



**ISE**

Industrial and  
Systems Engineering

# Gradient Sampling Methods with Inexact Subproblem Solves and Gradient Aggregation

FRANK E. CURTIS AND MINHAN LI

Department of Industrial and Systems Engineering, Lehigh University

COR@L Technical Report 20T-009-R1



**LEHIGH**  
UNIVERSITY.

**COR@L**  
COMPUTATIONAL OPTIMIZATION  
RESEARCH AT LEHIGH 

# Gradient Sampling Methods with Inexact Subproblem Solves and Gradient Aggregation\*

FRANK E. CURTIS<sup>†1</sup> AND MINHAN LI<sup>‡1</sup>

<sup>1</sup>Department of Industrial and Systems Engineering, Lehigh University

Original Publication: May 15, 2020

Last Revised: August 6, 2021

## Abstract

Gradient sampling (GS) has proved to be an effective methodology for the minimization of objective functions that may be nonconvex and/or nonsmooth. The most computationally expensive component of a contemporary GS method is the need to solve a convex quadratic subproblem in each iteration. In this paper, a strategy is proposed that allows the use of inexact solutions of these subproblems, which, as proved in the paper, can be incorporated without the loss of theoretical convergence guarantees. Numerical experiments show that by exploiting inexact subproblem solutions, one can consistently reduce the computational effort required by a GS method. Additionally, a strategy is proposed for aggregating gradient information after a subproblem is solved (potentially inexactly), as has been exploited in bundle methods for nonsmooth optimization. It is proved that the aggregation scheme can be introduced without the loss of theoretical convergence guarantees. Numerical experiments show that incorporating this gradient aggregation approach substantially reduces the computational effort required by a GS method.

## 1 Introduction

The gradient sampling (GS) methodology has proved to be effective for solving nonsmooth, nonconvex minimization problems. Based on the conceptually simple idea of computing an approximate  $\epsilon$ -steepest-descent direction at a point by finding the minimum-norm element of the convex hull of gradients evaluated at randomly generated nearby points, one can prove convergence to stationarity of a GS method under relatively loose assumptions. That said, here are two ways in which implementations of GS methods could be more efficient:

- Each iteration of a GS method requires the solution of a convex quadratic subproblem (QP) for computing a search direction. The overall computational expense of a GS method could be reduced if one could terminate each call to a QP solver early, then employ the inexact QP solution as the search direction in the “outer” GS method. Such an inexact solution might cause a search direction to be less productive than if an exact QP solution were computed, meaning that more “outer” iterations may be required. However, as in other optimization algorithms that exploit inexact subproblem solutions, one might still obtain overall computational savings through consistently reduced per-iteration costs.

---

\*This material is based upon work supported by the National Science Foundation under grant numbers CCF-1618717 and CCF-1740796.

<sup>†</sup>E-mail: frank.e.curtis@lehigh.edu

<sup>‡</sup>E-mail: mil417@lehigh.edu

- Bundle methods represent another important class of algorithms for nonsmooth minimization. Implementations of bundle methods can be made significantly more efficient through the use of *subgradient aggregation*, wherein one can compress the information from a QP solution such that a subsequent QP can be solved more rapidly. Implementations of GS methods could be made more efficient if such an idea could be incorporated.

In this paper, we propose enhancements to the GS methodology such that one can exploit inexact subproblem solutions and *gradient* aggregation. (We do not refer to *subgradient* aggregation since the GS methodology requires the identification of points at which the objective function is differentiable or even continuously differentiable, at which gradients are to be evaluated, when search directions are being computed.) We show techniques for exploiting these ideas that maintain the convergence guarantees of previously proposed GS methods. Implementations of our ideas in a C++ software package show that exploiting both inexact subproblem solutions and gradient aggregation can lead to consistently noticeable reductions in required computational effort.

## 1.1 Literature Review

The GS methodology was introduced by [4]; see also [3]. Shortly after, [24] showed an elegant convergence analysis of a GS method, and showed how the convergence guarantees could be maintained by multiple variations of the basic approach. Later, [10, 11] showed how one could sample gradients adaptively and introduce quasi-Newton Hessian approximations to maintain convergence guarantees while improving practical performance. (Here, as is common in the literature on quasi-Newton methods for solving nonsmooth optimization problems, we use the term “Hessian approximation” loosely; rather than as an approximate second-derivative matrix, it should merely be thought of as a matrix that approximates local changes in the gradient at points at which  $f$  is differentiable.) See also [12] for how to loosen the restrictions on the Hessian approximation scheme. A feature of the algorithms in all of these articles is that the analyses require that the convex QP subproblems for computing search directions must be solved exactly in every iteration.

A method for reducing the costs associated with solving QPs in a GS method was proposed by [31]. In this work, the authors argue that an “ideal” direction, which can be computed using a relatively inexpensive procedure, can be used in place of a QP solution when it is found to be sufficiently large in norm. The authors argue that convergence guarantees are maintained with this replacement, and show empirically that fewer QPs need to be solved. Our proposed approach is different from this one in two main respects. First, rather than prescribe a formula for a particular direction that may be used, our algorithm involves conditions for an inexact QP solution that are more generic. This gives more computational flexibility to the algorithm. Second, whereas the algorithm by [31] still requires that some QPs be solved exactly—such as when the “ideal” direction is too small in norm, which occurs when approaching stationarity—our algorithm allows for inexact solutions of the QPs in all cases.

GS ideas have been extended in various ways, such as to attain good local convergence rate properties [17] and to solve constrained optimization problems [9, 19, 38]. See [5] for further discussion. Such extensions are beyond the scope of this article, wherein we focus on techniques for unconstrained minimization that ensure convergence from an arbitrarily chosen starting point. That said, our proposed enhancements could be employed in conjunction with these extensions.

Another prevailing methodology for solving nonsmooth optimization problems is the class of bundle methods, which have a long history [1, 21, 22, 23, 14, 15, 16, 18, 25, 29, 33, 34, 36, 37]. The technique employed in some bundle methods that is relevant for this paper is that of subgradient aggregation; see, e.g., [21]. The use of aggregation in this paper is similar, although the surrounding convergence analysis is different due to the distinct differences in the convergence analyses of bundle and GS methodologies. For one thing, convergence analyses of GS methods are inherently probabilistic due to the random sampling of points.

## 1.2 Notation

We write  $\mathbb{R}$  to denote the set of real numbers,  $\mathbb{R}^n$  to denote the set of  $n$ -dimensional real vectors, and  $\mathbb{R}^{m \times n}$  to denote the set of  $m$ -by- $n$ -dimensional real matrices. We write  $\mathbb{N} := \{0, 1, 2, \dots\}$  to denote the set of nonnegative integers and use  $\mathbf{1}$  to denote a vector of ones whose length is determined by the context in which it appears (e.g., through an inner product with a vector of known length).

Throughout the paper, we consider the minimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

where the objective function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfies the following assumption.

**Assumption 1.1.** *The objective function  $f$  is bounded below over  $\mathbb{R}^n$ , locally Lipschitz on  $\mathbb{R}^n$ , and continuously differentiable on an open set  $\mathcal{D}$  with full measure in  $\mathbb{R}^n$ .*

We propose GS algorithms, each of which is designed to produce an iterate sequence—i.e.,  $\{x_k\}$  with  $x_k \in \mathbb{R}^n$  for all  $k \in \mathbb{N}$ —converging to stationarity of  $f$ , which is to say that any cluster point of  $\{x_k\}$  is stationary for  $f$ . Throughout, we refer to stationarity in the sense of [7]. Such stationarity for  $f$  can be defined as follows. By Rademacher’s theorem under Assumption 1.1, the (Clarke) set of generalized gradients of  $f$  at  $x \in \mathbb{R}^n$  is given by

$$\partial f(x) = \text{conv} \left\{ \lim_{k \rightarrow \infty} \nabla f(x_k) : \{x_k\} \rightarrow x \text{ with } x_k \in \mathcal{D} \text{ for all } k \in \mathbb{N} \right\}; \tag{2}$$

see [7, Theorem 2.5.1]. For  $\epsilon \in [0, \infty)$ , the set of  $\epsilon$ -generalized gradients of  $f$  at  $x \in \mathbb{R}^n$  is

$$\partial_\epsilon f(x) = \text{conv} \partial f(\mathbb{B}(x, \epsilon)), \quad \text{where } \mathbb{B}(x, \epsilon) := \{\bar{x} \in \mathbb{R}^n : \|\bar{x} - x\|_2 \leq \epsilon\}. \tag{3}$$

One finds that  $\partial_0 f(x) \equiv \partial f(x)$ ; see [13, Corollary 2.5]. A point  $x \in \mathbb{R}^n$  is said to be  $\epsilon$ -stationary for  $f$  if  $0 \in \partial_\epsilon f(x)$  and is said to be stationary for  $f$  if  $0 \in \partial f(x)$ .

The first algorithm that we propose (see Algorithm 1 on page 7) has a nested loop (with the “inner” loop being stated in Algorithm 2 on page 8). Iterations for the “outer” loop are indexed by  $k \in \mathbb{N}$ . We apply this iteration number subscript to other values—in addition to  $x_k$ —computed in the outer loop of the algorithm. Iterations for the inner loop are indexed by  $j \in \mathbb{N}$ . Quantities computed during the inner loop are denoted with a double-subscript; e.g.,  $d_{k,j}$ .

One could remove from Assumption 1.1 the assumption that  $f$  is bounded below, in which case the methods that we propose would terminate finitely at a stationary point for  $f$  or, with probability one, generate iterates that either converge to ( $\epsilon$ -)stationarity for  $f$  (see Theorems 2.1 and 3.1) or have objective values that diverge to  $-\infty$ . However, to focus on the more interesting setting, we include in Assumption 1.1 that  $f$  is bounded below, meaning this latter case cannot occur. For the algorithms that we propose to be well posed, one only needs to assume that  $f$  is (not necessarily continuously) differentiable in an open set with full measure in  $\mathbb{R}^n$ . However, a theoretical guarantee of convergence to stationarity requires that  $f$  be continuously differentiable over such a set, as we have included in Assumption 1.1. See [5] for further discussion.

## 1.3 Outline

In Section 2, we propose and analyze an algorithm that employs inexact subproblem solutions. In Section 3, we propose gradient aggregation within a GS method and show that it can be used while maintaining the same guarantees as the method from Section 2. Numerical experiments employing both techniques are presented in Section 4. Concluding remarks are given in Section 5.

## 2 GS Algorithm with Inexact Subproblem Solutions

We propose a GS algorithm that allows for the use of inexact subproblem solutions in each iteration. In this section, we present the proposed algorithm, then prove that iterates generated by it converge to ( $\epsilon$ -)stationarity with probability one. In our presentation, we focus on the components of the algorithm and analysis that are distinct from previous GS methods. Components that are not unique are provided in an online companion to this article.

### 2.1 Algorithm Description

In iteration  $k \in \mathbb{N}$  of our proposed algorithm, an iterate  $x_k \in \mathcal{D}$  is available along with a sampling radius  $\epsilon_k \in (0, \infty)$ , a set of sample points

$$\mathcal{X}_k := \{x_{k,0}, x_{k,1}, \dots, x_{k,p_k}\} \subset \mathbb{B}(x_k, \epsilon_k) \cap \mathcal{D} \text{ where } x_{k,0} \equiv x_k \text{ for some } p_k \in \mathbb{N},$$

and the corresponding matrix of gradients

$$G_k := [\nabla f(x_{k,0}) \quad \nabla f(x_{k,1}) \quad \cdots \quad \nabla f(x_{k,p_k})] \in \mathbb{R}^{n \times (p_k+1)}. \quad (4)$$

Given this matrix of gradients, a symmetric positive definite Hessian approximation  $H_k$ , and the corresponding inverse  $W_k := H_k^{-1}$ , the search direction is computed by *approximately* solving the primal-dual pair of quadratic optimization problems (QPs) given by

$$(P) := \left\{ \begin{array}{l} \min_{(d,z) \in \mathbb{R}^{n+1}} z + \frac{1}{2} \|d\|_{H_k}^2 \\ \text{s.t. } G_k^T d \leq z \mathbf{1} \end{array} \right\} \text{ and } (D) := \left\{ \begin{array}{l} \max_{y \in \mathbb{R}^{p_k+1}} -\frac{1}{2} \|G_k y\|_{W_k}^2 \\ \text{s.t. } \mathbf{1}^T y = 1, y \geq 0 \end{array} \right\}. \quad (5)$$

We assume that both  $H_k$  and  $W_k$  are available for all  $k \in \mathbb{N}$ . It is straightforward to maintain both approximations through the use of quasi-Newton techniques.

