2017 IEEE 56th Annual Conference on Decision and Control (CDC)
December 12-15, 2017, Melbourne, Australia

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

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

Abstract— **We develop a second order primal-dual algorithm for nonsmooth convex composite optimization problems in which the objective function is given by the sum of a twice differentiable term and a possibly non-differentiable regularizer. After introducing an auxiliary variable, we utilize the proximal operator of the nonsmooth regularizer to transform the associated augmented Lagrangian into a continuously differentiable function, the proximal augmented Lagrangian. We employ a generalization of the Jacobian to define second order updates on this function which locally converge quadratically/superlinearly to the optimal solution. We then use a merit function to develop a customized algorithm in which the search direction can be computed efficiently for large values of the regularization parameter. Finally, we illustrate the utility of our method using the ℓ_1-regularized least squares problem.**

I. Introduction

We study a class of composite optimization problems in which the objective function is 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. They often arise in structured feedback synthesis problems where it is desired to balance controller performance (e.g., the closed-loop $\mathcal{H}_2$ or $\mathcal{H}_\infty$ norm) with structural complexity [1], [2].

The lack of differentiability in the regularization term precludes the use of standard descent methods for smooth objective functions. Proximal gradient methods [3] and their accelerated variants [4] 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) [5]. 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 [6], we introduced the *proximal* augmented Lagrangian to enable the use of the standard method of multipliers for this reformulation, which leads to more robust convergence guarantees and better practical performance.

Although first order approaches are typically simple to implement, they tend to converge slowly to high-accuracy solutions. A generalization of Newton's method to nonsmooth problems was developed in [7]–[10], but it requires solving a regularized quadratic subproblem to determine a search direction. Related ideas have been successfully utilized in a number of applications, including sparse inverse covariance estimation in graphical models [11] and topology design in consensus networks [12].

Generalized Newton updates for identifying stationary points of (strongly) semismooth gradient mappings were first considered in [13]–[15] and employ a generalization of the Hessian for nonsmooth gradients. In [16]–[18] 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.

For smooth constrained optimization problems, recent work has extended the method of multipliers to incorporate second order updates of the primal and dual variables [19]–[21]. Since the optimal solution is the saddle point of the augmented Lagrangian, it is challenging to assess joint progress of the primal and dual iterates. In [19], Gill and Robinson introduced the primal-dual augmented Lagrangian which can serve as a merit function for measuring progress.

We draw on these advances to develop a second order primal-dual algorithm for nonsmooth composite optimization. Second order updates for the once-continuously differentiable proximal augmented Lagrangian are formed using a generalization of the Hessian. These updates have local (quadratic) superlinear convergence when the regularizer is associated with a (strongly) semismooth proximal operator. To guarantee convergence, we use the merit function employed in [20], [21] to assess algorithm progress.

Our paper is organized as follows. In Section II, we formulate the problem and provide background. In Section III, we derive the proximal augmented Lagrangian, the associated second order updates, and show fast asymptotic convergence rates. In Section IV, we provide a globally convergent algorithm. In Section V, we illustrate our approach with a LASSO problem and in Section VI we conclude the paper.

Financial support from the National Science Foundation under Awards ECCS-1739210 and CNS-1544887, the Air Force Office of Scientific Research under Award FA9550-16-1-0009, and the Institute for Mathematics and its Applications Postdoctoral Fellowship Program is acknowledged.

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

II. Problem formulation and background

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

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

where $T \in \mathbb{R}^{m\times n}$ has full row rank. We assume that the function f is strongly convex, twice continuously differentiable, that its gradient ∇f is Lipschitz continuous, and that g is convex, proper and lower semicontinuous. We now provide backround material and overview existing approaches.

