Computational Optimization and Applications manuscript No.
(will be inserted by the editor)

Inexact Successive Quadratic Approximation for Regularized Optimization

Ching-pei Lee · Stephen J. Wright

received: March 2018 / accepted: January 2019

Abstract Successive quadratic approximations, or second-order proximal methods, are useful for minimizing functions that are a sum of a smooth part and a convex, possibly nonsmooth part that promotes regularization. Most analyses of iteration complexity focus on the special case of proximal gradient method, or accelerated variants thereof. There have been only a few studies of methods that use a second-order approximation to the smooth part, due in part to the difficulty of obtaining closed-form solutions to the subproblems at each iteration. In fact, iterative algorithms may need to be used to find inexact solutions to these subproblems. In this work, we present global analysis of the iteration complexity of inexact successive quadratic approximation methods, showing that an inexact solution of the subproblem that is within a fixed multiplicative precision of optimality suffices to guarantee the same order of convergence rate as the exact version, with complexity related in an intuitive way to the measure of inexactness. Our result allows flexible choices of the second-order term, including Newton and quasi-Newton choices, and does not necessarily require increasing precision of the subproblem solution on later iterations. For problems exhibiting a property related to strong convexity, the algorithms converge at global linear rates. For general convex problems, the convergence rate is linear in early stages, while the overall rate is $O(1/k)$. For nonconvex problems, a first-order optimality criterion converges to zero at a rate of $O(1/\sqrt{k})$.

This work was supported by NSF awards 1447449, 1628384, 1634579, and 1740707; Subcontracts and 3F-30222 and 8F-30039 from Argonne National Laboratory; and Award N660011824020 from the DARPA Lagrange Program.

Ching-pei Lee
E-mail: ching-pei@cs.wisc.edu
Stephen J. Wright
E-mail: swright@cs.wisc.edu
Computer Sciences Department and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA

Keywords Convex optimization · Nonconvex optimization · Regularized optimization · Variable metric · Proximal method · Second-order approximation · Inexact method

1 Introduction

We consider the following regularized optimization problem:

$$\min_x F(x) \coloneqq f(x) + \psi(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz-continuously differentiable, and $\psi : \mathbb{R}^n \to \mathbb{R}$ is convex, extended-valued, proper, and closed, but might be nondifferentiable. Moreover, we assume that F is lower-bounded and the solution set Ω of (1) is non-empty. Unlike the many other works on this topic, we focus on the case in which ψ does *not* necessarily have a simple structure, such as (block) separability, which allows a prox-operator to be calculated economically, often in closed form. Rather, we assume that subproblems that involve ψ explicitly are solved inexactly, by an iterative process.

Problems of the form (1) arise in many contexts. The function ψ could be an indicator function for a trust region or a convex feasible set. It could be a multiple of an ℓ_1 norm or a sum-of-ℓ_2 norms. It could be the nuclear norm for a matrix variable, or the sum of absolute values of the elements of a matrix. It could be a smooth convex function, such as $\|\cdot\|_2^2$ or the squared Frobenius norm of a matrix. Finally, it could be a combination of several of these elements, as happens when different types of structure are present in the solution. In some of these situations, the prox-operator involving ψ is expensive to calculate exactly.

We consider algorithms that generate a sequence $\{x^k\}_{k=0,1,\dots}$ from some starting point x^0, and solve the following subproblem inexactly at each iteration, for some symmetric matrix H_k:

$$\arg\min_{d \in \mathbb{R}^n} Q_{H_k}^{x^k}(d) \coloneqq \nabla f\left(x^k\right)^T d + \frac{1}{2} d^T H_k d + \psi\left(x^k + d\right) - \psi\left(x^k\right). \tag{2}$$

We abbreviate the objective in (2) as $Q_k(\cdot)$ (or as $Q(\cdot)$ when we focus on the inner workings of iteration k). In some results, we allow H_k to have zero or negative eigenvalues, provided that Q_k itself is strongly convex. (Strong convexity in ψ may overcome any lack of strong convexity in the quadratic part of (2).)

In the special case of the proximal-gradient algorithm [8,34], where H_k is a positive multiple of the identity, the subproblem (2) can often be solved cheaply, particularly when ψ is (block) separable, by means of a prox-operator involving ψ. For more general choices of H_k, or for more complicated regularization functions ψ, it may make sense to solve (2) by an iterative process, such as accelerated proximal gradient or coordinate descent. Since it may be too expensive to run this iterative process to obtain a high-accuracy solution of (2), we consider the possibility of an inexact solution. In this paper, we assume that the inexact solution satisfies the following condition, for some constant $\eta \in [0,1)$:

$$Q\left(d\right) - Q^* \leq \eta \left(Q\left(0\right) - Q^*\right) \quad \Leftrightarrow \quad Q(d) \leq (1-\eta) Q^*, \tag{3}$$

where $Q^* := \inf_d Q(d)$ and $Q(0) = 0$. The value $\eta = 0$ corresponds to exact solution of (2). Other values $\eta \in (0,1)$ indicate solutions that are inexact to within a *multiplicative* constant.

The condition (3) is studied in [2, Section 4.1], which applies a primal-dual approach to (2) to satisfy it. In this connection, note that if we have access to a lower bound $Q_{LB} \leq Q^*$ (obtained by finding a feasible point for the dual of (2), or other means), then any d satisfying $Q(d) \leq (1-\eta) Q_{LB}$ also satisfies (3).

In practical situations, we need not enforce (3) explicitly for some chosen value of η. In fact, we do not necessarily require η to be known, or (3) to be checked at all. Rather, we can take advantage of the convergence rates of whatever solver is applied to (2) to ensure that (3) holds for *some* value of $\eta \in (0,1)$, possibly unknown. For instance, if we apply an iterative solver to the strongly convex function Q in (2) that converges at a global linear rate $(1-\tau)$, then the "inner" iteration sequence $\{d^{(t)}\}_{t=0,1,\dots}$ (starting from some $d^{(0)}$ with $Q(d^{(0)}) \leq 0$) satisfies

$$Q(d^{(t)}) - Q^* \leq (1-\tau)^t \left(Q\left(0\right) - Q^*\right), \quad t = 0, 1, 2, \dots. \tag{4}$$

If we fix the number of inner iterations at T (say), then $d^{(T)}$ satisfies (3) with $\eta = (1-\tau)^T$. Although τ might be unknown as well, we can implicitly tune the accuracy of the solution by adjusting T. On the other hand, if we wish to attain a certain target accuracy η and have an estimate of rate τ, we can choose the number of iterations T large enough that $(1-\tau)^T \leq \eta$. Note that τ depends on the extreme eigenvalues of H_k in some algorithms; we can therefore choose H_k to ensure that τ is restricted to a certain range for all k.

Empirically, we observe that Q-linear methods for solving (2) often have rapid convergence in their early stages, with slower convergence later. Thus, a moderate value of η may be preferable to a smaller value, because moderate accuracy is attainable in disproportionately fewer iterations than high accuracy.

A practical stopping condition for the subproblem solver in our framework is just to set a fixed number of iterations, provided that a linearly convergent method is used to solve (2). This guideline can be combined with other more sophisticated approaches, possibly adjusting the number of inner iterations (and hence implicitly η) according to some heuristics. For simplicity, our analysis assumes a fixed choice of $\eta \in (0,1)$. We examine in particular the number of outer iterations required to solve (1) to a given accuracy ϵ. We show that the dependence of the iteration complexity on the inexactness measure η is benign, increasing only modestly with η over approaches that require exact solution of (2) for each k.

1.1 Quadratic Approximation Algorithms

To build complete algorithms around the subproblem (2), we either do a step size line search along the inexact solution d^k, or adjust H_k and recompute d^k, seeking in both cases to satisfy a familiar "sufficient decrease" criterion. We present two algorithms that reflect each of these approaches. The first uses a line search approach on the step size with a modified Armijo rule, as presented in [32]. We consider a backtracking line-search procedure for simplicity; the analysis could be adapted for more sophisticated procedures. Given the current point x^k, the update direction d^k and parameters $\beta, \gamma \in (0,1)$, backtracking finds the smallest nonnegative integer i such that the step size $\alpha_k = \beta^i$ satisfies

$$F\left(x^k + \alpha_k d^k\right) \leq F\left(x^k\right) + \alpha_k \gamma \Delta_k, \tag{5}$$

where

$$\Delta_k \coloneqq \nabla f\left(x^k\right)^T d^k + \psi\left(x^k + d^k\right) - \psi\left(x^k\right). \tag{6}$$

This version appears as Algorithm 1. The exact version of this algorithm can be considered as a special case of the block-coordinate descent algorithm of [32].[1] In [2], Algorithm 1 (with possibly a different criterion on d^k) is called the "variable metric inexact line-search-based method". (We avoid the term "metric" because we consider the possibility of indefinite H_k in some of our results.) More complicated metrics, not representable by a matrix norm, were also considered in [2]. Since our analysis makes use only of the smallest and largest eigenvalues of H_k (which correspond to the strong convexity and Lipschitz continuity parameters of the quadratic approximation term), we could also generalize our approach to this setting. We present only the matrix-representable case, however, as it allows a more direct comparison with the second algorithm presented next.

Algorithm 1 Inexact Successive Quadratic Approximation with Backtracking Line Search

```
Given $\beta, \gamma \in (0,1)$, $x^0 \in \mathbb{R}^n$;
for $k = 0, 1, 2, \ldots$ do
    Choose a symmetric $H_k$ that makes $Q_k$ strongly convex;
    Obtain from (2) a vector $d^k$ satisfying (3), for some fixed $\eta \in [0,1)$;
    Compute $\Delta_k$ by (6);
    $\alpha_k \leftarrow 1$;
    while (5) is not satisfied do
        $\alpha_k \leftarrow \beta\alpha_k$;
    $x^{k+1} \leftarrow x^k + \alpha_k d^k$;
```

The second algorithm uses the following sufficient decrease criterion from [29,12]:

$$F(x) - F(x+d) \geq -\gamma Q_H^x(d) \geq 0, \tag{7}$$

[1] The definition of Δ in [32] contains another term $\omega d^T H d/2$, where $\omega \in [0,1)$ is a parameter. We take $\omega = 0$ for simplicity, but our analysis can be extended in a straightforward way to the case of $\omega \in (0,1)$.

for a given parameter $\gamma \in (0,1]$. If this criterion is not satisfied, the algorithm modifies H and recomputes d^k. The criterion (7) is identical to that used by trust-region methods (see, for example, [27, Chapter 4]), in that the ratio between the actual objective decrease and the decrease predicted by Q is bounded below by γ; that is,

$$\frac{F(x) - F(x+d)}{Q_H^x(0) - Q_H^x(d)} \geq \gamma.$$

We consider two variants of modifying H such that (7) is satisfied. The first successively increases H by a factor β^{-1} (for some parameter $\beta \in (0,1)$) until (7) holds. We require in this variant that the initial choice of H is positive definite, so that all eigenvalues grow by a factor of β^{-1} at each multiplication. The second variant uses a similar strategy, except that H is modified by adding a successively larger multiple of the identity, until (7) holds. (This algorithm allows negative eigenvalues in the initial estimate of H.) These two approaches are defined as the first and the second variants of Algorithm 2, respectively.