Letting  $(d_{k,*}, z_{k,*})$  denote the optimal solution of  $(P)$  for each  $k \in \mathbb{N}$ , one finds that the solution component  $d_{k,*}$  can be viewed as the minimizer of the piecewise quadratic function

$$\max_{i \in \{0, \dots, p_k\}} \{\nabla f(x_{k,i})^T d\} + \frac{1}{2} \|d\|_{H_k}^2.$$

The optimal solution  $y_{k,*}$  of  $(D)$  can be viewed as the vector such that  $G_k y_{k,*}$  is the least  $W_k$ -norm element of the convex hull of the columns of  $G_k$ , i.e., the  $W_k$ -projection of the origin onto this hull. The following lemma reveals important properties of these solutions.

**Lemma 2.1.** *For all  $k \in \mathbb{N}$ , either  $(d_{k,*}, z_{k,*}) = (0, 0)$  and the origin lies in the convex hull of the columns of  $G_k$ , or  $d_{k,*}$  is a direction of strict descent for  $f$  at  $x_k$  with*

$$\nabla f(x_k)^T d_{k,*} \leq -d_{k,*}^T H_k d_{k,*} < 0. \quad (6)$$

*In all cases,  $d_{k,*} = -W_k G_k y_{k,*}$  and  $\|G_k y_{k,*}\|_{W_k} = \|d_{k,*}\|_{H_k}$ .*

*Proof.* Proof. The properties follow from the Karush-Kuhn-Tucker conditions for (5); see, e.g., [10, Eq. (27)] and [11, Lemma 2.2].  $\square$   $\square$

As our focus is on an algorithm that solves (5) *approximately* for all  $k \in \mathbb{N}$ , the statement of our algorithm is facilitated by defining, in each “outer” iteration, sequences of inner iterates of a solver for the primal-dual subproblems (5). Let  $\{(d_{k,j}, z_{k,j})\}$  and  $\{y_{k,j}\}$  be sequences of primal and dual iterates, respectively, generated when (5) is solved iteratively. Our algorithm requires that *both* primal and dual QP iterate sequences are generated. However, this is not an expensive requirement. After all, motivated by Lemma 2.1, one may choose for a given  $y_{k,j} \in \mathbb{R}^{p_k+1}$  to set

$$d_{k,j} \leftarrow -W_k G_k y_{k,j} \text{ and } z_{k,j} \leftarrow \max_{i \in \{0, \dots, p_k\}} \nabla f(x_{k,i})^T d_{k,j}, \quad (7)$$

in which case one only needs to generate a dual iterate sequence and a corresponding sequence of primal-feasible solutions is obtained through (7). In addition, to reduce expense, one does not need to evaluate (7) in each inner iteration; one might only evaluate it and check for termination periodically and/or after an initial number of inner iterations. In any case, for the sake of generality, we define our algorithm to allow  $\{d_{k,j}\}$  and  $\{-W_k G_k y_{k,j}\}$  to differ.

With respect to the QP solver, we merely assume that the following holds.

**Assumption 2.1.** *For all  $k \in \mathbb{N}$ , the primal and dual iterates when solving (5) satisfy  $\{(d_{k,j}, z_{k,j}, y_{k,j})\} \rightarrow (d_{k,*}, z_{k,*}, y_{k,*})$ . In addition, for all  $(k, j) \in \mathbb{N} \times \mathbb{N}$ , one has*

$$G_k^T d_{k,j} \leq z_{k,j} \mathbf{1}, \quad \mathbf{1}^T y_{k,j} = 1, \quad \text{and} \quad y_{k,j} \geq 0,$$

*i.e.,  $(d_{k,j}, z_{k,j}, y_{k,j})$  is primal-dual feasible for all  $(k, j) \in \mathbb{N} \times \mathbb{N}$ .*

Under Assumption 2.1, the primal and dual iterates satisfy weak duality with respect to (5) for all  $(k, j) \in \mathbb{N} \times \mathbb{N}$ . In particular, defining the QP primal and dual objective functions  $q_k : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$  and  $\theta_k : \mathbb{R}^{p_k+1} \rightarrow \mathbb{R}$ , respectively, where

$$q_k(d, z) = z + \frac{1}{2} \|d\|_{H_k}^2 \quad \text{and} \quad \theta_k(y) = -\frac{1}{2} \|G_k y\|_{W_k}^2,$$

one has that  $q_k(d_{k,j}, z_{k,j}) \geq \theta_k(y_{k,j})$  for all  $(k, j) \in \mathbb{N} \times \mathbb{N}$ .

Our algorithm with inexact subproblem solutions is stated as Algorithm 1 on page 7, for which the details of the search direction computation are stated in Algorithm 2 on page 8. The statement of Algorithm 1 focuses on its unique aspects related to the conditions that we require of inexact QP solutions. Other subroutines that we employ for the line search, iterate perturbation strategy (a feature required by the theoretical convergence analyses of all GS methods), sample set updates, and quasi-Newton updates are similar to those used in [11, 12]. Hence, we relegate them to the online companion. The algorithm also requires a subroutine for setting parameters related to the quasi-Newton updates that influence the line search subroutine. The approach is derived from properties of the self-correcting nature of quasi-Newton updating; see [6]. This subroutine is also provided in the online companion.

Each iteration of Algorithm 2 takes a new approximate subproblem solution from the QP solver. The loop terminates in one of two situations. If (10) holds, then one has obtained a dual iterate such that the corresponding convex combination of columns of  $G_k$  is sufficiently small in appropriate norms. In this case, one has identified that the current iterate may be sufficiently close to  $\epsilon_k$ -stationarity, in which case Algorithm 1 reduces the sampling radius. On the other hand, if (11) holds along with either (12) or (13), then our analysis in the following subsection reveals that a sufficiently accurate QP solution yielding a direction of sufficient descent has been obtained. The condition (11) is motivated by Lemma 2.1, specifically (6), since  $(d_k, y_k) = (-W_k G_k y_{k,j_\theta}, y_{k,j_\theta})$  yields

$$\nabla f(x_k)^T d_k = -\nabla f(x_k)^T W_k G_k y_k \stackrel{(11)}{\leq} -\kappa y_k^T G_k^T W_k G_k y_k = -\kappa d_k^T H_k d_k.$$

The role played by conditions (12) and (13), which make use of the values defined in (8) and (9), is explained in the following subsection.

Notice that an implementation of Algorithm 2 does not require storage and a search through all previous subproblem solutions when determining the indices in Line 3. One only needs to store the best (in terms of objective values) primal and dual solution estimates during the loop and employ these values when checking for termination of the loop. Line 3 is only written in this manner for ease of exposition, and to allow us to consider situations in which these inner iterations do not necessarily produce primal and dual subproblem solutions that have objective values that converge monotonically to the optimal value.

## 2.2 Inexactness conditions for the QP solver

Convergence analyses of GS methods rely on a fundamental property of any compact, convex set, call it  $\mathcal{S} \subseteq \mathbb{R}^n$ , that does not contain the origin. Intuitively, this property is that if  $u \in \mathcal{S}$  is sufficiently close to the

---

**Algorithm 1** Gradient Sampling Algorithm with Inexact Subproblem Solutions
 

---

**Require:**  $(\sigma, \underline{\alpha}) \in (0, \infty)^2$ ;  $(\iota, \underline{\phi}) \in (0, 1)^2$ ;  $\bar{\phi} \in (1, \infty)$ ;  $\psi \in (0, 1]$ ;  $p \in \mathbb{N}$  with  $p \geq n + 1$ ;  $x_0 \in \mathcal{D}$ ;  $H_0 \succ 0$ ;  $\epsilon_0 \in (0, \infty)$ .

- 1: Set  $W_0 \leftarrow H_0^{-1}$ ,  $\mathcal{X}_0 \leftarrow \{x_0\}$ ,  $p_0 \leftarrow 0$ ,  $G_0$  by (4), and  $\sigma_0 \leftarrow \sigma$ .
- 2: Set  $(\eta, \mu) \in (0, 1) \times (1, \infty)$  by Algorithm 4.
- 3: Choose  $\bar{\eta} \in (\eta, 1)$ .
- 4: **for all**  $k \in \mathbb{N}$  **do**
- 5: **if**  $\|\nabla f(x_k)\|_2 = 0$  **then**
- 6: **terminate** and **return** the stationary point  $x_k$ .
- 7: **end if**
- 8: Set  $y_{k,j_\theta}$  by Algorithm 2 (page 8).
- 9: Set  $(d_k, y_k) \leftarrow (-W_k G_k y_{k,j_\theta}, y_{k,j_\theta})$ .
- 10: Set  $\alpha_k \geq 0$  by Algorithm 5.
- 11: **if** (10) holds (with  $y_{k,j_\theta} \equiv y_k$ )
- 12: set  $\epsilon_{k+1} \leftarrow \psi \epsilon_k$  and  $\sigma_{k+1} \leftarrow \sigma$ ;
- 13: **else if**  $\alpha_k \geq \underline{\alpha}$
- 14: set  $\epsilon_{k+1} \leftarrow \epsilon_k$  and  $\sigma_{k+1} \leftarrow \sigma_k$ ;
- 15: **else**
- 16: set  $\epsilon_{k+1} \leftarrow \epsilon_k$  and  $\sigma_{k+1} \leftarrow \iota \sigma_k$ .
- 17: **end if**
- 18: Set  $x_{k+1} \in \mathcal{D}$  by Algorithm 6.
- 19: Set  $(H_{k+1}, W_{k+1})$  by Algorithm 7.
- 20: Set  $(\mathcal{X}_{k+1}, p_{k+1})$  by Algorithm 8 and  $G_{k+1}$  by (4).
- 21: **end for**

---

projection of the origin onto  $\mathcal{S}$ , then  $u$  makes a sufficiently acute angle (with respect to a given metric) with any  $v \in \mathcal{S}$ . Such a lemma appears as [4, Lemma 3.1] and [24, Lemma 3.1], and is proved in a more general setting as [11, Lemma 3.5]. The conditions that we impose on inexact subproblem solutions are motivated by trying to ensure a property of this type, but in an even slightly more general setting. Specifically, the lemma that we use is the following. In the lemma, we refer to the concept of a  $W$ -projection (with  $W \succ 0$ ) of the origin onto a compact, convex set  $\mathcal{S}$ , i.e.,

$$P_W(\mathcal{S}) := \arg \min_{s \in \mathcal{S}} \|s\|_W. \quad (14)$$

Our new generalization of the lemma can be seen in the inequality (15), which does not require that a given vector  $u \in \mathcal{S}$  is sufficiently close to the  $W$ -projection of the origin, but merely sufficiently close to a small enough neighborhood of this projection.

**Lemma 2.2.** *Suppose  $\mathcal{S} \subseteq \mathbb{R}^n$  is a compact and convex set with  $0 \notin \mathcal{S}$ . For any  $\beta \in (0, 1)$  and  $W \succ 0$ , there exists  $(\varsigma, \delta) \in (0, \infty)^2$  such that, for any  $(u, v, \bar{\varsigma}, \bar{\delta}) \in \mathcal{S} \times \mathcal{S} \times (0, \varsigma] \times (0, \delta]$  with*

$$\|u\|_W \leq (1 + \varsigma) \|P_W(\mathcal{S})\|_W + \delta \quad (15)$$

(where  $P_W(\mathcal{S})$  is defined in (14)), it follows that  $v^T W u > \beta \|u\|_W^2$ .

*Proof.* Proof. Consider arbitrary  $\beta \in (0, 1)$  and  $W \succ 0$ . To derive a contradiction, suppose the implication is false, that is, for any  $(\varsigma, \delta) \in (0, \infty)^2$  there exists  $(u, v, \bar{\varsigma}, \bar{\delta}) \in \mathcal{S} \times \mathcal{S} \times (0, \varsigma] \times (0, \delta]$  with

$$\|u\|_W \leq (1 + \bar{\varsigma}) \|P_W(\mathcal{S})\|_W + \bar{\delta} \quad \text{and} \quad v^T W u \leq \beta \|u\|_W^2.$$

This means that there exist infinite sequences  $\{u_i\} \subset \mathcal{S}$  and  $\{v_i\} \subset \mathcal{S}$  such that

$$\|u_i\|_W \leq (1 + 1/i) \|P_W(\mathcal{S})\|_W + 1/i \quad \text{and} \quad v_i^T W u_i \leq \beta \|u_i\|_W^2 \quad \text{for all } i \in \mathbb{N}. \quad (16)$$

---

**Algorithm 2** Search Direction Computation for Algorithm 1
 

---

**Require:**  $\nu \in (0, \infty)$ ;  $(\rho, \kappa) \in (0, 1)^2$

1: Set

$$\tau_k \leftarrow \sigma_k^2 + 2\sigma_k \in (0, 1). \quad (8)$$

2: **for all**  $j \in \mathbb{N}$  **do**

3: Set  $j_q \leftarrow \arg \min_{i \in \{0, \dots, j\}} q_k(d_{k,i}, z_{k,i})$  and  $j_\theta \leftarrow \arg \max_{i \in \{0, \dots, j\}} \theta_k(y_{k,i})$ .

