A second order primal-dual method for nonsmooth convex composite optimization

Neil K. Dhingra, Sei Zhen Khong, and Mihailo R. Jovanović

Abstract

We develop a second order primal-dual method for optimization problems in which the objective function is given by the sum of a strongly convex twice differentiable term and a possibly nondifferentiable convex regularizer. After introducing an auxiliary variable, we utilize the proximal operator of the nonsmooth regularizer to transform the associated augmented Lagrangian into a function that is once, but not twice, continuously differentiable. The saddle point of this function corresponds to the solution of the original optimization problem. We employ a generalization of the Hessian to define second order updates on this function and prove global exponential stability of the corresponding differential inclusion. Furthermore, we develop a globally convergent customized algorithm that utilizes the primal-dual augmented Lagrangian as a merit function. We show that the search direction can be computed efficiently and prove quadratic/superlinear asymptotic convergence. We use the ℓ_1-regularized least squares problem and the problem of designing a distributed controller for a spatially-invariant system to demonstrate the merits and the effectiveness of our method.

I. INTRODUCTION

We study a class of composite optimization problems in which the objective function is given by the sum of a differentiable strongly convex component and a nondifferentiable convex component. Problems of this form are encountered in diverse fields including compressive sensing, machine learning, statistics, image processing, and control [1]–[5]. They often arise in structured feedback synthesis where it is desired to balance controller performance (e.g., the closed-loop $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm) with structural complexity [5], [6].

The lack of differentiability in the regularization term precludes the use of standard descent methods for smooth objective functions. Proximal gradient methods [7]–[9] and their accelerated variants [10] generalize gradient descent, but typically require the nonsmooth term to be separable.

An alternative approach introduces an auxiliary variable to split the smooth and nonsmooth components of the objective function. The reformulated problem facilitates the use of splitting methods such as the alternating direction method of multipliers (ADMM) [11]. This augmented-Lagrangian-based method divides the optimization problem into simpler subproblems, allows for a broader class of regularizers than proximal gradient, and it is convenient for distributed implementation. In [12], we exploited the structure of proximal operators associated with nonsmooth regularizers and introduced the *proximal augmented Lagrangian*. This continuously differentiable

N. K. Dhingra is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455. M. R. Jovanović is with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089. S. Z. Khong is with the Institute for Mathematics and its Applications, Minneapolis, MN 55455. E-mails: dhin0008@umn.edu, mihailo@usc.edu, szkhong@ima.umn.edu.

function enables the use of the standard method of multipliers (MM) for nonsmooth optimization and it is convenient for distributed implementation via the Arrow-Hurwicz-Uzawa gradient flow dynamics. Recent work has extended MM to incorporate second order updates of the primal and dual variables [13]–[15] for nonconvex problems with twice continuously differentiable objective functions.

Since first order approaches tend to converge slowly to high-accuracy solutions, much work has focused on developing second order methods for nonsmooth composite optimization. A generalization of Newton's method was developed in [16]–[19] and it requires solving a regularized quadratic subproblem to determine a search direction. Related ideas have been utilized for sparse inverse covariance estimation in graphical models [20], [21] and topology design in consensus networks [22].

Generalized Newton updates for identifying stationary points of (strongly) semismooth gradient mappings were first considered in [23]–[25]. In [26]–[28], the authors introduce the once-continuously differentiable Forward-Backward Envelope (FBE) and solve composite problems by minimizing the FBE using line search, quasi-Newton methods, or second order updates based on an approximation of the generalized Hessian.

We develop a second order primal-dual algorithm for nonsmooth composite optimization by leveraging these advances. Second order updates for the once continuously differentiable proximal augmented Lagrangian are formed using a generalized Hessian. We employ the merit function introduced in [13] to assess progress towards the optimal solution and develop a globally convergent customized algorithm with fast asymptotic convergence rate. When the proximal operator associated with a nonsmooth regularizer is (strongly) semismooth, our algorithm exhibits local (quadratic) superlinear convergence.

Our presentation is organized as follows. In Section II, we formulate the problem and provide necessary background material. In Section III, we define second order updates to find the saddle points of the proximal augmented Lagrangian. In Section IV, we prove global exponential stability of a continuous-time differential inclusion. In Section V, we develop a customized algorithm that converges globally to the saddle point and exhibits superlinear or quadratic asymptotic convergence rates. In Section VI, we provide examples to illustrate the utility of our method. We discuss interpretations of our approach and connections to the alternative algorithms in Section VII and conclude the paper in Section VIII.

II. Problem formulation and background

We consider the problem of minimizing the sum of two convex functions over an optimization variable $x \in \mathbb{R}^n$,

$$\underset{x}{\text{minimize}} \quad f(x) \, + \, g(Tx) \tag{1}$$

where $T \in \mathbb{R}^{m \times n}$. Problem (1) was originally formulated in the context of compressive sensing to incorporate structural considerations into traditional sensing or regression [1]–[3]. As specified in Assumption 1, the differentiable part of the objective function f, which quantifies loss or performance, is strictly convex with a Lipschitz continuous gradient. In contrast, the regularization function g may be nondifferentiable and is used to incorporate structural requirements on the optimization variable. In structured feedback synthesis [5], [6], f typically quantifies the closed-loop performance, e.g., the $\mathcal{H}_2$ norm, and g imposes structural requirements on the controller, e.g., by penalizing the amount of network traffic [29]–[32]. Although our strongest results require strong convexity of f, our theory and techniques are applicable as long as the Hessian of f is positive definite.

Assumption 1: The function f is twice continuously differentiable, has an L_f Lipschitz continuous gradient ∇f, and is strictly convex with $\nabla^2 f \succ 0$; the function g is proper, lower semicontinuous, and convex; and the matrix T has full row rank.

The matrix T is important when the desired structure has a simple representation in the co-domain of T, but it makes the problem more challenging. One approach is to reformulate (1) by introducing an auxiliary optimization variable $z \in \mathbb{R}^m$,

$$
\begin{aligned}
\underset{x,\,z}{\text{minimize}} \quad & f(x) \;+\; g(z) \\
\text{subject to} \quad & Tx \;-\; z \;=\; 0.
\end{aligned}
\tag{2}
$$

Problem (2) is convenient for constrained optimization algorithms based on the augmented Lagrangian,

$$
\mathcal{L}_\mu(x, z; y) \;:=\; f(x) \;+\; g(z) \;+\; y^T(Tx - z) \;+\; \tfrac{1}{2\mu}\|Tx - z\|^2,
$$

where $y \in \mathbb{R}^m$ is the Lagrange multiplier and μ is a positive parameter. Relative to the standard Lagrangian, $\mathcal{L}_\mu$ contains an additional quadratic penalty on the linear constraint in (2).

In the remainder of this section, we provide background on proximal operators and describe generalizations of the gradient for nondifferentiable functions. We also briefly overview existing approaches for solving (1).

A. Background

1) Proximal operators: The proximal operator of the function g is the minimizer of the sum of g and a proximal term,

$$
\mathbf{prox}_{\mu g}(v) \;:=\; \underset{z}{\arg\min}\; g(z) \;+\; \tfrac{1}{2\mu}\|z - v\|^2
\tag{3a}
$$

where μ is a positive parameter and v is a given vector. When g is convex, its proximal operator is Lipschitz continuous with parameter 1, differentiable almost everywhere, and firmly non-expansive [9]. The value function associated with (3a) specifies the Moreau envelope of g,

$$
\begin{aligned}
M_{\mu g}(v) \;:=\; & \inf_z\; g(z) \;+\; \tfrac{1}{2\mu}\|z - v\|^2 \\
\;=\; & g(\mathbf{prox}_{\mu g}(v)) \;+\; \tfrac{1}{2\mu}\|\mathbf{prox}_{\mu g}(v) - v\|^2.
\end{aligned}
\tag{3b}
$$

The Moreau envelope is continuously differentiable, even when g is not, and its gradient

$$
\nabla M_{\mu g}(v) \;=\; \tfrac{1}{\mu}\big(v \;-\; \mathbf{prox}_{\mu g}(v)\big)
\tag{3c}
$$

is Lipschitz continuous with parameter $1/\mu$.

For example, the proximal operator associated with the ℓ_1 norm, $g(z) = \sum |z_i|$, is determined by soft-thresholding,

$$
\mathcal{S}_\mu(v_i) \;:=\; \text{sign}(v_i)\max\{|v_i| \;-\; \mu,\; 0\},
$$

the associated Moreau envelope is the Huber function,

$$
M_{\mu g}(v_i) \;=\;
\begin{cases}
\tfrac{1}{2\mu}v_i^2, & |v_i| \leq \mu \\
|v_i| - \tfrac{\mu}{2}, & |v_i| \geq \mu
\end{cases}
$$

and its gradient is the saturation function,

$$\nabla M_{\mu g}(v_i) \;=\; \operatorname{sign}(v_i)\min\{|v_i|/\mu,\,1\}.$$

Other regularizers with efficiently computable proximal operators include indicator functions of simple convex sets as well as the nuclear and Frobenius norms of matricial variables. Such regularizers can enforce bounds on x, promote low rank solutions, and enhance group sparsity, respectively.

2) Generalization of the gradient and Jacobian: Although $\mathbf{prox}_{\mu g}$ is typically not differentiable, it is Lipschitz continuous and therefore differentiable almost everywhere [33]. One generalization of the gradient for such functions is given by the B-subdifferential set [34], which applies to locally Lipschitz continuous functions $h\colon \mathbb{R}^m \to \mathbb{R}$. Let C_h be a set at which h is differentiable. Each element in the set $\partial_B h(\bar{z})$ is the limit point of a sequence of gradients $\{\nabla h(z_k)\}$ evaluated at a sequence of points $\{z_k\} \subset C_h$ whose limit is $\bar{z}$,

$$\partial_B h(\bar{z}) \;:=\; \left\{ J \mid \exists\{z_k\} \subset C_h,\; z_k \to \bar{z},\; \nabla h(z_k) \to J \right\}. \tag{4}$$

If h is continuously differentiable in the neighborhood of a point z, the B-subdifferential set becomes single valued and it is given by the gradient, $\partial_B h(z) = \nabla h(z)$. In general, $\partial_B h(\bar{z})$ is not a convex set; e.g., if $h(z) = |z|$, $\partial_B h(0) = \{-1, 1\}$.

The Clarke subdifferential set of $h\colon \mathbb{R}^m \to \mathbb{R}$ at $\bar{z}$ is the convex hull of the B-subdifferential set [35],

$$\partial_C h(\bar{z}) \;:=\; \operatorname{conv}\left(\partial_B h(\bar{z}) \right).$$

When h is a convex function, the Clarke subdifferential set is equal to the subdifferential set $\partial h(\bar{z})$ which defines the supporting hyperplanes of h at $\bar{z}$ [36, Chapter VI]. For a function $G\colon \mathbb{R}^m \to \mathbb{R}^n$, the B-generalization of the Jacobian at a point $\bar{z}$ is given by

$$\partial_B G(\bar{z}) \;:=\; \left[\begin{array}{ccc} J_1^T & \dots & J_n^T \end{array} \right]^T$$

where each $J_i \in \partial_B G_i(\bar{z})$ is a member of the B-subdifferential set of the ith component of G evaluated at $\bar{z}$. The Clarke generalization of the Jacobian at a point $\bar{z}$, $\partial_C G(\bar{z})$, has the same structure where each $J_i \in \partial_C G_i(\bar{z})$ is a member of the Clarke subdifferential of $G_i(\bar{z})$.

3) Semismoothness: The mapping $G\colon \mathbb{R}^m \to \mathbb{R}^n$ is semismooth at $\bar{z}$ if for any sequence $z_k \to \bar{z}$, the sequence of Clarke generalized Jacobians $J_{G_k} \in \partial_C G(z_k)$ provides a first order approximation of G,

$$\|G(z_k) - G(\bar{z}) + J_{G_k}(\bar{z} - z_k)\| \;=\; o(\|z_k - \bar{z}\|). \tag{5}$$

where $\phi(k) = o(\psi(k))$ denotes that $\phi(k)/\psi(k) \to 0$ as k tends to infinity [37]. The function G is strongly semismooth if this approximation satisfies the stronger condition,

$$\|G(z_k) - G(\bar{z}) + J_{G_k}(\bar{z} - z_k)\| \;=\; O(\|z_k - \bar{z}\|^2),$$

where $\phi(k) = O(\psi(k))$ signifies that $|\phi(k)| \leq L\psi(k)$ for some positive constant L and positive $\psi(k)$ [37].

Remark 1: (Strong) semismoothness of the proximal operator leads to fast asymptotic convergence of the differential inclusion (see Section IV) and the efficient algorithm (see Section V). Proximal operators associated

with many typical regularization functions (e.g., the ℓ_1 and nuclear norms [38], piecewise quadratic functions [39], and indicator functions of affine convex sets [39]) are strongly semismooth. In general, semismoothness of $\mathbf{prox}_{\mu g}$ follows from semismoothness of the projection onto the epigraph of g [39]. However, there are convex sets onto which projection is not directionally differentiable [40]. The indicator functions associated with such sets or functions whose epigraph is described by such sets may induce proximal operators which are not semismooth.