A. Generalization of the gradient and Jacobian

The B-subdifferential set of a function $g\colon \mathbb{R}^m \to \mathbb{R}$ at a point $\bar{x}$ generalizes the notion of a gradient to functions which are nondifferentiable outside of a set C_g. Each element in the subdifferential set $\partial_B g(\bar{x})$ is the limit point of a sequence of gradients $\{\nabla g(x_k)\}$ evaluated at a sequence of points $\{x_k\} \subset C_g$ whose limit is $\bar{x}$,

$$\partial_B g(\bar{x}) := \{J_g \,|\, \exists \{x_k\} \subset C_g,\, \{x_k\} \to \bar{x}, \{\nabla g(x_k)\} \to J_g\}\,.$$

The Clarke subgradient of $g\colon \mathbb{R}^m \to \mathbb{R}$ is the convex hull of the B-subdifferential set [22], $\partial_C g(\bar{x}) := \mathrm{conv}(\partial_B g(\bar{x}))$. When g is a convex function, the Clarke subgradient is equal to the subdifferential set $\partial g(\bar{x})$ which defines the supporting hyperplanes of g at $\bar{x}$. For a function $G\colon \mathbb{R}^m \to \mathbb{R}^n$, the generalization of the Jacobian at a point $\bar{x}$ is given by $J_G = \left[\begin{array}{ccc} J_1 & \ldots & J_n \end{array}\right]^T$ where each $J_i \in \partial_C G_i(\bar{x})$ is a member of the Clarke subgradient of the ith component of G evaluated at $\bar{x}$.

The mapping $G\colon \mathbb{R}^m \to \mathbb{R}^n$ is semismooth at $\bar{x}$ if for any sequence $x^k \to \bar{x}$, the sequence of generalized Jacobians $J_{G_k} \in \partial_C G(x_k)$ provide a first order approximation of G,

$$\|G(x_k) \,-\, G(\bar{x}) \,+\, J_{G_k}(\bar{x} \,-\, x_k)\| \;=\; o(\|x_k \,-\, \bar{x}\|).$$

The function G is strongly semismooth if this approximation satisfies the condition,

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

where $o(\cdot)$ and $O(\cdot)$ denote, for some positive $\psi(k)$, that $\phi(k) = o(\psi(k))$ if $\phi(k)/\psi(k) \to 0$ and $\phi(k) = O(\psi(k))$ if $|\phi(k)| \le L\psi(k)$ for some constant L.

B. Proximal operators

The proximal operator of the function g is given by,

$$\mathbf{prox}_{\mu g}(v) \;:=\; \underset{z}{\mathrm{argmin}} \quad g(z) \,+\, 1/(2\mu)\,\|z \,-\, v\|^2 \tag{2a}$$

where μ is a positive parameter. It is Lipschitz continuous with parameter 1, differentiable almost everywhere, and firmly non-expansive [3]. Tthe proximal operators of many regularization functions are strongly semismooth, e.g., piecewise affine mappings such as the soft-thresholding operator, projection onto polyhedral sets, and projection onto symmetric cones. The value function associated with (2a) specifies the Moreau envelope of g,

$$M_{\mu g}(v) \;:=\; \inf_z \; g(z) \,+\, 1/(2\mu)\,\|z \,-\, v\|^2. \tag{2b}$$

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

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

For example, the proximal operator associated with the ℓ_1 norm, $g(z) = \sum |z_i|$, is determined by soft-thresholding, $\mathcal{S}_\mu(v) := \mathrm{sign}(v)\max\{|v|-\mu, 0\}$, the associated Moreau envelope is the Huber function, $M_{\mu g}(v) = \sum_i \{v_i^2/(2\mu), |v_i| \le \mu; |v_i| - \mu/2, |v_i| \ge \mu\}$ and its gradient is the saturation function, $\nabla M_{\mu g}(v) = \max\{-1, \min\{v/\mu, 1\}\}$.