4: **if**  $q_k(d_{k,j_q}, z_{k,j_q}) \geq 0$ , **then** set  $\lambda_{k,j_q} = \infty$

5: **else** set

$$\lambda_{k,j_q} \leftarrow \max \left\{ 1 - \frac{\sigma_k^2 + 2\sigma_k}{\frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} - 1}, \rho \right\}. \quad (9)$$

6: **end if**

7: **if**

$$\max\{\|W_k G_k y_{k,j_\theta}\|_2, \|G_k y_{k,j_\theta}\|_2\} \leq \nu \epsilon_k, \quad (10)$$

8: **then terminate** and **return**  $y_{k,j_\theta}$ ;

9: **else if**

$$-\nabla f(x_k)^T W_k G_k y_{k,j_\theta} \leq -\kappa y_{k,j_\theta}^T G_k^T W_k G_k y_{k,j_\theta} \quad (11)$$

10: **and either**

$$q_k(d_{k,j_q}, z_{k,j_q}) - \theta_k(y_{k,j_\theta}) \leq \tau_k (-q_k(d_{k,j_q}, z_{k,j_q})) \quad (12)$$

11: **or**

$$\theta_k(y_{k,j_\theta}) - \theta_k(y_{k,0}) \geq \lambda_{k,j_q} (q_k(d_{k,j_q}, z_{k,j_q}) - \theta_k(y_{k,0})) \quad (13)$$

12: **then terminate** and **return**  $y_{k,j_\theta}$ .

13: **end if**

14: **end for**

---

Since  $\mathcal{S}$  is compact, these sequences have convergent subsequences; hence, without loss of generality, one can assume that  $\{u_i\} \rightarrow u$  and  $\{v_i\} \rightarrow v$  for some  $(u, v) \in \mathcal{S} \times \mathcal{S}$  with

$$v^T W u \leq \beta \|u\|_W^2. \quad (17)$$

On the other hand, by the definition of  $\{u_i\}$ , it follows that  $u = P_W(\mathcal{S})$ , which is nonzero since  $\mathcal{S}$  does not include the origin. Moreover, by [2, Proposition 1.1.8] and the definition of  $P_W(\mathcal{S})$  (as the  $W$ -projection of the origin onto  $\mathcal{S}$ ), one finds that

$$(0 - u)^T W (v - u) \leq 0 \iff v^T W u \geq \|u\|_W^2,$$

which contradicts (17) since  $\beta \in (0, 1)$ . □ □

Our goal now is to prove two lemmas that motivate the use of (12) and (13) as stopping conditions for the loop in Algorithm 2. Specifically, if  $\theta_k(y_{k,*}) < 0$ , each lemma shows that these conditions—(12) and (13), respectively, in the two lemmas—imply that

$$\begin{aligned} 0 &> \theta_k(y_{k,j_\theta}) \geq (1 + \sigma_k)^2 \theta_k(y_{k,*}) \\ \iff 0 &< \|G_k y_{k,j_\theta}\|_{W_k} \leq (1 + \sigma_k) \|G_k y_{k,*}\|_{W_k}. \end{aligned} \quad (18)$$

Importantly, these algorithmic conditions imply that (18) holds *without knowledge of*  $y_{k,*}$ . The inequalities in (18) are important since they, along with Lemma 2.2 (c.f. (15)), play a central role in our convergence analysis in §2.3 for Algorithm 1.

**Lemma 2.3.** *Suppose that, in iteration  $k \in \mathbb{N}$  of Algorithm 1, one has  $\theta_k(y_{k,*}) < 0$ . In addition, suppose that, during iteration  $j \in \mathbb{N}$  of Algorithm 2 (during outer iteration  $k \in \mathbb{N}$ ), one finds that (12) holds. Then, (18) holds.*

*Proof.* Proof. By weak duality for (5), one has that

$$\begin{aligned} \theta_k(y_{k,*}) - \theta_k(y_{k,j_\theta}) &\leq q_k(d_{k,j_q}, z_{k,j_q}) - \theta_k(y_{k,j_\theta}) \\ \text{and } -q_k(d_{k,j_q}, z_{k,j_q}) &\leq -\theta_k(y_{k,*}). \end{aligned}$$

Combined with (12) and (8), it follows that

$$\theta_k(y_{k,*}) - \theta_k(y_{k,j_\theta}) \leq \tau_k(-\theta_k(y_{k,*})) = (\sigma_k^2 + 2\sigma_k)(-\theta_k(y_{k,*})),$$

which shows that (18) holds, as desired.  $\square$   $\square$

When  $\theta_k(y_{k,*}) < 0$ , weak duality for (5) implies that (12) can hold only if  $q_k(d_{k,j_q}, z_{k,j_q}) < 0$ . Hence, one does not need to check if  $q_k(d_{k,j_q}, z_{k,j_q}) < 0$  holds before employing (12) as a stopping condition for the QP solver. By contrast, the next lemma shows that (13) should be used as a stopping condition for the QP solver only if  $q_k(d_{k,j_q}, z_{k,j_q}) < 0$ . Algorithm 2 ensures this by setting  $\lambda_{k,j_q} \leftarrow \infty$  when  $q_k(d_{k,j_q}, z_{k,j_q}) \geq 0$ , and otherwise the lemma shows that  $\lambda_{k,j_q} \in (0, 1)$ .

**Lemma 2.4.** *Suppose that, in iteration  $k \in \mathbb{N}$  of Algorithm 1, one has  $\theta_k(y_{k,*}) < 0$ . In addition, suppose that, during iteration  $j \in \mathbb{N}$  of Algorithm 2 (during outer iteration  $k \in \mathbb{N}$ ), one finds that  $q_k(d_{k,j_q}, z_{k,j_q}) < 0$  and (13) holds. Then, (18) holds.*

*Proof.* Proof. By  $q_k(d_{k,j_q}, z_{k,j_q}) < 0$  and weak duality for (5), one finds in (9) that

$$\frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} \geq 1. \quad (19)$$

If  $\theta_k(y_{k,0}) = q_k(d_{k,j_q}, z_{k,j_q})$ , then  $(d_{k,j_q}, z_{k,j_q}, y_{k,0})$  is a primal-dual solution of (5) and  $\theta_k(y_{k,0}) = \theta_k(y_{k,j_\theta}) = \theta_k(y_{k,*})$ , which means that (18) holds. Hence, we may proceed under the assumption that  $\theta_k(y_{k,0}) < q_k(d_{k,j_q}, z_{k,j_q}) < 0$ , which implies that (19) holds strictly. Observing (9), one finds  $\lambda_{k,j_q} \in (0, 1)$ . This fact, (13), and weak duality for (5) imply

$$\begin{aligned} \theta_k(y_{k,j_\theta}) - \theta_k(y_{k,0}) &\geq \lambda_{k,j_q}(q_k(d_{k,j_q}, z_{k,j_q}) - \theta_k(y_{k,0})) \\ &\geq \lambda_{k,j_q}(\theta_k(y_{k,*}) - \theta_k(y_{k,0})) \geq 0, \end{aligned}$$

which along with  $\lambda_{k,j_q} \in (0, 1)$  and the facts that  $\theta_k(y_{k,*}) < 0$  and

$$\frac{\theta_k(y_{k,0})}{\theta_k(y_{k,*})} \leq \frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} \implies \theta_k(y_{k,0}) \geq \frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} \theta_k(y_{k,*})$$

implies that

$$\begin{aligned} \theta_k(y_{k,j_\theta}) &\geq \lambda_{k,j_q} \theta_k(y_{k,*}) + (1 - \lambda_{k,j_q}) \theta_k(y_{k,0}) \\ &\geq \left( \lambda_{k,j_q} + (1 - \lambda_{k,j_q}) \frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} \right) \theta_k(y_{k,*}). \end{aligned} \quad (20)$$

In addition, one finds that  $\lambda_{k,j_q}$  in (9) satisfies

$$\lambda_{k,j_q} \geq 1 - \frac{\sigma_k^2 + 2\sigma_k}{\frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} - 1} = \frac{\frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} - (1 + \sigma_k)^2}{\frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} - 1},$$

implying that

$$\lambda_{k,j_q} + (1 - \lambda_{k,j_q}) \frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} \leq (1 + \sigma_k)^2,$$

which along with (20) and the fact that  $\theta_k(y_{k,*}) < 0$  shows that

$$\theta_k(y_{k,j_\theta}) \geq \left( \lambda_{k,j_q} + (1 - \lambda_{k,j_q}) \frac{\theta_k(y_{k,0})}{q_k(d_{k,j_q}, z_{k,j_q})} \right) \theta_k(y_{k,*}) \geq (1 + \sigma_k)^2 \theta_k(y_{k,*}),$$

as desired.  $\square$   $\square$

### 2.3 Convergence Analysis

In this section, we show under Assumptions 1.1 and 2.1 that Algorithm 1 either terminates finitely with a stationary point for  $f$  or, with probability one, generates a sequence of iterates that converge to stationarity for  $f$ . Throughout this section, let  $\mathcal{K}$  be the indices of the outer iterations performed by the algorithm before termination (if the algorithm ever terminates) or the failure of a subroutine (if a subroutine ever fails). The subroutines that may fail are the iteration perturbation procedure (Algorithm 6 in the online companion) and the sample set update (Algorithm 8 in the online companion), wherein failure means that a loop does not terminate. If such an event occurs in iteration  $k$ , then  $\mathcal{K} = \{1, \dots, k\}$ . If the algorithm never terminates and no subroutine ever fails, then one simply has that the iterations performed are  $\mathcal{K} = \mathbb{N}$ .

We begin by showing that the algorithm is well posed along with important properties of the subroutines stated in the online companion.

**Lemma 2.5.** *Algorithm 1 is well posed; it either terminates finitely or, with probability one, it performs an infinite number of iterations. In any case, for all  $k \in \mathcal{K}$ , the following hold true.*

(a)  $H_k \succ 0$  and  $W_k = H_k^{-1} \succ 0$ .

(b)  $(d_k, y_k)$  satisfies  $\|d_k\|_{H_k} = \|G_k y_k\|_{W_k}$ .

(c) In Line 10, Algorithm 5 terminates finitely with  $\alpha_k \geq 0$ . If  $p_k < p$ , then  $\alpha_k = 0$  or  $\alpha_k \in [\underline{\alpha}, \bar{\alpha}]$ . Otherwise, if  $p_k = p$ , then  $\alpha_k \in (0, \bar{\alpha}]$ . In any case, if  $\alpha_k > 0$ , then

$$f(x_k) - f(x_k + \alpha_k d_k) > \underline{\eta} \alpha_k \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\} \quad (21a)$$

$$\text{and } v^T d_k \geq \bar{\eta} \nabla f(x_k)^T d_k, \text{ where } v \in \partial f(x_k + \alpha_k d_k), \quad (21b)$$

or at least (21a) holds (which is sufficient if deemed by Algorithm 5).

(d) In Line 18, Algorithm 6 yields, with probability one,  $x_{k+1} \in \mathcal{D}$  satisfying

$$f(x_k) - f(x_{k+1}) \geq \underline{\eta} \alpha_k \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\}, \quad (22a)$$

$$\nabla f(x_{k+1})^T d_k \geq \bar{\eta} \nabla f(x_k)^T d_k, \quad (22b)$$

$$\text{and } \|x_k + \alpha_k d_k - x_{k+1}\|_2 \leq \min\{\alpha_k, \epsilon_k\} \min\{\|d_k\|_2, \|G_k y_k\|_2\}, \quad (22c)$$

or at least satisfying (22a) and (22c) (which is sufficient if deemed by Algorithm 6).