Algorithm 2 Inexact Successive Quadratic Approximation with Modification of the Quadratic Term

```
Given $\beta, \gamma \in (0,1]$, $x^0 \in \mathbb{R}^n$;
for $k = 0, 1, 2, \ldots$ do
    if Variant 1 then Choose $H_k^0 \succ 0$;
    if Variant 2 then Choose a suitable $H_k^0$;
    $\alpha_k \leftarrow 1$, $H_k \leftarrow H_k^0$;
    Obtain from (2) a vector $d^k$ satisfying (3), for some fixed $\eta \in [0,1)$;
    while (7) is not satisfied do
        if Variant 1 then $\alpha_k \leftarrow \beta\alpha_k$, $H_k \leftarrow H_k^0/\alpha_k$;
        if Variant 2 then $H_k \leftarrow H_k^0 + \alpha_k^{-1} I$, $\alpha_k \leftarrow \beta\alpha_k$;
        Obtain from (2) a vector $d^k$ satisfying (3);
    $x^{k+1} \leftarrow x^k + d^k$;
```

Algorithm 1 and Variant 1 of Algorithm 2 are direct extensions of backtracking line search in the smooth case, in the sense that when ψ is not present, both approaches are identical to shrinking the step size. However, aside from the sufficient decrease criteria, the two differ when the regularization term is present.

The second variant of Algorithm 2 is similar to the method proposed in [29, 12], with the only difference being the inexactness criterion of the subproblem solution. This variant of modifying H can be seen as interpolating between the step from the original H and the proximal gradient step. It is also a generalization of the trust-region technique for smooth optimization. When ψ is not present, adding a multiple of the identity to H in (2) is equivalent to shrinking the trust region [23]. We can therefore think of Algorithm 2, Variant 2 as a generalized trust-region approach for regularized problems.

Rather than our multiplicative criterion (3), the works [29,12] use an *additive* criterion to measure inexactness of the solution. In the analysis of [29, 12], this tolerance must then be reduced to zero at a certain rate as the algorithm progresses, resulting in growth of the number of inner iterations per outer iteration as the algorithms progress. By contrast, we attain satisfactory performance (both in theory and practice) for a fixed value $\eta \in (0,1)$ in (3).

Which of the algorithms described above is "best" depends on the circumstances. When (2) is expensive to solve, Algorithm 1 may be preferred, as it requires approximate solution of this subproblem just once on each outer iteration. On the other hand, when ψ has special properties, such as inducing sparsity or low rank in x, Algorithm 2 might benefit from working with sparse iterates and solving the subproblem in spaces of reduced dimension.

Variants and special cases of the algorithms above have been discussed extensively in the literature. Proximal gradient algorithms have $H = \xi I$ for some $\xi > 0$ [8,34]; proximal-Newton uses $H = \nabla^2 f$ [17,28,19]; proximal-quasi-Newton and variable metric use quasi-Newton approximations for H_k [29, 12]. The term "successive quadratic approximation" is also used by [6]. Our methods can even be viewed as a special case of block-coordinate descent [32] with a single block. The key difference in this work is the use of the inexactness criterion (3), while existing works either assume exact solution of (2), or use a different criterion that requires increasing accuracy as the number of outer iterations grows. Some of these works provide only an asymptotic convergence guarantee and a local convergence rate, with a lack of clarity about when the fast local convergence rate will take effect. An exception is [2], which also makes use of the condition (3). However, [2] gives convergence rate only for convex f and requires existence of a scalar $\mu \geq 1$ and a sequence $\{\zeta_k\}$ such that

$$\sum_{k=0}^{\infty} \zeta_k < \infty, \quad \zeta_k \geq 0, \quad H_{k+1} \preceq (1+\zeta_k) H_k, \quad \mu I \succeq H_k \succeq \frac{1}{\mu} I, \quad \forall k, \tag{8}$$

where $A \succeq B$ means that $A - B$ is positive semidefinite. This condition may preclude such useful and practical choices of H_k as the Hessian and quasi-Newton approximations. We believe that our setting may be more general, practical, and straightforward in some situations.

1.2 Contribution

This paper shows that, when the initial value of H_k at all outer iterations k is chosen appropriately, and that (3) is satisfied for all iterations, then the objectives of the two algorithms converge at a global Q-linear rate under an "optimal set strong convexity" condition defined in (10), and at a sublinear rate for general convex functions. When F is nonconvex, we show sublinear convergence of the first-order optimality condition. Moreover, to discuss the relation between the subproblem solution precision and the convergence rate, we show that the iteration complexity is proportional to either $1/(1-\eta)$ or

$1/(2(1-\sqrt{\eta}))$, depending on the properties of f and ψ, and the algorithm parameter choices.[2]

In comparison to existing works, our major contributions are as follows.

- We quantify how the inexactness criterion (3) affects the step size of Algorithm 1, the norm of the final H in Algorithm 2, and the iteration complexity of these algorithms. We discuss why the process for finding a suitable value of α_k in each algorithm can potentially improve the convergence speed when the quadratic approximations incorporate curvature information, leading to acceptance of step sizes whose values are close to one.
- We provide a global convergence rate result on the first-order optimality condition for the case of nonconvex f in (1) for general choices of H_k, without assumptions beyond the Lipschitzness of ∇f.
- The global R-linear convergence case of a similar algorithm in [12] when F is strongly convex is improved to a global Q-linear convergence result for a broader class of problems.
- For general convex problems, in addition to the known sublinear ($1/k$) convergence rate, we show linear convergence with a rate independent of the conditioning of the problem in the early stages of the algorithm.
- Faster linear convergence in the early iterations also applies to problems with global Q-linear convergence, explaining in part the empirical observation that many methods converge rapidly in their early stages before settling down to a slower rate. This observation also allows improvement of iteration complexities.

1.3 Related Work

Our general framework and approach, and special cases thereof, have been widely studied in the literature. Some related work has already been discussed above. We give a broader discussion in this section.

When ψ is the indicator function of a convex constraint set, our approach includes an inexact variant of a constrained Newton or quasi-Newton method. There are a number of papers on this approach, but their convergence results generally have a different flavor from ours. They typically show only asymptotic convergence rates, together with global convergence results without rates, under weaker smoothness and convexity assumptions on f than we make here. For example, when ψ is the indicator function of a "box" defined by bound constraints, [9] applies a trust-region framework to solve (2) approximately, and shows asymptotic convergence. The paper [5] uses a line-search approach, with H_k defined by an L-BFGS update, and omits convergence results. For constraint sets defined by linear inequalities, or general convex constraints, [4] shows global convergence of a trust region method using the Cauchy point. A similar approach using the exact Hessian as H_k is considered in [20], proving local superlinear or quadratic convergence in the case of linear constraints.

[2] Note that for $\eta \in [0,1)$, $1/(1-\eta) > 1/(2(1-\sqrt{\eta}))$.

Turning to our formulation (1) in its full generality, Algorithm 1 is analyzed in [2], which refers to the condition (3) as "η-approximation." (Their η is equivalent to $1-\eta$ in our notation.) This paper shows asymptotic convergence of $Q_k(d)$ to zero without requiring convexity of F, Lipschitz continuity of ∇f, or a fixed value of η. The only assumptions are that $Q_k(d^k) < 0$ for all k and the sequence of objective function values converges (which always happens when F is bounded below). Under the additional assumptions that ∇f is Lipschitz continuous, F is convex, (8), and (3), they showed convergence of the objective value at a $1/k$ rate. The same authors considered convergence for nonconvex functions satisfying a Kurdyka-Łojasiewicz condition in [3], but the exact rates are not given. Our results differ in not requiring the assumption (8), and we are more explicit about the dependence of the rates on η. Moreover, we show detailed convergence rates for several additional classes of problems.

A version of Algorithm 2 without line search but requiring H_k to *overestimate* the Hessian, as follows:

$$f(x^k + d) \leq f(x^k) + \nabla f(x^k)^T d + \frac{1}{2} d^T H_k d$$

is considered in [7]. Asymptotic convergence is proved, but no rates are given.

Convergence of an inexact proximal-gradient method (for which $H_k = LI$ for all k) is discussed in [30]. With this choice of H_k, (7) always holds with $\gamma = 1$. They also discuss its accelerated version for convex and strongly convex problems. Instead of our multiplicative inexactness criterion, they assume an additive inexactness criterion in the subproblem, of the form

$$Q_k\left(d^k\right) \leq Q_k^* + \epsilon_k. \tag{9}$$

Their analysis also allows for an error e^k in the gradient term in (2). The paper shows that for general convex problems, the objective value converges at a $1/k$ rate provided that $\sum_k \sqrt{\epsilon_k}$ and $\sum_k \|e^k\|$ converge. For strongly convex problems, they proved R-linear convergence of $\|x^k - x^*\|$, provided that the sequence $\{\|e^k\|\}$ and $\{\sqrt{\epsilon_k}\}$ both decrease linearly to zero. When our approaches are specialized to proximal gradient ($H_k = LI$), our analysis shows a Q-linear rate (rather than R-linear) for the strongly convex case, and applies to the convergence of the objective value rather than the iterates. Additionally, our results shows convergence for nonconvex problems.

Variant 2 of Algorithm 2 is proposed in [29,12] for convex and strongly convex objectives, with inexactness defined additively as in (9). For convex f, [29] showed that if $\sum_{k=0}^{\infty} \epsilon_k / \|H_k\|$ and $\sum_{k=0}^{\infty} \sqrt{\epsilon_k / \|H_k\|}$ converge then a $1/k$ convergence rate is achievable. The same rate can be achieved if $\epsilon_k \leq (a/k)^2$ for any $a \in [0,1]$. When F is μ-strongly convex, [12] showed that if $\sum \epsilon_k / \rho^k$ is finite (where $\rho = 1 - (\gamma\mu)/(\mu + M)$, M is the upper bound for $\|H_k\|$, and γ is as defined in (7)), then a global R-linear convergence rate is attained. In both cases, the conditions require a certain rate of decrease for ϵ_k, a condition that can be achieved by performing more and more inner iterations as k increases. By contrast, our multiplicative inexactness criterion (3) can be attained with a fixed number of inner iterations. Moreover, we attain a Q-linear rather than an R-linear result.

Algorithm 1 is also considered in [17], with H_k set either to $\nabla^2 f(x^k)$ or a BFGS approximation. Asymptotic convergence and a local rate are shown for the exact case. For inexact subproblem solutions, local results are proved under the assumption that the unit step size is always taken (which may not happen for inexact steps). A variant of Algorithm 1 with a different step size criterion is discussed in [6], for the special case of $\psi(x) = \|x\|_1$. Inexactness of the subproblem solution is measured by the norm of a proximal-gradient step for Q. By utilizing specific properties of the ℓ_1 norm, this paper showed a global convergence rate on the norm of the proximal gradient step on F to zero, without requiring convexity of f — a result similar to our nonconvex result. However, the extension of their result to general ψ is not obvious and, moreover, our inexactness condition avoids the cost of computing the proximal gradient step on Q. When H_k is $\nabla^2 f(x^k)$ or a BFGS approximation, they obtain for the inexact version local convergence results similar to the exact case proved in [17].

For the case in which f is convex, thrice continuously differentiable, and self-concordant, and ψ is the indicator function of a closed convex set, [31] analyzed global and local convergence rates of inexact damped proximal Newton with a fixed step size. The paper [19] extends this convergence analysis to general convex ψ. However, generalization of these results beyond the case of $H_k = \nabla^2 f(x^k)$ and self-concordant f is not obvious.

Accelerated inexact proximal gradient is discussed in [30,33] for convex f to obtain an improved $O(1/k^2)$ convergence rate. The work [13] considers acceleration with more general choices of H under the requirement $H_k \succeq H_{k+1}$ for all k, which precludes many interesting choices of H_k. This requirement is relaxed by [12] to $\theta_k H_k \succeq \theta_{k+1} H_{k+1}$ for scalars θ_k that are used to decide the extrapolation step size. However, as shown in the experiment in [12], extrapolation may not accelerate the algorithm. Our analysis does not include acceleration using extrapolation steps, but by combining with the Catalyst framework [21], similar improved rates could be attained.