C. First order methods

When $T = I$ or is diagonal, the proximal gradient method generalizes gradient descent for (1) [3], [4]. When $g = 0$, it simplifies to gradient descent and when g is the indicator function of a convex set, it simplifies to projected gradient. For the ℓ_1-regularized least-squares (LASSO) problem,

$$\underset{x}{\text{minimize}} \quad (1/2)\,\|Ax \,-\, b\|^2 \,+\, \gamma\,\|x\|_1 \tag{3}$$

the proximal gradient method is given by the Iterative Soft-Thresholding Algorithm (ISTA). Acceleration techniques can also be used to improve the convergence rate [4].

By introducing an auxiliary optimization variable z, optimization problem (1) can be rewritten as

$$\begin{array}{ll} \underset{x,\,z}{\text{minimize}} & f(x) \,+\, g(z) \\ \text{subject to} & Tx \,-\, z \,=\, 0. \end{array} \tag{4}$$

This reformulation is convenient for constrained optimization algorithms based on the augmented Lagrangian,

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

where y 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 (4).

The alternating direction method of multipliers (ADMM) is appealing for (4) because it leads to tractable subproblems [5], but it is strongly influenced by μ. The standard method of multipliers (MM) [23], [24] is more robust and has effective μ-update rules, but requires *joint* minimization of $\mathcal{L}_\mu$ over x and z. This nondifferentiable optimization problem has been recently cast in a differentiable form [6].

D. Second order methods

The potentially slow convergence of first order methods to high-accuracy solutions motivates the development of second order methods. A generalization of Newton's method, developed in [7]–[10], derives a search direction via a regularized quadratic subproblem. When g is the ℓ_1 norm, this amounts to solving a LASSO problem. In [16]–[18], the authors introduce the once-continuously differentiable Forward-Backward Envelope (FBE) and employ an approximate generalized Hessian to obtain second order updates for (1) when $T = I$. Second order methods have also been developed to find the saddle point of the augmented Lagrangian [19]–[21], but assume twice differentiability and Lipschitz continuous gradients/Hessians and thus cannot be applied to (4).

In this paper, we employ a generalized Hessian to form second order updates to the proximal augmented Lagrangian and employ a merit function to assess progress towards the solution of (4).

III. A SECOND ORDER PRIMAL-DUAL METHOD

We now derive the proximal augmented Lagrangian, the associated generalized second order updates, and show that they lead to fast local convergence.

A. Proximal augmented Lagrangian

Following [6], completion of squares can be used to write

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

Minimization of the augmented Lagrangian over z yields an explicit expression in terms of the proximal operator,

$$\operatorname*{argmin}_z \;\mathcal{L}_\mu(x,z;y) \;=\; z^\star_\mu(x,y) \;=\; \mathbf{prox}_{\mu g}(Tx \,+\, \mu y), \tag{5}$$

and substitution of (5) into the augmented Lagrangian provides an expression for $\mathcal{L}_\mu$ in terms of the Moreau envelope,

$$\begin{aligned}\mathcal{L}_\mu(x;y) \;&:=\; \mathcal{L}_\mu(x,z^\star_\mu(x,y);y)\\ &=\; f(x)+M_{\mu g}(Tx+\mu y)-(\mu/2)\|y\|^2.\end{aligned} \tag{6}$$

This expression, the *proximal* augmented Lagrangian, collapses $\mathcal{L}_\mu(x,z;y)$ onto the manifold resulting from the explicit minimization over z. Continuity of the gradient of the Moreau envelope (2c) guarantees continuous differentiability of the proximal augmented Lagrangian $\mathcal{L}_\mu(x;y)$,

$$\nabla\mathcal{L}_\mu(x;y) \;=\; \left[\begin{array}{c} \nabla f(x) \,+\, T^T\nabla M_{\mu g}(Tx+\mu y) \\ \mu\nabla M_{\mu g}(Tx+\mu y) \,-\, \mu y \end{array}\right] \tag{7}$$

and ensures that when $x^\star$ minimizes $\mathcal{L}_\mu(x;y)$ over x, $(x^\star, z^\star_\mu(x^\star,y))$ minimizes $\mathcal{L}_\mu(x,z;y)$ over (x,z), i.e.,

$$\operatorname*{argmin}_{x,\,z} \;\mathcal{L}_\mu(x,z;y^k) \;=\; \operatorname*{argmin}_x \;\mathcal{L}_\mu(x,z^\star_\mu(x,y^k);y^k)$$

which facilitates MM. Instead of fixing y and minimizing $\mathcal{L}_\mu(x;y)$ over x, the Arrow-Hurwicz-Uzawa method can be used to jointly update both variables [6] using the gradient (7). In this paper, we form second order updates to both x and y to achieve fast convergence to high accuracy solutions.

