Technical Notes and Correspondence

The Proximal Augmented Lagrangian Method for Nonsmooth Composite Optimization

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

Abstract—**We study a class of optimization problems in which the objective function is given by the sum of a differentiable but possibly nonconvex component and a nondifferentiable convex regularization term. We introduce an auxiliary variable to separate the objective function components and utilize the Moreau envelope of the regularization term to derive the proximal augmented Lagrangian— a continuously differentiable function obtained by constraining the augmented Lagrangian to the manifold that corresponds to the explicit minimization over the variable in the nonsmooth term. The continuous differentiability of this function with respect to both primal and dual variables allows us to leverage the method of multipliers (MM) to compute optimal primal-dual pairs by solving a sequence of differentiable problems. The MM algorithm is applicable to a broader class of problems than proximal gradient methods and it has stronger convergence guarantees and a more refined step-size update rules than the alternating direction method of multipliers (ADMM). These features make it an attractive option for solving structured optimal control problems. We also develop an algorithm based on the primal-descent dual-ascent gradient method and prove global (exponential) asymptotic stability when the differentiable component of the objective function is (strongly) convex and the regularization term is convex. Finally, we identify classes of problems for which the primal-dual gradient flow dynamics are convenient for distributed implementation and compare/contrast our framework to the existing approaches.**

Index Terms—**Augmented Lagrangian, control for optimization, global exponential stability, method of multipliers, non-smooth optimization, primal-dual dynamics, proximal algorithms, proximal augmented Lagrangian, regularization for design, structured optimal control.**

I. Introduction

We study a class of composite optimization problems in which the objective function is the sum of a differentiable but possibly a nonconvex component and a convex nondifferentiable component. Problems of this form are encountered in diverse fields including compressive sensing [1], statistics [2], and control [3]. In feedback synthesis, they typically arise when a traditional performance metric (such as the $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm) is augmented with a regularization function to promote certain structural properties in the optimal controller. For example, the ℓ_1 norm and the nuclear norm are commonly used nonsmooth convex regularizers that encourage sparse and low-rank optimal solutions, respectively.

The lack of a differentiable objective function precludes the use of standard descent methods for smooth optimization. Proximal gradient methods [4] and their accelerated variants [5] generalize gradient descent, but typically require the nonsmooth term to be separable over the optimization variable. An alternative approach is to split the smooth and nonsmooth components in the objective function over separate variables that are coupled via an equality constraint. Such a reformulation facilitates the use of the alternating direction method of multipliers (ADMM) [6]. This augmented-Lagrangian-based method splits the optimization problem into subproblems that are either smooth or easy to solve. It also allows for a broader class of regularizers than the proximal gradient and it is convenient for distributed implementation. However, there are limited convergence guarantees for nonconvex problems and parameter tuning greatly affects its convergence rate.

The method of multipliers (MM) is the most widely used algorithm for solving constrained nonlinear programming problems [7]–[9]. In contrast to ADMM, it is guaranteed to converge for nonconvex problems and there are systematic ways to adjust algorithmic parameters. However, MM is not a splitting method and it requires *joint* minimization of the augmented Lagrangian with respect to *all* primal optimization variables. This subproblem is typically nonsmooth and as difficult to solve as the original optimization problem.

To make this difficult subproblem tractable, we transform the augmented Lagrangian into the continuously differentiable proximal augmented Lagrangian by exploiting the structure of proximal operators associated with nonsmooth regularizers. This new form is obtained by constraining the augmented Lagrangian to the manifold that corresponds to the explicit minimization over the variable in the nonsmooth term. The resulting expression is given in terms of the Moreau envelope of the nonsmooth regularizer and is continuously differentiable. This allows us to take advantage of standard optimization tools, including gradient descent and quasi-Newton methods, and enjoy the convergence guarantees of the standard MM.

The proximal augmented Lagrangian also enables Arrow–Hurwicz–Uzawa primal-dual gradient flow dynamics. Such dynamics can be used to identify saddle points of the Lagrangian [11] and have enjoyed recent renewed interest in the context of networked optimization because, in many cases, the gradient can be computed in a distributed manner [12]. Our approach yields a dynamical system with a continuous right-hand side for a broad class of nonsmooth optimization problems. This is in contrast to existing techniques that employ subgradient methods [13] or use discontinuous projected dynamics [14]–[16] to handle inequality constraints. Furthermore, since the proximal augmented Lagrangian is not strictly convex–concave, we make additional developments relative to [17] to show asymptotic convergence. Finally, inspired by

recent advances [18], [19], we employ the theory of integral quadratic constraints (IQCs) [20] to prove global exponential stability when the differentiable component of the objective function is strongly convex with a Lipschitz continuous gradient.

The rest of this paper is structured as follows. In Section II, we formulate the nonsmooth composite optimization problem and provide a brief background on proximal operators. In Section III, we exploit the structure of proximal operators to introduce the proximal augmented Lagrangian and provide an efficient algorithmic implementation of MM. In Section IV, we prove global (exponential) asymptotic stability of primal-descent dual-ascent gradient flow dynamics under a (strong) convexity assumption. In Section V, we use the problems of edge addition in directed consensus networks and optimal placement to illustrate the effectiveness of our approach. We close this paper in Section VI with concluding remarks.

II. PROBLEM FORMULATION AND BACKGROUND

We consider a composite optimization problem

$$\underset{x}{\text{minimize}} \quad f(x) + g\left(\mathcal{T}(x)\right) \tag{1}$$

where the optimization variable x belongs to a finite-dimensional Hilbert space (e.g., $\mathbb{R}^n$ or $\mathbb{R}^{m \times n}$) equipped with an inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. The function f is continuously differentiable but possibly nonconvex, the function g is convex but potentially nondifferentiable, and $\mathcal{T}$ is a bounded linear operator. We further assume that g is proper and lower semicontinuous, that (1) is feasible, and that its minimum is finite.

Problem (1) is often encountered in structured controller design [21]–[23], where f is a measure of the closed-loop performance, e.g., the $\mathcal{H}_2$ norm, and the regularization term g is introduced to promote certain structural properties of $\mathcal{T}(x)$. For example, in wide-area control of power systems, f measures the quality of synchronization between different generators and g penalizes the amount of communication between them [24]–[26].

In particular, for $z := \mathcal{T}(x) \in \mathbb{R}^m$, the ℓ_1 norm, $\|z\|_1 := \sum |z_i|$, is a commonly used convex proxy for promoting sparsity of z. For $z \in \mathbb{R}^{m \times n}$, the nuclear norm, $\|z\|_* := \sum \sigma_i(z)$, can be used to obtain low-rank solutions to (1), where $\sigma_i(z)$ is the ith singular value. The indicator function, $I_\mathcal{C}(z) := \{0, \ z \in \mathcal{C}; \ \infty, \ z \notin \mathcal{C}\}$ associated with the convex set $\mathcal{C}$ is the proper regularizer for enforcing $z \in \mathcal{C}$.

Regularization of $\mathcal{T}(x)$ instead of x is important in the situation where the desired structure has a simple characterization in the codomain of $\mathcal{T}$. For example, such problems arise in spatially invariant systems, where it is convenient to perform standard control design in the spatial frequency domain [27] but necessary to promote structure in the physical space, and in consensus/synchronization networks, where the objective function is expressed in terms of the deviation of node values from the network average but it is desired to impose structure on the network edge weights [22], [23].

A. Background on Proximal Operators

Problem (1) is difficult to solve directly because f is, in general, a nonconvex function and g is typically not differentiable. Since existing approaches and our method utilize proximal operators, we first provide a brief overview; for additional information, see [4].

