(x)$ satisfies Assumptions 1 and 2, Then, the primal-dual sequence $\{y_k, \nu_k\}_{k\geq 0}$ generated by Algorithm 2 satisfies

$$\frac{m_f}{2\sigma_{\max}^2(A)}\|Ay_{k+1} - b\|_2^2 \leq p^\star - d(\nu_k) \leq \mathcal{O}(1/k^3). \quad (36)$$

Proof: The right inequality comes from applying Corollary 3.5 of [21] with the negative of the dual function, $-d(\lambda)$, used in place of the convex function to be minimized. The left inequality follows from (47) in the proof of Lemma VII-B. ■

In Theorem 3 we establish that a second-order accelerated dual approach to solving (1) yields a convergent solution to a dual optimal point (8) and, hence, by strong duality, we converge to a saddle point of the Lagrangian in (6). The rate at which we converge to the primal-dual optimal pair is $\mathcal{O}(1/k^3)$, which is the state-of-the-art of existing dual methods for problems with affine constraints. This result requires strong convexity of the primal objective (Assumption 1) and Lipschitz continuity of its Hessian (Assumption 2).

Finally, we discuss some practical aspects of Algorithm 2. First, to implement the algorithm, we need to compute the Hessian of the dual–see (32b)–which requires inverting the Hessian of the primal objective function. Therefore, for large-scale problems, this Algorithm is not practical. Second, finding the cubic Newton step (32c) requires an iterative subroutine to solve the cubic problem in (32c). In particular, to obtain an ϵ-suboptimal solution to (32c), the gradient method solves (32c) in $\mathcal{O}(1/(\epsilon\log(1/\epsilon)))$ steps [30]. These steps, however, are relatively inexpensive since they do not require any oracle query from the dual function.

In the next section, we illustrate the performance properties of the developed methods in Sections III and IV via a synthetic example, as well as an application in model predictive control (MPC).

V. Empirical Evaluation

A. Simple Continuous-Time Example

We consider the following synthetic optimization problem

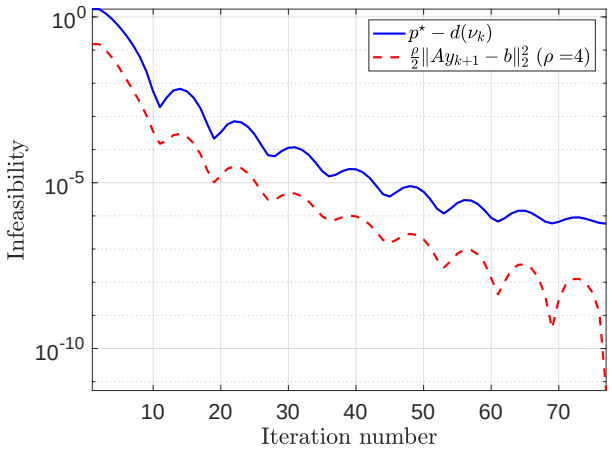
$$p^\star = \min_{x\in\mathbb{R}^n} \ \frac{1}{2}x^\top Px + \mu\exp(1_n^\top x) \ \text{ s.t. } \ Ax = b. \quad (37)$$

The above problem is convex with a twice-differentiable strongly convex objective when $\mu \geq 0$ and $P \succ 0$. For simulations, we choose $\mu = 0.01$, $n = 100$, $m = 50$, $P = QQ^\top$ where elements of $Q \in \mathbb{R}^{n\times n}$ are independently drawn from the standard normal distribution. The elements of $A \in \mathbb{R}^{m\times n}$ are also drawn from the standard normal distribution. Finally, we select $b = Az$, where $z \in \mathbb{R}^n$ is drawn from the standard normal distribution. For the selected problem data, the condition number of P is 4.8902×10^4. We then consider different selection of the scaling parameters: *(i)* the exponential regime, where $\beta_t = t, e^{\alpha_t} = t$ (which corresponds to an $\mathcal{O}(e^{-t})$ convergence rate), and *(ii)* polynomial regime, where $e^{\beta_t} = t^p, e^{\alpha_t} = p/t, p = 2, 3, 4$ (which corresponds to an $\mathcal{O}(1/t^p)$ convergence rate). We then consider a modified version of (21):