1.4 Outline: Remainder of the Paper

In Section 2, we introduce notation and prove some preliminary results. Convergence analysis appears in Section 3 for Algorithms 1 and 2, covering both convex and nonconvex problems. Some interesting and practical choices of H_k are discussed in Section 4 to show that our framework includes many existing algorithms. We provide some preliminary numerical results in Section 5, and make some final comments in Section 6.

2 Notations and Preliminaries

The norm $\|\cdot\|$, when applied on vectors, denotes the Euclidean norm. When applied to a symmetric matrix A, it denotes the corresponding induced norm, which is equivalent to the spectral radius of A. For any symmetric matrix A, $\lambda_{\min}(A)$ denotes its smallest eigenvalue. For any two symmetric matrices A and B, $A \succeq B$ (respectively $A \succ B$) denotes that $A-B$ is positive semidefinite (respectively positive definite). For our nonsmooth function F, ∂F denotes the set of generalized gradient defined as

$$\partial F(x) \coloneqq \nabla f(x) + \partial \varPsi(x),$$

where $\partial \varPsi$ denotes the subdifferential (as $\varPsi$ is convex). When the minimum F^* of $F(x)$ is attainable, we denote the solution set by $\varOmega \coloneqq \{x \mid F(x) = F^*\}$, and define $P_\varOmega(x)$ as the (Euclidean-norm) projection of x onto $\varOmega$.

In some results, we use a particular strong convexity assumption to obtain a faster rate. We say that F satisfies the *optimal set strong convexity* condition with modulus $\mu \geq 0$ if for any x and any $\lambda \in [0,1]$, we have

$$F(\lambda x + (1-\lambda)P_\varOmega(x)) \leq \lambda F(x) + (1-\lambda) F^* - \frac{\mu\lambda(1-\lambda)}{2} \|x - P_\varOmega(x)\|^2 . \tag{10}$$

This condition does not require the strong convexity to hold globally, but only between the current point and its projection onto the solution set. Examples of functions that are not strongly convex but satisfy (10) include:

- $F(x) = h(Ax)$ where h is strongly convex, and A is any matrix;
- $F(x) = h(Ax) + \mathbf{1}_X(x)$, where X is a polyhedron;
- Squared-hinge loss: $F(x) = \sum \max(0, a_i^T x - b_i)^2$.

A similar condition is the "quasi-strong convexity" condition proposed by [24], which always implies (10), and can be implied by optimal set strong convexity if F is differentiable. However, since we allow ψ (and therefore F) to be nonsmooth, we need a different definition here.

Turning to the subproblem (2) and the definition of $\varDelta_k$ in (6), we find a condition for d to be a descent direction.

Lemma 1 *If $\varPsi$ is convex and f is differentiable, then d is a descent direction for F at x if $\varDelta < 0$.*

Proof We know that d is a descent direction for F at x if the directional derivative

$$F'(x;d) \coloneqq \lim_{\alpha \to 0} \quad \frac{F(x+\alpha d) - F(x)}{\alpha}$$

is negative. Note that since f is differentiable and $\varPsi$ is convex,

$$F'(x;d) = \max_{s \in \partial F(x)} s^T d = \nabla f(x)^T d + \max_{\hat{s} \in \partial \varPsi(x)} \hat{s}^T d$$

is well-defined. Now from the convexity of Ψ,

$$\Psi(x+d) \geq \Psi(x) + \hat{s}^T d, \quad \forall \hat{s} \in \partial\Psi(x),$$

so

$$\max_{\hat{s} \in \partial\Psi(x)} \hat{s}^T d + \nabla f(x)^T d \leq \Psi(x+d) - \Psi(x) + \nabla f(x)^T d = \Delta.$$

Therefore, when $\Delta < 0$, the directional derivative is negative and d is a descent direction. □

The following lemma motivates our algorithms.

Lemma 2 *If Q and Ψ are convex and f is differentiable, then $Q(d) < 0$ implies that d is a descent direction for F at x.*

Proof Note that $Q(0) = 0$. Therefore, if Q is convex, we have

$$\lambda \nabla f\left(x\right)^T d + \frac{\lambda^2}{2} d^T H d + \psi\left(x + \lambda d\right) - \psi\left(x\right) = Q_H^x\left(\lambda d\right) \leq \lambda Q_H^x\left(d\right) < 0,$$

for all $\lambda \in (0,1]$. It follows that $\nabla f(x)^T(\lambda d) + \psi(x+\lambda d) - \psi(x) < 0$ for all sufficiently small λ. Therefore, from Lemma 1, λd is a descent direction, and since d and λd only differ in their lengths, so is d. □

Positive semidefiniteness of H suffices to ensure convexity of Q. However, Lemma 2 may be used even when H has negative eigenvalues, as ψ may have a strong convexity property that ensures convexity of Q. Lemma 2 then suggests that no matter how coarse the approximate solution of (2) is, as long as it is better than $d = 0$ for a convex Q, it results in a descent direction. This fact implies finite termination of the backtracking line search procedure in Algorithm 1.

3 Convergence Analysis

We start our analysis for both algorithms by showing finite termination of the line search procedures. We then discuss separately three classes of problems involving different assumptions on F, namely, that F is convex, that F satisfies optimal set strong convexity (10), and that F is nonconvex. Different iteration complexities are proved in each case. The following condition is assumed throughout our analysis in this section.

Assumption 1 *In* (1)*, f is L-Lipschitz-continuously differentiable for some $L > 0$; ψ is convex, extended-valued, proper, and closed; F is lower-bounded; and the solution set Ω of* (1) *is nonempty.*

3.1 Line Search Iteration Bound

We show that the line search procedures have finite termination. The following lemma for the backtracking line search in Algorithm 1 does not require H to be positive definite, though it does require strong convexity of Q (2).

Lemma 3 *If Assumption 1 holds, Q is σ-strongly convex for some $\sigma > 0$, and the approximate solution d to* (2) *satisfies* (3) *for some $\eta < 1$, then for Δ defined in* (6)*, we have*

$$\begin{aligned}\Delta &\leq -\frac{1}{2}\left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\sigma \left\|d\right\|^2 + d^T H d\right)\\ &\leq -\frac{1}{2}\left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\sigma + \lambda_{\min}\left(H\right)\right)\left\|d\right\|^2. \qquad (11)\end{aligned}$$

Moreover, if

$$(1-\sqrt{\eta})\sigma + (1+\sqrt{\eta})\lambda_{\min}(H) > 0,$$

then the backtracking line search procedure in Algorithm 1 terminates in finite steps and produces a step size α that satisfies the following lower bound:

$$\alpha \geq \min\left\{1, \beta\left(1-\gamma\right)\frac{\left(1-\sqrt{\eta}\right)\sigma + \left(1+\sqrt{\eta}\right)\lambda_{\min}\left(H\right)}{L\left(1+\sqrt{\eta}\right)}\right\}. \qquad (12)$$

Proof From (3) and strong convexity of Q, we have that for any $\lambda \in [0,1]$,

$$\begin{aligned}\frac{1}{1-\eta}\left(Q\left(0\right) - Q\left(d\right)\right) &\geq Q\left(0\right) - Q^*\\ &\geq Q\left(0\right) - Q\left(\lambda d\right) \qquad (13)\\ &\geq Q\left(0\right) - \left(\lambda Q\left(d\right) + \left(1-\lambda\right)Q\left(0\right) - \frac{\sigma\lambda\left(1-\lambda\right)}{2}\left\|d\right\|^2\right).\end{aligned}$$

Since $Q(0) = 0$, we obtain by substituting from the definition of Q that

$$\begin{aligned}&\frac{1}{1-\eta}\left(\nabla f\left(x\right)^T d + \frac{1}{2}d^T H d + \psi\left(x+d\right) - \psi\left(x\right)\right)\\ &\leq \lambda\left(\nabla f\left(x\right)^T d + \frac{1}{2}d^T H d + \psi\left(x+d\right) - \psi\left(x\right)\right) - \frac{\sigma\lambda\left(1-\lambda\right)}{2}\left\|d\right\|^2.\end{aligned}$$

Since $1/(1-\eta) \geq 1 \geq \lambda$, we have

$$\begin{aligned}\left(\frac{1}{1-\eta} - \lambda\right)\Delta &\leq -\frac{\sigma\lambda\left(1-\lambda\right)}{2}\left\|d\right\|^2 + \frac{1}{2}\left(\lambda - \frac{1}{1-\eta}\right)d^T H d\\ &\leq -\left(\frac{\sigma\lambda\left(1-\lambda\right)}{2} + \frac{1}{2}\left(\frac{1}{1-\eta} - \lambda\right)\lambda_{\min}\left(H\right)\right)\left\|d\right\|^2. \qquad (14)\end{aligned}$$

It follows immediately that the following bound holds for any $\lambda \in [0,1]$:

$$\Delta \leq -\frac{1}{2}\left(\frac{\sigma\lambda\left(1-\lambda\right)}{\left(\frac{1}{1-\eta}-\lambda\right)}+\lambda_{\min}\left(H\right)\right)\left\|d\right\|^2.$$

We make the following specific choice of λ:

$$\lambda = \frac{1-\sqrt{\eta}}{1-\eta} \in (0,1]. \tag{15}$$

for which

$$1-\lambda = \sqrt{\eta}\lambda, \quad \frac{1}{1-\eta}-\lambda = \frac{\sqrt{\eta}}{1-\eta}.$$

The result (11) follows by substituting these identities into (14).

If the right-hand side of (11) is negative, then we have from the Lipschitz continuity of ∇f, the convexity of ψ, and the mean value theorem that the following relationships are true for all $\alpha \in [0,1]$:

$$\begin{aligned}
&F\left(x+\alpha d\right)-F\left(x\right)\\
&= f\left(x+\alpha d\right)-f\left(x\right)+\psi\left(x+\alpha d\right)-\psi\left(x\right)\\
&\leq \alpha\nabla f\left(x\right)^T d-\alpha\left(\psi\left(x\right)-\psi\left(x+d\right)\right)+\alpha\int_0^1\left(\nabla f\left(x+t\alpha d\right)-\nabla f\left(x\right)\right)^T d\,dt\\
&\leq \alpha\Delta+\frac{L\alpha^2}{2}\left\|d\right\|^2\\
&\leq \alpha\Delta-\frac{L\alpha^2(1+\sqrt{\eta})}{\left(1-\sqrt{\eta}\right)\sigma+\left(1+\sqrt{\eta}\right)\lambda_{\min}\left(H\right)}\Delta.
\end{aligned}$$

Therefore, (5) is satisfied if

$$\alpha\Delta-\frac{L\alpha^2(1+\sqrt{\eta})}{\left(1-\sqrt{\eta}\right)\sigma+\left(1+\sqrt{\eta}\right)\lambda_{\min}\left(H\right)}\Delta \leq \alpha\gamma\Delta.$$

We thus get that (5) holds whenever

$$\alpha \leq (1-\gamma)\frac{\left(1-\sqrt{\eta}\right)\sigma+\left(1+\sqrt{\eta}\right)\lambda_{\min}\left(H\right)}{L\left(1+\sqrt{\eta}\right)}.$$

This leads to (12), when we introduce a factor β to account for possible undershoot of the backtracking procedure. □

Note that Lemma 3 allows indefinite H, and suggests that we can still obtain a certain amount of objective decrease as long as $\lambda_{\min}(H)$ is not too negative in comparison to the strong convexity parameter of Q. When the strong convexity of Q is accounted for completely by the quadratic part (that is, $\lambda_{\min}(H) = \sigma > 0$) we have the following simplification of Lemma 3.