(e) If Line 20 is reached and one finds that

$$\|d_k\|_{H_k}^2 \geq \xi \|d_k\|_2^2 \text{ and } \alpha_k \geq \underline{\alpha}, \quad (23)$$

then Algorithm 8 yields  $\mathcal{X}_{k+1} \leftarrow \{x_{k+1}\}$  and  $p_{k+1} \leftarrow 0$ ; otherwise, with probability one,

$$\mathcal{X}_{k+1} \leftarrow (\{x_{k+1}\} \cup \mathcal{S}_{k+1} \cup (\mathcal{X}_k \cap \mathbb{B}(x_{k+1}, \epsilon_{k+1}))) \subset \mathbb{B}(x_{k+1}, \epsilon_{k+1}) \text{ with } p_{k+1} \geq \min\{p_k + 1, p\}.$$

Finally, let  $\mathcal{K}_{H,W} := \{k \in \mathcal{K} : \alpha_k d_k =: s_k \neq 0\}$ , which are the indices for which Algorithm 7 may yield  $(H_{k+1}, W_{k+1}) \neq (H_k, W_k)$ . If  $\mathcal{K}_{H,W}$  is infinite, then for any  $\chi \in (0, 1)$  there exists  $(\underline{\mu}, \bar{\mu}) \in (0, \infty)^2$  with  $\underline{\mu} \leq \bar{\mu}$  such that, for every  $K \in \mathbb{N}$ , the following hold for at least  $\lceil \chi K \rceil$  values of  $k \in \mathcal{K}_{H,W}$ :

$$\underline{\mu} \|G_k y_k\|_2^2 \leq \|G_k y_k\|_{W_k}^2 \quad (24a)$$

$$\text{and } \|W_k G_k y_k\|_2^2 \leq \bar{\mu} \|G_k y_k\|_2^2. \quad (24b)$$

If  $\mathcal{K}_{H,W}$  is finite, then such constants exist satisfying (24) for all  $k \in \mathcal{K}$ .

*Proof.* Proof. If the algorithm reaches iteration  $k \in \mathcal{K}$  in which the condition in Line 5 holds, then the algorithm terminates finitely. In this case, all subroutines in iterations  $\{0, 1, \dots, k-1\}$  must have terminated successfully prior to termination. Moreover, in this case, (24) follows from the fact that only a finite number of iterations are performed and the following proof of part (a) of the lemma:

- (a) The facts that  $H_0 \succ 0$  and  $W_0 \succ 0$  follow from the initialization of the algorithm. Now suppose that iteration 1 is reached. If  $s_0 = 0$ , then  $H_1 \leftarrow H_0 \succ 0$  and  $W_1 \leftarrow W_0 \succ 0$ ; otherwise, positive definiteness of  $H_1$  and  $W_1$  follows the fact that (38) implies  $s_0^T v_0 > 0$  and from well-known properties of BFGS updating; see, e.g., [35, Chapter 6]. Inductively, positive definiteness of  $H_k$  and  $W_k$  for any  $k \in \mathbb{N}$  follows by the same arguments.

This completes the proof of the lemma for the case when the algorithm reaches  $k \in \mathcal{K}$  at which the condition in Line 5 holds. Hence, we may proceed under the assumption that this condition does not hold for any  $k \in \mathcal{K}$ .

Suppose that the algorithm reaches iteration  $k \in \mathcal{K}$ . To prove that, with probability one, it reaches iteration  $k+1$  (i.e., without failure of a subroutine), it suffices to prove parts (b)–(e) (since part (a) has been proved above).

- (b) By part (a), one has  $H_k \succ 0$  and  $W_k \succ 0$ , from which it follows that strong duality holds at the primal-dual optimal solution of (5). Since  $\theta_k(y) \leq 0$  for all  $y \in \mathbb{N}$ , there are two cases to consider, namely, whether  $\theta_k(y_{k,*}) = 0$  or  $\theta_k(y_{k,*}) < 0$ . First, suppose that  $\theta_k(y_{k,*}) = 0$ . Since  $W_k \succ 0$ , this implies that  $G_k y_{k,*} = 0$ . Under Assumption 2.1, we have that  $y_{k,j} \rightarrow y_{k,*}$ . This limit, the fact that  $G_k y_{k,*} = 0$ , and the facts that  $W_k \succ 0$  and  $\epsilon_k > 0$  together imply that (10) holds for some sufficiently large  $j \in \mathbb{N}$ . Now suppose that  $\theta_k(y_{k,*}) < 0$ . If (10) holds for any  $j \in \mathbb{N}$ , then the inner loop terminates and there is nothing left to prove; hence, we may proceed assuming that (10) does not hold for any  $j \in \mathbb{N}$ . Under Assumption 2.1, we have that  $(d_{k,j}, y_{k,j}) \rightarrow (d_{k,*}, y_{k,*})$ . This limit, continuity of  $q_k$  and  $\theta_k$ , the fact that  $\theta_k(y_{k,*}) < 0$ , strong duality for (5), Lemma 2.1, and the fact that  $\tau_k \in (0, 1)$  imply that (11) and (12) will be satisfied for some sufficiently large  $j \in \mathbb{N}$ . Finally, the fact that  $H_k = W_k^{-1}$  and at termination of the inner loop the algorithm yields  $d_k = -W_k G_k y_k$  implies that  $\|d_k\|_{H_k} = \|G_k y_k\|_{W_k}$ , as desired.

(c–d) The proof follows in the same manner as that for [11, Lemma 2.3].

(e) The proof follows in the same manner as that for [11, Lemma 2.5].

Since we have shown that if the algorithm reaches iteration  $k \in \mathcal{K}$ , then it reaches iteration  $k+1$  with probability one, it follows that, again with probability one, an infinite number of iterations are performed. Finally, with respect to the stated property of the sequence  $\{(H_k, W_k)\}_{k \in \mathcal{K}_{H,W}}$ , the proof follows in the same manner as that for [12, Corollary 3.2].  $\square$   $\square$

The next three lemmas are similar to results previously proved for GS methods. First, the following lemma is a simple consequence of the previous lemma (specifically, parts (c) and (e)) and the sample set update strategy, namely, Algorithm 8 (in the online companion). A similar result was proved as [11, Lemma 3.3].

**Lemma 2.6.** *If  $\mathcal{K} = \mathbb{N}$ , then  $\mathcal{K}_\alpha := \{k \in \mathbb{N} : \alpha_k > 0\}$  is infinite.*

*Proof.* Proof. Suppose  $\mathcal{K} = \mathbb{N}$  and observe by Lemma 2.5(c) that  $\alpha_k \geq 0$  for all  $k \in \mathbb{N}$ . In order to derive a contradiction, suppose that there exists an index  $k_\alpha \in \mathbb{N}$  such that  $\alpha_k = 0$  for all  $k \in \mathbb{N}$  with  $k \geq k_\alpha$ . By Lemma 2.5(c), this means that  $p_k \leq p - 1$  for all  $k \geq k_\alpha$ . However, with  $\alpha_k = 0$ , one finds that (23) does not hold, which by Lemma 2.5(e) implies that  $p_{k+1} \geq \min\{p_k + 1, p\}$ . This implies the existence of some  $k \geq k_\alpha$  such that  $p_k \geq p$ , which by Lemma 2.5(c) implies that  $\alpha_k > 0$ , a contradiction of the definition of the index  $k_\alpha$ .  $\square$   $\square$

The next lemma shows a useful upper bound on the objective function value at iteration  $k + 1 \in \mathcal{K}$ ; for a similar result, see, e.g., [11, Lemma 3.4].

**Lemma 2.7.** *If  $k + 1 \in \mathcal{K}$ , then*

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2}\underline{\eta}\|x_{k+1} - x_k\|_2 \max\{\|d_k\|_2, \|G_k y_k\|_2\}.$$

*Proof.* Proof. Suppose  $k + 1 \in \mathcal{K}$ , which implies that  $k \in \mathcal{K}$ . Lemma 2.5(a) and (b) imply that  $d_k = 0$  if and only if  $G_k y_k = 0$ . If  $d_k = 0$  and  $G_k y_k = 0$ , then  $x_{k+1} = x_k$  and the result follows trivially. Otherwise, in iteration  $k \in \mathcal{K}$ , Lemma 2.5(d) shows that  $x_{k+1}$  satisfies (22a) and (22c). The triangle inequality and (22c) imply

$$\begin{aligned} \|x_{k+1} - x_k\|_2 &\leq \min\{\alpha_k, \epsilon_k\} \min\{\|d_k\|_2, \|G_k y_k\|_2\} + \alpha_k \|d_k\|_2 \\ &\leq \alpha_k \|d_k\|_2 \min\{2, 1 + \|G_k y_k\|_2 / \|d_k\|_2\}. \end{aligned}$$

Hence, along with (22a), one finds that

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq -\underline{\eta}\alpha_k \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\} \\ &= -\underline{\eta}\alpha_k \|d_k\|_2 \max\{\|d_k\|_2, \|G_k y_k\|_2^2 / \|d_k\|_2\} \\ &\leq -\underline{\eta}\|x_{k+1} - x_k\|_2 \left( \frac{\max\{\|d_k\|_2, \|G_k y_k\|_2^2 / \|d_k\|_2\}}{\min\{2, 1 + \|G_k y_k\|_2 / \|d_k\|_2\}} \right) \\ &\leq -\frac{1}{2}\underline{\eta}\|x_{k+1} - x_k\|_2 \max\{\|d_k\|_2, \|G_k y_k\|_2\}, \end{aligned}$$

as desired.  $\square$   $\square$

Now we enter the core theory of GS methods. At its heart is the closure of the convex hull of gradients at points of differentiability in an  $\epsilon_k$ -neighborhood about a given point  $\bar{x} \in \mathbb{R}^n$ , namely,

$$\mathcal{G}(\bar{x}, \epsilon_k) := \text{cl conv } \nabla f(\mathbb{B}(\bar{x}, \epsilon_k) \cap \mathcal{D}). \quad (25)$$

along with, for any  $\omega \in (0, \infty)$ , the subset of the Cartesian product of  $\epsilon_k$ -balls about  $x_k$  given by

$$\mathcal{T}_k(\bar{x}, \omega) := \left\{ \mathcal{X}_k \in \prod_0^{p_k} (\mathbb{B}(x_k, \epsilon_k) \cap \mathcal{D}) : \right. \\ \left. \|P_{W_k}(\text{conv}(\{\nabla f(x)\}_{x \in \mathcal{X}_k}))\|_{W_k} \leq \|P_{W_k}(\mathcal{G}(\bar{x}, \epsilon_k))\|_{W_k} + \omega \right\},$$

both of which are defined with respect to each iteration number  $k \in \mathbb{N}$  and a point  $\bar{x} \in \mathbb{R}^n$ . (In the definition of  $\mathcal{T}_k(\bar{x}, \omega)$ , recall that  $P_{W_k}(\cdot)$  has been defined in (14).) The following lemma, which follows [24, Lemma 3.2(i)], [10, Lemma 4.7], and [11, Lemma 3.6], shows that if the sample set size inductor  $p_k$  is sufficiently large and  $x_k$  is sufficiently close to  $\bar{x}$ , then for any  $\omega \in (0, \infty)$  there exists a nonempty open subset of  $\mathcal{T}_k(\bar{x}, \omega)$ . This will be critical in our main result, where we need to show in certain situations that an element of this subset can be found through random sampling of points.

**Lemma 2.8.** *Let  $\bar{x} \in \mathbb{R}^n$  and  $\omega \in (0, \infty)$  be given. If  $k \in \mathcal{K}$  and  $p_k \geq n + 1$ , then there exists  $\zeta > 0$  such that with  $x_k \in \mathbb{B}(\bar{x}, \zeta)$  there is a nonempty open  $\mathcal{T} \subseteq \mathcal{T}_k(\bar{x}, \omega)$ .*

*Proof.* Proof. Using the metric defined by  $W_k$ , the proof follows the same argument of [24, Lemma 3.2(i)], which makes use of Carathéodory's theorem.  $\square$   $\square$

We now present a convergence theorem for Algorithm 1. Much of the proof follows similar arguments as that for [11, Theorem 3.1], which we present for completeness. The new features are two-fold: (1) Our algorithm is even less conservative about the Hessian and inverse Hessian updates than the method in [11], so our convergence result relies on arguments about self-correcting properties of BFGS updating that we have stated in Lemma 2.5, which borrows from [12]; and (2) our inexactness conditions and our Lemma 2.2, which have not appeared before for GS methods, play critical roles in the proof of the theorem.

**Theorem 2.1.** *Suppose  $\psi \in (0, 1)$ . Algorithm 1 either terminates finitely with a stationary point for  $f$  or, with probability one, it performs an infinite number of outer iterations. In the latter case, with probability one, the sampling radius sequence satisfies  $\{\epsilon_k\} \searrow 0$  and every cluster point of the iterate sequence  $\{x_k\}$  is stationary for  $f$ .*

*Proof.* Proof. If Algorithm 1 terminates finitely with a stationary point for  $f$ , then there is nothing left to prove. Otherwise, by Lemma 2.5, it follows with probability one that an infinite number of outer iterations are performed, meaning  $\mathcal{K} = \mathbb{N}$ . Since our desired conclusion only needs to hold with probability one, we may assume going forward that  $\mathcal{K} = \mathbb{N}$ . Under this assumption, our next aim is to prove that  $\{\epsilon_k\} \searrow 0$  with probability one. We consider two cases.