The proximal operator of the function g is given by

$$\mathbf{prox}_{\mu g}(v) := \underset{x}{\text{argmin}} \left(g(x) + \frac{1}{2\mu} \|x - v\|^2 \right) \tag{2a}$$

and the associated optimal value specifies its Moreau envelope

$$M_{\mu g}(v) := g(\mathbf{prox}_{\mu g}(v)) + \frac{1}{2\mu} \|\mathbf{prox}_{\mu g}(v) - v\|^2 \tag{2b}$$

where $\mu > 0$. The Moreau envelope is a continuously differentiable function, even when g is not, and its gradient [4] is given by

$$\nabla M_{\mu g}(v) = \frac{1}{\mu} \left(v - \mathbf{prox}_{\mu g}(v) \right). \tag{2c}$$

For example, when g is the ℓ_1 norm, $g(z) = \|z\|_1 = \sum |z_i|$, the proximal operator is determined by soft thresholding, $\mathbf{prox}_{\mu g}(v_i) = \mathcal{S}_\mu(v_i) := \text{sign}(v_i) \max\left(|v_i| - \mu, 0\right)$, the associated Moreau envelope is the Huber function, $M_{\mu g}(v_i) = \{\frac{1}{2\mu} v_i^2, \ |v_i| \le \mu; \ |v_i| - \frac{\mu}{2}, \ |v_i| \ge \mu\}$, and the gradient of this Moreau envelope is the saturation function, $\nabla M_{\mu g}(v_i) = \text{sign}(v_i) \min\left(|v_i|/\mu, 1\right)$.

B. Existing Algorithms

1) Proximal Gradient: The proximal gradient method generalizes standard gradient descent to certain classes of nonsmooth optimization problems. When $\mathcal{T} = I$, the proximal gradient method for problem (1) with step size α_l is given by $x^{l+1} = \mathbf{prox}_{\alpha_l g}(x^l - \alpha_l \nabla f(x^l))$. When $g = 0$, the proximal gradient method simplifies to standard gradient descent, and when g is indicator function of a convex set, it simplifies to projected gradient descent. The proximal gradient algorithm applied to the ℓ_1-regularized least-squares problem (LASSO)

$$\underset{x}{\text{minimize}} \ \frac{1}{2} \|Ax - b\|^2 + \gamma \|x\|_1 \tag{3}$$

where γ is a positive regularization parameter, yields the iterative soft-thresholding algorithm (ISTA) [5], $x^{l+1} = \mathcal{S}_{\gamma \alpha_l}(x^l - \alpha_l A^T(Ax^l - b))$. This method is effective only when the proximal operator of $g(\mathcal{T})$ is easy to evaluate. Except in special cases, e.g., when $\mathcal{T}$ is diagonal, efficient computation of $\mathbf{prox}_{\mu g(\mathcal{T})}$ does not necessarily follow from an efficiently computable $\mathbf{prox}_{\mu g}$. This makes the use of the proximal gradient method challenging for many applications and its convergence can be slow. Acceleration techniques improve the convergence rate [5], [28], but they do not perform well in the face of constraints such as closed-loop stability.

2) Augmented Lagrangian Methods: A common approach for dealing with a nondiagonal linear operator $\mathcal{T}$ in (1) is to introduce an additional optimization variable z as

$$\begin{aligned} \underset{x, z}{\text{minimize}} \quad & f(x) + g(z) \\ \text{subject to} \quad & \mathcal{T}(x) - z = 0. \end{aligned} \tag{4}$$

The augmented Lagrangian is obtained by adding a quadratic penalty on the violation of the linear constraint to the regular Lagrangian associated with (4) as

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

where y is the Lagrange multiplier and μ is a positive parameter.

ADMM solves (4) via an iteration that involves minimization of $\mathcal{L}_\mu(x, z; y)$ separately over x and z and a gradient ascent update (with step-size $1/\mu$) of y [6]

$$x^{k+1} = \underset{x}{\text{argmin}} \ \mathcal{L}_\mu(x, z^k; y^k) \tag{5a}$$

$$z^{k+1} = \underset{z}{\text{argmin}} \ \mathcal{L}_\mu(x^{k+1}, z; y^k) \tag{5b}$$

$$y^{k+1} = y^k + \frac{1}{\mu} \left(\mathcal{T}(x^{k+1}) - z^{k+1} \right). \tag{5c}$$

ADMM is appealing because, even when $\mathcal{T}$ is nondiagonal, the z-minimization step amounts to evaluating $\mathbf{prox}_{\mu g}$, and the x-minimization step amounts to solving a smooth (but possibly nonconvex) optimization problem. Although it was recently shown that ADMM is guaranteed to converge to a stationary point of (4) for some classes of nonconvex problems [29], its rate of convergence is strongly influenced by the choice of μ.

MM is the most widely used algorithm for solving constrained nonconvex optimization problems [7], [30] and it guarantees convergence to a local minimum. In contrast to ADMM, each MM iteration requires *joint* minimization of the augmented Lagrangian with respect to the primal variables x and z as

$$(x^{k+1}, z^{k+1}) = \underset{x, z}{\text{argmin}} \ \mathcal{L}_\mu(x, z; y^k) \tag{6a}$$

$$y^{k+1} = y^k + \frac{1}{\mu} \left(\mathcal{T}(x^{k+1}) - z^{k+1} \right). \tag{6b}$$

It is possible to refine MM to allow for inexact solutions to the (x, z)-minimization subproblem and adaptive updates of the penalty parameter μ. However, until now, MM has not been a feasible choice for solving (4) because the nonconvex and nondifferentiable (x, z)-minimization subproblem is as difficult as the original problem (1).

III. PROXIMAL AUGMENTED LAGRANGIAN

We next derive the proximal augmented Lagrangian, a continuously differentiable function resulting from explicit minimization of the augmented Lagrangian over the auxiliary variable z. This brings the (x, z)-minimization problem (6a) into a form that is continuously differentiable with respect to both x and y and facilitates the use of a wide suite of standard optimization tools for solving (1). In particular, as described below, our approach enables MM and the Arrow–Hurwicz–Uzawa gradient flow dynamics method.

A. Derivation of the Proximal Augmented Lagrangian

The first main result of this paper is provided in Theorem 1. We use the proximal operator of the function g to eliminate the auxiliary optimization variable z from the augmented Lagrangian and transform (6a) into a tractable continuously differentiable problem.

Theorem 1: For a proper, lower semicontinuous, and convex function function g, minimization of the augmented Lagrangian $\mathcal{L}_\mu(x, z; y)$ associated with problem (4) over (x, z) is equivalent to minimization of the *proximal augmented Lagrangian*

$$\mathcal{L}_\mu(x; y) := f(x) + M_{\mu g}(\mathcal{T}(x) + \mu y) - \frac{\mu}{2} \|y\|^2 \quad (7)$$

over x. Moreover, if f is continuously differentiable, $\mathcal{L}_\mu(x; y)$ is continuously differentiable over x and y, and if f has a Lipschitz continuous gradient, $\nabla \mathcal{L}_\mu(x; y)$ is Lipschitz continuous.

Proof: Through the completion of squares, the augmented Lagrangian $\mathcal{L}_\mu$ associated with (4) can be equivalently written as

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

Minimization with respect to z yields an explicit expression

$$z_\mu^\star(x, y) = \mathbf{prox}_{\mu g}(\mathcal{T}(x) + \mu y) \quad (8)$$

and substitution of $z_\mu^\star$ into the augmented Lagrangian provides (7), i.e., $\mathcal{L}_\mu(x; y) = \mathcal{L}_\mu(x, z_\mu^\star(x, y); y)$. Continuous differentiability of $\mathcal{L}_\mu(x; y)$ follows from continuous differentiability of $M_{\mu g}$ and Lipschitz continuity of $\nabla \mathcal{L}_\mu(x; y)$ follows from Lipschitz continuity of $\mathbf{prox}_{\mu g}$ and boundedness of the linear operator $\mathcal{T}$; see (2c). ∎

Expression (7), that we refer to as the *proximal augmented Lagrangian*, characterizes $\mathcal{L}_\mu(x, z; y)$ on the manifold corresponding to explicit minimization over the auxiliary variable z. Theorem 1 allows *joint* minimization of the augmented Lagrangian with respect to x and z, which is in general a nondifferentiable problem, to be achieved by minimizing differentiable function (7) over x. It thus facilitates the use of MM in Section III-B and the Arrow–Hurwicz–Uzawa gradient flow dynamics in Section IV.