Corollary 1 *If Assumption 1 holds, $\lambda_{\min}(H) = \sigma > 0$, and the approximate solution d to (2) satisfies (3) for some $\eta < 1$, we have*

$$\Delta \leq -\frac{1}{1+\sqrt{\eta}} d^T H d \leq -\frac{\sigma}{1+\sqrt{\eta}} \left\|d\right\|^2. \tag{16}$$

Moreover, the backtracking line search procedure in Algorithm 1 terminates in finite steps and produces a step size that satisfies the following lower bound:

$$\alpha \geq \bar{\alpha} := \min\left\{1, \frac{2\beta\left(1-\gamma\right)\sigma}{L\left(1+\sqrt{\eta}\right)}\right\}. \tag{17}$$

Proof Following (13), we have from convexity of ψ for any $\lambda \in [0,1]$ that

$$\begin{aligned}&\frac{1}{1-\eta}\left(\nabla f\left(x\right)^T d + \frac{1}{2} d^T H d + \psi\left(x+d\right) - \psi\left(x\right)\right)\\ \leq{}& \lambda\left(\nabla f\left(x\right)^T d + \frac{\lambda}{2} d^T H d + \psi\left(x+d\right) - \psi\left(x\right)\right).\end{aligned}$$

Therefore,

$$\left(\frac{1}{1-\eta} - \lambda\right)\Delta \leq \left(\lambda^2 - \frac{1}{1-\eta}\right)\frac{1}{2} d^T H d. \tag{18}$$

Using (15) in (18), we obtain (16). The bound (17) follows by substituting $\sigma = \lambda_{\min}(H)$ into (12). □

Note that the first inequality in (11) and the second inequality in (16) make use of the pessimistic lower bound $d^T H d \geq \lambda_{\min}(H)\|d\|^2$, in practice, we observe (see Section 5) that the unit step $\alpha_k = 1$ is often accepted in practice (significantly larger than the lower bounds (12) and (17)) when H_k is the actual Hessian $\nabla^2 f(x^k)$ or its quasi-Newton approximation.

Next we consider Algorithm 2.

Lemma 4 *If Assumption 1 holds, Q is σ-strongly convex for some $\sigma > 0$, and d is an approximate solution to (2) satisfying (3) for some $\eta \in [0,1)$, then (7) is satisfied if*

$$\left(1-\gamma\right)\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\sigma + \lambda_{\min}(H) \geq L. \tag{19}$$

Therefore, in Algorithm 2, if the initial H_k^0 satisfies

$$m_0 I \preceq H_k^0 \preceq M_0 I \tag{20}$$

for some $M_0 > 0$, $m_0 \leq M_0$, then for Variant 2, the final H_k satisfies

$$\|H_k\| \leq \tilde{M}_2(\eta) := M_0 + \max\left\{1, \frac{1}{\beta}\left(\frac{L\left(1+\sqrt{\eta}\right)}{2-\gamma\left(1-\sqrt{\eta}\right)} - m_0\right)\right\}. \tag{21}$$

For Variant 1, if we assume in addition that $m_0 > 0$, we have

$$\|H_k\| \leq \tilde{M}_1(\eta) := M_0 \max\left\{1, \frac{L\left(1+\sqrt{\eta}\right)}{\beta\left(2-\gamma\left(1-\sqrt{\eta}\right)\right)m_0}\right\}. \tag{22}$$

Proof From Lipschitz continuity of ∇f, we have that

$$\begin{aligned}
&F\left(x\right)-F\left(x+d\right)+\gamma Q_H^x\left(d\right)\\
=\ & f\left(x\right)-f\left(x+d\right)+\gamma\nabla f\left(x\right)^T d+\frac{\gamma}{2}d^THd+\left(1-\gamma\right)\left(\psi\left(x\right)-\psi\left(x+d\right)\right)\\
\geq\ & \left(\gamma-1\right)\nabla f\left(x\right)^T d-\frac{L}{2}\|d\|^2+\frac{\gamma}{2}d^THd+\left(1-\gamma\right)\left(\psi\left(x\right)-\psi\left(x+d\right)\right)\\
=\ & (\gamma-1)\varDelta-\frac{L}{2}\|d\|^2+\frac{\gamma}{2}d^THd \qquad (23)\\
\geq\ & \frac{1-\gamma}{2}\left(\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}\sigma\|d\|^2+d^THd\right)-\frac{L}{2}\|d\|^2+\frac{\gamma}{2}d^THd, \qquad (24)
\end{aligned}$$

where in (23) we used the definition (6), and in (24) we used Lemma 3. By noting $d^THd\geq\lambda_{\min}(H)\|d\|^2$, (24) shows that (19) implies (7).

Since ψ is convex, we have that $\sigma\geq\lambda_{\min}(H)$, so that a sufficient condition for (19) is that

$$\left((1-\gamma)\,\frac{1-\sqrt{\eta}}{1+\sqrt{\eta}}+1\right)\lambda_{\min}(H)\geq L,$$

which is equivalent to

$$\frac{2-\gamma(1-\sqrt{\eta})}{1+\sqrt{\eta}}\lambda_{\min}(H)\geq L.$$

Let the coefficient of $\lambda_{\min}(H)$ in the above inequality be denoted by C_1, this observation suggests that for Variant 1 the smallest eigenvalue of the final H is no larger than $L/(C_1\beta)$, and since the proportion between the largest and the smallest eigenvalues of H_k remains unchanged after scaling the whole matrix, we obtain (22).

For Variant 2, to satisfy $C_1H\succeq LI$, the coefficient for I must be at least L/C_1-m_0. Considering the overshoot, and that the difference between the largest and the smallest eigenvalues is fixed after adding a multiple of identity, we obtain the condition (21). □

By noting the simplification from $d^THd\geq\lambda_{\min}(H)\|d\|^2$, we rarely observe the worst-case bounds (22) or (21) in practice, unless H^0 is a multiple of the identity.

3.2 Iteration Complexity

Now we turn to the iteration complexity of our algorithms, considering three different assumptions on F: convexity, optimal set strong convexity, and the general (possibly nonconvex) case.

The following lemma is modified from some intermediate results in [12], which shows R-linear convergence of Variant 2 of Algorithm 2 for a strongly convex objective when the inexactness is measured by an additive criterion. A proof can be found in Appendix A.

Lemma 5 *Let F^* be the optimum of F. If Assumption 1 holds, f is convex and F is μ-optimal-set-strongly convex as defined in* (10) *for some $\mu \geq 0$, then for any given x and H, and for all $\lambda \in [0, 1]$, we have*

$$\begin{aligned} Q^* &\leq \lambda\left(F^* - F\left(x\right)\right) - \frac{\mu\lambda\left(1-\lambda\right)}{2}\left\|x - P_\Omega\left(x\right)\right\|^2 \\ &\quad + \frac{\lambda^2}{2}\left(x - P_\Omega\left(x\right)\right)^T H\left(x - P_\Omega\left(x\right)\right) \\ &\leq \lambda\left(F^* - F\left(x\right)\right) + \frac{1}{2}\left\|x - P_\Omega\left(x\right)\right\|^2\left(\left\|H\right\|\lambda^2 - \mu\lambda\left(1-\lambda\right)\right), \end{aligned} \tag{25}$$

where Q^* is the optimal objective value of (2). *In particular, by setting $\lambda = \mu/(\mu + \|H\|)$ (as in [12]), we have*

$$Q^* \leq \frac{\mu}{\mu + \|H\|}(F^* - F(x)). \tag{26}$$

Note that we allow $\mu = 0$ in Lemma 5.

3.2.1 Sublinear Convergence for General Convex Problems

We start with case of F convex, that is, $\mu = 0$ in the definition (10). In this case, the first inequality in (25) reduces to

$$Q_k^* \leq \lambda\left(F^* - F\left(x^k\right)\right) + \lambda^2\frac{\left(x^k - P_\Omega\left(x^k\right)\right)^T H_k\left(x^k - P_\Omega\left(x^k\right)\right)}{2}, \tag{27}$$

for all $\lambda \in [0, 1]$. We assume the following in this subsection.

Assumption 2 *There exists finite $R_0, M > 0$ such that*

$$\sup_{x : F(x) \leq F(x_0)} \|x - P_\Omega(x)\| = R_0 < \infty \quad \text{and} \quad \|H_k\| \leq M, \quad k = 0, 1, 2, \ldots. \tag{28}$$

Using this assumption, we can bound the second term in (27) by

$$\hat{A} := \sup_k \left(x^k - P_\Omega\left(x^k\right)\right)^T H_k\left(x^k - P_\Omega\left(x^k\right)\right) \leq MR_0^2. \tag{29}$$

The bound $\hat{A} \leq MR_0^2$ is quite pessimistic, but we use it for purposes of comparing with existing works.

The following lemma is inspired by [1, Lemma 4.4] but contains many nontrivial modifications, and will be needed in proving the convergence rate for general convex problems. Its proof can be found in Appendix B.

Lemma 6 *Assume we have three nonnegative sequences $\{\delta_k\}_{k\geq 0}$, $\{c_k\}_{k\geq 0}$, and $\{A_k\}_{k\geq 0}$, and a constant $A > 0$ such that for all $k = 0, 1, 2, \ldots$, and for all $\lambda_k \in [0, 1]$, we have*

$$0 < A_k \leq A, \quad \delta_{k+1} \leq \delta_k + c_k\left(-\lambda_k\delta_k + \frac{A_k}{2}\lambda_k^2\right). \tag{30}$$

Then for $\delta_k \geq A_k$, we have

$$\delta_{k+1} \leq \left(1 - \frac{c_k}{2}\right) \delta_k. \tag{31}$$

In addition, if we define $k_0 \coloneqq \arg\min\{k : \delta_k < A\}$, then

$$\delta_k \leq \frac{2A}{\sum_{t=k_0}^{k-1} c_t + 2}, \quad \text{for all } k \geq k_0. \tag{32}$$

By Lemma 6 together with Assumption 2, we can show that the algorithms converge at a global sublinear rate (with a linear rate in the early stages) for the case of convex F, provided that the final value of H_k for each iteration k of Algorithms 1 and 2 is positive semidefinite.

Theorem 1 *Assume that f is convex, Assumptions 1 and 2 hold, $H_k \succeq 0$ for all k, and there is some $\eta \in [0, 1)$ such that the approximate solution d^k of* (2) *satisfies* (3) *for all k. Then the following claims for Algorithm 1 are true.*

1. *When $F(x^k) - F^* \geq (x^k - P_\Omega(x^k))^T H_k (x^k - P_\Omega(x^k))$, we have a linear improvement of the objective error at iteration k, that is,*

$$F\left(x^{k+1}\right) - F^* \leq \left(1 - \frac{(1-\eta)\gamma\alpha_k}{2}\right)\left(F\left(x^k\right) - F^*\right). \tag{33}$$

2. *For any $k \geq k_0$, where $k_0 \coloneqq \arg\min\{k : F(x^k) - F^* < MR_0^2\}$, we have*

$$F\left(x^k\right) - F^* \leq \frac{2MR_0^2}{\gamma(1-\eta)\sum_{t=k_0}^{k-1} \alpha_t + 2}, \tag{34}$$

suggesting sublinear convergence of the objective error. If there exists $\bar{\alpha} > 0$ such that $\alpha_k \geq \bar{\alpha}$ for all k, we have

$$k_0 \leq \max\left\{0, 1 + \frac{2}{\gamma(1-\eta)\bar{\alpha}} \log \frac{F\left(x^0\right) - F^*}{MR_0^2}\right\}. \tag{35}$$

For Algorithm 2 under the condition (20), *the above results still hold, with $\bar{\alpha} = 1$, $\alpha_k \equiv 1$ for all k, and M replaced by $\tilde{M}_1(\eta)$ defined in* (22) *for Variant 1, and $\tilde{M}_2(\eta)$ defined in* (21) *for Variant 2.*

Proof Denoting $\delta_k \coloneqq F(x^k) - F^*$, we have for Algorithm 1 that the sufficient decrease condition (5) together with $H_k \succeq 0$ imply that

$$\delta_{k+1} - \delta_k \leq \alpha_k \gamma \Delta_k = \alpha_k \gamma \left(Q_k\left(d^k\right) - \frac{1}{2}\left(d^k\right)^T H_k d^k\right) \leq \alpha_k \gamma Q_k\left(d^k\right). \tag{36}$$

By defining

$$A_k \coloneqq \left(x^k - P_\Omega\left(x^k\right)\right)^T H_k \left(x^k - P_\Omega\left(x^k\right)\right), \quad A \coloneqq MR_0^2,$$

(note that $A_k \leq A$ follows from (29)) and using (3), (36), and (27), we obtain

$$\delta_{k+1} - \delta_k \leq \alpha_k \gamma \left(1 - \eta\right) \left(-\lambda_k \delta_k + \frac{A_k \lambda_k^2}{2}\right), \quad \forall \lambda_k \in [0, 1]. \tag{37}$$

We note that (37) satisfies (30) with

$$c_k = \alpha_k \gamma \left(1 - \eta\right).$$

The results now follow directly from Lemma 6.