- **Case 1:** Suppose that  $\mathcal{K}_d := \{k \in \mathbb{N} : d_k = 0\}$  is infinite. By Lemma 2.5(a) and (b), it follows that  $G_k y_k = 0$  for all  $k \in \mathcal{K}_d$ . This fact, the fact that  $|\mathcal{K}_d| = \infty$ , and (10) imply that  $\{\epsilon_k\} \searrow 0$ .
- **Case 2:** Suppose that  $\mathcal{K}_d := \{k \in \mathbb{N} : d_k = 0\}$  is finite. Let us proceed by supposing that there exists  $k_\epsilon \in \mathbb{N}$  and a sampling radius  $\epsilon \in (0, \infty)$  such that  $\epsilon_k = \epsilon$  for all  $k \in \mathbb{N}$  with  $k \geq k_\epsilon$ . Our aim is to show that the existence of such a pair  $(k_\epsilon, \epsilon)$  occurs with probability zero. From (10),

$$\max\{\|d_k\|_2, \|G_k y_k\|_2\} > \nu\epsilon \text{ for all } k \geq k_\epsilon. \quad (26)$$

On the other hand, Assumption 1.1, Lemma 2.7, and (22a) imply that

$$\sum_{k=k_\epsilon}^{\infty} \|x_{k+1} - x_k\|_2 \max\{\|d_k\|_2, \|G_k y_k\|_2\} < \infty \quad (27a)$$

$$\sum_{k=k_\epsilon}^{\infty} \alpha_k \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\} < \infty. \quad (27b)$$

In conjunction with (26), the bound in (27a) implies that the iterate sequence  $\{x_k\}$  is a Cauchy sequence, meaning  $\{x_k\} \rightarrow \bar{x}$  for some  $\bar{x} \in \mathbb{R}^n$ . At the same time, with (26), the bound in (27b) implies that  $\{\alpha_k\} \rightarrow 0$ . We claim that this implies that  $p_k = p$  for all sufficiently large  $k \in \mathcal{K}_\alpha$ , where  $\mathcal{K}_\alpha$  is defined as in Lemma 2.6. Indeed, since  $\{\alpha_k\} \rightarrow 0$ , it follows by Lemma 2.5(c) that for sufficiently large  $k \in \mathbb{N}$  either (i)  $p_k < p$  and  $\alpha_k = 0$  or (ii)  $p_k = p$  and  $\alpha_k > 0$ . Combined with the fact that  $|\mathcal{K}_d| < \infty$ , it follows along with Lemma 2.6 that there exists an infinite number of iterations indexed by  $k \geq k_\epsilon$  such that  $\alpha_k d_k \neq 0$  and  $p_k = p$ , whereas all other iterations for sufficiently large  $k \geq k_\epsilon$  yield  $\alpha_k = 0$ . Going forward, for ease of notation in the remainder of the proof of this case, since  $x_{k+1} \leftarrow x_k$  and  $(H_{k+1}, W_{k+1}) \leftarrow (H_k, W_k)$  whenever  $\alpha_k = 0$ , let us proceed without loss of generality under the assumption that  $k_\epsilon = 0$  and  $\epsilon = \epsilon_0$ , and that  $\alpha_k > 0$ ,  $d_k \neq 0$ , and  $p_k = p$  for all  $k \in \mathbb{N}$ . Notice that under these conditions the set  $\mathcal{K}_{H,W}$  defined in Lemma 2.5 equals  $\mathbb{N}$ . Correspondingly, for a given  $\chi \in (0, 1)$ , let  $\mathcal{K}_\chi$  be the indices of iterations for which (24) holds; in particular, for  $k \in \mathcal{K}_\chi$ , one has from (24a)–(24b) that

$$\max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\} \leq \mu \|G_k y_k\|_{W_k}^2, \text{ where } \mu := \max\left\{\frac{\bar{\mu}}{\underline{\mu}}, \frac{1}{\underline{\mu}}\right\}. \quad (28)$$

Since (10) does not hold for any  $k \geq k_\epsilon$ , it follows that either (12) or (13) holds for all  $k \geq k_\epsilon$ . Hence, by Lemmas 2.3 and 2.4, it follows that (18) holds for all  $k \geq k_\epsilon$ , meaning for all  $k \geq k_\epsilon$  that

$$\|G_k y_k\|_{W_k} \leq (1 + \sigma_k) \|P_{W_k}(\text{conv}(\{\nabla f(x)\}_{x \in \mathcal{X}_k}))\|_{W_k}. \quad (29)$$

- **Subcase 2a:** If  $\bar{x}$  is  $\epsilon$ -stationary, then  $\|P_{W_k}(\mathcal{G}(\bar{x}, \epsilon_k))\|_{W_k} = 0$  for any  $W_k \succ 0$ . Therefore, with  $\mu \in (0, \infty)$  defined in (28),  $\omega = \nu\epsilon/(\sqrt{\mu}(1 + \sigma))$ , and  $(\zeta, \mathcal{T})$  chosen as in Lemma 2.8, it follows that there exists  $k_\zeta \in \mathbb{N}$  with  $k_\zeta \geq k_\epsilon$  such that  $x_k \in \mathbb{B}(\bar{x}, \zeta)$  for all  $k \geq k_\zeta$  and, with (29),

$$\begin{aligned} \max\{\|d_k\|_2, \|G_k y_k\|_2\} &\leq \sqrt{\mu} \|G_k y_k\|_{W_k} \\ &\leq \sqrt{\mu}(1 + \sigma_k) \|P_{W_k}(\text{conv}(\{\nabla f(x)\}_{x \in \mathcal{X}_k}))\|_{W_k} \\ &\leq \sqrt{\mu}(1 + \sigma_k) \omega \leq \nu\epsilon \end{aligned} \quad (30)$$

whenever  $k \geq k_\zeta$ ,  $k \in \mathcal{K}_\chi$ , and  $\mathcal{X}_k \in \mathcal{T}$ . Combining (26) and (30), it follows that  $\mathcal{X}_k \notin \mathcal{T}$  for all  $k \geq k_\zeta$  with  $k \in \mathcal{K}_\chi$ . However, this is a probability zero event since for all such  $k$  the set  $\mathcal{X}_k$  will contain new points from  $\mathbb{B}(x_k, \epsilon_k)$  that are generated independently whether or not  $k \in \mathcal{K}_\chi$ , meaning that with probability one there exists sufficiently large such  $k$  with  $k \in \mathcal{K}_\chi$  and  $\mathcal{X}_k \in \mathcal{T}$ , which would yield (30).

- **Subcase 2b:** If  $\bar{x}$  is not  $\epsilon$ -stationary, then it follows from Lemma 2.5(c) that  $\alpha_k$  satisfies (21a) for all  $k \in \mathbb{N}$ . In particular, (21a) holds either with  $\alpha_k \geq \gamma\bar{\alpha}$  or with  $\alpha_k < \gamma\bar{\alpha}$  such that

$$f(x_k + \gamma^{-1}\alpha_k d_k) - f(x_k) \geq -\underline{\eta}\gamma^{-1}\alpha_k \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\}. \quad (31)$$

In the latter case, Lebourg's mean value theorem [7, Theorem 2.3.7] implies the existence of a point  $\tilde{x}_k \in [x_k, x_k + \gamma^{-1}\alpha_k d_k]$  and  $\tilde{g}_k \in \partial f(\tilde{x}_k)$  such that

$$f(x_k + \gamma^{-1}\alpha_k d_k) - f(x_k) = \gamma^{-1}\alpha_k \tilde{g}_k^T d_k. \quad (32)$$

Combining (31), (32), and the fact that  $d_k = -W_k G_k y_k$ , one finds that

$$\tilde{g}_k^T W_k G_k y_k \leq \underline{\eta} \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\}. \quad (33)$$

On the other hand, for any  $\omega \in (0, \infty)$  and  $(\zeta, \mathcal{T})$  as in Lemma 2.8, there exists  $k_\omega \geq k_\epsilon$  such that  $x_k \in \mathbb{B}(\bar{x}, \min\{\zeta, \epsilon/3\})$  for  $k \geq k_\omega$  and, with (29),

$$\begin{aligned} \|G_k y_k\|_{W_k} &\leq (1 + \sigma_k) \|P_{W_k}(\text{conv}(\{\nabla f(x)\}_{x \in \mathcal{X}_k}))\|_{W_k} \\ &\leq (1 + \sigma_k) \|P_{W_k}(\mathcal{G}(\bar{x}, \epsilon_k))\|_{W_k} + (1 + \sigma_k)\omega \end{aligned} \quad (34)$$

whenever  $k \geq k_\omega$ ,  $k \in \mathcal{K}_\chi$ , and  $\mathcal{X}_k \in \mathcal{T}$ . Hence, for such  $k$ , it follows by Lemma 2.2 with  $\mathcal{S} = \mathcal{G}(\bar{x}, \epsilon_k)$ ,  $\beta = \underline{\eta}\mu \in (0, 1)$  (where this inclusion is guaranteed by Algorithm 4), and  $W = W_k$  that for sufficiently small  $\sigma_k \in (0, \sigma)$  and  $\omega \in (0, \infty)$  one finds that (28) and (34) imply

$$\begin{aligned} v^T W_k G_k y_k &> \underline{\eta}\mu \|G_k y_k\|_{W_k}^2 \\ &\geq \underline{\eta} \max\{\|d_k\|_2^2, \|G_k y_k\|_2^2\} \quad \text{for all } v \in \mathcal{G}(\bar{x}, \epsilon_k). \end{aligned} \quad (35)$$

There exists  $k_\sigma \geq k_\omega$  such that  $\sigma_k$  is sufficiently small for all  $k \geq k_\sigma$  with  $k \in \mathcal{K}_\chi$  since the fact that  $\{\alpha_k\} \rightarrow 0$  and the construction of the algorithm implies that  $\{\sigma_k\} \rightarrow 0$ . Together, (33) and (35) imply that  $\tilde{g}_k \notin \mathcal{G}(\bar{x}, \epsilon_k)$  whenever  $k \geq k_\sigma$ ,  $k \in \mathcal{K}_\chi$ , and  $\mathcal{X}_k \in \mathcal{T}$ . However, by the facts that  $\mathbf{1}^T y_k = 1$  and  $y_k \geq 0$ , Assumption 1.1, and [7, Proposition 2.1.2], it follows for all  $k \geq k_\sigma$  with  $k \in \mathcal{K}_\chi$  that

$$\|d_k\|_2 = \|W_k G_k y_k\|_2 \leq \sqrt{\bar{\mu}} \|G_k y_k\|_2 \leq \sqrt{\bar{\mu}} L_{\mathbb{B}(\bar{x}, \epsilon)},$$

where  $L_{\mathbb{B}(\bar{x}, \epsilon)} \in (0, \infty)$  is a Lipschitz constant for  $f$  over  $\mathbb{B}(\bar{x}, \epsilon)$ . This shows that  $\{\|d_k\|_2\}_{k \in \mathcal{K}_\chi}$  is bounded. This fact, along with  $\{\alpha_k\} \rightarrow 0$ , implies that  $\alpha_k \leq \gamma\epsilon/(3\|d_k\|_2)$  for all sufficiently

large  $k \in \mathcal{K}_\chi$ , i.e.,  $\gamma^{-1}\alpha_k\|d_k\|_2 \leq \epsilon/3$  for all sufficiently large  $k \in \mathcal{K}_\chi$ . Along with the fact that  $x_k \in \mathbb{B}(\bar{x}, \min\{\zeta, \epsilon/3\})$  implies  $\|x_k - \bar{x}\|_2 \leq \epsilon/3$ , it follows that  $\tilde{x}_k \in \mathbb{B}(\bar{x}, 2\min\{\zeta, \epsilon/3\}/3)$  and hence  $\tilde{g}_k \in \mathcal{G}(\bar{x}, \epsilon_k)$  for all sufficiently large  $k \in \mathbb{N}$ . Overall, since  $\tilde{g}_k \notin \mathcal{G}(\bar{x}, \epsilon_k)$  whenever  $k \geq k_\sigma$ ,  $k \in \mathcal{K}_\chi$ , and  $\mathcal{X}_k \in \mathcal{T}$ , yet  $\tilde{g}_k \in \mathcal{G}(\bar{x}, \epsilon_k)$  for all sufficiently large  $k$ , it follows that  $\mathcal{X}_k \notin \mathcal{T}$  for all sufficiently large  $k \in \mathcal{K}_\chi$ . However, this is a probability zero event since  $|\mathcal{K}_\chi| = \infty$  and the sample points are generated independently of whether  $k \in \mathcal{K}_\chi$ .