Remark 1: The proximal augmented Lagrangian can be derived even in the presence of a more general linear constraint

$$\begin{aligned} \underset{x_1, x_2}{\text{minimize}} \quad & f(x_1) + g(x_2) \\ \text{subject to} \quad & \mathcal{T}_1(x_1) + \mathcal{T}_2(x_2) = 0. \end{aligned} \quad (9a)$$

The introduction of an additional auxiliary variable z can be used to recast problem (9a) as

$$\begin{aligned} \underset{x_1, x_2, z}{\text{minimize}} \quad & f(x_1) + g(z) \\ \text{subject to} \quad & \mathcal{T}_1(x_1) + \mathcal{T}_2(x_2) = 0, \quad x_2 - z = 0. \end{aligned} \quad (9b)$$

Via an analogous procedure to that described in Theorem 1, explicit minimization with respect to z can be employed to eliminate it from

the augmented Lagrangian and obtain a continuously differentiable function of both primal (x_1, x_2) and dual (y_1, y_2) variables

$$\mathcal{L}_\mu(x_1, x_2; y_1, y_2) = f(x_1) + \frac{1}{2\mu} \|\mathcal{T}_1(x_1) + \mathcal{T}_2(x_2) + \mu y_1\|^2$$
$$+ M_{\mu g}(x_2 + \mu y_2) - \frac{\mu}{2} \|y_1\|^2 - \frac{\mu}{2} \|y_2\|^2.$$

Here, y_1 and y_2 are the Lagrange multipliers associated with the respective linear constraints in (9b). This approach has numerous advantages over the standard ADMM; e.g., it can be readily extended to multiblock optimization problems for which ADMM is not guaranteed to converge in general [31]. These extensions are outside of the scope of this study and will be reported elsewhere.

B. MM Using the Proximal Augmented Lagrangian

Theorem 1 allows us to solve nondifferentiable subproblem (6a) by minimizing the continuously differentiable proximal augmented Lagrangian $\mathcal{L}_\mu(x; y^k)$ over x. We note that a similar approach was also applied to MM in [32] for the particular case in which g is the indicator function of a convex set. Relative to ADMM, our customized MM algorithm guarantees convergence to a local minimum and offers systematic update rules for the parameter μ. Relative to proximal gradient, we can solve (1) with a general bounded linear operator $\mathcal{T}$ and can incorporate second-order information about f.

Using reformulated expression (7) for the augmented Lagrangian, MM minimizes $\mathcal{L}_\mu(x; y^k)$ over the primal variable x and updates the dual variable y using the gradient ascent with step size $1/\mu$ as

$$x^{k+1} = \underset{x}{\text{argmin}} \ \mathcal{L}_\mu(x; y^k) \quad (\text{MMa})$$

$$y^{k+1} = y^k + \frac{1}{\mu} \nabla_y \mathcal{L}_\mu(x^{k+1}; y^k) \quad (\text{MMb})$$

where $\nabla_y \mathcal{L}_\mu(x^{k+1}; y^k) := \mathcal{T}(x^{k+1}) - z_\mu^\star(x^{k+1}, y^k) = \mathcal{T}(x^{k+1}) - \mathbf{prox}_{\mu g}(\mathcal{T}(x^{k+1}) + \mu y^k)$ denotes the primal residual.

In contrast to ADMM, our approach does not attempt to avoid the lack of differentiability of g by fixing z to minimize over x. By constraining $\mathcal{L}_\mu(x, z; y)$ to the manifold resulting from explicit minimization over z, we guarantee continuous differentiability of the proximal augmented Lagrangian $\mathcal{L}_\mu(x; y)$. MM is a gradient ascent algorithm on the Lagrange dual of a version of (4) in which the objective function is replaced by $f(x) + g(z) + \frac{1}{2\mu} \|\mathcal{T}(x) - z\|^2$; see [8, Sec. 2.3] and [33]. Since its closed-form expression is typically unavailable, MM uses (6a) to evaluate this dual computationally, and then, take a gradient ascent step (6b) in y. ADMM avoids this issue by solving simpler, separate subproblems (5a) and (5b), but, unlike (6b) in MM, the y-update (5c) in ADMM is not a gradient ascent step on the "strengthened dual." MM, thus, offers stronger convergence results [6], [7] and may lead to fewer y-update steps.

Remark 2: The proximal augmented Lagrangian enables MM because the x-minimization subproblem (MMa) in MM is not more difficult than its counterpart (5a) in ADMM. For LASSO problem (3), the z-update (5b) in ADMM is given by soft thresholding, $z^{k+1} = \mathcal{S}_{\gamma_\mu}(x^{k+1} + \mu y^k)$, and the x-update (5a) requires minimization of the quadratic function [6]. In contrast, the x-update (MMa) in MM requires minimization of $(1/2) \|Ax - b\|^2 + M_{\mu_k g}(x + \mu_k y^k)$, where $M_{\mu_k g}(v)$ is the Moreau envelope associated with the ℓ_1 norm; i.e., the Huber function. Although in this case, the solution to (5a) can be characterized explicitly by a matrix inversion, this is not true in general. The computational cost associated with solving either (5a) or (MMa) using first-order methods scales at the same rate.

Algorithm: The procedure outlined in [11, Algorithm 17.4] allows minimization subproblem (MMa) to be inexact, provides a method for adaptively adjusting μ_k, and describes a more refined update of the Lagrange multiplier y. We have incorporated these refinements into our proximal augmented Lagrangian algorithm for solving (4); see Algorithm 1 in [10]. Because of the equivalence established in Theorem 1, convergence to a local minimum follows from the convergence results for standard MM [9]. As discussed in [10], since $\mathcal{L}_\mu(x; y)$

is once continuously differentiable, many techniques can be used to solve subproblem (MMa), including gradient descent, proximal gradient methods that exploit the proximal operator associated with the Moreau envelope, and quasi-Newton methods such as the limited-memory Broyden-Fletcher-Goldfarb-Shannon (L-BFGS) [11, Algorithm 7.4] method, which is guaranteed to converge for convex functions with Lipschitz continuous gradients [36]. For example, in [34], proximal gradient methods were used for subproblem (MMa) to solve a sparse feedback synthesis problem introduced in [3]. Computational savings were shown relative to standard proximal gradient method and ADMM.

Remark 3: For regularization functions that do not admit simply computable proximal operators, $\mathbf{prox}_{\mu g}$ has to be evaluated numerically by solving (2a). If this is expensive, the primal-descent dual-ascent algorithm of Section IV offers an appealing alternative because it requires one evaluation of $\mathbf{prox}_{\mu g}$ per iteration.

IV. ARROW–HURWICZ–UZAWA GRADIENT FLOW

We now consider an alternative approach to solving (1). Instead of minimizing over the primal variable and performing gradient ascent in the dual, we simultaneously update the primal and dual variables to find the saddle point of the augmented Lagrangian. The continuous differentiability of $\mathcal{L}_\mu(x; y)$ established in Theorem 1 enables the use of Arrow–Hurwicz–Uzawa gradient flow dynamics [11]

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\nabla_x \mathcal{L}_\mu(x; y) \\ \nabla_y \mathcal{L}_\mu(x; y) \end{bmatrix}. \tag{GF}$$

In Section IV-A, we show that the gradient flow dynamics (GF) globally converge to the set of saddle points of the proximal augmented Lagrangian $\mathcal{L}_\mu(x; y)$ for a convex f with a Lipschitz continuous gradient. In Section IV-B, we employ the theory of integral quadratic constraints (IQCs) to establish global exponential stability for a strongly convex f with a Lipschitz continuous gradient and estimate convergence rates. Finally, in Section IV-C, we identify classes of problems for which dynamics (GF) are convenient for distributed implementation and compare/contrast our framework to the existing approaches.