B. Existing methods

Problem (1) is encountered in a host of applications and it has been the subject of extensive study. Herein, we provide a brief overview of existing approaches to solving it.

1) First order methods: When T is identity or a diagonal matrix, the proximal gradient method, which generalizes gradient descent to certain classes of nonsmooth composite optimization problems [9], [10], can be used to solve (1),

$$x^{k+1} \;=\; \mathbf{prox}_{\alpha^k g}\big(x^k \,-\, \alpha^k \nabla f(x^k)\big)$$

where x^k is the current iterate and α^k is the step size. When $g = 0$, we recover gradient descent, when g is the indicator function $I_{\mathcal{C}}(x)$ of the convex set $\mathcal{C}$, it simplifies to projected gradient descent, and when g is the ℓ_1 norm, it corresponds to the Iterative Soft-Thresholding Algorithm (ISTA). Nesterov-style techniques can also be employed for acceleration [10].

When the matrix T is not diagonal, the alternating direction method of multipliers (ADMM) provides an appealing option for solving (1) via (2) by alternating between minimization of $\mathcal{L}_\mu(x, z; y)$ over x (a continuously differentiable problem), minimization of $\mathcal{L}_\mu(x, z; y)$ over z (amounts to evaluating $\mathbf{prox}_{\mu g}$), and a gradient ascent step in y [11],

$$
\begin{aligned}
x^{k+1} &= \underset{x}{\operatorname{argmin}} \, \mathcal{L}_\mu(x, z^k; y^k) \\
z^{k+1} &= \mathbf{prox}_{\mu g}(Tx^{k+1} + \mu y^k) \\
y^{k+1} &= y^k + \tfrac{1}{\mu}\big(Tx^{k+1} - z^{k+1}\big).
\end{aligned}
\tag{6}
$$

Although each step in ADMM is conveniently computable, its convergence rate is strongly influenced by the parameter μ.

2) Second order methods: The slow convergence of first order methods to high-accuracy solutions motivates the development of second order methods for solving (1). A generalization of Newton's method for nonsmooth problems (1) with $T = I$ was developed in [16]–[19]. A sequential quadratic approximation of the smooth part of the objective function is utilized and a search direction $\tilde{x}$ is obtained as the solution of a regularized quadratic subproblem,

$$\underset{\tilde{x}}{\operatorname{minimize}} \quad \tfrac{1}{2}\,\tilde{x}^T H \tilde{x} \;+\; \nabla f(x^k)^T \tilde{x} \;+\; g(x^k + \tilde{x}) \tag{7}$$

where x^k is the current iterate and H is the Hessian of f. This method generalizes the projected Newton method [41] to a broader class of regularizers. For example, when g is the ℓ_1 norm, this amounts to solving a LASSO problem [42], which can be a challenging task. Coordinate descent is often used to solve this subproblem [19] and it has been observed to perform well in practice [20]–[22].

The Forward-Backward Envelope (FBE) was introduced in [26]–[28]. FBE is once-continuously differentiable nonconvex function of x and its minimum corresponds to the solution of (1) with $T = I$. As demonstrated in Section VII, FBE can be obtained from the proximal augmented Lagrangian (that we introduce in Section III). Since the generalized Hessian of FBE involves third-order derivatives of f (which may be expensive to compute), references [26]–[28] employ either truncated- or quasi-Newton methods to obtain a second order update to x.

III. THE PROXIMAL AUGMENTED LAGRANGIAN AND SECOND ORDER UPDATES

In this section, we transform $\mathcal{L}_\mu(x, z; y)$ into a form that is once but not twice continuously differentiable. For the resulting function, which we call the proximal augmented Lagrangian, we define second order updates to find its saddle points, show that they are always well defined, and prove that they are locally (quadratically) superlinearly convergent when $\mathbf{prox}_{\mu g}$ is (strongly) semismooth.

A. Proximal augmented Lagrangian

The continuously differentiable proximal augmented Lagrangian was recently introduced in [12]. This was done by rewriting $\mathcal{L}_\mu(x, z; y)$ via completion of squares,

$$\mathcal{L}_\mu(x, z; y) = f(x) + g(z) + \tfrac{1}{2\mu} \|z - (Tx + \mu y)\|^2 - \tfrac{\mu}{2} \|y\|^2$$

and restricting it to the manifold that corresponds to explicit minimization over the auxiliary variable z,

$$\mathcal{L}_\mu(x; y) := \mathcal{L}_\mu(x, z_\mu^\star(x, y); y)$$

where the minimizer of $\mathcal{L}_\mu(x, z; y)$ over z is determined by the proximal operator of the function g,

$$z_\mu^\star(x, y) = \underset{z}{\text{argmin}} \; \mathcal{L}_\mu(x, z; y) = \mathbf{prox}_{\mu g}(Tx + \mu y)$$

and it defines the aforementioned manifold.

Theorem 1 (Theorem 1 in [12]): Let Assumption 1 hold. Then, minimization of the augmented Lagrangian $\mathcal{L}_\mu(x, z; y)$ associated with problem (2) over (x, z) is equivalent to minimization of the *proximal augmented Lagrangian*

$$\mathcal{L}_\mu(x; y) = f(x) + M_{\mu g}(Tx + \mu y) - \tfrac{\mu}{2} \|y\|^2 \tag{8}$$

over x. Moreover, $\mathcal{L}_\mu(x; y)$ is continuously differentiable over x and y and its gradient $\nabla \mathcal{L}_\mu(x; y)$,

$$\nabla \mathcal{L}_\mu(x; y) = \begin{bmatrix} \nabla_x \mathcal{L}_\mu(x; y) \\ \nabla_y \mathcal{L}_\mu(x; y) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + T^T \nabla M_{\mu g}(Tx + \mu y) \\ \mu \nabla M_{\mu g}(Tx + \mu y) - \mu y \end{bmatrix} \tag{9}$$

is Lipschitz continuous.

The proximal augmented Lagrangian $\mathcal{L}_\mu(x; y)$ contains the Moreau envelope of g and its introduction allows the use of the method of multipliers (MM) to solve problem (2). MM requires *joint* minimization of $\mathcal{L}_\mu(x, z; y)$ over x

and z which is, in general, challenging because the (x, z)-minimization subproblem is nondifferentiable. However, Theorem 1 enables an equivalent implementation of MM

$$
\begin{aligned}
x^{k+1} &= \underset{x}{\operatorname{argmin}} \; \mathcal{L}_\mu(x; y^k) \\
y^{k+1} &= y^k + \tfrac{1}{\mu}\left(Tx^{k+1} - z_\mu^\star(x^{k+1}, y^k)\right)
\end{aligned}
\tag{10}
$$

which improves performance relative to ADMM and has guaranteed convergence to a local minimum even when f is nonconvex [12].

Continuous differentiability of $\mathcal{L}_\mu(x; y)$ also enables a joint update of x and y via the primal-dual Arrow-Hurwicz-Uzawa gradient flow dynamics,

$$
\begin{aligned}
\dot{x} &= -\nabla_x \mathcal{L}_\mu(x; y) \\
\dot{y} &= \nabla_y \mathcal{L}_\mu(x; y)
\end{aligned}
$$

where $\nabla_x \mathcal{L}_\mu$ and $\nabla_y \mathcal{L}_\mu$ are given by (9). When ∇f and T are sparse mappings, this method is convenient for distributed implementation and it is guaranteed to converge at an exponential rate for strongly convex f and sufficiently large μ [12].

In what follows, we extend the primal-dual algorithm to incorporate second order information of $\mathcal{L}_\mu(x; y)$ and thereby achieve fast convergence to high-accuracy solutions.

B. Second order updates

Even though Newton's method is primarily used for solving minimization problems in modern optimization, it was originally formulated as a root-finding technique and it has long been employed for finding stationary points [43]. In [23], a generalized Jacobian was used to extend Newton's method to semismooth problems. We employ this generalization of Newton's method to $\nabla \mathcal{L}_\mu(x; y)$ in order to compute the saddle point of the proximal augmented Lagrangian. The unique saddle point of $\mathcal{L}_\mu(x; y)$ is given by the optimal primal-dual pair $(x^\star, y^\star)$ and it thus provides the solution to (1).

1) Generalized Newton updates: Let $H := \nabla^2 f(x)$. We use the B-generalized Jacobian of the proximal operator $\mathbf{prox}_{\mu g}$, $\mathbb{P}_B := \partial_B \mathbf{prox}_{\mu g}(Tx + \mu y)$, to define the set of B-generalized Hessians of the proximal augmented Lagrangian,

$$
\partial_B^2 \mathcal{L}_\mu := \left\{ \begin{bmatrix} H + \tfrac{1}{\mu} T^T(I - P)T & T^T(I - P) \\ (I - P)T & -\mu P \end{bmatrix}, P \in \mathbb{P}_B \right\}
\tag{11a}
$$

and the Clarke generalized Jacobian $\mathbb{P}_C := \partial_C \mathbf{prox}_{\mu g}(Tx + \mu y)$ to define the set of Clarke generalized Hessians of the proximal augmented Lagrangian,

$$
\partial_C^2 \mathcal{L}_\mu := \left\{ \begin{bmatrix} H + \tfrac{1}{\mu} T^T(I - P)T & T^T(I - P) \\ (I - P)T & -\mu P \end{bmatrix}, P \in \mathbb{P}_C \right\}.
\tag{11b}
$$

Note that $\partial_B^2 \mathcal{L}_\mu(x; y) \subset \partial_C^2 \mathcal{L}_\mu(x; y)$ because $\mathbb{P}_B \subset \mathbb{P}_C$.

In the rest of the paper, we introduce the composite variable, $w := [\, x^T \; y^T \,]^T$, use $\mathcal{L}_\mu(w)$ interchangeably with $\mathcal{L}_\mu(x; y)$, and suppress the dependance of H and P on w to reduce notational clutter. For simplicity of exposition, we assume that $\mathbf{prox}_{\mu g}$ is semismooth and state the results for the Clarke generalized Hessian (11b),

i.e., $\partial^2 \mathcal{L}_\mu(w) = \partial_C^2 \mathcal{L}_\mu(w)$. As described in Remark 4 in Section IV, analogous convergence results for non-semismooth $\mathbf{prox}_{\mu g}$ can be obtained for the B-generalized Hessian (11a), i.e., $\partial^2 \mathcal{L}_\mu(w) = \partial_B^2 \mathcal{L}_\mu(w)$.

We use the Clarke generalized Hessian (11b) to obtain a second order update $\tilde{w}$ by linearizing the stationarity condition $\nabla \mathcal{L}_\mu(w) = 0$ around the current iterate w^k,

$$\partial_C^2 \mathcal{L}_\mu(w^k)\, \tilde{w}^k \;=\; -\nabla \mathcal{L}_\mu(w^k). \tag{12}$$

Since $\mathbf{prox}_{\mu g}$ is firmly nonexpansive, $0 \preceq P \preceq I$. In Lemma 2 we use this fact to prove that the second order update $\tilde{w}$ is well-defined for any generalized Hessian (11) of the proximal augmented Lagrangian $\mathcal{L}_\mu(x; y)$ as long as f is strictly convex with $\nabla^2 f(x) \succ 0$ for all $x \in \mathbb{R}^n$.

Lemma 2: Let $H \in \mathbb{R}^{n \times n}$ be symmetric positive definite, $H \succ 0$, let $P \in \mathbb{R}^{m \times m}$ be symmetric positive semidefinite with eigenvalues less than one, $0 \preceq P \preceq I$, let $T \in \mathbb{R}^{m \times n}$ be full row rank, and let $\mu > 0$. Then, the matrix

$$\begin{bmatrix} H + \frac{1}{\mu} T^T (I - P) T & T^T (I - P) \\ (I - P) T & -\mu P \end{bmatrix}$$

is invertible and it has n positive and m negative eigenvalues.

Proof: The Haynsworth inertia additivity formula [44] implies that the inertia of matrix (11) is determined by the sum of the inertias of matrices,

$$H \;+\; \tfrac{1}{\mu} T^T (I \,-\, P) T \tag{13a}$$

and

$$- \mu P \;-\; (I - P)\, T \left(H + \tfrac{1}{\mu} T^T (I - P) T \right)^{-1} T^T (I - P). \tag{13b}$$

Matrix (13a) is positive definite because $H \succ 0$ and both P and $I - P$ are positive semidefinite. Matrix (13b) is negative definite because the kernels of P and $I - P$ have no nontrivial intersection and T has full row rank. ∎

2) Fast local convergence: The use of generalized Newton updates for solving the nonlinear equation $G(x) = 0$ for nondifferentiable G was studied in [23]. We apply this framework to the stationarity condition $\nabla \mathcal{L}_\mu(w) = 0$ when $\mathbf{prox}_{\mu g}$ is (strongly) semismooth and show that second order updates (12) converge (quadratically) superlinearly within a neighborhood of the optimal primal-dual pair.