For Algorithm 2, from (7) and (3), we get that for any $k \geq 0$,

$$\delta_{k+1} - \delta_k \leq \gamma \left(1 - \eta\right) Q_k^*, \tag{38}$$

and the remainder of the proof follows the above procedure starting from the right-hand side of (36) with $\alpha_k \equiv 1$. □

The conditions of Parts 1 and 2 of Theorem 1 bear further consideration. When the regularization term ψ is not present in F, and M is a global bound on the norm of the true Hessian $\nabla^2 f(x)$, the condition in Part 2 of Theorem 1 is satisfied for $k_0 = 0$, since $f(x^0) - f^* \leq \frac{1}{2} M \|x^0 - P_\Omega(x^0)\|^2 \leq \frac{1}{2} M R_0^2$. Under these circumstances, the linear convergence result of Part 1 may appear not to be interesting. We note, however, that the contribution from ψ may make a significant difference in the general case (in particular, it may result in $F(x^0) - F^* > M R_0^2$) and, moreover, a choice of H_k with $\|H_k\|$ significantly less than M may result in the condition of Part 1 being satisfied intermittently during the computation. In particular, Part 1 lends some support to the empirical observation of rapid convergence on the early stages of the algorithms, as we discuss further below. Note that [26, Theorem 4] suggests that when the algorithm is exact proximal gradient, we get $F(x^k) - F^* \leq M R_0^2$ for all $k \geq 1$, but this is not always the case when a different H is picked or when (2) is solved only approximately.

By combining Theorem 1 with Lemma 3 and Corollary 1 (which yield lower bounds on α_k), we obtain the following results for Algorithm 1.

Corollary 2 *Assume the conditions of Theorem 1 are all satisfied. Then we have the following.*

1. *If there exists $\sigma > 0$ such that $\lambda_{\min}(H_k) \geq \sigma$ for all k, then* (33) *becomes*

$$\frac{F\left(x^{k+1}\right) - F^*}{F\left(x^k\right) - F^*} \leq 1 - \frac{\gamma}{2} \min\left\{\left(1 - \eta\right), \frac{2\left(1 - \sqrt{\eta}\right)\beta(1 - \gamma)\sigma}{L}\right\}, \tag{39}$$

(34) *becomes*

$$F\left(x^k\right) - F^* \leq \frac{2 M R_0^2}{\gamma (k - k_0) \min\left\{1 - \eta, \frac{2\left(1 - \sqrt{\eta}\right)\beta(1 - \gamma)\sigma}{L}\right\} + 2},$$

and (35) *becomes*

$$k_0 < 1 + \frac{2}{\gamma} \max\left\{0, \log \frac{F\left(x^0\right) - F^*}{M R_0^2}\right\} \cdot \max\left\{\frac{1}{(1 - \eta)}, \frac{L}{2(1 - \sqrt{\eta})\beta\left(1 - \gamma\right)\sigma}\right\}.$$

2. If Q_k is σ-strongly convex and $H_k \succeq 0$ for all k, then (33) *becomes*

$$\frac{F\left(x^{k+1}\right)-F^*}{F\left(x^k\right)-F^*} \leq 1-\frac{\gamma}{2} \min \left\{1-\eta, \frac{\left(1-\sqrt{\eta}\right)^2 \beta(1-\gamma)\sigma}{L}\right\},$$

(34) *becomes*

$$F\left(x^k\right)-F^* \leq \frac{2MR_0^2}{\gamma(k-k_0) \min \left\{1-\eta, \frac{(1-\sqrt{\eta})^2\beta(1-\gamma)\sigma}{L}\right\}+2},$$

and (35) *becomes*

$$k_0 < 1+\frac{2}{\gamma} \max\left\{0, \log\frac{F\left(x^0\right)-F^*}{MR_0^2}\right\} \max\left\{\frac{1}{(1-\eta)}, \frac{L}{(1-\sqrt{\eta})^2\beta\left(1-\gamma\right)\sigma}\right\}.$$

We make some remarks on the results above.

Remark 1 For any $\eta \in [0,1)$, we have

$$\frac{1}{2(1-\sqrt{\eta})} < \frac{1}{1-\eta} < \frac{1}{(1-\sqrt{\eta})^2}.$$

Therefore, Algorithm 1 with positive definite H_k has better dependency on η than the case in which we set $\lambda_{\min}(H_k) = 0$ and rely on ψ to make Q_k strongly convex. If ψ is strongly convex, we can move some of its curvature to H_k without changing the subproblems (2). This strategy may require us to increase M, but this has only a slight effect on the bounds in Corollary 2. These bounds give good reasons to capture the curvature of Q_k in the Hessian H_k alone, so henceforth we focus our discussion on this case.

Remark 2 For Algorithm 2, when we use the bounds (22) and (21) for M in (28), the dependency of the global complexity on η becomes

$$\max\left\{\frac{1}{1-\eta}, \frac{1}{\left(2-\gamma\left(1-\sqrt{\eta}\right)\right)\left(1-\sqrt{\eta}\right)}\right\} \leq \max\left\{\frac{1}{1-\eta}, \frac{1}{(2-\gamma)(1-\sqrt{\eta})}\right\},$$

This result is slightly worse than that of using positive definite H in Algorithm 1 if we compare the second part in the max operation.

Remark 3 The bound in (29) is not tight for general H, unless $H_k \equiv MI$, as in standard prox-gradient methods. This observation gives further intuition for why second-order methods tend to perform well even though their iteration complexities (which are based on the bound (29)) tend to be worse than first-order methods. Moreover, when H_k incorporates curvature information for f, step sizes α_k are often much larger than the worst-case bounds that are used in Corollary 2. Theorem 1, which shows how the convergence rates are related directly to the α_k, would give tighter bounds in such cases. Line search on H_k in Algorithm 2 does not improve the rate directly, but we note that using H_k with smaller norm whenever possible gives more chances of switching to the intermittent linear rate (33).

Part 1 of Theorem 1 also explains why solving the *subproblem* (2) approximately can save the running time significantly, since because of fast early convergence rate, a solution of moderate accuracy can be attained relatively quickly.

3.2.2 Linear Convergence for Optimal Set Strongly Convex Functions

We now consider problems that satisfy the μ-optimal-set-strong-convexity condition (10) for some $\mu > 0$, and show that our algorithms have a global linear convergence property.

Theorem 2 *If Assumption 1 holds, f is convex, F is μ-optimal-set-strongly convex for some $\mu > 0$, there is some $\eta \in [0, 1)$ such that at every iteration of Algorithm 1, the approximate solution d of* (2) *satisfies* (3)*, and*

$$\sigma I \preceq H_k \preceq MI, \quad \text{for some } M \geq \sigma > 0, \quad \forall k. \tag{40}$$

Then for $k = 0, 1, 2, \dots$, we have

$$\frac{F\left(x^{k+1}\right) - F^*}{F\left(x^k\right) - F^*} \leq 1 - \frac{\alpha_k \gamma (1-\eta) \mu}{\mu + \|H_k\|} \tag{41a}$$

$$\leq 1 - \frac{\gamma \mu}{\mu + M} \min\left\{(1-\eta), \frac{2\left(1-\sqrt{\eta}\right)\beta(1-\gamma)\sigma}{L}\right\}. \tag{41b}$$

Moreover, on iterates k for which $F(x^k) - F^* \geq (x^k - P_\Omega(x^k))^T H_k (x^k - P_\Omega(x^k))$, these per-iteration contraction rates can be replaced by the faster rates (33) *and* (39)*.*

Proof By rearranging (36), we have

$$\begin{aligned} F\left(x^{k+1}\right) - F^* &\leq F\left(x^k\right) - F^* + \alpha_k \gamma Q_k\left(d^k\right) \\ &\leq F\left(x^k\right) - F^* + \alpha_k \gamma (1-\eta) Q_k^* && \text{(42a)} \\ &\leq F\left(x^k\right) - F^* - \alpha_k \gamma (1-\eta) \frac{\mu}{\mu + \|H_k\|} \left(F\left(x^k\right) - F^*\right) && \text{(42b)} \\ &= \left(1 - \alpha_k \gamma (1-\eta) \frac{\mu}{\mu + \|H_k\|}\right) \left(F\left(x^k\right) - F^*\right), \end{aligned}$$

where in (42a) we used the inexactness condition (3) and in (42b) we used (26). Using the result in Corollary 1 to lower-bound α_k, we obtain (41b).

To show that the part for the early fast rate in (33) and (39) can be applied, we show that Assumption 2 holds. Then because f is assumed to be convex as well here, Theorem 1 and Corollary 2 apply as well. Consider (10), by rearranging the terms, we get

$$\begin{aligned} \lambda\left(F(x) - F^*\right) &\geq \frac{\mu\lambda(1-\lambda)}{2} \left\|x - P_\Omega(x)\right\|^2 + F\left(\lambda x + (1-\lambda) P_\Omega(x)\right) - F^* \\ &\geq \frac{\mu\lambda(1-\lambda)}{2} \left\|x - P_\Omega(x)\right\|^2, \quad \forall \lambda \in [0,1], \end{aligned} \tag{43}$$

as $F(\lambda x + (1-\lambda) P_\Omega(x)) \geq F^*$ from optimality. By dividing both sides of (43) by λ and letting $\lambda \to 0$, we get the bound

$$F(x^0) - F^* \geq F(x) - F^* \geq \frac{\sigma}{2}\|x - P_\Omega(x)\|, \forall x : F(x) \leq F(x^0), \tag{44}$$

validating Assumption 2. □

Note that the parameter μ in the theorem above is decided by the problem and cannot be changed, while σ can be altered according to the algorithm choice. We have a similar result for Algorithm 2.