A. Global Asymptotic Stability for Convex f

We first characterize the optimal primal-dual pairs of optimization problem (4) with the Lagrangian, $f(x) + g(z) + \langle y, \mathcal{T}(x) - z \rangle$. The associated first-order optimality conditions are given by

$$0 = \nabla f(x^\star) + \mathcal{T}^\dagger(y^\star) \tag{10a}$$

$$0 \in \partial g(z^\star) - y^\star \tag{10b}$$

$$0 = \mathcal{T}(x^\star) - z^\star \tag{10c}$$

where ∂g is the subgradient of g. Clearly, these are equivalent to the optimality condition for (1), i.e., $0 \in \nabla f(x^\star) + \mathcal{T}^\dagger(\partial g(\mathcal{T}(x^\star)))$. Even though we state the result for $x \in \mathbb{R}^n$ and $\mathcal{T}(x) = Tx$ where $T \in \mathbb{R}^{m \times n}$ is a given matrix, the proof for x in a Hilbert space and a bounded linear operator $\mathcal{T}$ follows from similar arguments.

Theorem 2: Let f be a continuously differentiable convex function with a Lipschitz continuous gradient and let g be a proper, lower semicontinuous, convex function. Then, the set of optimal primal-dual pairs $(x^\star, y^\star)$ of (4) for the gradient flow dynamics (GF)

$$\begin{bmatrix} \dot{x} \\ \dot{y} \end{bmatrix} = \begin{bmatrix} -\left(\nabla f(x) + T^T \nabla M_{\mu g}(Tx + \mu y)\right) \\ \mu\left(\nabla M_{\mu g}(Tx + \mu y) - y\right) \end{bmatrix} \tag{GF1}$$

is globally asymptotically stable (GAS) and each $x^\star$ is a solution of (1).

Proof: We introduce a change of variables $\tilde{x} := x - x^\star$, $\tilde{y} := y - y^\star$ and a Lyapunov function candidate, $V(\tilde{x}, \tilde{y}) = \frac{1}{2}\langle \tilde{x}, \tilde{x} \rangle + \frac{1}{2}\langle \tilde{y}, \tilde{y} \rangle$, where $(x^\star, z^\star; y^\star)$ is an optimal solution to (4) that satisfies (10). The

dynamics in the $(\tilde{x}, \tilde{y})$-coordinates are given by

$$\begin{bmatrix} \dot{\tilde{x}} \\ \dot{\tilde{y}} \end{bmatrix} = \begin{bmatrix} -(\nabla f(x) - \nabla f(x^\star) + (1/\mu)T^T\tilde{m}) \\ \tilde{m} - \mu\tilde{y} \end{bmatrix} \tag{11}$$

where $\tilde{m} = \mu\left(\nabla M_{\mu g}(Tx + \mu y) - \nabla M_{\mu g}(Tx^\star + \mu y^\star)\right)$ can be expressed as

$$\begin{aligned} \tilde{m} &:= \tilde{v} - \tilde{z} \\ \tilde{v} &:= T\tilde{x} + \mu\tilde{y} = (Tx + \mu y) - (Tx^\star + \mu y^\star) \\ \tilde{z} &:= \mathbf{prox}_{\mu g}(Tx + \mu y) - \mathbf{prox}_{\mu g}(Tx^\star + \mu y^\star). \end{aligned} \tag{12}$$

The derivative of V along the solutions of (11) is given by

$$\begin{aligned} \dot{V} &= -\langle \tilde{x}, \nabla f(x) - \nabla f(x^\star) \rangle - \tfrac{1}{\mu}\|T\tilde{x}\|^2 + \tfrac{1}{\mu}\langle T\tilde{x} - \mu\tilde{y}, \tilde{z} \rangle \\ &= -\langle \tilde{x}, \nabla f(x) - \nabla f(x^\star) \rangle - \tfrac{1}{\mu}\left(\|T\tilde{x}\|^2 - 2\langle T\tilde{x}, \tilde{z} \rangle + \langle \tilde{v}, \tilde{z} \rangle\right). \end{aligned}$$

Since f is convex with an L_f-Lipschitz continuous gradient and since $\mathbf{prox}_{\mu g}$ is firmly nonexpansive [4], i.e., $\langle \tilde{v}, \tilde{z} \rangle \geq \|\tilde{z}\|^2$, we have

$$\dot{V}(\tilde{x}, \tilde{y}) \leq -\tfrac{1}{L_f}\|\nabla f(x) - \nabla f(x^\star)\|^2 - \tfrac{1}{\mu}\|T\tilde{x} - \tilde{z}\|^2. \tag{13}$$

Thus, $\dot{V} \leq 0$ and each point in the set of optimal primal-dual pairs $(x^\star, y^\star)$ is stable in the sense of Lyapunov.

The right-hand side in (13) becomes zero when $\nabla f(x) = \nabla f(x^\star)$ and $T\tilde{x} = \tilde{z}$. Under these conditions, we have $\dot{V} = -\langle T^T\tilde{y}, \tilde{x} \rangle$ and the set of points for which $\dot{V} = 0$ is given by $\mathcal{D} = \{(x, y); \nabla f(x) = \nabla f(x^\star), T\tilde{x} = \tilde{z}, \langle T^T\tilde{y}, \tilde{x} \rangle = 0\}$. Furthermore, substitution of $T\tilde{x} = \tilde{z}$ into (12) yields $\tilde{m} = \mu\tilde{y}$ and (11) simplifies to, $\dot{\tilde{x}} = -T^T\tilde{y}$, $\dot{\tilde{y}} = 0$. For (11), the largest invariant set $\Omega := \{(x, y); \nabla f(x) = \nabla f(x^\star), T\tilde{x} = \tilde{z}, T^T\tilde{y} = 0\} \subseteq \mathcal{D}$ is obtained from

$$\langle T^T\tilde{y}, \tilde{x} \rangle \equiv 0 \;\Rightarrow\; \langle T^T\dot{\tilde{y}}, \tilde{x} \rangle + \langle T^T\tilde{y}, \dot{\tilde{x}} \rangle = -\|T^T\tilde{y}\|^2 \equiv 0$$

and LaSalle's invariance principle implies that Ω is GAS.

To complete the proof, we need to show that any x and y that lie in Ω also satisfy optimality conditions (10) of problem (4) with $z = z_\mu^\star(x, y) = \mathbf{prox}_{\mu g}(Tx + \mu y)$, and thus, that x solves problem (1). For any $(x, y) \in \Omega$, $\nabla f(x) = \nabla f(x^\star)$ and $T^T y = T^T y^\star$. Since $x^\star$ and $y^\star$ are optimal primal-dual points, we have $\nabla f(x) + T^T y = \nabla f(x^\star) + T^T y^\star = 0$ which implies that every $(x, y) \in \Omega$ satisfies (10a). Optimality condition (10b) for $(x^\star, y^\star)$, $Tx^\star = z^\star$, together with $T\tilde{x} = \tilde{z}$, imply that $Tx = z$, i.e., x and $z = \mathbf{prox}_{\mu g}(Tx + \mu y)$ satisfy (10c). Finally, the optimality condition of the problem (2a) that defines $\mathbf{prox}_{\mu g}(v)$ is $\partial g(z) + \frac{1}{\mu}(z - v) \ni 0$. Letting $v = Tx + \mu y$ from the expression (8) that characterizes the $z_\mu^\star$-manifold and noting $Tx = z$ by (10c) leads to (10b). Thus, every $(x, y) \in \Omega$ satisfies (10), implying that the set of primal-dual optimal points is GAS. ∎

B. Global Exponential Stability for Strongly Convex f

We express (GF), or equivalently (GF1), as a linear system G connected in feedback with nonlinearities that correspond to the gradients of f and of the Moreau envelope of g; see Fig. 1. These nonlinearities can be conservatively characterized by IQCs. Exponential stability of G connected in feedback with *any* nonlinearity that satisfies these IQCs implies exponential convergence of (GF) to the primal-dual optimal solution of (4). In what follows, we adjust the tools of [18] and [19] to our setup and establish global exponential stability by evaluating the feasibility of an LMI. We assume that the function f is m_f-strongly convex with an L_f-Lipschitz continuous gradient. Characterizing additional structural restrictions on f and g with IQCs may lead to tighter bounds on the rate of convergence.
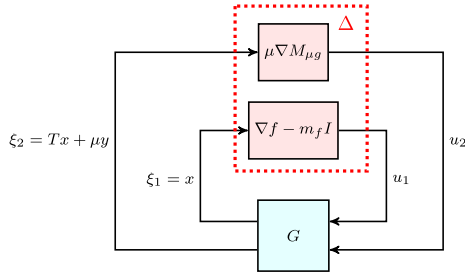
Fig. 1. Block diagram of gradient flow dynamics (GF1) where G is a linear system in feedback with nonlinearities that satisfy (15).