B. Second order updates

We use the Clarke subgradient set of the proximal operator, $\mathbb{P} := \partial_C\,\mathbf{prox}_{\mu g}(Tx+\mu y)$, to define

$$\partial^2_P\mathcal{L}_\mu = \left[\begin{array}{cc} H+(1/\mu)\,T^T(I\,-\,P)T & T^T(I\,-\,P) \\ (I\,-\,P)T & -\mu P \end{array}\right], \tag{8}$$

the set of generalized Hessians of $\mathcal{L}_\mu(x;y)$ where $H = \nabla^2 f(x)$ and $P \in \mathbb{P}$. Our generalization of the Hessian is inspired by [16]. In contrast to that work, however, we do not discard higher-order terms in forming (8).

When g is (block) separable, the matrix P is (block) diagonal and, since the proximal operator is firmly nonexpansive, $0 \preceq P \preceq I$. We introduce the composite variable, $w := [x^T\,y^T]^T$, use $\mathcal{L}_\mu(x;y)$ interchangeably with $\mathcal{L}_\mu(w)$, and suppress the dependance of P on w to reduce clutter.

We use (8) to obtain an update $\tilde{w}$ to the stationarity condition $\nabla\mathcal{L}_\mu(w)=0$ around the current iterate w^k,

$$\partial^2_P\mathcal{L}_\mu(w^k)\,\tilde{w}^k \;=\; -\nabla\mathcal{L}_\mu(w^k). \tag{9}$$

We next show that this update is well-defined.

Lemma 1: The generalized Hessian (8) of the proximal augmented Lagrangian is invertible for any choice of $P \in \mathbb{P}$ and it has n positive and m negative eigenvalues.

Proof: By the Haynsworth inertia additivity formula [25], the inertia of matrix (8) is determined by the sum of the inertias of matrices,

$$H \;+\; (1/\mu)T^T(I\,-\,P)T \tag{10a}$$

and

$$-\mu P-(I\,-\,P)T\left(H+(1/\mu)T^T(I-P)T\right)^{-1}T^T(I-P). \tag{10b}$$

Since $\mathbf{prox}_{\mu g}$ is firmly nonexpansive, both P and $I-P$ are positive semidefinite. The strong convexity of f implies that H and therefore (10a) are positive definite. Matrix (10b) is negative definite because the kernels of P and $I-P$ have no nontrivial intersection and T has full row rank. ■

Remark 1: The KKT conditions for problem (4) are,

$$0 \;=\; \nabla f(x) \,+\, T^Ty, \quad 0 \;=\; Tx \,-\, z, \quad 0 \;\ni\; \partial g(z) \,-\, y$$

Substituting $z^\star_\mu(x;y)$ makes the last two conditions redundant and renders (9) equivalent to a first order correction to the first and third condition. By premultiplying with,

$$\Pi \;:=\; \left[\begin{array}{cc} I & -(1/\mu)\,T^T \\ 0 & I \end{array}\right] \tag{11}$$

the second order update (9) can be expressed as,

$$\left[\begin{array}{cc} H & T^T \\ (I\,-\,P)T & -\mu P \end{array}\right]\left[\begin{array}{c}\tilde{x}^k\\ \tilde{y}^k\end{array}\right] = -\left[\begin{array}{c}\nabla f(x^k)+T^Ty^k\\ r^k\end{array}\right] \tag{12}$$

where $r^k := Tx^k - z^\star_\mu(x^k;y^k) = Tx^k - \mathbf{prox}_{\mu g}(Tx^k+\mu y)$ is the primal residual of (4).