Theorem 3 *If Assumption 1 holds, f is convex, F is μ-optimal-set-strongly convex for some $\mu > 0$, there exists some $\eta \in [0,1)$ such that at every iteration of Algorithm 2, the approximate solution d of* (2) *satisfies* (3)*, and the conditions for H_k^0 in Lemma 4 are satisfied for all k. Then we have*

$$\frac{F(x^{k+1}) - F^*}{F(x^k) - F^*} \leq 1 - \gamma \frac{\mu(1-\eta)}{\mu + \|H_k\|}, \quad k = 0, 1, 2, \ldots, \tag{45}$$

and the right-hand side of (45) *can be further bounded by*

$$1 - \gamma \frac{\mu(1-\eta)}{\mu + \tilde{M}_1(\eta)} \quad \textit{and} \quad 1 - \gamma \frac{\mu(1-\eta)}{\mu + \tilde{M}_2(\eta)} \tag{46}$$

for Variant 1 and Variant 2, respectively, where $\tilde{M}_1(\eta)$ and $\tilde{M}_2(\eta)$ are defined in Lemma 4. Moreover, when $F(x^k) - F^* \geq (x^k - P_\Omega(x^k))^T H_k (x^k - P_\Omega(x^k))$, the faster rate (33) *(with $\alpha_k \equiv 1$ and the modification for Algorithm 2 mentioned in Theorem 1) can be used to replace* (45)*.*

Proof From (26) and (38), we have

$$\begin{aligned} F(x^{k+1}) - F^* &\leq F(x^k) - F^* + \gamma Q_k(d^k) \\ &\leq F(x^k) - F^* + \gamma(1-\eta) Q_k^* \\ &\leq \left(1 - \gamma \frac{\mu}{\mu + \|H_k\|}(1-\eta)\right)\left(F(x^k) - F^*\right), \end{aligned}$$

proving (45). From Lemma 4, we ensure that $\|H_k\|$ is upper-bounded by $\tilde{M}_1(\eta)$ and $\tilde{M}_2(\eta)$ for the two variants respectively, leading to (46). The statement concerning (33) follows from the same reasoning as in the proof for Theorem 2. □

By reasoning with the extreme eigenvalues of H_k, we can see that the convergence rates still depend on the conditioning of f. For Algorithm 1, if we select $M \leq L$, then backtracking may be necessary, and the bound (41b) (in which a factor μ/L appears) is germane. This same factor appears in both (41a) and (41b) when $M > L$. Often, however, the backtracking line search chooses a value of α_k that is not much less than 1, which is why we believe that the bounds (33), (34), and (41a) (which depend explicitly on α_k) have some value in revealing the actual performance of the algorithm. Similar comments apply to Algorithm 2, because (7) may be satisfied with $\|H_k\|$ much smaller than the bounds for properly chosen H_k^0.

In the interesting case in which we choose $H_k \equiv LI$ and $\eta = 0$, we have $m_0 = \|H_k\| = L$ in Algorithm 2, and modification of H_k is not needed, since (7) always holds for $\gamma = 1$. The bound (34) becomes $(F(x^k)-F^*) \leq 2LR_0^2/(k+2)$, which matches the known convergence rates of proximal gradient [26] and gradient descent [25]. The global linear rate in Theorem 3 also matches that of existing proximal gradient analysis for strongly convex problems, but the intermittent linear rate (33) that applies to both cases is new. For the case of accelerated proximal gradient covered in [26], although not covered directly by our framework studied in this work, one can combine our algorithm and analysis with the Catalyst framework [21] to obtain similar accelerated rates for both the strongly convex and the general convex cases.

3.2.3 Sublinear Convergence of the First-order Optimality Condition for Nonconvex Problems

We consider now the case of nonconvex F. In this situation, Lemma 5 cannot be used, so we consider other properties of Q. We can no longer guarantee the convergence of the objective value to the global minimum. Instead, we consider the norm of the exact solution of the subproblem as the indicator of closeness to the first-order optimality condition $0 \in \partial F(x)$ for (1) (see, for example, [11, (14.2.16)]). In particular, it is known that $0 \in \partial F(x)$ if and only if

$$0 = \arg\min_d Q_I^x(d) = \arg\min_d \nabla f\left(x\right)^T d + \frac{1}{2} d^T d + \psi\left(x+d\right) - \psi\left(x\right). \tag{47}$$

This is a consequence of the following lemma.

Lemma 7 *Given any $H \succ 0$, and Q_H^x defined as in (2), the following are true.*

1. *A point x satisfies the first-order optimality condition $0 \in \partial F(x)$ if and only if*
$$0 = \arg\min_d Q_H^x(d).$$
2. *For any x, defining d^* to be the minimizer of $Q_H^x(\cdot)$, we have*
$$Q_H^x(d^*) \leq -\frac{1}{2}\lambda_{\min}\left(H\right)\|d^*\|^2. \tag{48}$$

Proof Part 1 is well known. For Part 2, we have from the optimality conditions for d^* that $-\nabla f(x) - Hd^* \in \partial\psi(x+d^*)$. By convexity of ψ, we thus have

$$\psi(x) \geq \psi(x+d^*) + (d^*)^T(\nabla f(x) + Hd^*) \quad \Rightarrow \quad 0 \geq Q_H^x(d^*) + \frac{1}{2}(d^*)^T Hd^*,$$

from which the result follows. □

As in (47), we consider the following measure of closeness to a stationary point:

$$G_k := \arg\min_d Q_I^{x^k}(d). \tag{49}$$

We show that the minimum value of the norm of this measure over the first k iterations converges to zero at a sublinear rate of $O(1/\sqrt{k})$. The first step is to show that the minimum of $|Q_k|$ converges at a $O(1/k)$ rate.

Lemma 8 *Assume that there is an $\eta \in [0,1)$ such that (3) is satisfied at all iterations. For Algorithm 1, if Assumption 1 holds and $H_k \succeq \sigma I$ for some $\sigma > 0$ and all k, we have*

$$\min_{0\le t\le k} \left|Q_t\left(d^t\right)\right| \le \frac{F\left(x^0\right) - F^*}{\gamma\left(k+1\right)\min_{0\le t\le k}\alpha_t} \le \frac{F\left(x^0\right) - F^*}{\gamma\left(k+1\right)} \max\left\{1, \frac{(1+\sqrt{\eta})L}{2\beta\left(1-\gamma\right)\sigma}\right\}. \tag{50}$$

For Algorithm 2 (requires $H_k^0 \succ 0$ for the first variant), we have

$$\min_{0\le t\le k} \left|Q_t\left(d^t\right)\right| \le \frac{F\left(x^0\right) - F^*}{\gamma\left(k+1\right)}.$$

Proof From (36), we have that for any $k \ge 0$,

$$F^* - F\left(x^0\right) \le F\left(x^{k+1}\right) - F\left(x^0\right) \le \gamma \sum_{t=0}^{k} \alpha_t Q_t\left(d^t\right) \le \gamma \min_{0\le t\le k} \alpha_t \sum_{t=0}^{k} Q_t\left(d^t\right). \tag{51}$$

From Corollary 1, we have that α_t for all t is lower bounded by a positive value. Therefore, using $\left|Q_t\left(d^t\right)\right| = -Q_t\left(d^t\right)$ for all t, we obtain

$$\min_{0\le t\le k} \left|Q_t\left(d^t\right)\right| \le -\frac{1}{k+1}\sum_{t=0}^{k} Q_t\left(d^t\right) \le \frac{F\left(x^0\right) - F^*}{\gamma\left(k+1\right)\min_{0\le t\le k}\alpha_t}.$$

Substituting the lower bound for α from Corollary 1 gives the desired result (50). The result for Algorithms 2 follows from the same reasoning applied to (7). □

The following lemma is from [32]. (Its proof is omitted.)

Lemma 9 ([32, Lemma 3]) *Given H_k satisfying (40) for all k, we have*

$$\|G_k\| \le \frac{1 + \frac{1}{\sigma} + \sqrt{1 - 2\frac{1}{M} + \frac{1}{\sigma^2}}}{2} M \left\|d^{k*}\right\|,$$

where

$$d^{k*} := \arg\min Q_k.$$

We are now ready to show the convergence of $\|G_k\|$.

Corollary 3 *Assume that* (3) *holds at all iterations for some* $\eta \in [0,1)$ *and that Assumption 1 holds. Let* $\tilde{M}_1(\eta)$ *and* $\tilde{M}_2(\eta)$ *be as defined in Lemma 4. For Algorithm 1, suppose that* H_k *satisfies* (40) *for all* $k \geq 0$. *We then have for all* $k = 0, 1, , 2, \ldots$ *that*

$$\begin{aligned}\min_{0\leq t\leq k} \|G_t\|^2 \leq& \frac{F\left(x^0\right) - F^*}{\gamma\left(k+1\right)} \frac{M^2\left(1 + \frac{1}{\sigma} + \sqrt{1 - \frac{2}{M} + \frac{1}{\sigma^2}}\right)^2}{2(1-\eta)\sigma \min_{0\leq t\leq k} \alpha_t}\\ \leq& \frac{F\left(x^0\right) - F^*}{\gamma\left(k+1\right)} \frac{M^2\left(1 + \frac{1}{\sigma} + \sqrt{1 - \frac{2}{M} + \frac{1}{\sigma^2}}\right)^2}{2\sigma}\\ &\max\left\{\frac{1}{1-\eta}, \frac{L}{2\left(1-\sqrt{\eta}\right)\left(1-\gamma\right)\sigma\beta}\right\}.\end{aligned}$$

For Algorithm 2, if the initial H_k^0 *satisfies* $M_0 I \succeq H_k^0 \succeq m_0 I$ *with* $M_0 \geq m_0 > 0$ *then for Variant 1 we have:*

$$\min_{0\leq t\leq k} \|G_t\|^2 \leq \frac{F\left(x^0\right)) - F^*}{\gamma\left((k+1)\right)} \frac{\tilde{M}_1(\eta)^2\left(1 + \frac{1}{m_0} + \sqrt{1 - \frac{2}{\tilde{M}_1(\eta)} + \frac{1}{m_0^2}}\right)^2}{2\left(1-\eta\right) m_0}.$$

For Variant 2, we have under the same assumptions on H_k^0 *that the same bound is satisfied,*[3] *with* $\tilde{M}_1(\eta)$ *replaced by* $\tilde{M}_2(\eta)$.

Proof Let $\bar{k} := \arg\min_{0\leq t\leq k} |Q_t(d^t)|$, the condition (3) and Lemmas 7 and 9 imply

$$\begin{aligned}-Q_{\bar{k}}\left(d^{\bar{k}}\right) &\geq -\left(1-\eta\right) Q_{\bar{k}}^*\\ &\geq \frac{\sigma\left(1-\eta\right)}{2}\left\|d^{\bar{k}*}\right\|^2\\ &\geq \frac{2\sigma\left(1-\eta\right)}{M^2\left(1 + \frac{1}{\sigma} + \sqrt{1 - \frac{2}{M} + \frac{1}{\sigma^2}}\right)^2}\left\|G_{\bar{k}}\right\|^2. \qquad (52)\end{aligned}$$

Finally, we note that $\|G_{\bar{k}}\| \geq \min_{0\leq t\leq k} \|G_t\|$. The proof is finished by combining (52) with Lemma 8. □

If we replace the definition of G_k in (49) by the solution of (2), the inequality in Lemma 9 is not needed. In particular, when we use the proximal gradient algorithm with $H_k = LI$ and $\eta = 0$ (so that (7) holds with $\gamma = 1$, and $M = L$) we obtain a bound of $2(F(x^0) - F^*)/(L(k+1))$ on $\|d^k\|^2$, matching the result shown in [26, 10].

[3] We could instead require only $H_k^0 \succeq 0$ and start with $H_k + I$ instead.

3.2.4 Comparison Among Different Approaches

Algorithms 1 and 2 both require evaluation of the function F for each choice of the parameter α_k, to check whether the decrease conditions (5) and (7) (respectively) are satisfied. The difference is that Algorithm 2 may also require solution of the subproblem (2) for each α_k. This additional computation comes with two potential benefits. First, the second variant of Algorithm 2 allows the initial choice of approximate Hessian H_k^0 to be indefinite, although the final value H_k at each iteration needs to be positive semidefinite for our analysis to hold. (There is a close analogy here to trust-region methods for nonconvex smooth optimization, where an indefinite Hessian is adjusted to be positive semidefinite in the process of solving the trust-region subproblem.) Second, because full steps are always taken in Algorithm 2, any structure induced in the iterates x^k by the regularizer ψ (such as sparsity) will be preserved. This fact in turn may lead to faster convergence, as the algorithm will effectively be working in a low-dimensional subspace.