We have shown that  $\{\epsilon_k\} \searrow 0$  with probability one. If  $\{\epsilon_k\} \searrow 0$ , then by (10) there exists an infinite index set  $\mathcal{K}_\epsilon := \{k \in \mathbb{N} : \epsilon_{k+1} \leftarrow \psi\epsilon_k\}$  where

$$\max\{\|d_k\|_2, \|G_k y_k\|_2\} \leq \epsilon_k \quad \text{for all } k \in \mathcal{K}_\epsilon.$$

The same argument as in [10, Theorem 4.2, Case 2], which borrows from [24, Theorem 3.3, part (iii)], shows all cluster points of  $\{x_k\}$  are stationary for  $f$ .  $\square$   $\square$

Our second convergence result, presented as the following corollary, considers the case when one chooses  $\psi = 1$  so that the sampling radius remains that  $\epsilon_0 \in (0, \infty)$  for all  $k \in \mathcal{K}$ . Similar results have appeared in the literature to prove a similar property of other GS methods; see, e.g., [24, Theorem 3.5].

**Corollary 2.1.** *Suppose  $\psi = 1$ . Algorithm 1 either terminates finitely with a stationary point for  $f$  or, with probability one, it performs an infinite number of outer iterations. In the latter case, with probability one, it either reaches iteration  $k \in \mathbb{N}$  such that  $0 \in \mathcal{G}(x_k, \epsilon_k)$  or every cluster point of the iterate sequence  $\{x_k\}$  is  $\epsilon_0$ -stationary for  $f$ .*

*Proof.* Proof. As in the proof of Theorem 2.1, if Algorithm 1 terminates finitely with a stationary point for  $f$ , then there is nothing left to prove. Otherwise, by Lemma 2.5, it follows with probability one that an infinite number of outer iterations are performed, meaning  $\mathcal{K} = \mathbb{N}$ . If the algorithm reaches iteration  $k \in \mathbb{N}$  in which  $0 \in \mathcal{G}(x_k, \epsilon_k)$ , then there is nothing left to prove. Otherwise, following the arguments in the proof of Theorem 2.1, it follows that  $\inf\{\|G_k y_k\|_2 : k \in \mathbb{N}\} > 0$  is a probability zero event. In the probability one event that  $\inf\{\|G_k y_k\|_2 : k \in \mathbb{N}\} = 0$ , the conclusion follows from the fact that  $\bar{\partial}_{\epsilon_0} f$  is closed.  $\square$   $\square$

### 3 GS Algorithm with Gradient Aggregation

Our second algorithm adds a conceptually straightforward, but practically significant enhancement to Algorithm 1. In particular, we add a procedure for exploiting gradient aggregation that can significantly reduce the size of the subproblems to be solved approximately in each “outer” iteration of the algorithm. We remark that this enhancement to the GS methodology is only possible when one is able to employ inexact subproblem solutions. This is the case since the exact solution of a subproblem involving a “gradient aggregation vector” does not offer the exact solution of a subproblem involving individual gradients and no aggregation.

In this section, we present a statement of the proposed algorithm, then show that it offers the same convergence guarantees as does Algorithm 1.

#### 3.1 Algorithm Description

Our algorithm with inexact subproblem solutions and gradient aggregation is stated as Algorithm 3. The algorithm borrows much from Algorithm 1; we have written it in such a manner that only its unique steps are stated. The main idea of the enhancement is the following. For any  $k+1 \in \mathcal{K}$  such that  $\alpha_k > 0$ , the matrix of gradients  $G_{k+1}$  contains all points in the set  $\mathcal{X}_{k+1}$ , as in Algorithm 1. However, for any  $k+1 \in \mathcal{K}$  such that  $x_{k+1} = x_k$  since  $\alpha_k = 0$ , rather than solve a subproblem defined by gradients at all points in  $\mathcal{X}_{k+1}$ , the algorithm considers a subproblem in which the gradients defining the matrix  $G_k$  (which compose a submatrix of  $G_{k+1}$ ) have been *aggregated* into a single “gradient aggregation vector”  $G_k y_k$ . The following lemma shows that a feasible point for the subproblem *that the algorithm considers* in iteration  $k+1$  corresponds to a feasible point for the subproblem *that would be defined by all gradients in  $G_{k+1}^{\text{full}}$ .*

---

**Algorithm 3** GS with Inexact Subproblem Solutions and Gradient Aggregation
 

---

**Require:** [... same parameters and initial values as in Algorithm 1, except  $G_0$  ...]

- 1: Set  $G_0^{\text{full}}$  by (4),  $G_0^{\text{agg}}$  by (4), and  $\alpha_{-1} \leftarrow 0$ .
  - 2: **for all**  $k \in \mathbb{N}$  **do**
  - 3:   **if**  $\alpha_{k-1} > 0$  or  $p_k \geq p$  **then**
  - 4:     set  $G_k \leftarrow G_k^{\text{full}}$ ;
  - 5:   **else**
  - 6:     set  $G_k \leftarrow G_k^{\text{agg}}$ .
  - 7:   **end if**
  - 8:   [... same as Line 5 through Line 19 of Algorithm 1 ...]
  - 9:   Set  $(\mathcal{X}_{k+1}, p_{k+1})$  by Algorithm 8 and  $G_{k+1}^{\text{full}}$  by (4).
  - 10:   **if**  $\alpha_k > 0$  **then**
  - 11:     set  $G_{k+1}^{\text{agg}} \leftarrow G_{k+1}^{\text{full}}$ ;
  - 12:   **else**
  - 13:     set  $G_{k+1}^{\text{agg}} \leftarrow [\nabla f(x_{k+1}) \quad G_k y_k \quad \nabla f(x)]_{x \in \mathcal{X}_{k+1} \setminus (x_{k+1} \cup \mathcal{X}_k)}$ .
  - 14:   **end if**
  - 15: **end for**
- 

**Lemma 3.1.** Consider  $k \in \mathcal{K}$  such that  $k \geq 1$  and  $\alpha_{k-1} = 0$ , meaning  $G_k = G_k^{\text{agg}}$ . For any  $j \in \mathbb{N}$  such that  $y_{k,j}$  is computed, this vector, which is feasible for the dual problem in (5), corresponds uniquely to a feasible point for the dual problem in (5) if  $G_k^{\text{full}}$  were used in place of  $G_k = G_k^{\text{agg}}$ .

*Proof.* Proof. Consider any  $j \in \mathbb{N}$  such that  $y_{k,j}$  is computed. Let  $[y_{k,j}]_1$  and  $[y_{k,j}]_2$  denote the first and second elements of  $y_{k,j}$ , respectively, with the subvector of all remaining elements of  $y_{k,j}$  being denoted as  $[y_{k,j}]_{>2}$ . One finds that

$$\begin{aligned}
 & G_k^{\text{agg}} y_{k,j} \\
 &= \nabla f(x_k) [y_{k,j}]_1 + (G_{k-1} y_{k-1}) [y_{k,j}]_2 + [\nabla f(x)]_{x \in \mathcal{X}_k \setminus (x_k \cup \mathcal{X}_{k-1})} [y_{k,j}]_{>2} \\
 &= [\nabla f(x_k) \quad G_{k-1} \quad \nabla f(x)]_{x \in \mathcal{X}_k \setminus (x_k \cup \mathcal{X}_{k-1})} \begin{bmatrix} [y_{k,j}]_1 \\ y_{k-1} [y_{k,j}]_2 \\ [y_{k,j}]_{>2} \end{bmatrix} = G_k^{\text{full}} \begin{bmatrix} [y_{k,j}]_1 \\ y_{k-1} [y_{k,j}]_2 \\ [y_{k,j}]_{>2} \end{bmatrix},
 \end{aligned}$$

where—since  $\mathbf{1}^T y_{k-1} = 1$ ,  $\mathbf{1}^T y_{k,j} = 1$ ,  $y_{k-1} \geq 0$ , and  $y_{k,j} \geq 0$ —it follows that

$$\mathbf{1}^T \begin{bmatrix} [y_{k,j}]_1 \\ y_{k-1} [y_{k,j}]_2 \\ [y_{k,j}]_{>2} \end{bmatrix} = 1 \quad \text{and} \quad \begin{bmatrix} [y_{k,j}]_1 \\ y_{k-1} [y_{k,j}]_2 \\ [y_{k,j}]_{>2} \end{bmatrix} \geq 0,$$

which proves the desired result. □ □

**Theorem 3.1.** Suppose  $\psi \in (0, 1)$ . Algorithm 3 either terminates finitely with a stationary point for  $f$  or, with probability one, it performs an infinite number of outer iterations. In the latter case, with probability one, the sampling radius sequence satisfies  $\{\epsilon_k\} \searrow 0$  and every cluster point of the iterate sequence  $\{x_k\}$  is stationary for  $f$ .

*Proof.* Proof. For all  $k \in \mathbb{N}$ , the result of Lemma 2.1 holds regardless of whether  $G_k = G_k^{\text{agg}}$  or  $G_k = G_k^{\text{full}}$  due to the fact that  $G_k$  has  $\nabla f(x_k)$  as its first column in either case. The results of Lemmas 2.3 and 2.4 also continue to hold regardless of whether  $G_k = G_k^{\text{agg}}$  or  $G_k = G_k^{\text{full}}$ , implying that the inner loop terminates finitely for all  $k \in \mathcal{K}$ . Now consider the pair  $(d_k, y_k) = (-W_k G_k y_k, y_k)$  upon termination of the inner loop in iteration  $k \in \mathcal{K}$ . If  $G_k = G_k^{\text{full}}$ , then the properties of  $(d_k, y_k)$  are the same as that in Algorithm 1. Otherwise,

when  $G_k = G_k^{\text{agg}}$ , one may consider

$$\begin{bmatrix} [y_k]_1 \\ y_{k-1}[y_k]_2 \\ [y_k]_{>2} \end{bmatrix} \quad (36)$$

as the dual vector, as shown by Lemma 3.1. The arguments of Lemmas 2.5–2.8 and Theorem 2.1 now follow in the same manner as in Section 2 using  $G_k^{\text{full}}$  in place of  $G_k$  and  $y_k$  or (36) in place of the dual vector for all  $k \in \mathcal{K}$ . Crucial in these arguments is that, if the sample set size indicator  $p_k$  ever exceeds  $p$ , then  $G_k = G_k^{\text{full}}$  and the algorithm behaves as Algorithm 1 for such  $k \in \mathcal{K}$ .  $\square$   $\square$

**Corollary 3.1.** *Suppose  $\psi = 1$ . Algorithm 3 either terminates finitely with a stationary point for  $f$  or, with probability one, it performs an infinite number of outer iterations. In the latter case, with probability one, it either reaches iteration  $k \in \mathbb{N}$  such that  $0 \in \mathcal{G}(x_k, \epsilon_k)$  or every cluster point of the iterate sequence  $\{x_k\}$  is  $\epsilon_0$ -stationary for  $f$ .*

*Proof.* Proof. The proof follows from that of Theorem 3.1 in the same manner as the proof of Corollary 2.1 follows from that of Theorem 2.1.  $\square$   $\square$

## 4 Numerical Experiments

In this section, we present the results of numerical experiments with implementations of our proposed algorithms. The main purpose of these experiments is to show that the introduction of inexactness and gradient aggregation can reduce the computational expense of an adaptive GS algorithm consistently and often substantially. As a sanity check, we also provide a comparison between our implementation of Algorithm 3 and a state-of-the-art code. All experiments were run on a Macbook Air with a 2.2 GHz Dual-Core Intel Core i7 processor running macOS 11.4.

We implemented our algorithms in the C++ software package `NonOpt` [8]. For the parameters used in the algorithms and subroutines, we employed the values stated in Table 1. These values are used consistently across all of our experiments. As is typical in implementations of GS methods, our implementations assume that  $x_k + \alpha_k d_k \in \mathcal{D}$  for all  $k \in \mathbb{N}$ , meaning that the loop in Algorithm 6 always terminates in the first iteration; hence, the parameter  $\bar{\ell}$  is not used. The initial point  $x_0 \in \mathbb{R}^n$  in each run of the algorithm was chosen in a problem-dependent manner; see the references given below in our discussion of the test problems used.