C. Efficient computation of the Newton direction

We next demonstrate that the solution to (12) can be efficiently computed when $T=I$ and P is a sparse diagonal matrix whose entries are either 0 or 1. For example, when $g(z)=\gamma\|z\|_1$, larger values of γ are more likely to yield sparse P. The extension to P with entries between 0 and 1 or a general diagonal T follow from similar arguments.

We write the system of equations (12) as,

$$\left[\begin{array}{cc} H & I \\ I-P & -\mu P \end{array}\right]\left[\begin{array}{c}\tilde{x}\\ \tilde{y}\end{array}\right] \;=\; \left[\begin{array}{c}\vartheta\\ \theta\end{array}\right], \tag{13}$$

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

$$H \;=\; \left[\begin{array}{cc} H_{11} & H_{12} \\ H_{12}^T & H_{22}\end{array}\right], \qquad P \;=\; \left[\begin{array}{cc} I & 0 \\ 0 & 0\end{array}\right].$$

We use v_1 (v_2) to denote the subvector of v corresponding to the entries of P which are equal to 1 (0). Here, v is used

to denote either the primal variable x or the dual variable y.

Note that $(I-P)v = 0$ when $v_2 = 0$, and $Pv = 0$ when $v_1 = 0$. As a result, the bottom row of (13) is $-\mu\tilde{y}_1 = \theta_1$ and $\tilde{x}_2 = \theta_2$. Substituting $\tilde{x}_2$ and $\tilde{y}_1$ into (13) yields,

$$H_{11}\tilde{x}_1 \;=\; \vartheta_1 \;+\; H_{12}\tilde{x}_2 \;+\; \tilde{y}_1 \tag{14}$$

which must be solved via a matrix inversion. The rest can be computed using only matrix-vector products, $\tilde{y}_2 = -\left(\vartheta_2 + H_{21}\tilde{x}_1 + H_{22}\tilde{x}_2\right)$. The major computational burden in solving (12) thus lies in (14) and is more efficient when P is sparse. For example, in ℓ_1 regularized or box-constrained problems, P is sparser when the weight on the ℓ_1 norm is larger or the box constraints are tighter.

D. Asymptotic rate of convergence

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

Proposition 2: Let the proximal operator associated with the regularization function g in (4) be (strongly) semismooth, and let $\tilde{w}^k$ be defined by (9). Then, for any $P \in \mathbb{P}$, 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 1 establishes that $\partial_P^2\,\mathcal{L}_\mu$ is nonsingular for any $P \in \mathbb{P}$. The gradient of the proximal augmented Lagrangian, $\nabla\mathcal{L}_\mu$, is Lipschitz continous because ∇f and the $\mathbf{prox}_{\mu g}$ are Lipschitz continuous. Nonsingularity of $\partial_P^2\mathcal{L}_\mu$, Lipschitz continuity of $\nabla\mathcal{L}_\mu$, and (strong) semismoothness of the proximal operator establish (quadratic) superlinear convergence of the iterates by [14, Theorem 3.2]. ∎

IV. A PRIMAL-DUAL ALGORITHM

Proposition 2 establishes fast local convergence of the second order iterates to the saddle point of $\mathcal{L}_\mu$. To establish global convergence, it is necessary to limit the stepsize α_k in $w^{k+1} \;=\; w^k + \alpha_k\tilde{w}^k$. This is difficult for saddle point problems because standard notions, such as sufficient descent of $\mathcal{L}_\mu$, cannot be employed to assess progress of the iterates.

We employ the primal-dual augmented Lagrangian introduced in [19] as a merit function to evaluate progress towards the saddle point. Drawing upon recent advancements for constrained optimization [19]–[21], we show global convergence under a boundedness assumption on the sequence of gradients. This assumption is standard for augmented Lagrangian based methods [21], [24].