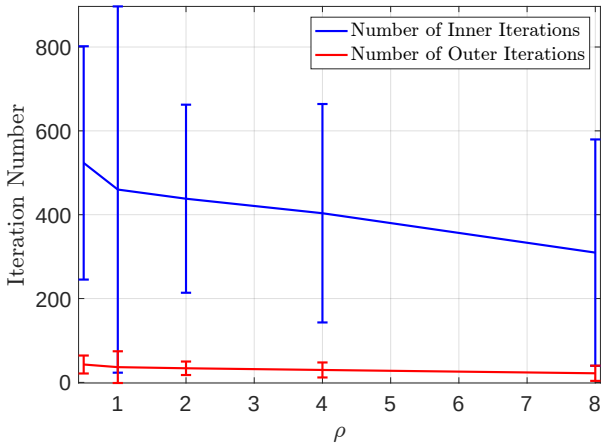
$$\dot{\nabla}\psi(\lambda_t + e^{-\alpha_t}\dot{\lambda}_t) = e^{\alpha_t+\beta_t}(A\overline{x}_t - b), \quad (38a)$$

$$\dot{\overline{x}}_t = -[\nabla^2 f(\overline{x}_t)]^{-1}(\nabla f(\overline{x}_t) + A^\top\lambda_t + A^\top\dot{\lambda}_t). \quad (38b)$$

The initial conditions are $\lambda_0 = \dot{\lambda}_0 = 0$, and $\overline{x}_0 = \operatorname{argmin}_{x\in\mathbb{R}^n}\mathcal{L}(x, \lambda_0)$ (so that $\nabla_x\mathcal{L}(\overline{x}_0, \lambda_0) = 0$), and the Bregman distance function ψ is chosen to be the Euclidean norm. Comparing to (21), the above ODE includes an additional correction term $\nabla_x\mathcal{L}(\overline{x}_t, \lambda_t) = \nabla f(\overline{x}_t) + A^\top\lambda_t$. Theoretically, the solutions to (38) satisfy $\nabla_x\mathcal{L}(\overline{x}_t, \lambda_t) = 0$ for all $t \geq t_0$. To see this, it is easy to verify that the derivative $\frac{d}{dt}\nabla_x\mathcal{L}(\overline{x}_t, \lambda_t)$ is zero for all $t \geq t_0$ and, hence, $\nabla_x\mathcal{L}(\overline{x}_t, \lambda_t) = \nabla_x\mathcal{L}(\overline{x}_0, \lambda_0) = 0$. Therefore, the solutions of (21) and (38) with the same initialization are identical.



(a)



(b)

Fig. 2: Fig. 2a: The evolution of the primal infeasibility ($\rho/2\|Ay_{k+1} - b\|_2^2$) and the dual suboptimality ($p^\star - d(\nu_k)$) for the numerical example in Subsection V-B–see (29). Both quantities decrease at an $\mathcal{O}(1/\rho k^2)$ rate. Fig. 2b: Plot of total number of inner and outer iterations for Algorithm 1 applied to the numerical example in Section V-B, for different values of the augmentation constant ρ. The error bar for each ρ represents the standard deviation obtained from 200 random realizations.

However, the dual feasibility condition $\nabla_x \mathcal{L}(\overline{x}_t, \lambda_t) = 0$ might be violated over time due to accumulated numerical errors arising from discretization. Therefore, we include the additional term $\nabla_x \mathcal{L}(\overline{x}_t, \lambda_t) = \nabla f(\overline{x}_t) + A^\top \lambda_t$ to correct the trajectory against numerical errors–see [31] for more details. We simulate (38) over the time interval $t \in [10^{-1}, 10]$ using MATLAB ODE23 solver. Note that the method must start at $t > 0$ due to the presence of a singularity at $t = 0$ in (23). The results are depicted in Figure 1. We plot the dual sub-optimality $p^\star - d(\lambda_t)$ over t for various selections of the scaling parameters. We also plot $\|A\overline{x}_t - b\|_2$ as a barometer for attaining primal feasibility. Notice that the convergence improves as we increase p. Also, exponential convergence is attained by selecting the scaling parameters as $\beta_t = t, e^{\alpha_t} = t$.

B. Linear Model Predictive Control with State-Input Constraints

As an application of the developed framework in this paper, we consider predictive control of discrete-time linear systems. More explicitly, we consider the following state-space model

$$x_{k+1} = A_x x_k + B_u u_k, \tag{39}$$

where at each time index k, $x_k \in \mathbb{R}^{n_x}$ is the state vector and $u_k \in \mathbb{R}^{n_u}$ is the input vector. The matrices $A_x \in \mathbb{R}^{n_x \times n_x}$ and $B_u \in \mathbb{R}^{n_x \times n_u}$ are the state and input matrices, respectively. We consider the following classical MPC problem:

$$\begin{aligned} &\text{minimize} \sum_{k=0}^{K-1} \frac{1}{2} x_k^\top Q x_k + \frac{1}{2} u_k^\top R u_k + \frac{1}{2} x_K^\top Q_f x_K, \qquad (40)\\ &\text{subject to}\\ &x_0 = x,\ x_{k+1} = A_x x_k + B_u u_k\ k = 0, \dots, K-1,\\ &(x_k, u_k) \in \mathcal{X} \times \mathcal{U},\ k = 0, \dots, K-1,\ x_K \in \mathcal{X}_f, \end{aligned}$$