Proposition 3: Let $\mathbf{prox}_{\mu g}$ be (strongly) semismooth, and let $\tilde{w}^k$ be defined by (12). Then, there is a neighborhood of the optimal solution $w^\star$ in which the second order iterates $w^{k+1} = w^k + \tilde{w}^k$ converge (quadratically) superlinearly to $w^\star$.

Proof: Lemma 2 establishes that $\partial_C^2 \mathcal{L}_\mu(w)$ is nonsingular for any $P \in \mathbb{P}_C$. Since the gradient $\nabla \mathcal{L}_\mu(w)$ of the proximal augmented Lagrangian is Lipschitz continuous by Theorem 1, nonsingularity of $\partial_C^2 \mathcal{L}_\mu(w)$ and (strong) semismoothness of the proximal operator guarantee (quadratic) superlinear convergence of the iterates by [23, Theorem 3.2]. ∎

IV. A GLOBALLY CONVERGENT DIFFERENTIAL INCLUSION

Since we apply a generalization of Newton's method to a saddle point problem and the second order updates are set valued, convergence to the optimal point is not immediate. Although we showed local convergence rates in

Proposition 3 by leveraging the results of [23], proof of the global convergence is more subtle and it is established next.

To justify the development of a discrete-time algorithm based on the search direction resulting from (12), we first examine the corresponding differential inclusion,

$$\dot{w} \in -(\partial_C^2 \mathcal{L}_\mu(w))^{-1} \nabla \mathcal{L}_\mu(w) \tag{14}$$

where $\partial_C^2 \mathcal{L}_\mu$ is the Clarke generalized Hessian (11b) of $\mathcal{L}_\mu$. We assume existence of a solution and prove asymptotic stability of (14) under Assumption 1 and global exponential stability under an additional assumption that f is strongly convex.

Assumption 2: Differential inclusion (14) has a solution.

A. Asymptotic stability

We first establish asymptotic stability of differential inclusion (14).

Theorem 4: Let Assumptions 1 and 2 hold and let $\mathbf{prox}_{\mu g}$ be semismooth. Then, differential inclusion (14) is asymptotically stable. Moreover,

$$V(w) := \tfrac{1}{2} \|\nabla \mathcal{L}_\mu(w)\|^2 \tag{15}$$

provides a Lyapunov function and

$$\dot{V}(t) = -2V(t). \tag{16}$$

Proof: Lyapunov function candidate (15) is a positive function of w everywhere apart from the optimal primal-dual pair $w^\star$ where it is zero. It remains to show that V is decreasing along the solutions $w(t)$ of (14), i.e., that $\dot{V}$ is strictly negative for all $w(t) \neq w^\star$,

$$\dot{V}(t) := \tfrac{\mathrm{d}}{\mathrm{d}t} V(w(t)) = -2V(w(t))$$

For Lyapunov function candidates $\hat{V}(w)$ which are differentiable with respect to w, $\dot{\hat{V}} = \dot{w}^T \nabla \hat{V}$. Although (15) is *not* differentiable with respect to w, we show that $V(w(t))$ *is* differentiable *along the solutions* of (14). Instead of employing the chain rule, we use the limit that defines the derivative,

$$\dot{V}(t) := \tfrac{\mathrm{d}}{\mathrm{d}t} V(w(t)) = \lim_{s \to 0} \frac{V(w(t) + s\tilde{w}(t)) - V(w(t))}{s}$$

to show that $\dot{V}$ exists and is negative along the solutions of (14). Here, $\tilde{w} \in -(\partial_C^2 \mathcal{L}_\mu(w))^{-1} \nabla \mathcal{L}_\mu(w)$ is determined by the dynamics (14). We first introduce

$$h_s(t) := \frac{V(w(t) + s\tilde{w}(t)) - V(w(t))}{s}$$

which gives $\dot{V}$ in the limit $s \to 0$. We then rewrite $h_s(t)$ as the limit point of a sequence of functions $\{h_{s,k}(t)\}$ so that

$$\dot{V}(t) = \lim_{s \to 0} h_s(t) = \lim_{s \to 0} \lim_{k \to \infty} h_{s,k}(t) \tag{17}$$

and use the Moore-Osgood theorem [45, Theorem 7.11] to exchange the order of the limits and establish that $\dot{V} = -2V$.

Let C_g denote a subset of $\mathbb{R}^{n+m}$ over which $\mathbf{prox}_{\mu g}(Tx + \mu y)$ is differentiable (and therefore V is differentiable with respect to w) and let $\{w_k\}$ be a sequence of points in C_g that converges to w. We define the sequence of functions $\{h_{s,k}(t)\}$,

$$h_{s,k}(t) := \frac{V(w_k(t) + s\tilde{w}_k(t)) - V(w_k(t))}{s}$$

where $\tilde{w}_k(t) \in -(\partial_C^2 \mathcal{L}_\mu(w(t)))^{-1} \nabla \mathcal{L}_\mu(w_k(t))$, as we establish below, converges to $\tilde{w}$. To employ the Moore-Osgood theorem, it remains to show that $h_{s,k}(t)$ converges pointwise (for any k) as $s \to 0$ and that $h_{s,k}(t)$ converges uniformly on some interval $s \in [0, \bar{s}]$ as $k \to \infty$.

Since $\{w_k\} \subset C_g$, $V(w_k)$ is differentiable for every $k \in \mathbb{Z}_+$ and $\nabla V(w_k) = \nabla^2 \mathcal{L}_\mu(w_k) = \partial_C^2 \mathcal{L}_\mu(w_k)$. It thus follows that

$$\lim_{s \to 0} h_{s,k}(t) = \tilde{w}_k^T(t) \partial_C^2 \mathcal{L}_\mu(w_k(t)) \nabla \mathcal{L}_\mu(w_k(t)) \tag{18}$$

pointwise (for any k).

We now show that the sequence $\{h_{s,k}(t)\}$ converges uniformly to $h_s(t)$ as $k \to \infty$ implying that we can exchange the order of the limits in (17). Since $\mathbf{prox}_{\mu g}$ is semismooth, $\nabla \mathcal{L}_\mu$ is semismooth. By (5), $\nabla \mathcal{L}_\mu(w_k)$ can be written as,

$$\nabla \mathcal{L}_\mu(w_k) = \nabla \mathcal{L}_\mu(w) - \partial_C^2 \mathcal{L}_\mu(w_k)(w - w_k) + R_k$$

for sufficiently large k, where $\|R_k\| = o(\|w_k - w\|)$. Lemma 2 and [23, Proposition 3.1] imply that $\partial_C^2 \mathcal{L}_\mu(w_k)$ is bounded within some neighborhood of w and thus that $\tilde{w}_k$ can be written as $\tilde{w}_k = \tilde{w} + \hat{R}_k$ where $\|\hat{R}_k\| = O(\|w_k - w\|)$. This implies convergence of $\tilde{w}_k$ to $\tilde{w}$ and, combined with local Lipschitz continuity of V with respect to w_k, uniform convergence of $h_{s,k}(t)$ to $h_s(t)$ on $s \in (0, \bar{s}]$ where $\bar{s} > 0$.

Therefore, the Moore-Osgood theorem on exchanging limits [45, Theorem 7.11] in conjunction with (18) imply

$$\begin{aligned}
\dot{V}(t) &= \lim_{s \to 0} h_s(t) \\
&= \lim_{s \to 0} \lim_{k \to \infty} h_{s,k}(t) = \lim_{k \to \infty} \lim_{s \to 0} h_{s,k}(t) \\
&= \lim_{k \to \infty} \tilde{w}_k^T(t) \partial_C^2 \mathcal{L}_\mu(w_k(t)) \nabla \mathcal{L}_\mu(w_k(t)) \\
&= -\|\nabla \mathcal{L}_\mu(w(t))\|^2 = -2V(t)
\end{aligned}$$

which establishes (16) and thereby completes the proof. ∎

B. Global exponential stability

To establish global asymptotic stability, we show that the Lyapunov function (15) is radially unbounded, and to prove exponential stability we bound it with quadratic functions. We first provide two lemmas that characterize the mappings $\mathbf{prox}_{\mu g}$ and ∇f in terms of the spectral properties of matrices that describe the corresponding input-output relations at given points.

Lemma 5 (Lemma 2 in [12]): Let g be a proper, lower semicontinuous, convex function and let $\mathbf{prox}_{\mu g} \colon \mathbb{R}^m \to \mathbb{R}^m$ be the corresponding proximal operator. Then, for any $a, b \in \mathbb{R}^m$, there exists a symmetric matrix $D_{a,b}$ satisfying

$0 \preceq D_{a,b} \preceq I$ such that

$$\mathbf{prox}_{\mu g}(a) - \mathbf{prox}_{\mu g}(b) = D_{a,b}(a - b).$$

Lemma 6: Let f be strongly convex with parameter m_f and let its gradient ∇f be Lipschitz continuous with parameter L_f. Then, for any $a, b \in \mathbb{R}^n$ there exists a symmetric matrix $G_{a,b}$ satisfying $m_f I \preceq G_{a,b} \preceq L_f I$ such that

$$\nabla f(a) - \nabla f(b) = G_{a,b}(a - b).$$

Proof: Let $c := a - b$, $d := \nabla f(a) - \nabla f(b)$, $e := d - m_f c$, and

$$\hat{G}_{a,b} := \{ee^T/(e^T c),\ e \neq 0;\ 0,\ \text{otherwise}\} \tag{19a}$$

$$G_{a,b} := \hat{G}_{a,b} + m_f I. \tag{19b}$$

Clearly, by construction, $\hat{G}_{a,b} = \hat{G}_{a,b}^T \succeq 0$. It is also readily verified that $G_{a,b} c = d$ when $e^T c \neq 0$. It thus remains to show that (i) $G_{a,b} c = d$ when $e^T c = 0$; and (ii) $\hat{G}_{a,b} \preceq (L_f - m_f)I$.

(i) Since f is m_f strongly convex and ∇f is L_f Lipschitz continuous, $h(x) := f(x) - \frac{m_f}{2}\|x\|^2$ is convex and $\nabla h(x) = \nabla f(x) - m_f x$ is $L_f - m_f$ Lipschitz continuous. Furthermore, we have $e = \nabla h(a) - \nabla h(b)$, and [46, Proposition 5] implies

$$e^T c \geq \frac{1}{L_f - m_f}\|e\|^2, \quad \text{for all } c \in \mathbb{R}^n. \tag{20}$$

This shows that $e^T c = 0$ only if $e := d - m_f c = 0$ and, thus, $d = m_f c = G_{a,b} c$ when $e^T c = 0$. Therefore, there always exist a symmetric matrix $G_{a,b}$ such that $G_{a,b} c = d$.

(ii) When $e^T c \neq 0$, $\hat{G}_{a,b}$ is a rank one matrix and its only nonzero eigenvalue is $\|e\|^2/(e^T c)$; this follows from $\hat{G}_{a,b} e = (\|e\|^2/(e^T c))\, e$. In this case, inequality (20) implies $e^T c > 0$ and (20) is equivalent to $1/(e^T c) \leq (L_f - m_f)/\|e\|^2$. Thus, $\|e\|^2/(e^T c) \leq L_f - m_f$ and $\hat{G}_{a,b} \preceq (L_f - m_f)I$ when $e^T c \neq 0$. Since $\hat{G}_{a,b} = 0$ when $e^T c = 0$, $\hat{G}_{a,b} \preceq (L_f - m_f)I$ for all a and b. Finally, $\hat{G}_{a,b} \succeq 0$ and (19b) imply $m_f I \preceq G_{a,b} \preceq L_f I$. ∎

Remark 2: Although matrices $D_{a,b}$ and $G_{a,b}$ in Lemmas 5 and 6 depend on the operating point, their spectral properties, $0 \preceq D_{a,b} \preceq I$ and $m_f I \preceq G_{a,b} \preceq L_f I$, hold for all a and b. These lemmas can be interpreted as a combination between a generalization of the mean value theorem [45, Theorem 5.9] to vector-valued functions and spectral bounds on the operators $\mathbf{prox}_{\mu g} \colon \mathbb{R}^m \to \mathbb{R}^m$ and $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ arising from firm nonexpansiveness of $\mathbf{prox}_{\mu g}$, strong convexity of f, and Lipschitz continuity of ∇f.

We now combine Lemmas 5 and 6 to establish quadratic upper and lower bounds for Lyapunov function (15) and thereby prove global exponential stability of differential inclusion (14) for strongly convex f.

Theorem 7: Let Assumptions 1 and 2 hold, let $\mathbf{prox}_{\mu g}$ be semismooth, and let f be m_f strongly convex. Then, differential inclusion (14) is globally exponentially stable, i.e., there exists $\kappa > 0$ such that $\|w(t) - w^\star\| \leq \kappa\, e^{-t}\, \|w(0) - w^\star\|$.