4 Choosing $\boldsymbol{H_k}$

Here we discuss some ways to choose H_k so that the algorithms are well defined and practical, and our convergence theory can be applied.

When H_k are chosen to be positive multiples of identity ($H_k = \zeta_k I$, say), our algorithms reduce to variants of proximal gradient. If we set $\zeta_k \geq L$, then the unit step size is always accepted even if the problem is not solved exactly, because $Q_k(d^k)$ is an upper bound of $F(x^k) - F(x^k + d^k)$. When L is not known in advance, adaptive strategies can be used to find it. For Algorithm 2, we could define ζ_k^0 (such that $H_k^0 = \zeta_k^0 I$) to be the final value ζ_{k-1} from the previous iteration, possibly choosing a smaller value at some iterations to avoid being too conservative. For Algorithm 1, we could increase ζ_k^0 over ζ_{k-1} if backtracking was necessary at iteration $k-1$, and shrink it when a unit stepsize sufficed for several successive iterations.

The proximal Newton approach of setting $H_k = \nabla^2 f(x^k)$ is a common choice in the convex case [17], where we can guarantee that H_k is at least positive semidefinite. In [17], it is shown that in some neighborhood of the optimum, when d^k is the exact solution of (2), then unit step size is always taken, and superlinear or quadratic convergence to the optimum ensues. (A global complexity condition is not required for this result.) Generally, however, indefiniteness in $\nabla^2 f(x^k)$ may lead to the search direction d^k not being a descent direction, and the backtracking line search will not terminate in this situation. (Our convergence results for Algorithm 1 do not apply in the case of H_k indefinite.) A common fix is to use damping, setting $H_k = \nabla^2 f(x^k) + \zeta_k I$, for some $\zeta_k \geq 0$ that at least ensures positive definiteness of H_k. Strategies for choosing ζ_k adaptively have been the subject of much research in the context of smooth minimization, for example, in trust-region methods and the Levenberg-Marquardt method for nonlinear least squares (see [27]). Variant 2 of our Algorithm 2 uses this strategy. It is desirable to ensure that $\zeta_k \to 0$ as the iterates approach a solution at which local convexity holds, to ensure rapid local convergence.

An L-BFGS approximation of $\nabla^2 f(x^k)$ could also be used for H_k. When ψ is not present in (1) and f is strongly convex, it is shown in [22] that this approach has global linear convergence because the eigenvalues of H_k are restricted to a bounded positive interval. This proof can be extended to our algorithms, when a convex ψ is present in (1). When f is not strongly convex, one can apply safeguards to the L-BFGS update procedure (as described in [18]) to ensure that the upper and lower eigenvalues of H_k are bounded uniformly away from zero.

Another interesting choice of H_k is a block-diagonal approximation of the Hessian, which (when ψ can be partitioned accordingly) allows the subproblem (2) to be solved in parallel while still retaining some curvature information. Strategies like this one are often used in distributed optimization for machine learning problems (see, for example, [35, 14, 36]).

5 Numerical Results

We sketch some numerical simulations that support our theoretical results. We conduct experiments on two different problems: ℓ_1-regularized logistic regression, and the Lagrange dual problem of ℓ_2-regularized squared-hinge loss minimization. The algorithms are implemented in C/C++.

5.1 ℓ_1-regularized Logistic Regression

Given training data points $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}$, $i = 1, \dots, l$, and a specified parameter $C > 0$, we solve the following convex problem

$$\min_{x \in \mathbb{R}^n} C \sum_{i=1}^{l} \psi \left(1 + \exp\left(-b_i a_i^T x\right)\right) + \|x\|_1. \tag{53}$$

We define H_k to be the limited-memory BFGS approximation [22] based on the past 10 steps, with a safeguard mechanism proposed in [18] to ensure uniform boundedness of H_k. The subproblems (2) are solved with SpaRSA [34], a proximal-gradient method which, for bounded H_k, converges globally at a linear rate. We consider the publicly available data sets listed in Table 1,[4] and present empirical convergence results by showing the relative objective error, defined as

$$\frac{F(x) - F^*}{F^*}, \tag{54}$$

where F^* is the optimum, obtained approximately through running our algorithm with long enough time. For all variants of our framework, we used

[4] Downloaded from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Data set	l	n	#nonzeros
a9a	$32,561$	123	$451,592$
rcv1_test.binary	$677,399$	$47,236$	$49,556,258$
epsilon	$400,000$	$2,000$	$800,000,000$

Table 1: Properties of the Data Sets

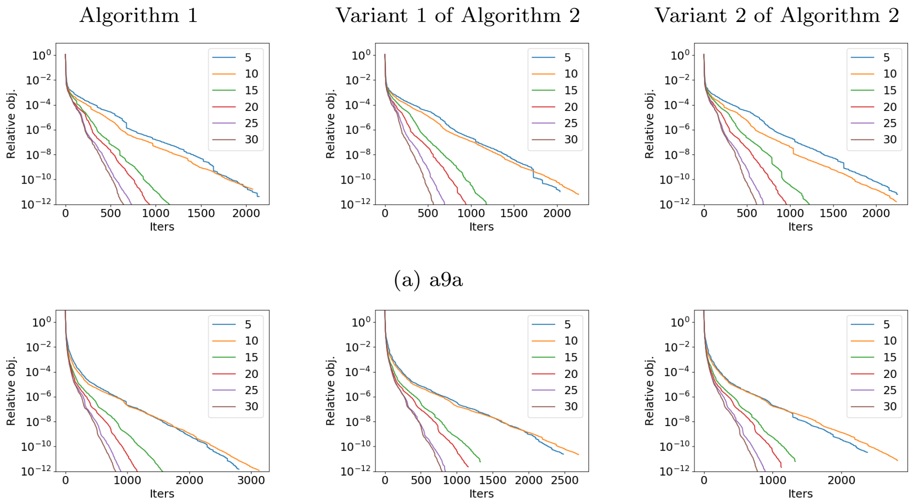
parameters $\beta = 0.5$, and $\gamma = 10^{-4}$. Further details of our implementation are described in [15].

We use the two smaller data sets a9a and rcv1 to quantify the relationship between accuracy of the subproblem solution and the number of outer iterations. We compare running SpaRSA with a fixed number of iterations $T \in \{5, 10, 15, 20, 25, 30\}$. Figure 1 shows that, in all cases, the number of outer iterations decreases monotonically as the (fixed) number of inner iterations is increased. For $T \geq 15$, the degradation in number of outer iterations resulting from less accurate solution of the subproblems is modest, as our theory suggests. We also observe the initial fast linear rates in the early stages of the method that are predicted by our theory, settling down to a slower linear rate on later iterations, but with sudden drops of the objective, possibly as a consequence of intermittent satisfaction of the condition in Part 1 of Theorem 1.

Next, we examine empirically the step size distribution for Algorithm 1 and how often in Algorithm 2 the matrix H_k needs to be modified. On both a9a and rcv1, the initial step estimate $\alpha = 1$ is accepted on over 99.5% of iterations in Algorithm 1, while in both variants of Algorithm 2, the initial choice of H_k is used without modification on over 99% of iterations. These statistics hold regardless of the value of T (the number of inner iterations), though in the case of Algorithm 2, we see a faint trend toward more adjustments for larger values of T. When adjustments are needed, they never number more than 4 at any one iteration, except for a single case (a9a for Variant 1 of Algorithm 2 with $T = 5$) for which up to 8 adjustments are needed.

We next compare our inexact method with an exact version, in which the subproblems (2) are solved to near-optimality at each iteration. Since the three algorithms behave similarly, we use Algorithm 1 as the representative for this investigation. We use a local cluster with 16 nodes for the two larger data sets rcv1 and epsilon, while for the small data set a9a, only one node is used. Iteration counts and running time comparisons are shown in Figure 2. The exact version requires fewer iterations, as expected, but the inexact version requires only modestly more iterations. In terms of runtime, the inexact versions with moderate amount of inner iterations (at least 30) has the advantage, due to the savings obtained by solving the subproblem inexactly.

We note that the approach of gradually increasing the number of inner iterations, suggested in [29,12], produces good results for this application, the number of iterations required being comparable to those for the exact solver while the running time is slightly faster than that of $T = 30$ for epsilon and competitive with it for the rest two data sets.


Algorithm 1

Variant 1 of Algorithm 2

Variant 2 of Algorithm 2

(a) a9a

(b) rcv1t

Fig. 1: Comparison of different subproblem solution exactness in solving (53). The y-axis is the relative objective error (54), and the x-axis is the iteration count.

5.2 Dual of ℓ_2-regularized Squared-Hinge Loss Minimization

Given the same binary-labelled data points as in the previous experiment and a parameter $C > 0$, the ℓ_2-regularized squared-hinge loss minimization problem is

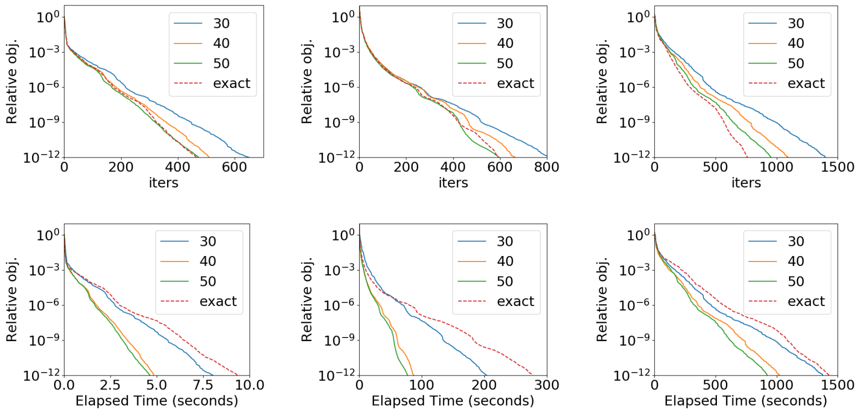
$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x\|_2^2 + C \sum_{i=1}^{l} \max(1 - b_i a_i^T x, 0)^2.$$

With the notation $A := (b_1 a_1, b_2 a_2, \dots, b_l a_l)$, the dual of this problem is

$$\min_{\alpha \geq 0} \frac{1}{2} \alpha^T A^T A \alpha - \mathbf{1}^T \alpha + \frac{1}{4C} \|\alpha\|_2^2, \tag{55}$$

which is $(1/2C)$-strongly convex. We consider the distributed setting such that the columns of A are stored across multiple processors. In this setup, only the block-diagonal parts (up to a permutation) of $A^T A$ can be easily formed locally on each processor. We take H_k to be the matrix formed by these diagonal blocks, so that the subproblem (2) can be decomposed into independent parts. We use cyclic coordinate descent with random permutation (RPCD) as the solver for each subproblem. (Note that this algorithm partitions trivially across processors, because of the block-diagonal structure of H_k.)