As illustrated in Fig. 1, (GF1) can be expressed as a linear system G connected via feedback to a nonlinear block Δ as

$$\dot{w} = A\,w + B\,u, \quad \xi = C\,w, \quad u = \Delta(\xi)$$

$$A = \begin{bmatrix} -m_f I & \\ & -\mu I \end{bmatrix}, \quad B = \begin{bmatrix} -I & -\frac{1}{\mu}T^T \\ 0 & I \end{bmatrix}, \quad C = \begin{bmatrix} I & 0 \\ T & \mu I \end{bmatrix}$$

where $w := [\,x^T\ y^T\,]^T$, $\xi := [\,\xi_1^T\ \xi_2^T\,]^T$, and $u := [\,u_1^T\ u_2^T\,]^T$. Nonlinearity Δ maps the system outputs $\xi_1 = x$ and $\xi_2 = Tx + \mu y$ to the inputs u_1 and u_2 via $u_1 = \Delta_1(\xi_1) := \nabla f(\xi_1) - m_f \xi_1$ and $u_2 = \Delta_2(\xi_2) := \mu \nabla M_{\mu g}(\xi_2) = \xi_2 - \mathbf{prox}_{\mu g}(\xi_2)$.

When the mapping $u_i = \Delta_i(\xi_i)$ is the L_i-Lipschitz continuous gradient of a convex function, it satisfies the IQC [19, Lemma 6]

$$\begin{bmatrix} \xi_i & -\xi_0 \\ u_i & -u_0 \end{bmatrix}^T \begin{bmatrix} 0 & \hat{L}_i I \\ \hat{L}_i I & -2I \end{bmatrix} \begin{bmatrix} \xi_i - \xi_0 \\ u_i - u_0 \end{bmatrix} \geq 0 \quad (14)$$

where $\hat{L}_i \geq L_i$, ξ_0 is some reference point, and $u_0 = \Delta_i(\xi_0)$. Since f is m_f-strongly convex, the mapping $\Delta_1(\xi_1)$ is the gradient of the convex function $f(\xi_1) - (m_f/2)\|\xi_1\|^2$. Lipschitz continuity of ∇f with parameter L_f implies Lipschitz continuity of $\Delta_1(\xi_1)$ with parameter $L_1 := L_f - m_f$; thus, Δ_1 satisfies (14) with $\hat{L}_1 \geq L_1$. Similarly, $\Delta_2(\xi_2)$ is the scaled gradient of the convex Moreau envelope and is Lipschitz continuous with parameter 1; thus, Δ_2 also satisfies (14) with $\hat{L}_2 \geq 1$. These two IQCs can be combined into

$$(\eta - \eta_0)^T \Pi\,(\eta - \eta_0) \geq 0, \quad \eta := [\,\xi^T\ u^T\,]^T. \quad (15)$$

For a linear system G connected in feedback with nonlinearities that satisfy IQC (15), [20, Th. 3] establishes ρ-exponential convergence, i.e., $\|w(t) - w^\star\| \leq \tau e^{-\rho t}\|w(0) - w^\star\|$ for some $\tau, \rho > 0$, by verifying the existence of a matrix $P \succ 0$ such that

$$\begin{bmatrix} A_\rho^T P + P A_\rho & PB \\ B^T P & 0 \end{bmatrix} + \begin{bmatrix} C^T & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} C & 0 \\ 0 & I \end{bmatrix} \preceq 0 \quad (16)$$

where $A_\rho := A + \rho I$. In Theorem 3, we determine a scalar condition that ensures global exponential stability when TT^T is full rank.

Theorem 3: Let f be strongly convex with parameter m_f, let its gradient be Lipschitz continuous with parameter L_f, let g be proper, lower semicontinuous, and convex, and let TT^T be full rank. Then, if $\mu \geq L_f - m_f$, there is a $\rho > 0$ such that the dynamics (GF) converge ρ-exponentially to the optimal point of (4).

Proof: Since any function that is Lipschitz continuous with parameter L is also Lipschitz continuous with parameter $\hat{L} \geq L$, we establish the result for $\mu = \hat{L}_1 := L_f - m_f$ and $\hat{L}_2 = 1$. We utilize [20, Th. 3] to show ρ-exponential convergence by verifying matrix inequality (16) through a series of equivalent expressions (17). We first apply the Kalman-Yakubovich-Popov (KYP) Lemma [39, Th. 1] to (16) to

obtain an equivalent frequency-domain characterization

$$\begin{bmatrix} G_\rho(\mathrm{j}\omega) \\ I \end{bmatrix}^* \Pi \begin{bmatrix} G_\rho(\mathrm{j}\omega) \\ I \end{bmatrix} \preceq 0, \quad \forall\,\omega \in \mathbb{R} \quad (17\text{a})$$

where $G_\rho(\mathrm{j}\omega) = C(\mathrm{j}\omega I - A_\rho)^{-1}B$. Evaluating the left-hand side of (17a) for $L = \mu$ and dividing by -2 yields the matrix inequality

$$\begin{bmatrix} \dfrac{\mu\hat{m} + \hat{m}^2 + \omega^2}{\hat{m}^2 + \omega^2}\,I & \dfrac{\hat{m}}{\hat{m}^2 + \omega^2}\,T^T \\[2mm] * & \dfrac{\hat{m}/\mu}{\hat{m}^2 + \omega^2}\,TT^T + \dfrac{\omega^2 - \rho\hat{\mu}}{\hat{\mu}^2 + \omega^2}\,I \end{bmatrix} \succ 0 \quad (17\text{b})$$

where $\hat{m} := m_f - \rho > 0$ and $\hat{\mu} := \mu - \rho > 0$ so that A_ρ is Hurwitz, i.e., the system G_ρ is stable. Since the $(1,1)$ block in (17b) is positive definite for all ω, the matrix in (17b) is positive definite if and only if the corresponding Schur complement is positive definite

$$\dfrac{\hat{m}/\mu}{\mu\hat{m} + \hat{m}^2 + \omega^2}\,TT^T + \dfrac{\omega^2 - \rho\hat{\mu}}{\hat{\mu}^2 + \omega^2}\,I \succ 0. \quad (17\text{c})$$

We exploit the symmetry of TT^T to diagonalize (17c) via a unitary coordinate transformation. This yields m scalar inequalities parameterized by the eigenvalues λ_i of TT^T. Multiplying the left-hand side of these inequalities by the positive quantity $(\hat{\mu}^2 + \omega^2)(\mu\hat{m} + \hat{m}^2 + \omega^2)$ yields a set of equivalent, quadratic in ω^2, conditions

$$\omega^4 + \left(\dfrac{\hat{m}\lambda_i}{\mu} + \hat{m}^2 + \mu\hat{m} - \rho\hat{\mu} \right)\omega^2 + \hat{m}\hat{\mu}\left(\dfrac{\hat{\mu}\lambda_i}{\mu} - \rho(\mu + \hat{m}) \right) > 0. \quad (17\text{d})$$

Condition (17d) is satisfied for all $\omega \in \mathbb{R}$ if there are no $\omega^2 \geq 0$ for which the left-hand side is nonpositive. When $\rho = 0$, both the constant term and the coefficient of ω^2 are strictly positive, which implies that the roots of (17d) as a function of ω^2 are either not real or lie in the domain $\omega^2 < 0$, which cannot occur for $\omega \in \mathbb{R}$. Finally, continuity of (17d) with respect to ρ implies the existence a positive ρ that satisfies (17d) for all $\omega \in \mathbb{R}$. ∎

Remark 4: Each eigenvalue λ_i of a full rank matrix TT^T is positive, and hence, to estimate the exponential convergence rate it suffices to check (17d) only for the smallest λ_i. A sufficient condition for (17d) to hold for each $\omega \in \mathbb{R}$ is positivity of the constant term and the coefficient multiplying ω^2. For $\rho < \min(m_f, \mu)$, these can be, respectively, expressed as the following quadratic inequalities in ρ:

$$\rho^2 - \gamma\rho + \lambda_{\min} > 0, \quad 2\rho^2 - (\gamma + \mu + m_f)\rho + \gamma\,m_f > 0$$

where $\gamma := \mu + m_f + \frac{\lambda_{\min}}{\mu}$. The solutions to these provide the following estimates of the exponential convergence rate: 1) $\rho < \rho_1$ when $m_f \geq \mu$; and 2) $\rho < \min(\rho_1, \rho_2)$ when $m_f < \mu$, where

$$\rho_1 = \tfrac{1}{2}\left(\gamma - \sqrt{\gamma^2 - 4\lambda_{\min}} \right)$$

$$\rho_2 = \tfrac{1}{4}\left(\gamma + \mu + m_f - \sqrt{(\gamma + \mu + m_f)^2 - 8\gamma\,m_f} \right).$$

Our explicit analytical expressions can be used to determine the optimal value of $\mu \geq L_f - m_f$ to maximize the aforementioned decay rates.