Proof: Given the assumptions, Theorem 4 establishes asymptotic stability of (14) with the dissipation rate $\dot{V}(w) = -2V(w)$. It remains to show the existence of positive constants κ_1 and κ_2 such that Lyapunov

function (15) satisfies

$$\tfrac{\kappa_1}{2}\,\|\tilde{w}\|^2 \;\leq\; V(w) \;\leq\; \tfrac{\kappa_2}{2}\,\|\tilde{w}\|^2 \tag{21}$$

where $\tilde{w} := w - w^\star$ and $w^\star := (x^\star, y^\star)$ is the optimal primal-dual pair. The upper bound in (21) follows from Lipschitz continuity of $\nabla\mathcal{L}_\mu(w)$ (see Theorem 1), with κ_2 determined by the Lipschitz constant of $\nabla\mathcal{L}_\mu(w)$.

To show the lower bound in (21), and thus establish radial unboundedness of $V(w)$, we construct matrices that relate $V(w)$ to $\tilde{w}$. Lemmas 5 and 6 imply the existence of symmetric matrices $D_{\tilde{w}}$ and $G_{\tilde{w}}$ such that $0 \preceq D_{\tilde{w}} \preceq I$, $m_f I \preceq G_{\tilde{w}} \preceq L_f I$, and

$$\mathbf{prox}_{\mu g}(Tx + \mu y) - \mathbf{prox}_{\mu g}(Tx^\star + \mu y^\star) \;=\; D_{\tilde{w}}\,(T\tilde{x} + \mu\tilde{y})$$
$$f(x) - f(x^\star) \;=\; G_{\tilde{w}}\,\tilde{x}.$$

As noted in Remark 2, although $D_{\tilde{w}}$ and $G_{\tilde{w}}$ depend on the operating point, their spectral properties hold for all $\tilde{w}$.

Since $\nabla\mathcal{L}_\mu(w^\star) = 0$, we can write

$$\nabla\mathcal{L}_\mu(w) \;=\; \nabla\mathcal{L}_\mu(w) \,-\, \nabla\mathcal{L}_\mu(w^\star) \;=\; Q_{\tilde{w}}\,\tilde{w}$$

and express Lyapunov function (15) as

$$V(w) \;=\; \tfrac{1}{2}\,\tilde{w}^T Q_{\tilde{w}}^T Q_{\tilde{w}}\,\tilde{w}$$

where

$$Q_{\tilde{w}} \;:=\; \begin{bmatrix} G_{\tilde{w}} + \tfrac{1}{\mu}\,T^T(I - D_{\tilde{w}})T & T^T(I - D_{\tilde{w}}) \\[2mm] (I - D_{\tilde{w}})T & -\mu D_{\tilde{w}} \end{bmatrix}$$

for some $(D_{\tilde{w}}, G_{\tilde{w}}) \in \Omega_{\tilde{w}}$,

$$\Omega_{\tilde{w}} \;:=\; \{(D_{\tilde{w}}, G_{\tilde{w}}) \mid 0 \preceq D_{\tilde{w}} \preceq I,\; m_f I \preceq G_{\tilde{w}} \preceq L_f I\}.$$

The set $\Omega_{\tilde{w}}$ is closed and bounded and the minimum eigenvalue of $Q_{\tilde{w}}^T Q_{\tilde{w}}$ is a continuous function of $G_{\tilde{w}}$ and $D_{\tilde{w}}$. Thus, the extreme value theorem [45, Theorem 4.14] implies that its infimum over $\Omega_{\tilde{w}}$,

$$\kappa_1 \;=\; \inf_{(D_{\tilde{w}}, G_{\tilde{w}}) \in \Omega_{\tilde{w}}} \lambda_{\min}\left(Q_{\tilde{w}}^T Q_{\tilde{w}}\right)$$

is achieved. By Lemma 2, $Q_{\tilde{w}}$ is a full rank matrix, which implies that $Q_{\tilde{w}}^T Q_{\tilde{w}} \succ 0$ for all $\tilde{w}$ and therefore that κ_1 is positive. Thus, $V(w) \geq \tfrac{\kappa_1}{2}\,\|\tilde{w}\|^2$, establishing condition (21).

Condition (16) and [47, Lemma 3.4] imply $V(w(t)) = \mathrm{e}^{-2t} V(w(0))$. It then follows from (21) that

$$\|w(t) - w^\star\|^2 \;\leq\; (\kappa_2/\kappa_1)\,\mathrm{e}^{-2t}\,\|w(0) - w^\star\|^2.$$

Taking the square root completes the proof and provides an upper bound for the constant κ, $\kappa \leq \sqrt{\kappa_2/\kappa_1}$. ∎

Remark 3: The *rate* of exponential convergence established by Theorem 7 is independent of m_f, L_f, and μ. This is a consequence of insensitivity of Newton-like methods to poor conditioning. In contrast, the first order primal-dual method considered in [12] requires a sufficiently large μ for exponential convergence. In our second

order primal-dual method, problem conditioning and parameter selection affect the multiplicative constant κ but not the rate of convergence.

Remark 4: When differential inclusion (14) is defined with the B-generalized Hessian (11a), Theorems 4 and 7 hold even for proximal operators which are not semismooth. This follows from defining $\tilde{w}_k = \tilde{w} \in -(\partial_B^2 \mathcal{L}_\mu(w))^{-1}\nabla\mathcal{L}_\mu(w)$ and choosing $\{w_k\} \subset C_g$ such that $\partial_B^2\mathcal{L}_\mu(w) = \lim_{k\to\infty}\nabla^2\mathcal{L}_\mu(w_k)$ in the proof of Theorem 4. Such a choice of $\{w_k\}$ is possible by the definition of the B-subdifferential (4); since the Clarke subgradient is the convex hull of the B-subdifferential, it contains points outside of $\partial_B^2\mathcal{L}_\mu(w)$ and thus such a sequence $\{w_k\} \subset C_g$ is not guaranteed to exist for any $\partial_C^2\mathcal{L}_\mu(w)$. When defined in this manner, $\tilde{w}_k$ is constant with respect to w_k and thus uniform convergence of $h_k(t,s)$ to $h(t,s)$ is immediate.

V. A SECOND ORDER PRIMAL-DUAL ALGORITHM

An algorithm based on the second order updates (12) requires step size selection to ensure global convergence. This is challenging for saddle point problems because standard notions, such as sufficient descent, cannot be applied to assess the progress of the iterates. Instead, it is necessary to identify a merit function whose minimum lies at the stationary point and whose sufficient descent can be used to evaluate progress towards the saddle point.

An approach based on discretization of differential inclusion (14) and Lyapunov function (15) as a merit function leads to Algorithm 2 in Appendix A. However, such a merit function is nonconvex and nondifferentiable in general which makes the utility of backtracking (e.g., the Armijo rule) unclear. Moreover, Algorithm 2 employs a fixed penalty parameter μ. *A priori* selection of this parameter is difficult and it has a large effect on the convergence speed.

Instead, we employ the primal-dual augmented Lagrangian introduced in [13] as a merit function and incorporate an adaptive μ update. This merit function is convex in both x and y and it facilitates an implementation with outstanding practical performance. Drawing upon recent advancements for constrained optimization of twice differentiable functions [14], [15], we show that our algorithm converges to the solution of (2). Finally, our algorithm exhibits local (quadratic) superlinear convergence for (strongly) semismooth $\mathbf{prox}_{\mu g}$.

A. Merit function

The primal-dual augmented Lagrangian,

$$\mathcal{V}_\mu(x,z;y,\lambda) := \mathcal{L}_\mu(x,z;\lambda) + \tfrac{1}{2\mu}\left\|Tx - z + \mu\left(\lambda - y\right)\right\|^2$$

was introduced in [13], where λ is an estimate of the optimal Lagrange multiplier $y^\star$. Following [13, Theorem 3.1], it can be shown that the optimal primal-dual pair $(x^\star, z^\star; y^\star)$ of optimization problem (2) is a stationary point of $\mathcal{V}_\mu(x,z;y,y^\star)$. Furthermore, for any fixed λ, $\mathcal{V}_\mu$ is a convex function of x, z, and y and it has a unique global minimizer.

In contrast to [13], we study problems in which a component of the objective function is not differentiable. As in Theorem 1, the Moreau envelope associated with the nondifferentiable component g allows us to eliminate the dependence of the primal-dual augmented Lagrangian $\mathcal{V}_\mu$ on z,

$$\begin{aligned}
\hat{z}_\mu^\star(x;y,\lambda) &= \operatorname*{argmin}_z \mathcal{V}_\mu(x,z;y,\lambda) \\
&= \mathbf{prox}_{\frac{\mu}{2}g}(Tx + \tfrac{\mu}{2}(2\lambda - y))
\end{aligned}$$

and to express $\mathcal{V}_\mu$ as a continuously differentiable function,

$$\mathcal{V}_\mu(x; y, \lambda) \; := \; \mathcal{V}_\mu(x, \hat{z}_\mu^\star(x; y, \lambda); y, \lambda) \; = \; f(x) \; + \; M_{\frac{\mu}{2}g}\big(Tx + \tfrac{\mu}{2}(2\lambda - y)\big) \; + \; \tfrac{\mu}{4}\|y\|^2 \; - \; \tfrac{\mu}{2}\|\lambda\|^2.$$

For notational compactness, we suppress the dependence on λ and write $\mathcal{V}_\mu(w)$ when λ is fixed.

Remark 5: The primal-dual augmented Lagrangian is not a Lyapunov function unless $\lambda = y^\star$. We establish convergence by minimizing $\mathcal{V}_\mu(x; y, \lambda)$ over $(x; y)$ – a convex problem – while adaptively updating the Lagrange multiplier estimate λ.

In [13], [15], the authors obtain a search direction using the Hessian of the merit function, $\nabla^2 \mathcal{V}_\mu$. Instead of implementing an analogous update using generalized Hessian $\partial^2 \mathcal{V}_\mu$, we take advantage of the efficient inversion of $\partial^2 \mathcal{L}_\mu$ (see Section V-B3) to define the update

$$\partial^2 \mathcal{L}_\mu(w^k)\, \tilde{w} \; = \; - \text{ blkdiag}(I, -I)\, \nabla \mathcal{V}_{2\mu}(w^k) \tag{22}$$

where the identity matrices are sized conformably with the dimensions of x and y, and

$$\nabla \mathcal{V}_{2\mu}(w) = \left[\begin{array}{c} \nabla f(x) + T^T \nabla M_{\mu g}(Tx + \mu(2\lambda - y)) \\ \mu(y - \nabla M_{\mu g}(Tx + \mu(2\lambda - y))) \end{array} \right]. \tag{23}$$

Multiplication by $\text{blkdiag}(I, -I)$ is used to ensure descent in the dual direction and $\mathcal{V}_{2\mu}$ is employed because $\hat{z}_\mu^\star(x; y, \lambda)$ is determined by the proximal operator associated with $(\mu/2)g$. When $\lambda = y$, $\nabla_x \mathcal{V}_{2\mu} = \nabla_x \mathcal{L}_\mu$, $\nabla_y \mathcal{V}_{2\mu} = -\nabla_y \mathcal{L}_\mu$, and (22) becomes equivalent to the second order update (12).

Lemma 8: Let $\tilde{w}$ solve (22). Then, for the fixed value of the Lagrange multiplier estimate λ and any $\sigma \in (0, 1]$,

$$d \; := \; (1 - \sigma)\, \tilde{w} \; - \; \sigma\, \nabla \mathcal{V}_{2\mu}(w, \lambda) \tag{24}$$

is a descent direction of the merit function $\mathcal{V}_{2\mu}(w, \lambda)$.

Proof: By multiplying (22) with the nonsingular matrix

$$\Pi \; := \; \left[\begin{array}{cc} I & -\tfrac{1}{\mu}T^T \\ 0 & I \end{array} \right] \tag{25}$$

we can express it as

$$\left[\begin{array}{cc} H & T^T \\ (I - P)T & -\mu P \end{array} \right] \left[\begin{array}{c} \tilde{x} \\ \tilde{y} \end{array} \right] = \left[\begin{array}{c} -(\nabla f(x) + T^T y) \\ \nabla_y \mathcal{V}_{2\mu}(x; y, \lambda) \end{array} \right] \tag{26}$$

where $H := \nabla^2 f(x) \succ 0$. Using (23) and (26), $\nabla \mathcal{V}_{2\mu}(w)$ can be expressed as,

$$\nabla \mathcal{V}_{2\mu}(w) = \left[\begin{array}{c} -(H + \tfrac{1}{\mu}T^T(I - P)T)\tilde{x} - T^T(I - P)\tilde{y} \\ (I - P)T\tilde{x} - \mu P\tilde{y} \end{array} \right].$$

Thus,

$$\tilde{w}^T \nabla \mathcal{V}_{2\mu}(w) \; = \; -\tilde{x}^T(H + \tfrac{1}{\mu}T^T(I - P)T)\tilde{x} - \mu \tilde{y}^T P \tilde{y}$$

is negative semidefinite, and the inner product

$$d^T \nabla \mathcal{V}_{2\mu}(w) = (1 - \sigma) \tilde{w}^T \nabla \mathcal{V}_{2\mu}(w) - \sigma \| \nabla \mathcal{V}_{2\mu}(w) \|^2$$

is negative definite when $\nabla \mathcal{V}_{2\mu}$ is nonzero. ∎

B. Second order primal-dual algorithm

We now develop a customized algorithm that alternates between minimizing the merit function $\mathcal{V}_\mu(x; y, \lambda)$ over $(x; y)$ and updating λ. Near the optimal solution, the algorithm approaches second order updates (12) with unit step size, leading to local (quadratic) superlinear convergence for (strongly) semismooth $\mathbf{prox}_{\mu g}$.