A. Merit function

The primal-dual augmented Lagrangian,

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

was introduced in [19], where λ is an estimate of the optimal Lagrange multiplier $y^\star$. It can be shown that the optimal primal-dual pair $(x^\star, z^\star; y^\star)$ of optimization problem (4) is a stationary point of $\mathcal{V}_\mu(x,z;y,y^\star)$ [19, Theorem 3.1]. Furthermore, $\mathcal{V}_\mu$ is convex with respect to $(x,z;y)$ and, for any fixed λ, there is a unique global minimizer.

In contrast to [19], we study problems in which a component of the objective function is not differentiable. The Moreau envelope associated with the non-differentiable component g allows us to eliminate the dependence of the primal-dual augmented Lagrangian $\mathcal{V}_\mu$ on z,

$$\hat{z}_\mu^\star(x;y,\lambda) \;=\; \mathbf{prox}_{(\mu/2)g}(Tx \;+\; (\mu/2)(2\lambda \;-\; y))$$

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

$$\begin{aligned}\mathcal{V}_\mu(x;y,\lambda) \;=\; & f(x) \;+\; M_{(\mu/2)g}(Tx \;+\; (\mu/2)(2\lambda \;-\; y)) \\ & +\; (\mu/4)\,\|y\|^2 \;-\; (\mu/2)\,\|\lambda\|^2.\end{aligned} \tag{15}$$

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

In [19], the authors obtain a search direction using the Hessian of the merit function, $\nabla^2\mathcal{V}_\mu$. Instead of implementing the analogous update using generalized Hessian $\partial_P^2\mathcal{V}_\mu$ of semismooth $\nabla\mathcal{V}_\mu$, we take advantage of the efficient inversion of $\partial_P^2\mathcal{L}_\mu$ to define the update

$$\partial_P^2\mathcal{L}_\mu(w)\,\tilde{w} \;=\; -\,\mathrm{blkdiag}(I,-I)\nabla\mathcal{V}_{2\mu}(w) \tag{16}$$

where the multiplication by $\mathrm{blkdiag}(I,-I)$ is used to ensure descent in the dual direction and $\mathcal{V}_{2\mu}$ is employed because $\hat{z}_\mu^\star$ is given by the proximal operator associated with $\mu/2$. When $\lambda = y$, (16) is equivalent to second order update (9).

Lemma 3: Let $\tilde{w}$ solve the system of equations obtained by premultiplying (16) by the matrix Π given by (11),

$$\begin{bmatrix} H & T^T \\ (I \;-\; P)T & -\mu P \end{bmatrix}\tilde{w} \;=\; -\begin{bmatrix} \nabla f(x) \;+\; T^T y \\ s \;+\; 2\mu\,(\lambda \;-\; y) \end{bmatrix} \tag{17}$$

where $H \;:=\; \nabla^2 f(x) \;\succ\; 0$ and $s \;:=\; Tx \;-\; \mathbf{prox}_{\mu g}(Tx + \mu(2\lambda - y))$. Then, for any $\sigma \in (0,1]$,

$$d \;:=\; (1 \;-\; \sigma)\,\tilde{w} \;-\; \sigma\,\nabla\mathcal{V}_{2\mu}(w) \tag{18}$$

is a descent direction of the merit function $\mathcal{V}_{2\mu}(w)$ for the fixed Lagrange multiplier estimate λ.