Remark 5: A similar convergence rate result can be obtained by applying [19, Th. 4] to a discrete-time implementation of the primal-descent dual-ascent dynamics that results from a forward Euler discretization of (GF); for details, see [38].

Remark 6: To the best of our knowledge, we are the first to establish global exponential stability of the primal-dual gradient flow dynamics for nonsmooth composite optimization problems (1) with a strongly convex f. Recent reference [39] proves similar result for a narrower class of problems (strongly convex and smooth objective function with either affine equality or inequality constraints). Both of these can be cast as (1) via introduction of suitable indicator functions and exponential stability follows immediately from our Theorem 3.

C. Distributed Implementation

Gradient flow dynamics (GF) are convenient for distributed implementation. If the state vector x corresponds to the concatenated states of individual agents, x_i, the sparsity pattern of T and the structure of the gradient map $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ dictate the communication topology required to form $\nabla \mathcal{L}_\mu$ in (GF). For example, if $f(x) = \sum f_i(x_i)$ is separable over the agents, then $\nabla f_i(x_i)$ can be formed locally. If in addition T^T is an incidence matrix of an undirected network with the graph Laplacian $T^T T$, each agent need only share its state x_i with its neighbors and maintain dual variables y_i that correspond to its edges. A distributed implementation is also natural when the mapping $\nabla f \colon \mathbb{R}^n \to \mathbb{R}^n$ is sparse.

Our approach provides several advantages over existing distributed optimization algorithms. Even for problems (1) with nondifferentiable regularizers g, a formulation based on the proximal augmented Lagrangian yields gradient flow dynamics (GF) with a continuous right-hand side. This is in contrast to existing approaches that employ subgradient methods [13] or use discontinuous projected dynamics [14]–[17]. Note that although the augmented Lagrangian $\mathcal{L}_\mu(x, y; z)$ contains a quadratic term $\frac{1}{2\mu}\|\mathcal{T}(x) - z\|^2$, it is not *jointly* strongly convex in x and z and the resulting proximal augmented Lagrangian (7) is not strictly convex–concave in x and y. Furthermore, when T is not diagonal, a distributed proximal gradient cannot be implemented because the proximal operator of $g(Tx)$ may not be separable. Finally, ADMM has been used for distributed implementation in the situations where f is separable and T is an incidence matrix. Relative to such a scheme, (GF) does not require solving an x-minimization subproblem in each iteration and provides a guaranteed rate of convergence.

Remark 7: Special instances of our framework have strong connections with the existing methods for distributed optimization on graphs; e.g., [12], [13], and [40]. The networked optimization problem of minimizing $f(\bar{x}) = \sum f_i(\bar{x})$ over a single variable $\bar{x}$ can be reformulated as $\sum f_i(x_i) + g(Tx)$ where the components f_i of the objective function are distributed over independent agents x_i, x is the aggregate state, T^T is the incidence matrix of a strongly connected and balanced graph, and g is the indicator function associated with the set $Tx = 0$. The $g(Tx)$ term ensures that at feasible points, $x_i = x_j = \bar{x}$ for all i and j. It is easy to show that $\nabla M_{\mu g}(v) = (1/\mu)\,v$ and that the dynamics (GF) are

$$\dot{x} = -\nabla f(x) - (1/\mu)\,L\,x - \tilde{y}, \quad \dot{\tilde{y}} = \beta\,L\,x \qquad (18)$$

where $\beta > 0$, $L := T^T T$ is the graph Laplacian of a connected undirected network, and $\tilde{y} := T^T y$ belongs to the orthogonal complement of the vector of all ones. The only difference relative to [13, Eq. (20)] and [42, Eq. (11)] is that $-\tilde{y}$ appears instead of $-L\tilde{y}$ in (18) for the dynamics of the primal variable x.

Remark 8: The forward Euler discretization of (18) is given by

$$\begin{aligned} x^{k+1} &= (I - (\alpha/\mu)L)\,x^k - \alpha\,\nabla f(x^k) - \alpha\,\tilde{y}^k \\ \tilde{y}^{k+1} &= \tilde{y}^k + \alpha\,\beta\,L\,x^k \end{aligned} \qquad (19)$$

where α is the step size, and the EXTRA algorithm [43, Eq. (2.13)], which has received significant recent attention

$$x^{k+1} = W x^k - \alpha\,\nabla f(x^k) + \frac{1}{2}\sum_{i=0}^{k-1}(W - I)\,x^i \qquad (20)$$

can be clearly recovered from (19) by setting $\beta = 1/(2\alpha\mu)$ and taking $W = I - (\alpha/\mu)L$ in (20).

V. EXAMPLES

We solve the problems of edge addition in directed consensus networks and optimal placement to illustrate the effectiveness of the proximal augmented Lagrangian method.

(a) Directed consensus network   (b) Performance-sparsity tradeoff
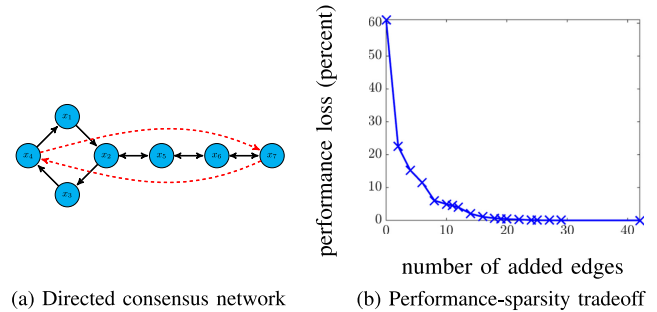
Fig. 2.   (a) Balanced plant graph with seven nodes and ten directed edges (solid black lines). A sparse set of two added edges (dashed red lines) is identified by solving (22) with $\gamma = 3.5$ and $R = I$. (b) Tradeoff between performance and sparsity resulting from the solution to (22) and (23) for the network shown in Fig. 2(a). Performance loss is measured relative to the optimal centralized controller (i.e., all edges are used).

A. Edge Addition in Directed Consensus Networks

A consensus network with N nodes converges to the average of the initial node values $\bar{\psi} = (1/N)\sum_i \psi_i(0)$ if and only if it is strongly connected and balanced [42]. Unlike for undirected networks [22], [23], the problem of edge addition in directed consensus networks is not known to be convex. The steady-state variance of the deviations from average is given by the square of the $\mathcal{H}_2$ norm of

$$\dot{\psi} = -(L_p + L_x)\,\psi + d, \quad \xi = \begin{bmatrix} Q^{1/2} \\ -R^{1/2} L_x \end{bmatrix}\psi$$

where d is a disturbance, L_p is a weighted directed graph Laplacian of a plant network, $Q := I - (1/N)\mathbb{1}\mathbb{1}^T$ penalizes the deviation from average, and $R \succ 0$ is the control weight. The objective is to optimize the $\mathcal{H}_2$ norm (from d to ξ) by adding a few additional edges, specified by the graph Laplacian L_x of a controller network.