where the decision variables are $u_0, \cdots, u_{K-1}, x_0, \cdots, x_K$. The constraint $x_0 = x$ is imposed by the given initial condition $x \in \mathbb{R}^{n_x}$, the second constraint is imposed by the dynamics of the system, and $(x_k, u_k) \in \mathcal{X} \times \mathcal{U}$ are hard constraints on the states and inputs. We assume that $Q \in \mathbb{S}^{n_x}$ and $Q_f \in \mathbb{S}^{n_x}$ are positive semidefinite, and that $R \in \mathbb{S}^{n_u}$ is positive definite. We further assume that projection onto $\mathcal{X}, \mathcal{U}$, and $\mathcal{X}_f$ is easy to perform; for instance, box constraints of the form $x_{\min} \leq x_k \leq x_{\max}$ and $u_{\min} \leq u \leq u_{\max}$. Finally, $K \geq 1$ is the horizon length. Given these elements, the MPC policy solves (40), executes $u_0 = u_0^\star$, updates the initial condition x to the measured (or estimated) state at time $k = 1$, and solves (40) again for the next control action u_1.

By introducing $z = [u_0^\top \ u_1^\top \ \dots \ u_{K-1}^\top \ x_1^\top \ \dots x_K^\top]$ as the stacked vector of the decision variables, the optimization problem (40) can be compactly written as

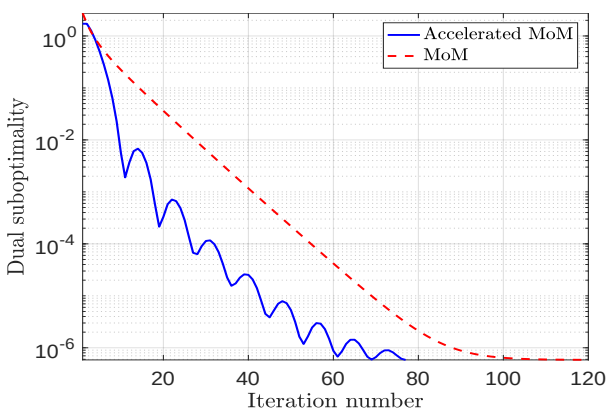
$$\text{minimize}_{z \in \mathbb{R}^{n_z}} \ \frac{1}{2} z^\top H z + \mathbb{I}_{\mathcal{Z}}(z) \text{ subject to } Az = b, \tag{41}$$

where $\mathcal{Z} = \mathcal{U}^K \times \mathcal{X}^{K-1} \times \mathcal{X}_f$ is the constraint set of z and $\mathbb{I}_{\mathcal{Z}}(z)$ is its indicator function. Moreover, the Hessian matrix H is given by $H = \text{blkdiag}(R, \dots, R, Q, \dots, Q, Q_f)$. Finally, A and b are given by

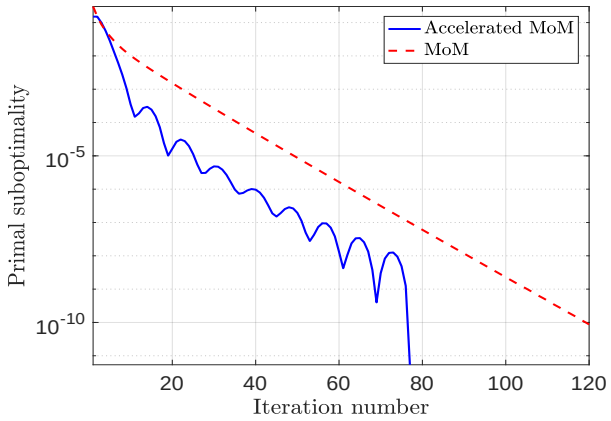
$$A = \left[\begin{array}{cccc|cccc} B & 0 & \cdots & 0 & -I & 0 & \cdots & 0 \\ 0 & B & \cdots & 0 & A & -I & \cdots & 0 \\ & & & & & & & \\ 0 & 0 & \ddots & 0 & 0 & A & \ddots & 0 \\ 0 & 0 & \dots & B & 0 & 0 & \dots & -I \end{array}\right], \ b = \begin{bmatrix} -Ax \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{42}$$

Notice that we could eliminate the equality constraints by making the substitution $x_k = A_x^k x_0 + \sum_{\ell=0}^{k-1} A_x^{k-\ell-1} B_x u_\ell$ in the cost function f and, therefore, remove the equality constraints $Az = b$ entirely. However, this formulation can potentially become ill-posed, especially for a large horizon length K, as high powers of A_x would appear in the cost function.

Formally, the objective function in (41) is non-smooth and is not necessarily strongly convex. This implies that the dual function is not differentiable. Following the discussion in



(a)



(b)

Fig. 3: Comparison of dual suboptimality (left) and primal infeasibility (right) obtained by AMM (Algorithm 1) and that of standard MM applied to the example in Section V-B. The total number of inner iterations for AMM and MM are 775 and 535, respectively.