`NonOpt` contains a dual active-set QP solver that we used for solving the QP subproblems arising in the implementations of our algorithms. To reduce CPU time, during the solve of a given QP, the termination conditions (10)–(13) are not checked in every iteration of the QP solver. Instead, these conditions are checked only after  $(p_k + 1)/4$  QP iterations have been performed, and after this threshold is reached, the conditions are checked only once every four QP solver iterations.

In our implementations, the outer iteration sequence terminates if

$$\max\{\|G_k y_k\|_\infty, \|W_k G_k y_k\|_\infty, \epsilon_k\} \leq 10^{-4} \quad (37)$$

or once an objective function value tolerance or CPU time limit is reached. These latter criteria are discussed in further detail in the subsequent subsections.

We consider the performance of three implementations, to which we refer as follows:

- **GS-exact:** An implementation of an adaptive GS method in which the QP subproblems are solved “exactly” in each iteration; in particular, every aspect of this implementation is the same as that of **GS-inexact** (below), except that, when tasked to solve each QP subproblem, the QP solver is run until the  $\ell_\infty$ -norm of the KKT error for the QP is reduced below  $10^{-10}$ .
- **GS-inexact:** An implementation of Algorithm 1.
- **GS-inexact-agg:** An implementation of Algorithm 3.

Table 1: User-specified parameters for our implemented algorithms and subroutines.

Parameter(s)	Range	Values	Description
$\nu$	$(0, \infty)$	1	Stationarity measure tolerance
$\underline{\alpha} \leq \bar{\alpha}$	$(0, \infty)$	$10^{-20} \leq 100$	Stepsize thresholds
$\alpha_{\text{init}}$	$(0, \infty)$	1	Initial stepsize
$\rho$	$(0, 1)$	0.01	Inexactness threshold bound
$\kappa$	$(0, 1)$	0.0001	Inexactness threshold
$\psi$	$(0, 1)$	0.1	Sampling radius reduction factor
$\iota$	$(0, 1)$	0.5	Inexactness parameter reduction factor
$\underline{\eta} < \bar{\eta}$	$(0, 1)$	$10^{-10} < 0.9$	Armijo–Wolfe line search parameters
$p$	$[n + 1, \infty) \cap \mathbb{N}$	$10n$	Sample set size threshold
$\sigma$	$(0, \infty)$	10	Inexactness threshold reset value
$\gamma$	$(0, 1)$	0.5	Stepsize modification factor
$\underline{\phi} < 1 < \bar{\phi}$	$(0, \infty)$	$10^{-20} < 1 < 10^8$	BFGS updating thresholds
$\xi$	$(0, \infty)$	$10^{-20}$	Curvature threshold
$\bar{p}$	$\mathbb{N}$	100	Size of addition to sample set
$H_0$	$\succ 0$	$I$	Initial Hessian approximation
$\epsilon_0$	$(0, \infty)$	$\max\{0.01, 0.1\ \nabla f(x_0)\ _\infty\}$	Initial stationarity radius

## 4.1 Randomly Generated Test Problems

Our algorithms are designed to minimize objectives that may be nonconvex and/or nonsmooth. However, in order to conduct a controlled comparison between the aforementioned implemented algorithms, our main experiment involves randomly generated convex test problems of the form

$$\min_{x \in \mathbb{R}^n} g^T x + \frac{1}{2} x^T H x + \max\{Ax + b\},$$

where  $g \in \mathbb{R}^n$ ,  $H \in \mathbb{R}^{n \times n}$  is symmetric and positive definite,  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ , and the max is taken element-wise. (By employing convex, as opposed to nonconvex test problems, we can be sure that the results of our experiments are not skewed by two algorithms converging to different local minimizers and other related circumstances.) The problems were constructed such that the unique global minimizer is always  $x_* = 0$ , the global minimum is always  $f(x_*) = 0$ , and the number of elements of the vector  $Ax_* + b = b$  yielding the max, call it  $m_{\mathcal{A}}$ , is always predetermined.

When solving potentially nonconvex and/or nonsmooth optimization problems, termination conditions can be sensitive in practice; e.g., one can find that the termination condition (37) may be satisfied relatively early for some problems, whereas for other problems the magnification of small numerical errors can cause (37) to take longer to be satisfied. Hence, we added a condition for these experiments that terminates an algorithm whenever the objective value is less than a prescribed threshold of  $10^{-3}$ . This is reasonable in these experiments since  $f(x_*) = 0$  for all problems.

For the purposes of these experiments, fifteen problems were generated; with  $n = 1000$  and  $m = 500$ , five problems were generated for each of the values  $m_{\mathcal{A}} \in \{125, 250, 375\}$ . In this manner, we provide results for a range of dimensions of the “ $\mathcal{U}$ -space” and “ $\mathcal{V}$ -space” at the minimizer; see, e.g., [27]. For each problem, each of the three implemented algorithms were run from the same randomly generated starting point; in particular, each element for the initial point was drawn from a standard normal distribution. Since GS methods are randomized, we ran each algorithm 10 times for each problem and provide averages of performance measures over these 10 runs.

Results for **GS-exact**, **GS-inexact**, and **GS-inexact-agg** are provided in Tables 2, 3, and 4, respectively. Averaged over the 10 runs for each algorithm and problem, we provide the required number of iterations (**iters**), required total number of QP solver iterations (**QP-iters**), required number of objective function evaluations (**funcs**), required number of objective gradient evaluations (**grads**), and final objective value

(f). Since the total computational effort is roughly proportional to the total number of QP iterations, for **GS-inexact** and **GS-inexact-agg**, we provide the relative change in the required total number of QP iterations as compared to **GS-exact**. (This is a rough proxy for computational effort since the cost for each QP solver iteration can differ depending on the number of nonzero variables in the dual solution estimate. That said, we found it to be the best measure for comparison, as opposed to CPU time which can vary despite the algorithm being run with the same initial conditions, random number generator seeds, and so on.) In these statistics, a negative percentage indicates that **GS-inexact** (or **GS-inexact-agg**) required fewer total QP solver iterations than **GS-exact**; e.g., a statistic of  $-z\%$  indicates that the algorithm lowered the required total number of QP solver iterations by  $z\%$ .

Table 2: Results for **GS-exact** averaged over 10 runs.

$n$	$m$	$m_A$	iters	QP-iters	funcs	grads	$f$
1000	500	125	964	4583	6486	21495	+9.948857e-04
1000	500	125	961	5199	6521	24482	+9.943792e-04
1000	500	125	898	3042	6023	19725	+9.956177e-04
1000	500	125	1029	3026	6782	20148	+9.940536e-04
1000	500	125	916	4040	6258	25183	+9.953463e-04
1000	500	250	1076	16090	7433	41892	+9.922383e-04
1000	500	250	869	10493	6303	36224	+9.893498e-04
1000	500	250	1210	37689	7563	43487	+1.194211e-03
1000	500	250	1088	13620	7189	38885	+1.409556e-03
1000	500	250	1080	13870	7204	41252	+1.092632e-03
1000	500	375	2063	47562	10405	55532	+2.080894e-03
1000	500	375	1861	56658	10348	74029	+1.329274e-03
1000	500	375	2193	87014	11563	74775	+1.756560e-03
1000	500	375	2061	71832	11111	79926	+2.028260e-03
1000	500	375	1882	65654	10059	62165	+2.326188e-03

Table 3: Results for **GS-inexact** averaged over 10 runs. The final column indicates the relative change in **QP-iters** compared to **GS-exact**.

$n$	$m$	$m_A$	iters	QP-iters	funcs	grads	$f$	change in QP-iters
1000	500	125	957	3024	6518	26306	+9.968597e-04	-34.003142%
1000	500	125	948	3131	6448	26093	+9.943444e-04	-39.767679%
1000	500	125	897	2818	6116	23888	+9.951412e-04	-7.350186%
1000	500	125	1025	2848	6837	22942	+9.967565e-04	-5.865829%
1000	500	125	905	4096	6351	32989	+9.944478e-04	+1.391089%
1000	500	250	1105	9444	7456	44674	+1.078713e-03	-41.305334%
1000	500	250	1036	8875	7015	44131	+1.076791e-03	-15.417969%
1000	500	250	915	3953	6656	39961	+9.926499e-04	-89.511723%
1000	500	250	924	3428	6717	39259	+9.904635e-04	-74.829662%
1000	500	250	1082	8180	7559	48848	+1.031102e-03	-41.021038%
1000	500	375	1362	17336	8145	55212	+1.509360e-03	-63.549472%
1000	500	375	1547	31698	9817	83939	+1.379415e-03	-44.053803%
1000	500	375	1382	16480	8545	63765	+1.559232e-03	-81.059730%
1000	500	375	1718	35042	9584	66791	+1.977902e-03	-51.216579%
1000	500	375	2031	56392	11059	92092	+1.801977e-03	-14.107727%

Let us start with a few general observations about these experiments. First, between the termination condition (37) and the condition that the algorithm terminates if the objective value fell below  $10^{-3}$ , one finds that the solutions obtained by all algorithms on all problems were comparable in quality with final objective values on the order of  $10^{-3}$ . Second, one finds that the computational effort required by the algorithms was directly proportional to  $m_A$ ; this is expected since increasing  $m_A$  increases the dimension of the  $\mathcal{V}$ -space of the objective function at the solution.

Most importantly, one finds in the results in Tables 2, 3, and 4 that inexactness and gradient aggregation reduces the total number of QP solver iterations consistently and often substantially. Interestingly, one also finds in many cases that **GS-inexact** and **GS-inexact-agg** also require fewer outer iterations. This was not

Table 4: Results for `GS-inexact-agg` averaged over 10 runs. The final column indicates the relative change in QP-iters compared to `GS-exact`.

$n$	$m$	$m_A$	iters	QP-iters	funcs	grads	$f$	change in QP-iters
1000	500	125	974	2130	6428	17515	+9.950228e-04	-53.521403%
1000	500	125	965	2147	6413	18405	+9.965422e-04	-58.709156%
1000	500	125	911	1682	6020	14611	+9.948074e-04	-44.679662%
1000	500	125	1012	2715	6733	22958	+9.941496e-04	-10.277594%
1000	500	125	929	2022	6180	17479	+9.948088e-04	-49.940594%
1000	500	250	935	2775	6675	32963	+9.948877e-04	-82.748634%
1000	500	250	1021	10064	6742	33816	+1.002834e-03	-4.092971%
1000	500	250	897	2316	6350	30019	+9.920159e-04	-93.854024%
1000	500	250	932	2541	6608	31248	+9.937333e-04	-81.339941%
1000	500	250	924	2302	6459	27277	+9.965494e-04	-83.398221%
1000	500	375	860	2679	6414	45224	+9.910356e-04	-94.367142%
1000	500	375	855	2702	6466	47196	+9.928050e-04	-95.229876%
1000	500	375	831	2676	6285	47392	+9.938719e-04	-96.924309%
1000	500	375	960	14692	6797	48239	+1.141195e-03	-79.545916%
1000	500	375	1015	19647	7155	54267	+1.005219e-03	-70.073932%

necessarily expected, and might not represent behavior that one should anticipate in general. That said, one explanation for this behavior is that requiring exact subproblem solutions may tend to produce shorter search directions, whereas by allowing inexactness in the subproblem solutions the algorithm is able to take longer steps in each iteration. In any case, due to the reduced number of QP solver iterations required per outer iteration, one may expect a reduction in total computational effort for `GS-inexact` and `GS-inexact-agg` even if these algorithms were to require the same number, or even more, outer iterations than `GS-exact`.

## 4.2 Test Set Problems

To demonstrate that our implementations can be competitive with a state-of-the-art solver, we performed an experiment to compare the performance of the state-of-the-art code `LMBM` [20] and `GS-inexact-agg`. The experiments with `GS-inexact-agg` in the previous subsection were performed with full BFGS approximations, but the experiments in this subsection were performed with a limited-memory BFGS strategy with a history of 50 so that the algorithm would be more similar to `LMBM`, which uses limited memory approximations (with a history of 7).

We chose a set of 20 test problems for which `LMBM` has been tuned, some of which are convex and some of which are nonconvex. The first ten problems come from [14] and the second ten come from [30]. (`LMBM` comes with implementations of the first ten problems; for the remaining test problems, we obtained Fortran implementations from [28].) In these sources, each problem is provided with an initial point  $x_0 \in \mathbb{R}^n$ , which were the initial points that we used in our experiments. All of the problems are scalable in the sense that they are defined for any value of  $n \in \mathbb{N}$ . We chose  $n = 1000$  for all problems.