Our experiment compares the strategy of performing a fixed number of RPCD iterations for each subproblem with one of increasing the number of inner iterations as the algorithm proceeds, as in [29, 12]. We use the data sets



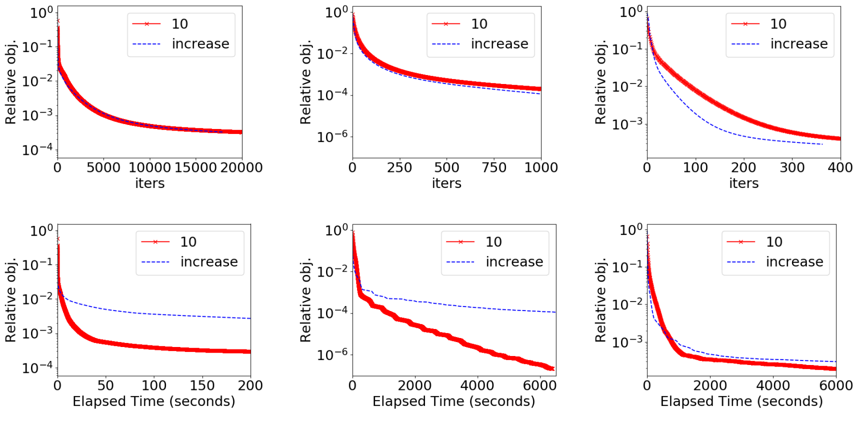
(a) a9a (b) rcv1 (c) epsilon

Fig. 2: Comparison between the exact version and the inexact version of Algorithm 1 for solving (53). Top: outer iterations; bottom: running time. The y-axis is the relative objective error (54).

in Table 1, and compare the two strategies on Algorithm 1, but use an exact line search to choose α_k rather than the backtracking approach. (An exact line search is made easy by the quadratic objective.) For the first strategy, we use ten iterations of RPCD on each subproblem, while for the second strategy, we perform $1 + \lfloor k/10 \rfloor$ iterations of RPCD at the kth outer iteration as suggested by [29, 12]. The implementation is a modification of the experimental code of [16]. We run the algorithms on a local cluster with 16 machines, so that H_k contains 16 diagonal blocks. Results are shown in Figure 3. Since the choice of H_k in this case does not capture global curvature information adequately, the strategy of increasing the accuracy of subproblem solution on later iterations does not reduce the number of iterations as significantly as in the previous experiment. The runtime results show a significant advantage for the first strategy of a fixed number of inner iterations, particularly on the a9a and rcv1 data sets. Judging from the trend in the approach of increasing inner iterations, we can expect that the exact version will show huger disadvantage for running time in this case. We also observe the faster linear rate on early iterations, matching our theory.

6 Conclusions

We have analyzed global convergence rates of three practical inexact successive quadratic approximation algorithms under different assumptions on the objective function, including the nonconvex case. Our analysis shows that inexact solution of the subproblems affects the rates of convergence in fairly benign ways, with a modest factor appearing in the bounds. When linearly convergent



(a) a9a (b) rcv1 (c) epsilon

Fig. 3: Comparison of two strategies for inner iteration count in Algorithm 1 applied to (2): Increasing accuracy on later iterations (blue) and a fixed number of inner iterations (red). Top: outer iterations; bottom: running time. Vertical axis shows relative objective error (54).

methods are used to solve the subproblems, the inexactness condition holds when a fixed number of inner iterations is applied at each outer iteration k.

References

1. Bach, F.: Duality between subgradient and conditional gradient methods. SIAM Journal on Optimization **25**(1), 115–129 (2015)
2. Bonettini, S., Loris, I., Porta, F., Prato, M.: Variable metric inexact line-search-based methods for nonsmooth optimization. SIAM Journal on Optimization **26**(2), 891–921 (2016)
3. Bonettini, S., Loris, I., Porta, F., Prato, M., Rebegoldi, S.: On the convergence of a line-search based proximal-gradient method for nonconvex optimization. Inverse Problems **33**(5) (2017)
4. Burke, J.V., Moré, J.J., Toraldo, G.: Convergence properties of trust region methods for linear and convex constraints. Mathematical Programming **47**(1-3), 305–336 (1990)
5. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing **16**, 1190–1208 (1995)
6. Byrd, R.H., Nocedal, J., Oztoprak, F.: An inexact successive quadratic approximation method for $L-1$ regularized optimization. Mathematical Programming **157**(2), 375–396 (2016)
7. Chouzenoux, E., Pesquet, J.C., Repetti, A.: Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. Journal of Optimization Theory and Applications **162**(1), 107–132 (2014)
8. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Modeling and Simulation **4**(4), 1168–1200 (2005)
9. Conn, A.R., Gould, N.I.M., Toint, P.L.: Global convergence of a class of trust region algorithms for optimization with simple bounds. SIAM Journal on Numerical Analysis **25**(2), 433–460 (1988)

10. Drusvyatskiy, D., Lewis, A.S.: Error bounds, quadratic growth, and linear convergence of proximal methods. Mathematics of Operations Research (2018)
11. Fletcher, R.: Practical Methods of Optimization. John Wiley and Sons (1987)
12. Ghanbari, H., Scheinberg, K.: Proximal quasi-Newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates. Computational Optimization and Applications **69**(3), 597–627 (2018)
13. Jiang, K., Sun, D., Toh, K.C.: An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. SIAM Journal on Optimization **22**(3), 1042–1064 (2012)
14. Lee, C.p., Chang, K.W.: Distributed block-diagonal approximation methods for regularized empirical risk minimization. Tech. rep. (2017)
15. Lee, C.p., Lim, C.H., Wright, S.J.: A distributed quasi-Newton algorithm for empirical risk minimization with nonsmooth regularization. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1646–1655. ACM, New York, NY, USA (2018)
16. Lee, C.p., Roth, D.: Distributed box-constrained quadratic optimization for dual linear SVM. In: Proceedings of the International Conference on Machine Learning (2015)
17. Lee, J.D., Sun, Y., Saunders, M.A.: Proximal Newton-type methods for minimizing composite functions. SIAM Journal on Optimization **24**(3), 1420–1443 (2014)
18. Li, D.H., Fukushima, M.: On the global convergence of the BFGS method for nonconvex unconstrained optimization problems. SIAM Journal on Optimization **11**(4), 1054–1064 (2001)
19. Li, J., Andersen, M.S., Vandenberghe, L.: Inexact proximal Newton methods for self-concordant functions. Mathematical Methods of Operations Research **85**(1), 19–41 (2017)
20. Lin, C.J., Moré, J.J.: Newton's method for large-scale bound constrained problems. SIAM Journal on Optimization **9**, 1100–1127 (1999)
21. Lin, H., Mairal, J., Harchaoui, Z.: Catalyst acceleration for first-order convex optimization: from theory to practice. Journal of Machine Learning Research **18**(212), 1–54 (2018)
22. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical programming **45**(1), 503–528 (1989)
23. Moré, J.J., Sorensen, D.C.: Computing a trust region step. SIAM Journal on Scientific and Statistical Computing **4**(3), 553–572 (1983)
24. Necoara, I., Nesterov, Y., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Mathematical Programming pp. 1–39 (2018)
25. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers (2004)
26. Nesterov, Y.: Gradient methods for minimizing composite functions. Mathematical Programming **140**(1), 125–161 (2013)
27. Nocedal, J., Wright, S.J.: Numerical optimization, second edn. Springer (2006)
28. Rodomanov, A., Kropotov, D.: A superlinearly-convergent proximal Newton-type method for the optimization of finite sums. In: Proceedings of the International Conference on Machine Learning, pp. 2597–2605 (2016)
29. Scheinberg, K., Tang, X.: Practical inexact proximal quasi-Newton method with global complexity analysis. Mathematical Programming **160**(1-2), 495–529 (2016)
30. Schmidt, M., Roux, N., Bach, F.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems, pp. 1458–1466 (2011)
31. Tran-Dinh, Q., Kyrillidis, A., Cevher, V.: An inexact proximal path-following algorithm for constrained convex minimization. SIAM Journal on Optimization **24**(4), 1718–1745 (2014)
32. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming **117**(1), 387–423 (2009)
33. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward-backward algorithms. SIAM Journal on Optimization **23**(3), 1607–1633 (2013)
34. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing **57**, 2479–2493 (2009)

35. Yang, T.: Trading computation for communication: Distributed stochastic dual coordinate ascent. In: Advances in Neural Information Processing Systems, pp. 629–637 (2013)
36. Zheng, S., Wang, J., Xia, F., Xu, W., Zhang, T.: A general distributed dual coordinate optimization framework for regularized loss minimization. Journal of Machine Learning Research **18**(115), 1–52 (2017)

A Proof of Lemma 5

Proof We have

$$
\begin{aligned}
Q^* &= \min_d \nabla f(x)^T d + \frac{1}{2} d^T H d + \psi(x+d) - \psi(x) \\
&\le \min_d f(x+d) + \psi(x+d) + \frac{1}{2} d^T H d - F(x) && \text{(56a)}\\
&\le F(x + \lambda(P_\Omega(x) - x)) + \frac{\lambda^2}{2}(P_\Omega(x) - x)^T H (P_\Omega(x) - x) - F(x) \quad \forall \lambda \in [0,1] && \text{(56b)}\\
&\le (1-\lambda) F(x) + \lambda F^* - \frac{\mu\lambda(1-\lambda)}{2}\|x - P_\Omega(x)\|^2 && \text{(56c)}\\
&\quad + \frac{\lambda^2}{2}(x - P_\Omega(x))^T H (x - P_\Omega(x)) - F(x) \quad \forall \lambda \in [0,1]\\
&\le \lambda(F^* - F(x)) - \frac{\mu\lambda(1-\lambda)}{2}\|x - P_\Omega(x)\|^2 + \frac{\lambda^2}{2}\|H\|\|x - P_\Omega(x)\|^2 \quad \forall \lambda \in [0,1],
\end{aligned}
$$

where in (56a) we used the convexity of f, in (56b) we set $d = \lambda(P_\Omega(x) - x)$, and in (56c) we used the optimal set strong convexity (10) of F. Thus we obtain (25). □

B Proof of Lemma 6

Proof Consider

$$\lambda_k = \arg\min_{\lambda \in [0,1]} -\lambda\delta_k + \frac{\lambda^2}{2} A_k, \tag{57}$$

then by setting the derivative to zero in (57), we have

$$\lambda_k = \min\left\{1, \frac{\delta_k}{A_k}\right\}. \tag{58}$$

When $\delta_k \ge A_k$, we have from (58) that $\lambda_k = 1$. Therefore, from (30) we get

$$\delta_{k+1} \le \delta_k + c_k\left(-\delta_k + \frac{A_k}{2}\right) \le \delta_k + c_k\left(-\delta_k + \frac{\delta_k}{2}\right) = \left(1 - \frac{c_k}{2}\right)\delta_k,$$

proving (31).

On the other hand, since $A \ge A_k > 0, c_k \ge 0$ for all k, (30) can be further upper-bounded by

$$\delta_{k+1} \le \delta_k + c_k\left(-\lambda_k\delta_k + \frac{A_k}{2}\lambda_k^2\right) \le \delta_k + c_k\left(-\lambda_k\delta_k + \frac{A}{2}\lambda_k^2\right), \quad \forall \lambda_k \in [0,1].$$

Now take

$$\lambda_k = \min\left\{1, \frac{\delta_k}{A}\right\}. \tag{59}$$

For $\delta_k \geq A \geq A_k$, (31) still applies. If $A > \delta_k$, we have from (59) that $\lambda_k = \delta_k / A$, hence

$$\delta_{k+1} \leq \delta_k - \frac{c_k}{2A} \delta_k^2. \tag{60}$$

This together with (31) imply that $\{\delta_k\}$ is a monotonically decreasing sequence. Dividing both sides of (60) by $\delta_{k+1}\delta_k$, and from the fact that δ_k is decreasing and nonnegative, we conclude

$$\delta_k^{-1} \leq \delta_{k+1}^{-1} - \frac{c_k \delta_k}{2\delta_{k+1} A} \leq \delta_{k+1}^{-1} - \frac{c_k}{2A}$$

Summing this inequality from k_0, and using $\delta_{k_0} < A$, we obtain

$$\delta_k^{-1} \geq \delta_{k_0}^{-1} + \frac{\sum_{t=k_0}^{k-1} c_t}{2A} \geq \frac{\sum_{t=k_0}^{k-1} c_t + 2}{2A} \Rightarrow \delta_k \leq \frac{2A}{\sum_{t=k_0}^{k-1} c_t + 2},$$

proving (32). □