Subsection IV-A, we adopt an augmented Lagrangian approach for which the dual function becomes differentiable, according to Lemma 3. In this setting, we can apply Algorithm 1 to solve the problem in the dual domain.

In our numerical experiments, we consider $K = 10$ and a randomly generated system with $n_x = 30$ states and $n_u = 10$ control inputs ($K \times (n_x + n_u) = 400$ decision variables), where the entries of the system's matrices A_x and B_x are chosen from the standard normal distribution. The matrix A_x is then rescaled so that its spectral norm is equal to one, yielding a marginally stable system. Further, the positive semidefinite matrices Q and Q_f are generated according to $Q = \tilde{Q}^\top \tilde{Q}$ and $Q_f = \tilde{Q}_f^\top \tilde{Q}_f$, where the entries of $\tilde{Q} \in \mathbb{R}^{n_q \times n_x}$ and $\tilde{Q}_f \in \mathbb{R}^{n_q \times n_x}$ (with $n_q < n_x$) are chosen from the standard normal distribution. Similarly, the positive definite matrix R is selected as $R = \tilde{R}^\top \tilde{R}$, where the entries of $\tilde{R} \in \mathbb{R}^{n_x \times n_x}$ are chosen from the standard normal distribution. Finally, we set $\psi(\cdot) = \frac{1}{2}\|\cdot\|_2^2$, $\rho = 4$, $N = 1$ and $C = 1/8$ for the parameters of the algorithm.

In Figure 2, we plot the quantities $\rho/2\|Ay_{k+1} - b\|_2^2$, which measures the primal infeasibility, and $p^\star - d(\nu_k)$, which is the dual suboptimality. Both quantities decrease approximately as $\mathcal{O}(1/\rho k^2)$ to the primal-dual optimal pair. Thus, we have an empirical validation of the theoretical rate established by Theorem 2.

Effect of Augmentation Constant. When the augmentation constant ρ increases, the inner minimizations typically become easier to solve, but the dual problem becomes harder to maximize, since the maximum allowable step size in the dual domain is bounded by ρ^{-1}. To examine the effect of ρ in practice, we compare the total number of inner and outer iterations for various values of ρ. To obtain these quantities for each ρ, we solve (40) for 200 random realization of the problem data and then take the average of the obtained values. In Figure 2b, we plot the average number of the inner and outer iterations for the values $\rho \in \{0.5, 1, 2, 4, 8\}$. We observe that the complexity of the dual problem is almost unaffected by ρ on average. However, the sample average number of inner iterations decreases by increasing ρ.

Finally, in Figure 3, we compare the performance of the algorithm against the standard Method of Multipliers. We can clearly observe the effect that acceleration has on reducing the number of required iterations to obtain a certain accuracy. We remark that the oscillations pattern in the trajectories of the accelerated algorithm is inherent to accelerated methods in general, as they relax the requirement that the objective function must decrease at each iteration.

VI. Conclusion

We develop a family of accelerated methods, inspired by [21], in order to solve 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 [21] to develop the discrete-time counterparts of this family. We specialize the algorithm to the accelerated method of multipliers and the 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, while the former is realizable under the less restrictive assumption of weak convexity and non-differentiability. We observe exponential as well as polynomial convergence rates in our synthetic example in continuous time of the Euler-Lagrange ODE (Section III-D) translates into practice for polynomial scaling selections. Moreover, the discretization of these polynomial scalings (Section IV) yields effective tools for solving MPC problems for linear systems.

VII. Appendix

A. Proof of Lemma 1

By Assumption 1, the objective function and, hence, the Lagrangian is twice differentiable with respect to x. Applying

the chain rule to the dual feasibility identity $\nabla_x \mathcal{L}(\overline{x}(\lambda), \lambda) = 0$ for all $\lambda \in \mathbb{R}^m$, we obtain

$$\nabla_{xx}\mathcal{L}(\overline{x}(\lambda),\lambda)\frac{d}{d\lambda^\top}\overline{x}(\lambda) + \nabla_{x\lambda}\mathcal{L}(\overline{x}(\lambda),\lambda) = 0,$$

where $[\frac{d}{d\lambda^\top}\overline{x}(\lambda)]_{ij} = \frac{d}{d\lambda_j}\overline{x}_i(\lambda)$. By rearranging the terms and recalling the definition of $\mathcal{L}(x,\lambda)$ in (6), we get

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

The last result requires the objective function $f(x)$ to have an invertible second derivative (Assumption 1). By taking the derivative of $\nabla d(\lambda) = A\overline{x}(\lambda) - b$ with respect to λ and using the last result, we obtain the Hessian $\nabla^2 d(\lambda)$, as follows,