Our approach builds on the sequential quadratic programming method described in [13]–[15] and it uses the primal-dual augmented Lagrangian as a merit function to assess progress of iterates to the optimal solution. Inspired by [48], we ensure sufficient progress with damped second order updates.

The following two quantities

$$\begin{aligned} r &:= Tx - \mathbf{prox}_{\mu g}(Tx + \mu y) \\ s &:= Tx - \mathbf{prox}_{\mu g}(Tx + \mu(2\lambda - y)) \end{aligned}$$

appear in the proof of global convergence. Note that r is the primal residual of optimization problem (2) and that $\nabla \mathcal{V}_{2\mu}$ can be equivalently expressed as

$$\nabla \mathcal{V}_{2\mu}(w) = \begin{bmatrix} \nabla f(x) + \frac{1}{\mu} T^T(s + \mu(2\lambda - y)) \\ -(s + 2\mu(\lambda - y)) \end{bmatrix}. \tag{27}$$

1) Global convergence: We now establish global convergence of Algorithm 1 under an assumption that the sequence of gradients generated by the algorithm, $\nabla f(x^k)$, is bounded. This assumption is standard for augmented Lagrangian based methods [15], [49] and it does not lead to a loss of generality when f is strongly convex.

Theorem 9: Let Assumption 1 hold and let the sequence $\{\nabla f(x^k)\}$ resulting from Algorithm 1 be bounded. Then, the sequence of iterates $\{w^k\}$ converges to the optimal primal-dual point of problem (2) and the Lagrange multiplier estimates $\{\lambda^k\}$ converge to the optimal Lagrange multiplier.

Proof: Since $\mathcal{V}_{2\mu}(w, \lambda)$ is convex in w for any fixed λ, condition (30) in Algorithm 1 will be satisfied after finite number of iterations. Combining (30) and (27) shows that $s^k + 2\mu^k(\lambda^k - y^k) \to 0$ and $\nabla f(x^k) + \frac{1}{\mu^k} T^T(s^k + \mu(2\lambda^k - y^k)) \to 0$. Together, these statements imply that the dual residual $\nabla f(x^k) + T^T y^k$ of (2) converges to zero.

To show that the primal residual r^k converges to zero, we first show that $s^k \to 0$. If Step $2a$ in Algorithm 1 is executed infinitely often, $s^k \to 0$ since it satisfies (29) at every iteration and $\eta \in (0, 1)$. If Step $2a$ is executed finitely often, there is k_0 after which $\lambda^k = \lambda^{k_0}$. By adding and subtracting $2\mu^k \nabla f(x^k) + T^T s^k + 4\mu^k T^T(\lambda^{k_0} - y^k)$ and rearranging terms, we can write

$$T^T s^k = 2\mu^k(\nabla f(x^k) + \frac{1}{\mu^k} T^T(s^k + \mu^k(2\lambda^{k_0} - y^k))) - 2\mu^k \nabla f(x^k) - T^T(s^k + 2\mu^k(\lambda^{k_0} - y^k)) - 2\mu^k T^T \lambda^{k_0}.$$

Taking the norm of each side and applying the triangle inequality, (30) and (27) yields

$$\|T^T s^k\| \ \leq \ 2\mu^k \epsilon^k + 2\mu^k \|\nabla f(x^k)\| + \|T^T\|\epsilon_k + 2\mu^k \|T^T \lambda^{k_0}\|. \tag{28}$$

This inequality implies that $T^T s^k \to 0$ because $\nabla f(x^k)$ is bounded, $\epsilon^k \to 0$, and $\mu^k \to 0$. Since T has full row rank, T^T has full column rank and it follows that $s^k \to 0$.

Substituting $s^k \to 0$ and $\nabla f(x^k) + T^T y^k \to 0$ into the first row of (27) and applying (30) implies $\lambda^k \to y^k$. Thus, $s^k \to r^k$, implying that the iterates asymptotically drive the primal residual r^k to zero, thereby completing the proof. ■

Remark 6: Despite the assumption that $\{\nabla f(x^k)\}$ is bounded, Theorem 9 can be used to ensure global convergence whenever f is strongly convex. We show in Lemma 12 in Appendix B that a bounded set $\mathcal{C}_f$ containing the optimal point can be identified *a priori*. One can thus artificially bound $\nabla f(x)$ for all $x \notin \mathcal{C}_f$ to satisfy the conditions of Theorem 9 and guarantee global convergence to the solution of (1).

Algorithm 1 Second order primal-dual algorithm for nonsmooth composite optimization.

---

input: Initial point $w^0 = (x^0, y^0)$, and parameters $\eta \in (0,1)$, $\beta \in (0,1)$, $\tau_a, \tau_b \in (0,1)$, $\epsilon^k \geq 0$ such that $\epsilon^k \to 0$.

initialize: Set $\lambda^0 = y^0$.

    Step 1: If

$$\|s^k\| \ \leq \ \eta \|s^{k-1}\| \tag{29}$$

        go to Step 2*a*. If not, go to Step 2*b*.

    Step 2*a*: Set

$$\mu^{k+1} \ = \ \tau_a \mu^k, \qquad \lambda^{k+1} \ = \ y^k$$

    Step 2*b*: Set

$$\mu^{k+1} \ = \ \tau_b \mu^k, \qquad \lambda^{k+1} \ = \ \lambda^k$$

    Step 3: Using a backtracking line search, perform a sequence of inner iterations to choose w^{k+1} until

$$\|\nabla \mathcal{V}_{2\mu^{k+1}}(w^{k+1}, \lambda^{k+1})\| \ \leq \ \epsilon^k \tag{30}$$

where the search direction d is obtained using (24)–(26) with

$$\sigma \ = \ 0, \qquad \frac{(\tilde{w}^k)^T \nabla \mathcal{V}_{2\mu^{k+1}}(w^k)}{\|\nabla \mathcal{V}_{2\mu^{k+1}}(w^k)\|^2} \ \leq \ -\beta, \tag{31a}$$

$$\sigma \ \in \ (0,1], \qquad\qquad\qquad \text{otherwise.} \tag{31b}$$

---

2) Asymptotic convergence rate: The invertibility of the generalized Hessian $\partial^2 \mathcal{L}_\mu(w)$ allows us to establish local convergence rates for the second order updates (12) when $\mathbf{prox}_{\mu g}$ is (strongly) semismooth.

We now show that the updates in Algorithm 1 are equivalent to the second order updates (12) as $k \to \infty$. Thus, if $\mathbf{prox}_{\mu g}$ is (strongly) semismooth, the sequence of iterates generated by Algorithm 1 converges (quadratically) superlinearly to the optimal point in some neighborhood of it.

Theorem 10: Let the conditions of Theorem 9 hold, let $\mathbf{prox}_{\mu g}$ be (strongly) semismooth, and let ϵ^k be such that $\|w^k - w^\star\| = O(\epsilon^k)$. Then, in a neighborhood of the optimal point $w^\star$, the iterates w^k converge (quadratically) superlinearly to $w^\star$.

Proof: From Theorem 9, $\lambda^k \to y^k$ and thus $\nabla \mathcal{V}_{2\mu}(w^k) \to \nabla \mathcal{L}_\mu(w^k)$. Descent of the Lyapunov function in Theorem 4 therefore implies that the update in Step 3 of Algorithm 1 is given by (31a), which is equivalent to (12) because $\lambda = y$. The assumption on $\{\epsilon^k\}$ in conjunction with Proposition 3, and [50, Theorem 3.2] imply that this update asymptotically satisfies (30) in one iteration with a unit step size. Therefore, Step 3 reduces to (12) in some neighborhood of the optimal solution and Proposition 3 implies that w^k converges to $w^\star$ (quadratically) superlinearly. ∎

3) Efficient computation of the Newton direction: When g is (block) separable, the matrix P in (11) is (block) diagonal. We next demonstrate that the solution to (26) can be efficiently computed when $T = I$ and P is a sparse diagonal matrix whose entries are either 0 or 1. The extensions to a low rank P, to a P with entries between 0 and 1, or to a general diagonal T follow from similar arguments.

These conditions occur, for example, when $g(z) = \gamma \|z\|_1$. The matrix P is sparse when $\mathbf{prox}_{\mu g}(x + \mu y) = \mathcal{S}_{\gamma\mu}(x + \mu y)$ is sparse. Larger values of γ are more likely to produce a sequence of iterates w^k for which P is sparse and thus the second order search directions (26) are cheaper to compute.

We can write (26) as

$$\begin{bmatrix} H & I \\ I - P & -\mu P \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} = \begin{bmatrix} \vartheta \\ \theta \end{bmatrix}, \tag{32a}$$

permute it according to the entries of P which are 1 and 0, respectively, and partition the matrices H, P, and $I - P$ conformably such that

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix}, \quad P = \begin{bmatrix} I & \\ & 0 \end{bmatrix}.$$

Let v denote either the primal variable x or the dual variable y. We use v_1 to denote the subvector of v corresponding to the entries of P which are equal to 1 and v_2 to denote the subvector corresponding to the zero diagonal entries of P.

Note that $(I - P)v = 0$ when $v_2 = 0$ and $Pv = 0$ when $v_1 = 0$. As a result, $\tilde{x}_2$ and $\tilde{y}_1$ are explicitly determined by the bottom row of the system of equations (32a),

$$\begin{bmatrix} 0 & -\mu I \\ I & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_2 \\ \tilde{y}_1 \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \tag{32b}$$

Substitution of the subvectors $\tilde{x}_2$ and $\tilde{y}_1$ into (32a) yields,

$$H_{11}\tilde{x}_1 = \vartheta_1 + H_{12}\tilde{x}_2 + \tilde{y}_1 \tag{32c}$$

which must be solved for $\tilde{x}_1$. Finally, the computation of $\tilde{y}_2$ requires only matrix-vector products,

$$\tilde{y}_2 = -(\vartheta_2 + H_{21}\tilde{x}_1 + H_{22}\tilde{x}_2). \tag{32d}$$

Thus, the major computational burden in solving (26) lies in performing a Cholesky factorization to solve (32c), where H_{11} is a matrix of a much smaller size than H.
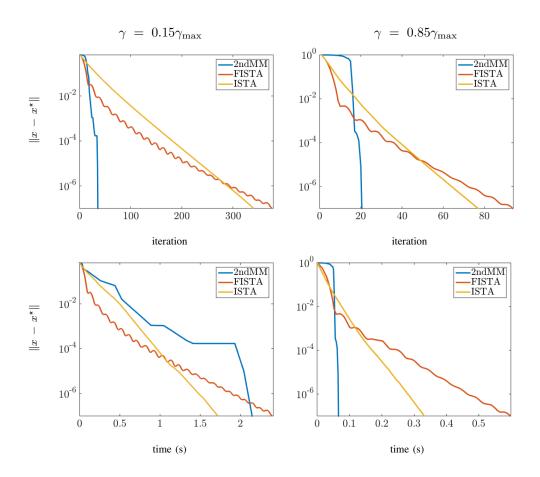
Fig. 1: Distance from optimal solution as a function of the iteration number and solve time when solving LASSO for two values of γ using ISTA, FISTA, and our algorithm (2ndMM).

VI. COMPUTATIONAL EXPERIMENTS

In this section, we illustrate the merits and the effectiveness of our approach. We first apply our algorithm to the ℓ_1-regularized least squares problem and then study a system theoretic problem of controlling a spatially-invariant system.

A. ℓ_1-regularized least squares

The LASSO problem (33) regularizes a least squares objective with a γ-weighted ℓ_1 penalty,

$$\underset{x}{\text{minimize}} \quad \tfrac{1}{2} \left\| Ax - b \right\|^2 + \gamma \left\| x \right\|_1. \tag{33}$$

As described in Section II-A1, the associated proximal operator is given by soft-thresholding $\mathcal{S}_{\gamma\mu}$, the Moreau envelope is the Huber function, and its gradient is the saturation function. Thus, $P \in \mathbb{P}$ is diagonal and P_{ii} is 0 when $|x_i + \mu y_i| < \gamma\mu$, 1 outside this interval, and between 0 and 1 on the boundary. Larger values of the regularization parameter γ induce sparser solutions for which one can expect a sparser sequence of iterates. Note that we require strong convexity of the least squares penalty; i.e., that $A^T A$ is positive definite.
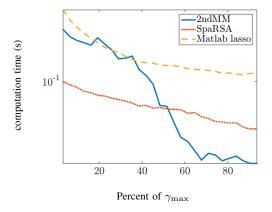
Fig. 2: Solve times for LASSO with $n = 1000$ obtained using ISTA, FISTA, and our algorithm (2ndMM) as a function of the sparsity-promoting parameter γ.
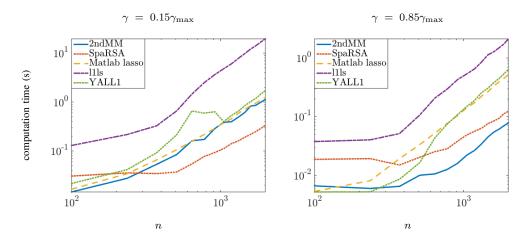

Fig. 3: Comparison of our algorithm (2ndMM) with state-of-the-art methods for LASSO with problem dimension varying from $n = 100$ to 2000.

