I. INTRODUCTION

Background: Recently, several researchers have concentrated on the minimization of a sum of two convex objective components, where the first component is usually assumed to be Lipschitz continuously differentiable and convex, and the second term is possibly nonsmooth and convex. Mathematically, this problem can be formulated as follows:

\[ \min_{x \in \mathbb{R}^n} \{ F(x) := f(x) + g(x) \}. \] (1)

In principle, diverse convex optimization methods can be used to solve (1). However, it has been recognized that exploiting specific structure of such problems would lead to more efficient methods than general ones. Two common approaches dominate in solving problem (1): (i) optimization using splitting techniques due to the composite structure of (1) and, (ii) methods based on augmented Lagrangian combined with decomposition techniques.

Contributions: In this work, we study a special case of (1), where \( f(x) \) is assumed to be self-concordant. Under this setting, the gradient of \( f \) is no longer Lipschitz continuous; thus, the fast gradient framework [1] is not directly applicable under this setting. Problem (1) under the self-concordant assumption also includes many important instances in machine learning and data analysis such as graph selection and data clustering [2], [3].

Our contribution can be summarized as follows:

(i) We propose a new proximal-gradient-type method with explicit step-size selection strategy at each iteration. The main advantage of this method is that it does not require any backtracking line search to determine the step-size. Moreover, we show that our step size is optimal in the gradient-based methods.

(ii) We show that, if \( g \) has bounded subgradients, the proximal-gradient-type method can be accelerated by using a similar framework as in [4], [5].

(iii) We extend our idea to propose a second-order proximal-Newton-type method and derive a two-phase algorithm. Similar to the first-order method, we do not require any globalization technique for step-size selection in this algorithm.

II. A SHORT DESCRIPTION OF METHODS

Let \( g \) be a proper, lower semicontinuous and convex. We define \( P_\gamma(x) = (I + \gamma \partial g)^{-1}(x) := \arg \min_{y \in \mathbb{R}^n} \left\{ g(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\} \). The proximal-descent-type method can be simply represented as:

\[
\begin{align*}
    y^{k+1} &:= P_{\alpha_k} (x^k - \alpha_k \nabla f(x^k)), \\
    x^{k+1} &:= \psi_k(y^{k+1}, y^k).
\end{align*}
\] (2)

Here \( \alpha_k \in (0, 1) \) is a given step size and \( H_k \) is a directional scaling matrix. Depending on the choice of \( H_k \) and the mapping \( \psi_k \) we obtain different methods.

Proximal-gradient method: If we choose \( \psi_k(y^{k+1}, y^k) = y^{k+1} \) and \( H_k = I_n \), we obtain a standard proximal-gradient method; cf. [5]. In the extended version of this abstract, the step size \( \alpha_k \) is given explicitly and satisfies the following descent condition.

Theorem 1. Suppose that \( \{x^k\}_{k \geq 0} \) is a sequence generated by our proximal-gradient method and the step size \( \alpha_k \) satisfies a explicit closed-form equation. Then

\[ F(x^{k+1}) \leq F(x^k) - c(x^k, x^{k+1}), \] (3)

where \( c(\cdot) \) is a given metric. Moreover, the selection strategy for step-size \( \alpha_k \) is optimal.

Our step-size selection does not require any backtracking line-search procedure. Moreover, in the case where \( \psi_k(y^{k+1}, y^k) = y^{k+1} + \tau_k(y^{k+1} - y^k) \), we obtain a fast proximal-gradient method.

Proximal-Newton method: In the case where \( H_k = \nabla^2 f(x^k)^{-1} \) and \( \psi_k(y^{k+1}, y^k) = y^{k+1} \), we obtain a proximal-Newton method with the following convergence guarantees.

Theorem 2. Let \( \alpha_k := (1 + \lambda_k)^{-1} \in (0, 1) \) where \( \lambda_k < 1 - 1/\sqrt{2} \).

Then, the proposed proximal-Newton schemes generates a new point \( x^{k+1} \) such that

\[ F(x^{k+1}) \leq F(x^k) - \omega(\lambda_k), \] (4)

where \( \omega(\lambda_k) > 0 \). Moreover, the step \( \alpha_k = (1 + \lambda_k)^{-1} \) is optimal.

The local convergence rate of this method is quadratic.

In the extended version of this work, we investigate the convergence and the computational-complexity of the proposed methods.

III. APPLICATIONS

To illustrate the effectiveness of our framework in practice, we consider the graph selection problem and the sparse minimization problem under max-norm constraints for clustering problems. These problems can be cast in two the framework of (1) by using matrix variables. For numerical experiments we refer the reader to our full paper at http://arxiv.org/abs/1301.1459.

REFERENCES