To ensure convergence of ψ to the average of the initial node values, we require that the closed-loop graph Laplacian, $L = L_p + L_x$, is balanced. This condition amounts to the linear constraint, $\mathbb{1}^T L = 0$. We express the directed graph Laplacian of the controller network as, $L_x = \sum_{i \neq j} L_{ij} z_{ij} =: \sum_l L_l z_l$ where $z_{ij} \geq 0$ is the added edge weight that connects node j to node i, $L_{ij} := e_i e_i^T - e_i e_j^T$, e_i is the ith basis vector in $\mathbb{R}^n$, and the integer l indexes the edges such that $z_l = z_{ij}$ and $L_l = L_{ij}$. For simplicity, we assume that the plant network L_p is balanced and connected. Thus, enforcing that L is balanced amounts to enforcing the linear constraint $\mathbb{1}^T L_x = \mathbb{1}^T(\sum L_l z_l) =: (Ez)^T = 0$ on z, where E is the incidence matrix [42] of the edges that may be added. Any vector of edge weights z that satisfies this constraint can be written as $z = Tx$ where the columns of T span the nullspace of the matrix E and provide a basis for the space of balanced graphs, i.e., the cycle space [42]. Each feasible controller Laplacian can thus be written as

$$L_x = \sum_l L_l\,[Tx]_l = \sum_l L_l\left[\sum_k (Te_k)\,x_k\right]_l =: \sum_k \hat{L}_k\,x_k \quad (21a)$$

where the matrices $\hat{L}_k$ are given by $\hat{L}_k = \sum_l L_l\,[Te_k]_l$.

Since the mode corresponding to $\mathbb{1}$ is marginally stable, unobservable, and uncontrollable, we introduce a change of coordinates to the deviations from average $\phi = V^T \psi$ where $V^T \mathbb{1} = 0$ and discard the average mode $\bar{\psi} = \mathbb{1}^T \psi$. The energy of the deviations from average is given by the the $\mathcal{H}_2$ norm squared of the reduced system

$$f(x) = \left\langle V^T(Q + L_x^T R L_x)V, X\right\rangle, \quad \hat{A}X + X\hat{A}^T + \hat{B}\hat{B}^T = 0 \qquad (21b)$$

where X is the controllability Gramian of the reduced system with $\hat{A} := -V^T(L_p + L_x)V$ and $\hat{B} := V^T$.
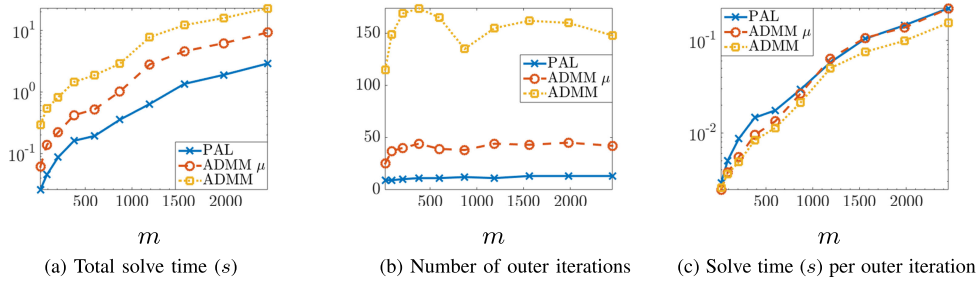
Fig. 3. (a) Total time; (b) number of outer iterations; and (c) average time per outer iteration required to solve (22) with $\gamma = 0.01, 0.1, 0.2$ for a cycle graph with $N = 5$ to 50 nodes as a function of $m = 20$ to 2450 potential added edges using PAL (–×–), ADMM (- -○- -), and ADMM with the adaptive μ-update heuristic [6] (···□···). PAL requires fewer outer iterations, and thus, a smaller total solving time.

To balance the closed-loop $\mathcal{H}_2$ performance with the number of added edges, we introduce a regularized optimization problem

$$x_\gamma = \underset{x}{\mathrm{argmin}} \; f(x) + \gamma \mathbb{1}^T T x + I_+(Tx). \tag{22}$$

Here, the regularization parameter $\gamma > 0$ specifies the emphasis on sparsity relative to the closed-loop performance f, and I_+ is the indicator function associated with the nonnegative orthant $\mathbb{R}_+^m$. When the desired level of sparsity for the vector of the added edge weights $z_\gamma = Tx_\gamma$ has been attained, optimal weights for the identified set of edges are obtained by solving

$$\underset{x}{\mathrm{minimize}} \; f(x) + I_{\mathcal{Z}_\gamma}(Tx) + I_+(Tx) \tag{23}$$

where $\mathcal{Z}_\gamma$ is the set of vectors with the same sparsity pattern as z_γ and $I_{\mathcal{Z}_\gamma}$ is the indicator function associated with this set.

1) Implementation: We next provide implementation details for solving (22) and (23). The proof of next lemma is omitted for brevity.

Lemma 4: Let a graph Laplacian of a directed plant network L_p be balanced and connected and let $\hat{A}, \hat{B}, L_x$, and V be as defined in (21a) and (21b). The gradient of $f(x)$ defined in (21b) is given by

$$\nabla f(x) = 2\mathrm{vec}\left(\left\langle (R L_x V - V P) X V^T, \hat{L}_k \right\rangle\right)$$

where X and P are the controllability and observability Gramians determined by (21b) and $\hat{A}^T P + P \hat{A} + V^T (Q + L_x^T R L_x) V = 0$.

The proximal operator associated with the regularization function $g_s(z) := \gamma \mathbb{1}^T z + I_+(z)$ in (22) is $\mathbf{prox}_{\mu g_s}(v_i) = \max\{0, v_i - \gamma\mu\}$, the Moreau envelope is given by $M_{\mu g_s}(v) = \sum_i \{v_i^2/(2\mu), v_i \leq \gamma\mu; \gamma(v_i - \gamma\mu/2), v_i > \gamma\mu\}$, and $\nabla M_{\mu g_s}(v) = \max\{v/\mu, \gamma\}$. The proximal operator of the regularization function in (23), $g_p(z) := I_{\mathcal{Z}_\gamma}(z) + I_+(z)$, is a projection onto the intersection of the set $\mathcal{Z}_\gamma$ and the nonnegative orthant, $\mathbf{prox}_{\mu g_p}(v) = \mathcal{P}_{\mathcal{E}}(v)$, the Moreau envelope is the distance to $\mathcal{E} := \mathcal{Z}_\gamma \cap \mathbb{R}_+^m$, $M_{\mu g_p}(v) = \frac{1}{2\mu}\|v - \mathcal{P}_{\mathcal{E}}(v)\|^2$ and $\nabla M_{\mu g_p}(v)$ is determined by a vector pointing from $\mathcal{E}$ to v, $\nabla M_{\mu g_p}(v) = \frac{1}{\mu}(v - \mathcal{P}_{\mathcal{E}}(v))$.

2) Computational Experiments: We solve (22) and (23) using Algorithm 1, where L-BFGS is employed in the x-minimization subproblem (MMa). For the plant network shown in Fig. 2(a), Fig. 2(b) illustrates the tradeoff between the number of added edges and the closed-loop $\mathcal{H}_2$ norm. The added edges are identified by computing the γ-parameterized homotopy path for problem (22), and the optimal edge weights are obtained by solving (23). The red dashed lines in Fig. 2(a) show the optimal set of two added edges. These are obtained for $\gamma = 3.5$ and they yield 23.91% performance loss relative to the setup in which all edges in the controller graph are used. We note that the same set of edges is obtained by conducting an exhaustive search. This suggests that the proposed convex regularizers may offer a good proxy for solving difficult combinatorial optimization problems.

We also consider simple directed cycle graphs with $N = 5$ to 50 nodes and $m = N^2 - N$ potential added edges. We solve (22) for $\gamma = 0.01, 0.1, 0.2$, and $R = I$ using the proximal augmented Lagrangian
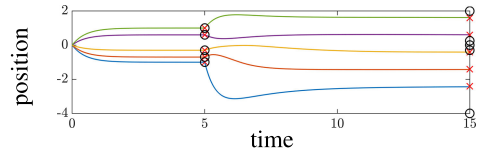

Fig. 4. Set of five distributed agents tracking targets (**black** ○) whose optimal positions are determined by the solution to (24) (**red** ×).