In Fig. 1, we show the distance of the iterates from the optimal for the standard proximal gradient algorithm ISTA, its accelerated version FISTA, and our customized second order primal-dual algorithm for a problem where $A^T A$ has condition number 3.26×10^4. We plot distance from the optimal point as a function of both iteration number and solve time. Although our method always requires much fewer iterations, it is most effective when γ is large. In this case the most computationally demanding step (32c) required to determine the second order search direction (26) involves a smaller matrix inversion; see Section V-B3 for details. In Fig. 2, we show the solve times for $n = 1000$ as the sparsity-promoting parameter γ ranges from 0 to $\gamma_{\max} = \|A^T b\|$, where $\gamma_{\max}$ yields a zero solution. All numerical experiments consist of 20 averaged trials.

In Fig. 3, we compare the performance of our algorithm with the LASSO function in Matlab (a coordinate descent method [51]), SpaRSA [52], an interior point method [53], and YALL1 [54]. Problem instances were randomly generated with $A \in \mathbb{R}^{m \times n}$, n ranging from 100 to 2000, $m = 3n$, and $\gamma = 0.15\gamma_{\max}$ or $0.85\gamma_{\max}$. The solve times

and scaling of our algorithm is competitive with these state-of-the-art methods. For larger values of γ, the second order search direction (26) is cheaper to compute and our algorithm is the fastest.

B. Distributed control of a spatially-invariant system

We now apply our algorithm to a structured control design problem aimed at balancing closed-loop $\mathcal{H}_2$ performance with spatial support of a state-feedback controller. Following the problem formulation of [5], ADMM was used in [32], [55] to design sparse feedback gains for spatially-invariant systems. Herein, we demonstrate that our algorithm provides significant computational advantage over both the ADMM algorithm and a proximal Newton scheme.

1) Spatially-invariant systems: Let us consider

$$
\begin{aligned}
\dot{\psi} &= A\,\psi \;+\; u \;+\; d \\
\zeta &= \left[\begin{array}{c} Q^{1/2}\,\psi \\ R^{1/2}\,u \end{array} \right]
\end{aligned}
\tag{34}
$$

where ψ, u, d, and ζ are the system state, control input, white stochastic disturbance, and performance output and A, $Q \succeq 0$, and $R \succ 0$ are $n \times n$ circulant matrices. Such systems evolve over a discrete spatially-periodic domain; they can be used to model spatially-invariant vehicular platoons [56] and can result from a spatial discretization of fluid flows [57].

Any circulant matrix can be diagonalized via the discrete Fourier transform (DFT). Thus, the coordinate transformation $\psi := T\hat{\psi}$, $u := T\hat{u}$, $d := T\hat{d}$, where T^{-1} is the DFT matrix, brings the state equation in (34) into,

$$
\dot{\hat{\psi}} \;=\; \hat{A}\,\hat{\psi} \;+\; \hat{u} \;+\; \hat{d}.
\tag{35}
$$

Here, $\hat{A} := T^{-1}AT$ is a diagonal matrix whose main diagonal $\hat{a}$ is determined by the DFT of the first row a of the matrix A,

$$
\hat{a}_k \;:=\; \sum_{i=0}^{n-1} a_i\, \mathrm{e}^{-\mathrm{j}\frac{2\pi i k}{n}}, \quad k \in \{0,\ldots,n-1\}.
$$

We are interested in designing a structured state-feedback controller, $u = -Z\psi$, that minimizes the closed-loop $\mathcal{H}_2$ norm, i.e., the variance amplification from the disturbance d to the regulated output ζ. Since the optimal unstructured Z for spatially-invariant system (34) is a circulant matrix [58], we restrict our attention to circulant feedback gains Z. Thus, Z can also be diagonalized via a DFT and we equivalently take $x := \hat{z}$ as our optimization variable where $T^{-1}ZT = \mathrm{diag}\,(\hat{z})$.

For simplicity, we assume that A and Z are symmetric. In this case, $\hat{a}$ and $\hat{z}$ are real vectors and the closed-loop $\mathcal{H}_2$ norm of system (34) takes separable form $f(x) = \sum_{k=0}^{n-1} f_k(x_k)$,

$$
f_k(x_k) \;=\; \left\{ \begin{array}{ll} \dfrac{\hat{q}_k \;+\; \hat{r}_k x_k^2}{2\,(x_k \;-\; \hat{a}_k)}, & x_k \;>\; \hat{a}_k \\[2ex] \infty, & \text{otherwise} \end{array} \right.
$$

where $x_k > \hat{a}_k$ guarantees closed-loop stability. To promote sparsity of Z, we consider a regularized optimization

problem (2),

$$
\begin{aligned}
\underset{x,\,z}{\text{minimize}} \quad & f(x) \; + \; \gamma \, \|z\|_1 \\
\text{subject to} \quad & Tx \, - \, z \; = \; 0
\end{aligned}
\tag{36}
$$

where γ is a positive regularization parameter, T is the inverse DFT matrix, and $z \in \mathbb{R}^n$ denotes the first row of the symmetric circulant matrix Z. Formulation (36) signifies that while it is convenient to quantify the $\mathcal{H}_2$ norm in the spatial frequency domain, sparsity has to be promoted in the physical space.

By solving (36) over a range of γ, we identify distributed controller structures which are specified by the sparsity pattern of the solutions $z_\gamma^\star$ to (36) at different values of γ. After selecting a controller structure associated with a particular value of γ, we solve a 'polishing' or 'debiasing' problem,

$$
\begin{aligned}
\underset{x,z}{\text{minimize}} \quad & f(x) \; + \; I_{\mathrm{sp}(z_\gamma^\star)}(z) \\
\text{subject to} \quad & Tx \, - \, z \; = \; 0
\end{aligned}
\tag{37}
$$

where $I_{\mathrm{sp}(z_\gamma^\star)}(z)$ is the indicator function associated with the sparsity pattern of $z_\gamma^\star$. The solution to this problem is the optimal controller for system (34) with the desired structure, i.e., the same sparsity pattern as $z_\gamma^\star$. This step is necessary because the ℓ_1 norm in (36) imposes an additional penalty on z that compromises closed-loop performance.

2) Implementation: The elements of the gradient of f are

$$
\frac{\mathrm{d}f_k(x_k)}{\mathrm{d}x_k} \;=\; \frac{\hat{r}_k x_k^2 \,-\, 2\hat{a}_k \hat{r}_k x_k \,-\, \hat{q}_k}{2\,(x_k \,-\, \hat{a}_k)^2}
$$

the Hessian is a diagonal matrix with non-zero entries,

$$
\frac{\mathrm{d}^2 f_k(x_k)}{\mathrm{d}x_k^2} \;=\; \frac{\hat{q}_k \,+\, \hat{r}_k \hat{a}_k^2}{(x_k \,-\, \hat{a}_k)^3}
$$

and the proximal operator associated with the nonsmooth regularizer in (36) is given by soft-thresholding $\mathcal{S}_{\gamma\mu}$.

While the optimal unstructured controller can be obtained by solving n uncoupled scalar quadratic equations for x_k, sparsity-promoting problem (36) is not in a separable form (because of the linear constraint) and computing the second order update (12) requires solving a system of equations

$$
\begin{bmatrix} H & T^* \\ (I - P)T & -\mu P \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \;=\; \begin{bmatrix} \vartheta \\ \theta \end{bmatrix}
\tag{38}
$$

Pre-multiplying by the nonsingular matrix $\mathrm{blkdiag}(T, I)$ and changing variables to solve for $\tilde{z} := T\tilde{x}$ brings (12) into

$$
\begin{bmatrix} T H T^{-1} & (1/n)\,I \\ I - P & -\mu P \end{bmatrix} \begin{bmatrix} \tilde{z} \\ \tilde{y} \end{bmatrix} \;=\; \begin{bmatrix} T\vartheta \\ \theta \end{bmatrix}
$$

which is of the same form as (32a). This equation can be solved efficiently when P is sparse (i.e., $\gamma\mu$ is large; cf. Section V-B3) and $\tilde{x}$ can be recovered from $\tilde{z}$ via FFT.

Since the Hessian H of a separable function f is a diagonal matrix, the search direction can also be efficiently computed when $I - P$ is sparse (i.e., $\gamma\mu$ is small). As in Section V-B3, the component of $\tilde{y}$ in the support of P is determined from the bottom row of (38). The top row of (38) implies $\tilde{x} = H^{-1}(\vartheta - T^*\tilde{y})$ and substitution into

the bottom row yields

$$(I - P) T H^{-1} T^* (I - P) \tilde{y} \ = \ \tilde{\theta}$$

where $\tilde{\theta} := (I - P)(TH^{-1}(\vartheta - T^*P\tilde{y}) - \theta)$ is a known vector. Thus, the component of $\tilde{y}$ in the support of $I - P$ can be determined by inverting a matrix whose size is determined by the support of $I - P$ and $\tilde{x}$ is readily obtained from $\tilde{y}$ and ϑ. The operations involving T and T^* can be performed via FFT; since H is diagonal, multiplication by these matrices is cheap and the computational burden in solving (12) again arises from a limited matrix inversion. In contrast to Section V-B3, the computation of the search direction using this approach is efficient when $I - P$ is sparse, i.e., $\gamma\mu$ is small.

3) Swift-Hohenberg equation: We consider the linearized Swift-Hohenberg equation [59],

$$\partial_t \psi(t,\xi) \ = \ (cI \ - \ (I \ + \ \partial_{\xi\xi})^2) \, \psi(t,\xi) \ + \ u(t,\xi) \ + \ d(t,\xi)$$

with periodic boundary conditions on a spatial domain $\xi \in [-\pi, \pi]$. Finite-dimensional approximation and diagonalization via the DFT (with an even number of Fourier modes n) yields (35) with $\hat{a}_k = c - (1 - k^2)^2$ where $k = \{-n/2 + 1, \ldots, n/2\}$ is a spatial wavenumber.

Figure 4a shows the optimal centralized controller and solutions to (36) for $c = -0.01$, $n = 64$, $Q = R = I$, and $\gamma = 4 \times 10^{-4}$, 4×10^{-3}, and 4. As further illustrated in Fig. 4b, the optimal solutions to (36) become sparser as γ is increased.

In Fig. 5, we demonstrate the utility of using regularized problems to navigate the tradeoff between controller performance and structure, an approach pioneered by [5]. The polished optimal structured controllers (—∘—) were designed by first solving (36) to identify an optimal structure and then solving (37) to further improve the closed-loop performance. To illustrate the importance of polishing step (37), we also show the closed-loop performance of unpolished optimal structured controllers (- -×- -) resulting from (36). Finally, to evaluate the controller *structures* identified by (36), we show the closed-loop performance of polished 'reference' structured controllers ($\cdots + \cdots$). Instead of solving (36), the reference structures are *a priori* specified as nearest neighbor symmetric controllers of the same cardinality as controllers resulting from (36). Among controllers with the same number of nonzero entries, the polished optimal structured controller consistently achieves the best closed-loop performance.

We compare the computational efficiency of our approach with the proximal Newton method [19] and ADMM [32], [55]. The proximal Newton method requires solving a LASSO subproblem (7), for which we employ SpaRSA [52]. Since A is circulant, the x-minimization step in ADMM (6) requires solving n uncoupled cubic scalar equations. In general, when A is block circulant, the DFT only block diagonalizes the dynamics and thus the x-minimization step has to be solved via an iterative procedure [32], [55].

Figure 6 shows the time to solve (36) with $\gamma = 0.004$ using our method, proximal Newton, and ADMM. Our algorithm and ADMM were stopped when the primal and dual residuals were below 1×10^{-8}. The proximal Newton method was stopped when the norm of the difference between two consecutive iterates was smaller than 1×10^{-8}. In Fig. 7, we show the per iteration cost and the total number of iterations required to find the optimal solution using each method. Our algorithm clearly outperforms proximal Newton and ADMM.