Proof: The gradient $\nabla\mathcal{V}_{2\mu}(w)$ of the primal-dual augmented Lagrangian with penalty parameter 2μ is,

$$\nabla\mathcal{V}_{2\mu}(w) = \begin{bmatrix} \nabla f(x) + (1/\mu)\,T^T(s + \mu(2\lambda - y)) \\ -(s + 2\mu(\lambda - y)) \end{bmatrix}. \tag{19}$$

Using (17), $\nabla\mathcal{V}_{2\mu}(w)$ can be expressed as,

$$\begin{bmatrix} -(H\tilde{x} + \tilde{y}) - (1/\mu)T^T(I - P)T\tilde{x} + T^T P\tilde{y} \\ (I \;-\; P)T\tilde{x} \;-\; \mu P\tilde{y} \end{bmatrix}.$$

Thus, the inner product,

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

is negative semidefinite, and

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

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

B. Algorithm

We now develop a customized algorithm that alternates between minimizing the merit function $\mathcal{V}_\mu(w,\lambda)$ over w and updating λ. Inspired by [26], we ensure sufficient progress

Algorithm 1 Second order primal-dual algorithm

input: Initial point 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 $\|s^k\| \leq \eta \|s^{k-1}\|$ (20)

go to Step 2a. If not, go to Step 2b.
Step 2a**:** Set $\mu_{k+1} = \tau_a \mu_k$, $\lambda^{k+1} = y^k$
Step 2b**:** Set $\mu_{k+1} = \tau_b \mu_k$, $\lambda^{k+1} = \lambda^k$
Step 3**:** Using a backtracking line search, perform a sequence of inner iterations to choose w^{k+1} until

$$\|\nabla \mathcal{V}_{2\mu_{k+1}}(w^{k+1}, \lambda^{k+1})\| \leq \epsilon_k \quad (21)$$

where search direction d is obtained using (17)–(18) with

$$\sigma = 0 \quad \frac{\langle \tilde{w}^k, \nabla \mathcal{V}_{2\mu_{k+1}}(w^k) \rangle}{\|\nabla \mathcal{V}_{2\mu_{k+1}}(w^k)\|^2} \leq -\beta, \quad (22a)$$

$$\sigma \in (0,1] \quad \text{otherwise.} \quad (22b)$$

with damped second order updates. Note that

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

appear in the proof and that r is the primal residual of (4).

Theorem 4: 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 (4) 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 (21) in Algorithm 1 will be satisfied after finite number of iterations. Combining (21) and (19) 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 (4) 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 (20) 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

$$\begin{aligned} T^T s^k = \; & 2\mu^k(\nabla f(x^k) + \tfrac{1}{\mu^k} T^T(s^k + \mu^k(2\lambda^{k_0} - y^k))) \\ & - 2\mu^k \nabla f(x^k) - T^T(s^k + 2\mu^k(\lambda^{k_0} - y^k)) \\ & - 2\mu^k T^T \lambda^{k_0}. \end{aligned}$$

Taking the norm of each side and applying the triangle inequality, (21) and (19) yield

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

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 (19) and applying (21) 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. ■

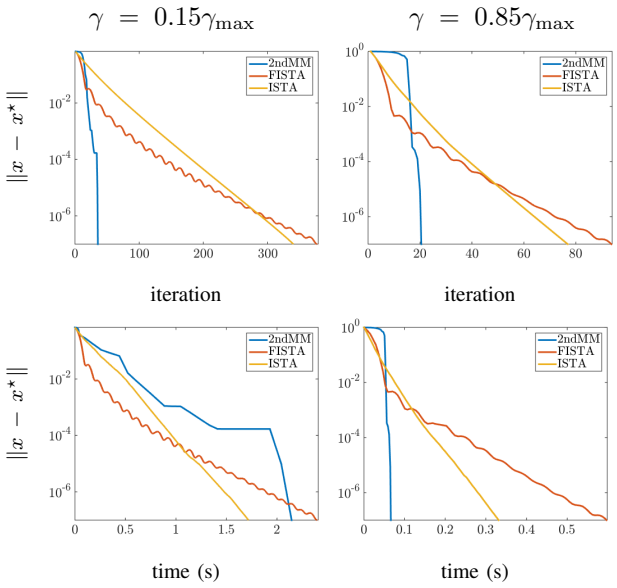

Fig. 1: Number of iterations and solve time required to compute solution to LASSO for two values of γ using ISTA, FISTA, and our algorithm (2ndMM).