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

We can use the strong convexity of $f(x)$, $\nabla^2 f(x) \succeq m_f \mathrm{I}_n$, to obtain a bound on $\nabla^2 d(\lambda)$ as follows,

$$\|\nabla^2 d(\lambda)\|_2 \leq \|A\|_2\|\nabla^2 f(\overline{x}(\lambda))^{-1}\|_2\|A^\top\|_2 = \frac{\sigma^2_{\max}(A)}{m_f}.$$

Finally, uniform boundedness of the dual Hessian is equivalent to Lipschitz continuity of the dual gradient, with the Lipschitz parameter being equal to the upper bound. ■

B. Proof of Lemma 2

Consider the following Lyapunov functional:

$$E_t = D_\psi(\lambda^*, \lambda_t + e^{-\alpha_t}\dot{\lambda}_t) + e^{\beta_t}(d(\lambda^*) - d(\lambda_t)). \quad (43)$$

Observe that (43) is non-negative: since the function ψ is convex, we have $D_\psi(\lambda,\nu) \geq 0$ for all $\lambda, \nu \in \mathbb{R}^m$. Moreover, the last term is positive due to the Saddle Point Theorem (see [23]): $d(\lambda^*) = \mathcal{L}(x^*,\lambda^*) \geq \mathcal{L}(x^*,\lambda_t) \geq d(\lambda_t)$, where the second inequality follows from the fact that $d(\lambda_t)$ is the minimum value of $\mathcal{L}(x,\lambda_t)$ with respect to x. This Lyapunov function may be identified as being in the form of that which is considered in the proof of Theorem 2.1 of [21] by noting that the negative of the dual function $-d(\cdot)$ is convex in λ_t. Stability follows from computing the derivative of (43), substituting in the dynamics (16), which in a manner analogous to [21] results in

$$\dot{E}_t \leq -e^{\alpha_t+\beta_t}D_{-d}(\lambda^*,\lambda_t) + (\dot{\beta}_t - e^{\alpha_t})e^{\beta_t}(d(\lambda^*) - d(\lambda_t)). \quad (44)$$

Then, we apply the ideal scaling (11) to the last term, and the fact that $-d(\cdot)$ is convex so that $D_{-d}(\cdot,\cdot) \geq 0$. Thus, $\dot{E}_t \leq 0$, implying that the dynamics defined by (16) satisfies the conditions for LaSalle's Invariance Principle via the energy function (43) and is, hence, stable. Exponential convergence may be obtained by noting that $e^{\beta_t}(d(\lambda^*) - d(\lambda_t)) \leq E_t$ due to the nonnegativity of the Bregman divergence term in (43), and invoking $\dot{E}_t \leq 0$ to write $E_t \leq E_0$. Therefore,

$$d(\lambda^*) - d(\lambda_t) \leq e^{-\beta_t}E_t \leq e^{-\beta_t}E_0 = \mathcal{O}(e^{-\beta_t}). \quad (45)$$

Note that $E_0 = D_\psi(\lambda^*,\lambda_0) + e^{\beta_0}(d(\lambda^\star) - d(\lambda_0))$, which is finite since $d(\lambda_0) \leq d(\lambda^\star) = p^\star < \infty$ by assumption. Finally, observe that since $\lambda \mapsto d(\lambda)$ has a bounded Hessian (see Lemma 1), for any $\lambda, \nu \in \mathbb{R}^m$ we have that,

$$d(\nu) \geq d(\lambda) + \nabla d(\lambda)^\top(\nu - \lambda) - \frac{\|A\|_2^2}{m_f}\|\nu - \lambda\|_2^2 \quad (46)$$

Maximizing both sides of (46) with respect to ν yields,

$$d^\star - d(\lambda) \geq \frac{m_f}{2\|A\|_2^2}\|\nabla d(\lambda)\|_2^2, \quad (47)$$

which then allows us to conclude the left inequality in (17). Finally, according to Lemma 1, the strong concavity constant of the dual function is $m_d = \sigma^2_{\min}(A)/L_f$, which is strictly positive if $L_f/\sigma^2_{\min}(A) < \infty$. Under this condition, we have that $-d(\lambda) \in \mathcal{F}(m_d, L_d)$. For this class of functions, we have that

$$\frac{m_d}{2}\|\lambda_t - \lambda_\star\|_2^2 \leq d(\lambda_\star) - d(\lambda_t). \quad (48)$$

By combining (45) and (48), we obtain (18). The proof is now complete.