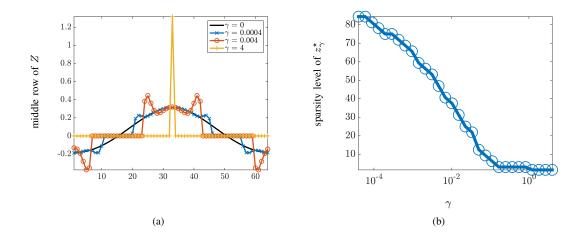Although proximal Newton requires a similar number of iterations, the LASSO subproblem (7) that determines its

(a)   (b)

Fig. 4: (a) The middle row of the circulant feedback gain matrix Z; and (b) the sparsity level of $z_\gamma^\star$ (relative to the sparsity level of the optimal centralized controller $z_0^\star$) resulting from the solutions to (36) for the linearized Swift-Hohenberg equation with $n = 64$ Fourier modes and $c = -0.01$.
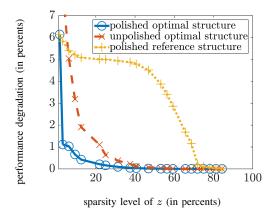

Fig. 5: Performance degradation (in percents) of structured controllers relative to the optimal centralized controller: polished optimal structured controller obtained by solving (36) and (37) (—∘—); unpolished optimal structured controller obtained by solving only (36) (- -×- -); and optimal structured controller obtained by solving (37) for an *a priori* specified nearest neighbor reference structure (···+···).

search direction is much more expensive; this increases the computation cost of each iteration and slows the overall algorithm. Moreover, for larger problem sizes, the proximal Newton method struggles with finding a stabilizing search direction because $\nabla^2 f$ seems to bring it away from the set of stabilizing feedback gains. It appears that our method circumvents this issue because its iterates lie in a larger lifted space in which stability is easier to enforce via backtracking.

On the other hand, while the x- and z- minimization steps in ADMM are quite efficient, as a first order method ADMM requires a large number of iterations to reach high-accuracy solutions. Our algorithm achieves better performance because its use of second order information leads to relatively few iterations and the structured matrix inversion leads to efficient computation of the search direction.
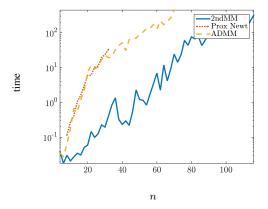
Fig. 6: Total time to compute the solution to (36) with $\gamma = 0.004$ using our algorithm (2ndMM), proximal Newton, and ADMM.
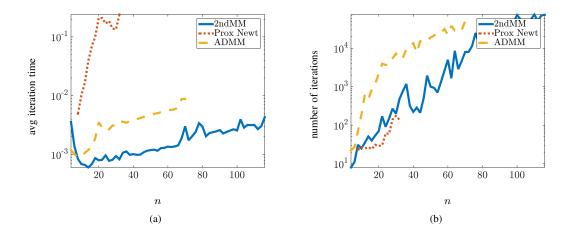

Fig. 7: Comparison of (a) times to compute an iteration (averaged over all iterations); and (b) numbers of iterations required to solve (36) with $\gamma = 0.004$.

VII. CONNECTIONS AND DISCUSSION

The proximal augmented Lagrangian $\mathcal{L}_\mu(x; y)$ is obtained by constraining $\mathcal{L}_\mu(x, z; y)$ to the manifold

$$\mathcal{Z} := \{(x, z_\mu^\star; y) \mid z_\mu^\star = \underset{z}{\operatorname{argmin}} \ \mathcal{L}_\mu(x, z; y)\}$$
$$= \{(x, z_\mu^\star; y) \mid Tx + \mu y \in z_\mu^\star + \mu \, \partial_C g(z_\mu^\star)\}$$

which results from the explicit minimization over the auxiliary variable z. Herein, we interpret the second order search direction as a linearized update to the KKT conditions for problem (2) and discuss connections to the alternative algorithms.

A. Second order updates as linearized KKT corrections

The second order update (12) can be viewed as a first order correction to the KKT conditions for optimization problem (2),

$$
\begin{aligned}
0 &= \nabla f(x) + T^T y \\
0 &= Tx - z \\
0 &\in \partial_C g(z) - y.
\end{aligned}
\tag{39}
$$

Substitution of $z_\mu^\star$ into (39) makes the last two conditions redundant; when combined with the definition of the manifold $\mathcal{Z}$, $Tx = z$ implies $y \in \partial_C g(z)$ and $y \in \partial_C g(z)$ implies $Tx = z$.

Multiplying equation (12) for the second order update with the nonsingular matrix (25) recovers the equivalent expression

$$
\begin{bmatrix} H & T^T \\ (I - P)T & -\mu P \end{bmatrix} \begin{bmatrix} \tilde{x}^k \\ \tilde{y}^k \end{bmatrix} = - \begin{bmatrix} \nabla f(x^k) + T^T y^k \\ r^k \end{bmatrix}
\tag{40}
$$

where $r^k := Tx^k - z_\mu^\star(x^k; y^k) = Tx^k - \mathbf{prox}_{\mu g}(Tx^k + \mu y^k)$ is the primal residual of (2). Thus, (40) describes a first order correction to the first and second KKT conditions in (39).

B. Connections with other methods

We now discuss broader implications of our framework and draw connections to the existing methods for solving versions of (1). Many techniques for solving composite minimization problems of the form (1) can be expressed in terms of functions embedded in the augmented Lagrangian; see Table I. Trivially, the objective function in (1) corresponds to $\mathcal{L}_\mu(x, z; y)$ over the manifold $z = Tx$. The Lagrange dual of a problem equivalent to (2),

$$
\begin{aligned}
\underset{x, z}{\text{minimize}} \quad & f(x) + g(z) + \tfrac{1}{2\mu} \|Tx - z\|^2 \\
\text{subject to} \quad & Tx - z = 0
\end{aligned}
\tag{41}
$$

is recovered by collapsing $\mathcal{L}_\mu(x, z; y)$ onto the intersection of the z-minimization manifold $\mathcal{Z}$ and the x-minimization manifold,

$$
\begin{aligned}
\mathcal{X} &:= \{(x^\star, z; y) \mid x^\star = \underset{x}{\text{argmin}} \ \mathcal{L}_\mu(x, z; y)\} \\
&= \{(x^\star, z; y) \mid \mu\left(\nabla f(x^\star) + T^T y\right) = T^T(z - Tx^\star)\}.
\end{aligned}
$$

The Lagrange dual of (2) is recovered from the Lagrange dual of (41) in the limit $\mu \to \infty$.

1) MM and ADMM: MM implements gradient ascent on the dual of (41) by collapsing $\mathcal{L}_\mu(x, z; y)$ computationally onto $\mathcal{X} \cap \mathcal{Z}$. The joint (x, z)-minimization step in (10) evaluates the Lagrange dual at discrete iterates y^k by finding the corresponding (x, z)-pair on $\mathcal{X} \cap \mathcal{Z}$; i.e., the iterate $(x^{k+1}, z^{k+1}; y^k)$ generated by (10) lies on the manifold $\mathcal{X} \cap \mathcal{Z}$. Note that, in this form, joint (x, z)-minimization is a challenging nondifferentiable optimization problem.

ADMM avoids this challenge by collapsing $\mathcal{L}_\mu(x, z; y)$ onto $\mathcal{X}$ and $\mathcal{Z}$ separately. While the underlying x- and z-minimization subproblems in ADMM are relatively simple, the iterate $(x^{k+1}, z^{k+1}; y^k)$ generated by (6) does not typically lie on the $\mathcal{X} \cap \mathcal{Z}$ manifold. Thus ADMM does not represent gradient ascent on the dual of (41), causing looser theoretical guarantees and often worse practical performance.

x	z	y	function	methods
$z = Tx$	-	-	objective function of (1)	subgradient iteration [60]
-	-	-	augmented Lagrangian	ADMM [5], [11]
$\mathcal{X}$	$\mathcal{Z}$	-	Lagrange dual of (41)	Dual ascent (if explicit expression available)
-	$\mathcal{Z}$	-	proximal augmented Lagrangian	MM [12], Arrow-Hurwicz-Uzawa [12], **2nd order primal-dual**
-	$\mathcal{Z}$	$\mathcal{Y}$	Forward-Backward Envelope	FB Truncated Newton [26], related to prox. gradient [26]

TABLE I: Summary of different functions embedded in the augmented Lagrangian of (2) and methods for solving (1) based on these functions.

By collapsing the augmented Lagrangian onto $\mathcal{Z}$, the proximal augmented Lagrangian (8) allows us to express the (x, z)-minimization step in MM as a tractable continuously differentiable problem (cf. Theorem 1). This avoids challenges associated with ADMM and it does not increase computational complexity in the x-minimization step in (10) relative to ADMM when using first order methods. We finally note that unlike the Rockafellar's *proximal method of multipliers* [61] which applies the proximal point algorithm to the primal-dual optimality conditions, our framework *reformulates* the *standard method of multipliers* and develops second order algorithm to solve nonsmooth composite optimization problems.

2) Second order methods: We identify the saddle point of $\mathcal{L}_\mu(x, z; y)$ by forming second order updates to x and y along the $\mathcal{Z}$ manifold. Just as Newton's method approximates an objective function with a convex quadratic function, we approximate $\mathcal{L}_\mu(x; y)$ with a quadratic saddle function.

Constraining the dual variable itself yields connections with other methods. When $T = I$, (39) implies that the optimal dual variable is given by $y^\star = -\nabla f(x^\star)$, so it is natural to collapse the augmented Lagrangian onto the manifold

$$\mathcal{Y} := \{(x, z; y^\star) \mid y^\star = -\nabla f(x)\}.$$

The augmented Lagrangian over the manifold $\mathcal{Z} \cap \mathcal{Y}$ corresponds to the Forward-Backward Envelope (FBE) introduced in [26]. The proximal gradient algorithm with step size μ can be recovered from a variable metric gradient iteration on the FBE [26]. In [26]–[28], the approximate line search and quasi-Newton methods based on the FBE were developed to solve (1) with $T = I$. Since the Hessian of the FBE involves third order derivatives of f, these techniques employ either truncated- or quasi-Newton methods.

The approach advanced in the current paper applies a second order method to the augmented Lagrangian that is constrained over the larger manifold $\mathcal{Z}$. Relative to alternatives, our framework offers several advantages. First, while the FBE is in general nonconvex function of x, $\mathcal{L}_\mu(x; y)$ is always convex in x and concave in y. Furthermore, we can compute the Hessian exactly using only second order derivatives of f and its structure allows for efficient computation of the search direction. Finally, our formulation allows us to leverage recent advances in second order methods for augmented Lagrangian methods, e.g., [13]–[15].

VIII. Concluding remarks

We have developed a second order primal-dual method for nonsmooth convex composite optimization. We establish global exponential convergence of the corresponding differential inclusion in continuous-time and use the

primal-dual augmented Lagrangian as a merit function in a discrete-time implementation. Our globally convergent customized algorithm exhibits asymptotic (quadratic) superlinear convergence rate when the proximal operator associated with nonsmooth regularizer is (strongly) semismooth. We use the ℓ_1-regularized least squares and spatially-invariant control problems to demonstrate competitive performance of our algorithm relative to the available state-of-the-art alternatives.

Future research will focus on nonconvex problems, the development of inexact second order methods, and application of our framework to problems in which the codomain of the regularizer has a larger dimensions than the domain of the optimization variable. In particular, inexact but structured approximations of the Hessian can lead to efficient methods that may even be convenient for distributed implementation.

APPENDIX

A. Algorithm based on $V(w)$ as a merit function

Since Theorem 7 establishes global convergence of the differential inclusion, one algorithmic approach is to implement a Forward Euler discretization of differential inclusion (14). A natural choice of merit function is the Lyapunov function V defined in (15). A simple corollary of Theorem 4 shows that the second order update (12) is a descent direction for V.

Corollary 11: The second order update (12) is a descent direction for the merit function V defined in (15).

Proof: Follows from (16) in Theorem 4. ∎

Corollary 11 enables the use of a backtracking Armijo rule for step size selection. A natural choice of stopping criterion for such an algorithm is a condition on the size of the primal and dual residuals. Moreover, proposition 3 suggests fast asymptotic convergence when $\mathbf{prox}_{\mu g}$ is semismooth. The LASSO example in Fig. 8 verifies this intuition when solving LASSO to a threshold of 1×10^{-8} for the primal and dual residuals.

Algorithm 2 Second order primal-dual algorithm for nonsmooth composite optimization based on discretizing (14).

input: Initial point x_0, y_0, backtracking constant $\alpha \in (0,1)$, Armijo parameter $\sigma \in (0,1)$, and stopping tolerances ε_1, ε_2.

While: $\|Tx^k - \mathbf{prox}_{\mu g}(Tx^k + \mu y^k)\| > \varepsilon_1$ or
$\qquad \|\nabla f(x^k) - T^T y^k\| > \varepsilon_2$

    Step 1: Compute $\tilde{w}^k$ as defined in (12)
    Step 2: Choose the smallest $j \in \mathbb{Z}_+$ such that
$$V(w^k + \alpha^j \tilde{w}^k) \leq V(w^k) - \sigma \alpha^j \|\nabla \mathcal{L}_\mu(w^k)\|^2$$
    Step 3: Update the primal and dual variables
$$w^{k+1} = w^k + \alpha^j \tilde{w}^k$$

However, such an implementation would require a fixed penalty parameter μ, which typically has a large effect on the convergence speed of augmented Lagrangian algorithms and is difficult to select *a priori*. Moreover, stability of the solution to a differential equation does not always imply stability of its discretization.