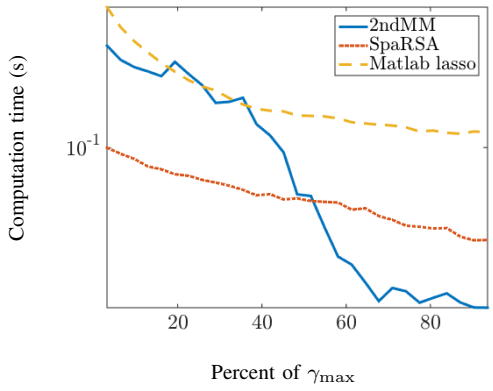

Fig. 2: Time to compute LASSO using ISTA, FISTA, and our algorithm (2ndMM) for different values of sparsity-promoting parameter γ.

V. Computational Experiments

The LASSO problem (3) regularizes a least squares objective with a γ-weighted ℓ_1 penalty. The proximal operator of g 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 $x_i + \mu y_i \in (-\gamma\mu, \gamma\mu)$, 1 outside the interval, and between 0 and 1 on the boundary. Larger values of γ induce sparser solutions for which one can expect a sparser sequence of iterates. Note that we require strong convexity, i.e. $A^T A \succ 0$.

In Fig. 1, we show the distance of the iterates from the optimal for ISTA, FISTA, and our algorithm for a represen-

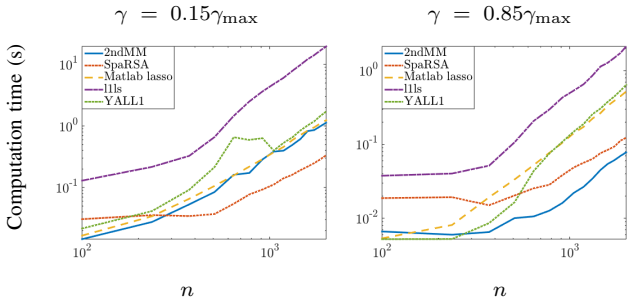

Fig. 3: Comparison of our algorithm (2ndMM) with state-of-the-art methods for LASSO problems with problem dimension varying from $n = 100$ to 2000.

tative LASSO 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 clock time. Although our method always requires much fewer iterations, it is most effective when γ is large and thus the search direction is cheap to compute; see Section III-C for details. In Fig. 2, we show the computation speed for $n = 1000$ as γ ranges from 0 to $\gamma_{\max}$ where $\gamma_{\max} = \|A^T b\|_\infty$ induces 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 [27]), SpaRSA [28], an interior point method [29], and YALL1 [30]. Problem instances were randomly generated with $A \in \mathbb{R}^{m\times n}$, n ranging from 100 to 2000, $m = 3n$, and $\gamma = 0.15\gamma_{\max}$ and $0.85\gamma_{\max}$. Our algorithm is competitive with these state-of-the-art methods, and is the fastest for larger values of γ when the second order search direction (17) is cheaper to compute.

VI. Concluding Remarks

We have developed a second order primal-dual algorithm for nonsmooth convex composite optimization problems. After introducing an auxiliary variable, we transform the associated augmented Lagrangian into the once continuously differentiable proximal augmented Lagrangian and form second order primal and dual updates using the generalized Hessian. These updates can be efficiently computed when the emphasis on the nonsmooth term is large. We establish global convergence using the primal-dual augmented Lagrangian as a merit function. Finally, an ℓ_1 regularized least squares example demonstrates the competitive performance of our algorithm relative to the available state-of-the-art alternatives.

References

[1] 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, 2013.

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

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

[4] 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.

[5] 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.

[6] N. K. Dhingra, S. Z. Khong, and M. R. Jovanović, "The proximal augmented Lagrangian method for nonsmooth composite optimization," 2016, arXiv:1610.04514.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

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

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

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

[18] 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.

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

[20] 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.

[21] 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.

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

[23] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. New York: Academic Press, 1982.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.