■

C. Proof of Theorem 1

First, observe that by strong duality, we have $\mathcal{L}(x^*,\lambda^*) = d(\lambda^*)$. Next, we use the chain rule to compute the time derivative of $\nabla_x \mathcal{L}(\overline{x}_t, \lambda_t)$:

$$\dot{\nabla}_x \mathcal{L}(\overline{x}_t,\lambda_t) = \nabla_{xx}\mathcal{L}(\overline{x}_t,\lambda_t)\dot{\overline{x}}_t + \nabla_{x\lambda}\mathcal{L}(\overline{x}_t,\lambda_t)\dot{\lambda}_t. \quad (49)$$

Now, substitute $\dot{\overline{x}}_t$ from (21b) back into (49), which yields $\dot{\nabla}_x \mathcal{L}(\overline{x}_t,\lambda_t) = 0$. Therefore, we have that $\nabla_x \mathcal{L}(\overline{x}_t,\lambda_t) = \nabla_x \mathcal{L}(\overline{x}_0,\lambda_0) = 0$ for all $t \geq 0$, i.e., we have that $\overline{x}_t = \arg\min_x \mathcal{L}(x,\lambda_t)$ and, hence, $\mathcal{L}(\overline{x}_t,\lambda_t) = d(\lambda_t)$. Then, the result follows by applying Lemma 2, which yields $0 \leq \mathcal{L}(x^\star,\lambda^\star) - \mathcal{L}(\overline{x}_t,\lambda_t) \leq \mathcal{O}(e^{-\beta_t})$. Finally, the left inequality follows from the left inequality in (17) and noting that $\nabla d(\lambda_t) = A\overline{x}_t - b$, by (13). The proof becomes complete. ■

D. Proof of Lemma 3

Consider the two points $x \in \operatorname{argmin}_{\xi\in\mathbb{R}^n}\mathcal{L}(\xi,\lambda)$ and $y \in \operatorname{argmin}_{\xi\in\mathbb{R}^n}\mathcal{L}(\xi,\nu)$. We then have

$$\begin{aligned} T_f(x) + A^\top\lambda + \rho A^\top(Ax - b) &= 0, \\ T_f(y) + A^\top\nu + \rho A^\top(Ay - b) &= 0, \end{aligned} \quad (50)$$

where $T_f \in \partial f$ is a subgradient of f, so that when $L_f < \infty$, we have that $T_f = \nabla f$. By co-coercivity of T_f, we can write for $0 < L_f \leq \infty$ that

$$(T_f(x) - T_f(y))^\top(x - y) \geq \frac{1}{L_f}\|T_f(x) - T_f(y)\|_2^2. \quad (51)$$

Substituting (50) into (51) allows us to write

$$\begin{aligned} -(\lambda - \nu + \rho(Ax - Ay))^\top(Ax - Ay) \geq \\ \frac{1}{L_f}\|A^\top(\lambda - \nu + \rho(Ax - Ay))\|_2^2. \end{aligned} \quad (52)$$

Considering the quadratic expression on the right-hand side of (52), we can write the following inequality,

$$\|A^\top(\lambda-\nu+\rho(Ax-Ay))\|_2^2 \geq \sigma_{\min}^2(A)\|\lambda-\nu+\rho(Ax-Ay)\|_2^2. \tag{53}$$

From the last two inequalities, we obtain

$$-(\lambda-\nu+\rho(Ax-Ay))^\top(Ax-Ay) \geq \frac{1}{L_f}\sigma_{\min}^2(A)\|\lambda-\nu+\rho(Ax-Ay)\|_2^2. \tag{54}$$

Next, we complete the squares and rearrange terms to obtain

$$p\|Ax-Ay\|_2^2 + q(\lambda-\nu)^\top(Ax-Ay) + r\|\lambda-\nu\|_2^2 \leq 0, \tag{55}$$

where in (55) we have defined $p, q, r \in \mathbb{R}_+$ as

$$p = \rho + \frac{\rho^2\sigma_{\min}^2(A)}{L_f},\ q = 1 + \frac{2\rho\sigma_{\min}^2(A)}{L_f},\ r = \frac{\sigma_{\min}^2(A)}{L_f}.$$

Since the third term, $r\|\lambda-\nu\|_2^2$, in the left-hand side of (55) is nonnegative, the sum of the first two terms must be nonpositive, from where we obtain

$$\frac{p}{q}\|Ax-Ay\|_2^2 + (\lambda-\nu)^\top(Ax-Ay) \leq 0. \tag{56}$$

In particular, by setting $\lambda = \nu$, we conclude that $Ax = Ay$. This implies that the subdifferential set $\partial d(\lambda) = \{Ax - b\colon x \in \operatorname{argmin}_{x\in\mathbb{R}^n} \mathcal{L}(\xi,\lambda)\}$ is a singleton and, hence, the dual function d is differentiable. Denoting $\nabla d(\lambda) = Ax - b$ and $\nabla d(\nu) = Ay - b$, the inequality in (56) can be equivalently written as

$$\frac{p}{q}\|\nabla d(\lambda)-\nabla d(\nu)\|_2^2 + (\lambda-\nu)^\top(\nabla d(\lambda)-\nabla d(\nu)) \leq 0, \tag{57}$$

which establishes the co-coercivity of $\lambda \mapsto \nabla d(\lambda)$ with parameter L_d equal to

$$L_d = \frac{q}{p} = \frac{2\rho\sigma_{\min}^2(A) + L_f}{\rho^2\sigma_{\min}^2(A) + \rho L_f}.$$