MM algorithm (PAL), ADMM, and ADMM with an adaptive heuristic for updating μ [6] (ADMM μ). The x-update in each algorithm is obtained using L-BFGS. Since $g_s(Tx)$ and $g_p(Tx)$ are not separable in x, proximal gradient cannot be used here.

Fig. 3(a) shows the time required to solve problem (22) in terms of the total number of potential added edges; Fig. 3(b) demonstrates that PAL requires fewer outer iterations; and Fig. 3(c) illustrates that the average computation time per outer iteration is roughly equivalent for all three methods. Even with an adaptive update of μ, ADMM requires more outer iterations, which increases overall solving time relative to the proximal augmented Lagrangian method. Thus, compared to ADMM, PAL provides computational advantage by reducing the number of outer iterations [indexed by k in Algorithm 1 and in (5)].

B. Optimal Placement Problem

To illustrate the utility of our primal-descent dual-ascent approach, we consider an example in which mobile agents aim to minimize their Euclidean distances relative to a set of targets $\{b_i\}$ while staying within a desired distance from their neighbors in a network with the incidence matrix T as

$$\underset{x}{\mathrm{minimize}} \; \sum_i (x_i - b_i)^2 + I_{[-1,1]}(Tx). \tag{24}$$

Here, Tx is a vector of interagent distances that must be kept within an interval $[-1, 1]$. Applying primal-descent dual-ascent update rules to (24) achieves path planning for first-order agents $\dot{x} = u$ with $u = -\nabla_x \mathcal{L}_\mu(x; y)$. The proximal operator is projection onto a box, $\mathbf{prox}_{\mu I_{[-1,1]}}(z) = \max(\min(z, 1), -1)$, the Moreau envelope is the distance squared to that set, $M_{\mu I_{[-1,1]}}(z) = \frac{1}{2\mu}\sum \mathcal{S}_1^2(z_i)$, and $\nabla M_{\mu I_{[-1,1]}}(z) = \frac{1}{\mu}\mathcal{S}_1(z)$. To update its state, each agent x_i needs information from its neighbors in a network with a Laplacian $T^T T$.

Methods based on the subderivative are not applicable because the indicator function is not subdifferentiable. Proximal methods are hindered because the proximal operator of $I_{[-1,1]}(Tx)$ is difficult to compute due to T. Since $f(x) = \sum(x_i - b_i)^2$ is separable, a distributed ADMM implementation can be applied; however, it may require large discrete jumps in agent positions, which could be unsuitable for vehicles. Moreover, when f is not separable, a distributed implementation of the x-minimization step (5a) in ADMM would not be possible.

Fig. 4 shows an implementation for a problem with five agents whose set of targets changes position at time 5. The primal-descent dual-ascent dynamics (GF) are simulated in MATLAB using ode45.

VI. CONCLUDING REMARKS

For a class of nonsmooth composite optimization problems that arise in structured optimal control, we have introduced the continuously differentiable proximal augmented Lagrangian function. This function is obtained by collapsing the associated augmented Lagrangian onto the manifold resulting from explicit minimization over the variable in the nonsmooth part of the objective function. Our approach facilitates development of customized algorithms based on MM and the primal-descent dual-ascent method.

MM based on the proximal augmented Lagrangian is applicable to a broader class of problems than proximal gradient methods, and it has more robust convergence guarantees, more rigorous parameter update rules, and better practical performance than ADMM. The primal-descent dual-ascent gradient dynamics we proposed are suitable for distributed implementation and have a continuous right-hand side. When the differentiable component of the objective function is (strongly) convex, we establish global (exponential) asymptotic stability. Finally, we illustrate the efficacy of our algorithms using the edge addition and optimal placement problems. Future work will focus on developing second-order updates for the primal and dual variables and on providing an extension to nonconvex regularizers.

REFERENCES

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

[2] P. J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *Ann. Statist.*, vol. 36, pp. 199–227, 2008.

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

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

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

[6] 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. Learn.*, vol. 3, no. 1, pp. 1–124, 2011.

[7] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York, NY, USA: Academic, 1982.

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

[9] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York, NY, USA: Springer, 2006.

[10] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," arXiv:1610.04514.

[11] K. J. Arrow, L. Hurwicz, and H. Uzawa, *Studies in Linear and Non-Linear Programming*. Stanford, CA, USA: Stanford Univ. Press, 1958.

[12] J. Wang and N. Elia, "A control perspective for centralized and distributed convex optimization," in *Proc. IEEE 50th Conf. Decis. Control*, 2011, pp. 3800–3805.

[13] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[14] D. Feijer and F. Paganini, "Stability of primal–dual gradient dynamics and applications to network optimization," *Automatica*, vol. 46, no. 12, pp. 1974–1981, 2010.

[15] A. Cherukuri, E. Mallada, and J. Cortés, "Asymptotic convergence of constrained primal–dual dynamics," *Syst. Control Lett.*, vol. 87, pp. 10–15, 2016.

[16] A. Cherukuri, E. Mallada, S. Low, and J. Cortes, "The role of convexity on saddle-point dynamics: Lyapunov function and robustness," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2449–2464, Aug. 2018.

[17] A. Cherukuri, B. Gharesifard, and J. Cortés, "Saddle-point dynamics: Conditions for asymptotic stability of saddle points," *SIAM J. Control Optim.*, vol. 55, no. 1, pp. 486–511, 2017.

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

[19] B. Hu and P. Seiler, "Exponential decay rate conditions for uncertain linear systems using integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 61, no. 11, pp. 3631–3637, Nov. 2016.

[20] A. Megretski and A. Rantzer, "System analysis via integral quadratic constraints," *IEEE Trans. Autom. Control*, vol. 42, no. 6, pp. 819–830, Jun. 1997.

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

[22] 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.

[23] 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.

[24] F. Dörfler, M. R. Jovanović, M. Chertkov, and F. Bullo, "Sparse and optimal wide-area damping control in power networks," in *Proc. Amer. Control Conf.*, 2013, pp. 4295–4300.

[25] 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, Sep. 2014.

[26] 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.

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

[28] H. Li and Z. Lin, "Accelerated proximal gradient methods for nonconvex programming," in *Proc. Adv. Neural Inf. Process Syst.*, 2015, pp. 379–387.

[29] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.

[30] A. R. Conn, N. I. M. Gould, and P. L. Toint, "A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds," *SIAM J. Numer. Anal.*, vol. 28, pp. 545–572, 1991.

[31] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent," *Math. Program.*, vol. 155, no. 1–2, pp. 57–79, 2016.

[32] R. T. Rockafellar, "Augmented Lagrangians and applications of the proximal point algorithm in convex programming," *Math. Oper. Res.*, vol. 1, no. 2, pp. 97–116, 1976.

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

[34] N. K. Dhingra and M. R. Jovanović, "A method of multipliers algorithm for sparsity-promoting optimal control," in *Proc. Amer. Control Conf.*, Boston, MA, USA, 2016, pp. 1942–1947.

[35] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317. Berlin, Germany: Springer, 2009.

[36] J.-F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal, *Numerical Optimization: Theoretical and Practical Aspects*. Berlin, Germany: Springer, 2013.

[37] A. Rantzer, "On the Kalman-Yakubovich-Popov lemma," *Syst. Control Lett.*, vol. 28, no. 1, pp. 7–10, 1996.

[38] D. Ding, B. Hu, N. K. Dhingra, and M. R. Jovanović, "An exponentially convergent primal-dual algorithm for nonsmooth composite minimization," in *Proc. IEEE 57th Conf. Decis. Control*, Miami, FL, USA, 2018.

[39] G. Qu and N. Li, "On the exponential stability of primal-dual gradient dynamics," *IEEE Control Syst. Lett.*, vol. 3, no. 1, pp. 43–48, 2019.

[40] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Trans. Autom. Control*, vol. 59, no. 3, pp. 781–786, Mar. 2014.

[41] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, pp. 944–966, 2015.

[42] M. Mesbahi and M. Egerstedt, *Graph Theoretic Methods in Multiagent Networks*. Princeton, NJ, USA: Princeton Univ. Press, 2010.