An example: We implement Algorithm 2 to solve the LASSO problem (33) studied in Section VI. The LASSO problem was randomly generated with $n = 500$, $m = 1000$, and $\gamma = 0.85\gamma_{\max}$. Figure 8 illustrates the quadratic asymptotic convergence of Algorithm 2 and a strong influence of μ.
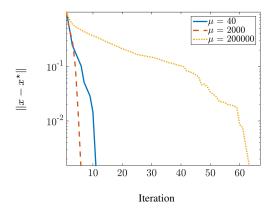
Fig. 8: Distance from the optimal solution as a function of iteration number when solving LASSO using Algorithm 2 for different values of μ.

B. Bounding $\nabla f(x)$ for Algorithm 1

A bounded set $\mathcal{C}_f$ containing the solution to (1) can always be identified from any point $\bar{x}$, $\nabla f(\bar{x})$, any element of the subgradient $\partial g(\bar{x})$, and a lower bound on the parameter of strong convexity.

Lemma 12: Let Assumption 1 hold and let f be m_f-strongly convex. Then, for any $\bar{x}$, the optimal solution to (1) lies within the ball of radius $\frac{2}{m_f}\|\nabla f(\bar{x}) + T^T \partial g(T\bar{x})\|$ centered at $\bar{x}$.

Proof: Given any point $\bar{x}$, strong convexity of f and convexity of g imply that,

$$f(x) + g(Tx) - f(\bar{x}) - g(T\bar{x}) \geq \langle \nabla f(\bar{x}) + T^T \partial g(T\bar{x}), x - \bar{x} \rangle + \tfrac{m_f}{2}\|x - \bar{x}\|^2 \tag{42}$$

for all x where $\partial g(T\bar{x})$ is any member of the subgradient of $g(T\bar{x})$. For any x with $\|x - \bar{x}\| \geq \frac{2}{m_f}\|\nabla f(\bar{x}) + T^T \partial g(T\bar{x})\|$, the right-hand side of (42) must be nonnegative which implies that $f(x) + g(Tx) \geq f(\bar{x}) + g(T\bar{x})$ and thus that x cannot solve (1). ∎

For any strongly convex function f and convex function g, a function $\hat{f}$ can be selected such that

$$\operatorname{argmin} f(x) + g(Tx) = \operatorname{argmin} \hat{f}(x) + g(Tx).$$

Here, $\hat{f}$ is identical to f in some closed set $\mathcal{C}_f$ containing $x^\star$,

$$\hat{f}(x) := \begin{cases} f(x), & x \in \mathcal{C}_f \\ \tilde{f}(x), & x \notin \mathcal{C}_f \end{cases}$$

and $\tilde{f}(x)$ is chosen such that $\hat{f}(x)$ is convex and twice differentiable, $\nabla \hat{f}(x)$ is uniformly bounded, and $\nabla^2 \hat{f}(x) \succ 0$. Lemma 12 implies that a set $\mathcal{C}_f$ that contains the optimal solution $x^\star$ of (1) can be identified *a priori* from any given point $\bar{x}$. Since $\hat{f}(x)$ satisfies all the conditions of Theorem 9, Algorithm 1 can be used to solve (1) and, since $\mathcal{C}_f$ contains $x^\star$, its optimal solution will also solve (1).

An example: When $x \in \mathbb{R}$, $f(x) = \frac{1}{2}x^2$, and $\mathcal{C}_f = [-1, 1]$, a potential choice of $\tilde{f}(x)$ is given by,

$$\tilde{f}(x) = \begin{cases} -2x + e^{x+1} - 2.5, & x \leq -1 \\ 2x + e^{-x+1} - 2.5, & x \geq 1. \end{cases}$$

For this choice of $\tilde{f}$, the gradient of $\hat{f}$ is continuous and bounded,

$$\nabla \hat{f}(x) = \begin{cases} -2 + e^{x+1}, & x \leq -1 \\ x, & x \in [-1, 1] \\ 2 - e^{-x+1}, & x \geq 1 \end{cases}$$

and the Hessian of $\hat{f}$ is determined by

$$\nabla^2 \hat{f}(x) = \begin{cases} e^{x+1}, & x \leq -1 \\ 1, & x \in [-1, 1] \\ e^{-x+1}, & x \geq 1. \end{cases}$$

REFERENCES

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[2] T. Goldstein and S. Osher, "The split Bregman method for ℓ_1-regularized problems," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 323–343, 2009.

[3] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, no. 6, pp. 805–849, 2012.

[4] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Stat.*, pp. 199–227, 2008.

[5] F. Lin, M. Fardad, and M. R. Jovanović, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *IEEE Trans. Automat. Control*, vol. 58, no. 9, pp. 2426–2431, September 2013.

[6] M. R. Jovanović and N. K. Dhingra, "Controller architectures: tradeoffs between performance and structure," *Eur. J. Control*, vol. 30, pp. 76–91, July 2016.

[7] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM J. Numer. Anal.*, vol. 16, no. 6, pp. 964–979, 1979.

[8] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*, 2011, pp. 185–212.

[9] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 123–231, 2013.

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learning*, vol. 3, no. 1, pp. 1–124, 2011.

[12] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," *IEEE Trans. Automat. Control*, 2016, submitted; also arXiv:1610.04514.

[13] P. E. Gill and D. P. Robinson, "A primal-dual augmented Lagrangian," *Comput. Optim. Appl.*, vol. 51, no. 1, pp. 1–25, 2012.

[14] P. Armand, J. Benoist, R. Omheni, and V. Pateloup, "Study of a primal-dual algorithm for equality constrained minimization," *Comput. Optim. Appl.*, vol. 59, no. 3, pp. 405–433, 2014.

[15] P. Armand and R. Omheni, "A globally and quadratically convergent primal-dual augmented Lagrangian algorithm for equality constrained optimization," *Optim. Method. Softw.*, pp. 1–21, 2015.

[16] S. Becker and J. Fadili, "A quasi-Newton proximal splitting method," in *Adv. Neural Inf. Process. Syst.*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., 2012, pp. 2618–2626.

[17] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Math. Program.*, vol. 117, no. 1-2, pp. 387–423, 2009.

[18] R. H. Byrd, J. Nocedal, and F. Oztoprak, "An inexact successive quadratic approximation method for ℓ_1 regularized optimization," *Math. Program.*, pp. 1–22, 2015.

[19] J. D. Lee, Y. Sun, and M. A. Saunders, "Proximal Newton-type methods for minimizing composite functions," *SIAM J. Optimiz.*, vol. 24, no. 3, pp. 1420–1443, 2014.

[20] C.-J. Hsieh, I. S. Dhillon, P. K. Ravikumar, and M. A. Sustik, "Sparse inverse covariance matrix estimation using quadratic approximation," in *Adv. Neural Inf. Process. Syst.*, 2011, pp. 2330–2338.

[21] C.-J. Hsieh, M. A. Sustik, I. S. Dhillon, and P. Ravikumar, "QUIC: Quadratic approximation for sparse inverse covariance estimation," *J. Mach. Learn. Res.*, vol. 15, pp. 2911–2947, 2014.

[22] S. Hassan-Moghaddam and M. R. Jovanović, "Topology design for stochastically-forced consensus networks," *IEEE Trans. Control Netw. Syst.*, 2017, doi:10.1109/TCNS.2017.2674962.

[23] L. Qi and J. Sun, "A nonsmooth version of Newton's method," *Math. Program.*, vol. 58, no. 1, pp. 353–367, 1993.

[24] L. Qi and D. Sun, "A survey of some nonsmooth equations and smoothing Newton methods," in *Progress in Optimization*. Springer, 1999, pp. 121–146.

[25] R. Mifflin, "Semismooth and semiconvex functions in constrained optimization," *SIAM J. Control Opt.*, vol. 15, no. 6, pp. 959–972, 1977.

[26] P. Patrinos, L. Stella, and A. Bemporad, "Forward-backward truncated Newton methods for large-scale convex composite optimization," 2014, arXiv:1402.6655.

[27] L. Stella, A. Themelis, and P. Patrinos, "Forward-backward quasi-Newton methods for nonsmooth optimization problems," 2016, arXiv:1604.08096.

[28] A. Themelis, L. Stella, and P. Patrinos, "Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone line-search algorithms," 2016, arXiv:1606.06256.

[29] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparse and optimal wide-area damping control in power networks," in *Proceedings of the 2013 American Control Conference*, 2013, pp. 4295–4300.

[30] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparsity-promoting optimal wide-area control of power networks," *IEEE Trans. Power Syst.*, vol. 29, no. 5, pp. 2281–2291, September 2014.

[31] X. Wu, F. Dörfler, and M. R. Jovanović, "Input-output analysis and decentralized optimal control of inter-area oscillations in power systems," *IEEE Trans. Power Syst.*, vol. 31, no. 3, pp. 2434–2444, May 2016.

[32] X. Wu and M. R. Jovanović, "Sparsity-promoting optimal control of systems with symmetries, consensus and synchronization networks," *Syst. Control Lett.*, vol. 103, pp. 1–8, May 2017.

[33] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer Science & Business Media, 2009, vol. 317.

[34] S. M. Robinson, "Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity," in *Nonlinear Analysis and Optimization*. Springer, 1987, pp. 45–66.

[35] F. H. Clarke, *Optimization and nonsmooth analysis*. SIAM, 1990, vol. 5.

[36] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex analysis and minimization algorithms I: Fundamentals*. Springer Science & Business Media, 2013, vol. 305.

[37] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*. Oxford University Press, 1979, vol. 5.

[38] K. Jiang, D. Sun, and K.-C. Toh, "A partial proximal point algorithm for nuclear norm regularized matrix least squares problems," *Math. Program. Comp.*, vol. 6, no. 3, pp. 281–325, 2014.

[39] F. Meng, D. Sun, and G. Zhao, "Semismoothness of solutions to generalized equations and the Moreau-Yosida regularization," *Math. Programming*, vol. 104, no. 2, pp. 561–581, 2005.

[40] J. Kruskal, "Two convex counterexamples: A discontinuous envelope function and a nondifferentiable nearest-point mapping," *Proceedings of the American Mathematical Society*, vol. 23, no. 3, pp. 697–703, 1969.

[41] D. P. Bertsekas, "Projected Newton methods for optimization problems with simple constraints," *SIAM J. Control Opt.*, vol. 20, no. 2, pp. 221–246, 1982.

[42] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal. Statist. Soc B.*, vol. 58, no. 1, pp. 267–288, 1996.

[43] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. SIAM, 2000.

[44] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[45] W. Rudin, *Principles of mathematical analysis*. McGraw-Hill, Inc., 1964, vol. 3.

[46] L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *SIAM J. Optimiz.*, vol. 26, no. 1, pp. 57–95, 2016.

[47] H. Khalil, *Nonlinear Systems*. Prentice Hall, New Jersey, 2002.

[48] T. De Luca, F. Facchinei, and C. Kanzow, "A semismooth equation approach to the solution of nonlinear complementarity problems," *Math. Program.*, vol. 75, no. 3, pp. 407–439, 1996.

[49] A. R. Conn, N. I. Gould, and P. Toint, "A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds," *SIAM J. Numer. Anal.*, vol. 28, no. 2, pp. 545–572, 1991.

[50] F. Facchinei, "Minimization of SC^1 functions and the Maratos effect," *Oper. Res. Lett.*, vol. 17, no. 3, pp. 131–137, 1995.

[51] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, p. 1, 2010.

[52] S. J. Wright, R. D. Nowak, and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, 2009.

[53] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, 2007.

[54] J. Yang and Y. Zhang, "Alternating direction algorithms for ℓ_1-problems in compressive sensing," *SIAM J. Sci. Comput.*, vol. 33, no. 1, pp. 250–278, 2011.

[55] D. M. Zoltowski, N. K. Dhingra, F. Lin, and M. R. Jovanović, "Sparsity-promoting optimal control of spatially-invariant systems," in *Proceedings of the 2014 American Control Conference*, 2014, pp. 1261–1266.

[56] B. Bamieh, M. R. Jovanović, P. Mitra, and S. Patterson, "Coherence in large-scale networks: dimension dependent limitations of local feedback," *IEEE Trans. Automat. Control*, vol. 57, no. 9, pp. 2235–2249, September 2012.

[57] M. R. Jovanović and B. Bamieh, "Componentwise energy amplification in channel flows," *J. Fluid Mech.*, vol. 534, pp. 145–183, July 2005.

[58] B. Bamieh, F. Paganini, and M. A. Dahleh, "Distributed control of spatially-invariant systems," *IEEE Trans. Automat. Control*, vol. 47, no. 7, pp. 1091–1107, 2002.

[59] J. Swift and P. C. Hohenberg, "Hydrodynamic fluctuations at the convective instability," *Phys. Rev. A*, vol. 15, no. 1, p. 319, 1977.

[60] D. P. Bertsekas, *Nonlinear programming*. Athena Scientific, 1999.

[61] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM J. Control Optim.*, vol. 14, no. 5, pp. 877–898, 1976.