Next, by applying an analogous logic to the first term in the left-hand side of (55), we can omit this term and then divide both sides by q to obtain

$$(\lambda-\nu)^\top(Ax-Ay) + \frac{r}{q}\|\lambda-\nu\|_2^2 \leq 0.$$

Recalling $\nabla d(\nu) = Ay - b$ and that $d(\lambda)$ is concave, the last inequality establishes the strong concavity of the dual function with parameter

$$m_d = \frac{r}{q} = \frac{\sigma_{\min}^2(A)}{2\rho\sigma_{\min}^2(A) + L_f}.$$

Notice that the analysis simply carries over to the case $L_f = \infty$ (non-differentiable f) by setting $1/L_f = 0$ in all the steps above. The proof is now complete. ■

E. Proof of Lemma 4

Twice continuous differentiability of the dual function follows from Lemma 1. We need to show the Lipschitz continuity of the Hessian. To do so, we consider the difference of two Hessians of the dual function and substitute in its definition in (14b),

$$\begin{aligned}\|\nabla^2 d(\lambda) - \nabla^2 d(\nu)\|_2 &= \|A\Big([\nabla^2 f(\overline{x}(\lambda))]^{-1} - [\nabla^2 f(\overline{x}(\nu))]^{-1}\Big)A^\top\|_2 \\ &\leq \|A\|_2^2\|[\nabla^2 f(\overline{x}(\lambda))]^{-1} - [\nabla^2 f(\overline{x}(\nu))]^{-1}\|_2,\end{aligned} \tag{58}$$

where in the first equality we have factored out A and $A^\top$ from left and right, respectively, and to obtain the inequality, we have applied the Cauchy-Schwartz inequality. We proceed to analyze the norm difference of Hessian inverses on the right-hand side of (58) and establish that the inverse Hessian is Lipschitz with parameter C_f/m_f^2. To do so, we write

$$\begin{aligned}&\|[\nabla^2 f(x)]^{-1} - [\nabla^2 f(y)]^{-1}\|_2 \\ &= \|[\nabla^2 f(x)]^{-1}\Big(\nabla^2 f(y) - \nabla^2 f(x)\Big)[\nabla^2 f(y)]^{-1}\|_2 \\ &\leq \|[\nabla^2 f(x)]^{-1}\|_2\|\nabla^2 f(y) - \nabla^2 f(x)\|_2\|[\nabla^2 f(y)]^{-1}\|_2 \\ &\leq \frac{C_f}{m_f^2}\|x-y\|_2.\end{aligned} \tag{59}$$

In the first inequality, we use the Cauchy-Schwartz inequality, whereas the last inequality applies the strong convexity of $f(x)$, i.e., $\nabla^2 f(x) \succeq m_f \mathrm{I}_n$, and Lipschitz continuity of the Hessian, i.e., $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq C_f\|x-y\|_2$. By using (59) back in (58), we obtain

$$\|\nabla^2 d(\lambda) - \nabla^2 d(\nu)\| \leq \frac{C_f}{m_f^2}\|A\|_2^2\|\overline{x}(\lambda) - \overline{x}(\nu)\|. \tag{60}$$

On the other hand, using the integral form of Taylor's Theorem, we have that

$$\|\overline{x}(\lambda)-\overline{x}(\nu)\|_2 = \|\int_0^1 \frac{d\overline{x}}{d\lambda}(\nu + t(\lambda-\nu))(\lambda-\nu)dt\|_2. \tag{61}$$

From (19), we have that $\|\frac{d}{d\lambda}\overline{x}(\lambda)\|_2 \leq \frac{\|A\|_2}{m_f}$ for all λ. Therefore, by applying the Cauchy-Schwartz on the right-hand side of (61), we can write

$$\begin{aligned}\|\overline{x}(\lambda)-\overline{x}(\nu)\|_2 &\leq \|\int_0^1 \frac{d\overline{x}}{d\lambda}(\nu + t(\lambda-\nu))dt\|_2\|\lambda-\nu\|_2 \\ &\leq \frac{\|A\|_2}{m_f}\|\lambda-\nu\|_2.\end{aligned} \tag{62}$$

Combining (60) and (62) yields the inequality

$$\|\nabla^2 d(\lambda) - \nabla^2 d(\nu)\|_2 \leq \frac{C_f}{m_f^3}\|A\|_2^3\|\lambda-\nu\|_2,$$

which completes the proof. ■

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. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *Signal Processing, IEEE Transactions on*, vol. 58, no. 12, pp. 6369–6386, Dec 2010.
[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] J. Goodman, "Newton's method for constrained optimization," *Mathematical Programming*, vol. 33, no. 2, pp. 162–171, 1985.
[11] 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.
[12] R. Tutunov, H. B. Ammar, and A. Jadbabaie, "A distributed newton method for large scale consensus optimization," *arXiv preprint arXiv:1606.06593*, 2016.
[13] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.
[14] ——, "A method of solving a convex programming problem with convergence rate o (1/k2)," vol. 27, no. 2, 1983, pp. 372–376.
[15] 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.
[16] A. Koppel, F. Jakubiec, and A. Ribeiro, "A saddle point algorithm for networked online convex optimization," *IEEE Trans. Signal Process.*, p. 15, Oct 2015.
[17] Y. Xu, "Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming," *arXiv preprint arXiv:1606.09155*, 2016.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] M. Fazlyab, A. Koppel, A. Ribeiro, and V. M. Preciado., "Accelerated Dual Methods for Constrained Convex Optimization," in *Proceedings of the American Control Conference*, Seattle, WA, USA, May 2017.
[23] D. P. Bertsekas, A. Nedi, A. E. Ozdaglar *et al.*, "Convex analysis and optimization," 2003.
[24] C. Bailey, "Hamilton's principle and the calculus of variations," *Acta Mechanica*, vol. 44, no. 1-2, pp. 49–57, 1982.
[25] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1999.
[26] Y. Nesterov, "Accelerating the cubic regularization of newton?s method on convex problems," *Mathematical Programming*, vol. 112, no. 1, pp. 159–181, 2008.
[27] M. Baes, "Estimate sequence methods: extensions and approximations," 2009.
[28] D. P. Bertsekas, *Nonlinear programming*. Athena scientific Belmont, 1999.
[29] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[30] Y. Carmon and J. C. Duchi, "Gradient descent efficiently finds the cubic-regularized non-convex newton step," *arXiv preprint arXiv:1612.00547*, 2016.
[31] 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.

Mahyar Fazlyab received his B.Sc. and M.Sc. degrees in Mechanical Engineering from Sharif University of Technology, Tehran, Iran, in 2010 and 2013, respectively. He has been a Ph.D. student with the Department of Electrical and Systems Engineering at the University of Pennsylvania since September 2013. His research interests include the analysis, optimization, and control of (networked) dynamical systems.

Alec Koppel Alec Koppel began as a Research Scientist in the Computational and Information Sciences Directorate (CISD) of the U.S. Army Research Laboratory (ARL) in Adelphi, Maryland in September of 2017. Previously, he completed his A.M. in Statistics and PhD in Electrical and Systems Engineering, both from the University of Pennsylvania in August of 2017. His doctoral work was part of the Science, Mathematics, and Research for Transformation (SMART) Scholarship Program sponsored by the American Society of Engineering Education with ARL as his sponsoring facility, where he spent doctoral summers. Before coming to UPenn, he completed his M.S. degree in Systems Science and Mathematics and B.A. degree in Mathematics at Washington University in St. Louis, MO (WashU). His research focuses on developing new methods for both supervised and reinforcement learning, especially from the perspective of stochastic optimization. Applications include adaptive signal processing, learning-based control systems, and autonomous robotics. At the 2017 IEEE Asilomar Conference in Signals, Systems, and Computers, a paper he co-authored was selected as a Best Paper Award Finalist.

Alejandro Ribeiro received the B.Sc. degree in electrical engineering from the Universidad de la Republica Oriental del Uruguay, Montevideo, in 1998 and the M.Sc. and Ph.D. degrees in electrical engineering from the Department of Electrical and Computer Engineering, the University of Minnesota, Minneapolis, MN, USA, in 2005 and 2007, respectively. From 1998 to 2003, he was a member of the technical staff at Bellsouth Montevideo. After his M.Sc. and Ph.D. studies, in 2008, he joined the University of Pennsylvania, Philadelphia, PA. USA, where he is currently an Assistant Professor at the Department of Electrical and Systems Engineering. His research interests are in the applications of statistical signal processing to the study of networks and networked phenomena. His current research focuses on wireless networks, network optimization, learning in networks, networked control, robot teams, and structured representations of networked data structures. Dr. Ribeiro received the 2012 S. Reid Warren, Jr. Award presented by the University of Pennsylvanias undergraduate student body for outstanding teaching and the NSF CAREER Award in 2010. He is also a Fulbright scholar and the recipient of student paper awards at the 2013 American Control Conference (as adviser), as well as the 2005 and 2006 International Conferences on Acoustics, Speech and Signal Processing.

Victor M. Preciado received his Ph.D. degree in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 2008. He is currently the Raj and Neera Singh Assistant Professor of Electrical and Systems Engineering at the University of Pennsylvania. He is a member of the Networked and Social Systems Engineering (NETS) program and the Warren Center for Network and Data Sciences. His research interests include network science, dynamic systems, control theory, and convex optimization with applications in socio-technical systems, technological infrastructure, and biological networks. Dr. Preciado received the NSF CAREER Award in 2017 and best student paper (as advisor) at the 2016 Stochastic Network Conference.