`LMBM` and `GS-inexact-agg` have many differences. For example, `LMBM` employs a bundle method while `GS-inexact-agg` employs a GS method. The termination criteria of the two codes are also very different; e.g., besides observing termination criteria related to detecting stationarity, `LMBM` may terminate due to various reasons related to the iterate and/or objective value not changing significantly between iterations. Hence, in order to offer a fair and illustrative comparison, we ran `LMBM` for each problem 10 times and found the average CPU time required per problem. (The solve for most problems terminated within a few seconds; the only exceptions were that for `Test29_5`, which required approximately 30 seconds, and that for `Test29_13`, which required approximately 15 seconds.) We then ran `GS-inexact-agg` with a CPU time limit of the maximum of the average time required by `LMBM` and one second. This caused `GS-inexact-agg` to terminate in a large majority of the runs due to the CPU time limit, even though it would have continued to iterate to obtain better solutions if it were allowed to do so. Due to the random behavior of `GS-inexact-agg`, we ran the solver 10 times for each problem and present averages over these runs.

The results obtained by the codes are shown in Table 5. `LMBM` reports the number of iterations (`iters`),

function evaluations (**funcs**), and final objective value ( $f$ ). For **GS-inexact-agg**, we additionally provide (averaged over all runs) the number of QP subproblem solver iterations (**QP-iters**) and gradient evaluations (**grads**). Since the algorithms have various differences, it is not necessarily informative to compare the number of iterations or function evaluations required by the two methods. On the other hand, one can compare final objective values, with respect to which one finds that the results are generally comparable. LBM yields lower values for some problems while **GS-inexact-agg** yields lower values for a few others.

Table 5: Results for LBM and **GS-inexact-agg** averaged over 10 runs

name	iters	funcs	$f$	iters	QP-iters	funcs	grads	$f$
MaxQ	21940	22808	+4.987830e-06	5234	5253	15000	5267	+4.021445e-04
MxHilb	441	861	+6.166410e-03	140	140	687	200	+1.134878e-03
Chained_LQ	300	1824	-1.412780e+03	66	150	446	130	-1.412639e+03
Chained_CB3.1	291	1690	+1.998000e+03	72	143	501	145	+2.012761e+03
Chained_CB3.2	66	150	+1.998000e+03	61	80	300	93	+1.998000e+03
ActiveFaces	523	569	+1.376680e-14	15	17	391	344	+3.961526e-05
Brown_Function.2	493	4217	+2.136910e-09	27	141	201	107	+9.041635e-01
Chained_Mifflin.2	546	3892	-7.064510e+02	40	136	316	168	-7.062611e+02
Chained_Crescent.1	177	817	+3.681010e-08	37	44	187	52	+4.728769e-08
Chained_Crescent.2	903	9626	+1.369240e-04	34	141	262	101	+1.353061e-01
Test29.2	62	63	+9.815390e-01	307	354	1643	358	+7.212000e-01
Test29.5	1230	4563	+6.434430e-06	185	369	948	374	+8.359626e-07
Test29.6	44	48	+2.000000e+00	37	142	292	119	+2.007236e+00
Test29.11	283	1336	+1.203580e+04	13	147	170	121	+1.208292e+04
Test29.13	3747	7092	+5.665460e+02	92	1451	1219	1134	+6.140429e+02
Test29.17	962	2247	+3.574260e-03	8	152	148	114	+1.096233e-03
Test29.19	143	1012	+1.000000e+00	29	144	274	123	+1.013406e+00
Test29.20	277	3087	+5.000010e-01	46	140	394	142	+5.008065e-01
Test29.22	21	172	+1.966970e-06	10	153	163	118	+3.056232e-04
Test29.24	315	1945	+4.232150e-02	32	138	352	119	+1.102612e-01

## 5 Conclusion

We have proposed, analyzed, and tested two algorithms for minimizing locally Lipschitz objective functions. The algorithms are based on the gradient sampling methodology. The unique feature of the first algorithm is that it can allow *inexactness* in the subproblem solutions while maintaining convergence guarantees, which is new to the literature on gradient sampling methods. The unique feature of the second algorithm is that it can use inexact subproblem solutions and *aggregated* gradients in place of individual gradients in the subproblem definitions. Our numerical experiments show that employing inexactness and aggregation can each reduce computational effort.

## References

- [1] P. Apkarian, D. Noll, and O. Prot. A trust region spectral bundle method for nonconvex eigenvalue optimization. *SIAM Journal on Optimization*, 19(1):281–306, 2008.
- [2] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, Nashua, NH, USA, 2009.
- [3] J. V. Burke, A. S. Lewis, and M. L. Overton. Approximating Subdifferentials by Random Sampling of Gradients. *Mathematics of Operations Research*, 27(3):567–584, 2002.
- [4] J. V. Burke, A. S. Lewis, and M. L. Overton. A Robust Gradient Sampling Algorithm for Nonsmooth, Nonconvex Optimization. *SIAM Journal on Optimization*, 15(3):751–779, 2005.

- [5] James V. Burke, Frank E. Curtis, Adrian S. Lewis, Michael L. Overton, and Lucas E. A. Simões. Gradient Sampling Methods for Nonsmooth Optimization. In *Numerical Nonsmooth Optimization*, chapter 6, pages 201–225. Springer, 2020.
- [6] R. H. Byrd and J. Nocedal. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM Journal on Numerical Analysis*, 26(3):727–739, 1989.
- [7] F. H. Clarke. *Optimization and Nonsmooth Analysis*. Canadian Mathematical Society Series of Monographs and Advanced Texts. John Wiley & Sons, New York, NY, USA, 1983.
- [8] F. E. Curtis. NonOpt. <https://coral.ise.lehigh.edu/frankecurtis/nonopt/>, 2021.
- [9] F. E. Curtis and M. L. Overton. A Sequential Quadratic Programming Algorithm for Nonconvex, Nonsmooth Constrained Optimization. *SIAM Journal on Optimization*, 22(2):474–500, 2012.
- [10] Frank E. Curtis and Xiaocun Que. An Adaptive Gradient Sampling Algorithm for Nonsmooth Optimization. *Optimization Methods and Software*, 28(6):1302–1324, 2013.
- [11] Frank E. Curtis and Xiaocun Que. A Quasi-Newton Algorithm for Nonconvex, Nonsmooth Optimization with Global Convergence Guarantees. *Mathematical Programming Computation*, 7(4):399–428, 2015.
- [12] Frank E. Curtis, Daniel P. Robinson, and Baoyu Zhou. A Self-Correcting Variable-Metric Algorithm Framework for Nonsmooth Optimization. *IMA Journal of Numerical Analysis*, 40(2):1154–1187, 2020.
- [13] A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.
- [14] N. Haaraala, K. Miettinen, and M. M. Mäkelä. New limited memory bundle method for large-scale nonsmooth optimization. *Optimization Methods and Software*, 19(6):673–692, 2004.
- [15] N. Haaraala, K. Miettinen, and M. M. Mäkelä. Globally convergent limited memory bundle method for large-scale nonsmooth optimization. *Mathematical Programming*, 109(1):181–205, 2007.
- [16] W. Hare and C. Sagastizábal. A redistributed proximal bundle method for nonconvex optimization. *SIAM Journal on Optimization*, 20(5):2442–2473, 2010.
- [17] E. S. Helou, S. A. Santos, and L. E. A. Simões. On the Local Convergence Analysis of the Gradient Sampling Method for Finite Max-Functions. *Journal of Optimization Theory and Applications*, 175(1):137–157, 2017.
- [18] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms II*. A Series of Comprehensive Studies in Mathematics. Springer-Verlag, New York, NY, USA, 1993.
- [19] S. Hosseini and A. Uschmajew. A Riemannian Gradient Sampling Algorithm for Nonsmooth Optimization on Manifolds. *SIAM Journal on Optimization*, 27(1):173–189, 2017.
- [20] N. Karmitsa. LMBM. <http://napsu.karmitsa.fi/lmbm>, accessed 2021.
- [21] K. C. Kiwiel. A linearization algorithm for nonsmooth minimization. *Mathematics of Operations Research*, 10(2):185–194, 1985.
- [22] K. C. Kiwiel. *Methods of Descent for Nondifferentiable Optimization*. Lecture Notes in Mathematics. Springer-Verlag, New York, NY, USA, 1985.
- [23] K. C. Kiwiel. Restricted step and Levenberg-Marquardt techniques in proximal bundle methods for nonconvex nondifferentiable optimization. *SIAM Journal on Optimization*, 6(1):227–249, 1996.
- [24] K. C. Kiwiel. Convergence of the Gradient Sampling Algorithm for Nonsmooth Nonconvex Optimization. *SIAM Journal on Optimization*, 18(2):379–388, 2007.

- [25] C. Lemaréchal, A. Nemirovskii, and Yu. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69(1):111–147, 1995.
- [26] A. S. Lewis and M. L. Overton. Nonsmooth Optimization via Quasi-Newton Methods. *Mathematical Programming*, 141(1–2):135–163, 2013.
- [27] Shuai Liu and Claudia Sagastizábal. *Beyond First Order:  $\mathcal{VU}$ -Decomposition Methods*, pages 297–329. Springer International Publishing, Cham, 2020.
- [28] L. Luksan. Test Problems in Fortran. <http://www.cs.cas.cz/~luksan/test.html>, accessed 2021.
- [29] L. Lukšan and J. Vlček. A bundle-Newton method for nonsmooth unconstrained minimization. *Mathematical Programming*, 83(1):373–391, 1998.
- [30] L. Lukšan, M. Tůma, M. Šiška, J. Vlček, and N. Ramešová. UFO 2002: Interactive System for Universal Functional Optimization. Technical Report 883, Institute of Computer Science, Academy of Sciences of the Czech Republic, 2002.
- [31] Morteza Maleknia and Mostafa Shamsi. A gradient sampling method based on ideal direction for solving nonsmooth optimization problems. *Journal of Optimization Theory and Applications*, 187(3):181–204, 2020.
- [32] R. Mifflin. An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research*, 2(2):191–207, 1977.
- [33] R. Mifflin. A modification and an extension of Lemaréchal’s algorithm for nonsmooth minimization. In D. C. Sorensen and R. J.-B. Wets, editors, *Nondifferential and Variational Techniques in Optimization*, pages 77–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 1982.
- [34] R. Mifflin and C. Sagastizábal. A  $\mathcal{VU}$ -algorithm for convex minimization. *Mathematical Programming*, 104(2):583–608, 2005.
- [35] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer, New York, NY, USA, 2nd edition, 2006.
- [36] A. Ruszczyński. *Nonlinear Optimization*. Princeton University Press, Princeton, NJ, USA, 2006.
- [37] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: Conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 2(1):121–152, 1992.
- [38] C.-M. Tang, S. Liu, J.-B. Jian, and J.-L. Li. A Feasible SQP-GS Algorithm for Nonconvex, Nonsmooth Constrained Optimization. *Numerical Algorithms*, 65(1):1–22, Jan 2014.

## A Online Companion: GS Algorithm Subroutines

In this online companion, we present subroutines needed for the GS algorithms with inexact subproblem solutions and gradient aggregation proposed in the paper. Besides the first, these subroutines have been motivated and presented in previous articles, as mentioned in the following subsections. We include them here, along with pointers to lemmas that prove their properties, for ease of reference for the main body of the paper.

## A.1 Sufficient decrease parameter selection

For our proposed algorithm, the sufficient decrease parameter  $\underline{\eta} \in (0, 1)$  employed in the Armijo-Wolfe conditions, specifically in (21a), needs to be set sufficiently small relative to an upper bound on the condition number of (at least a subset of) the generated Hessian approximations. A similar relationship was required in [11, eq. (3.1)]. Generally speaking, the Hessian approximations in a BFGS updating strategy can become arbitrarily ill-conditioned, but it is sufficient for our purposes to employ an upper bound that holds for a fraction of *good* iterations that will provably occur; see [6].

---

### Algorithm 4 Sufficient Decrease Parameter Selection

---

**Require:**  $\underline{\phi} \in (0, 1)$ ,  $\bar{\phi} \in (1, \infty)$ , and  $H_0 \succ 0$  from outer algorithm;  $\chi \in (0, 1)$ .

1: Set

$$c_0 \leftarrow \frac{1}{1 - \chi} (\text{tr}(H_0) - \ln \det(H_0) + \bar{\phi} - 1 - \ln \underline{\phi}) \in (0, \infty).$$

2: Set  $c_1 \leftarrow e^{-c_0/2}$ .

3: Set  $c_2$  (resp.,  $c_3$ ) as the smallest (resp., largest) value in  $(0, \infty)$  such that

$$1 - r + \ln r \geq -c_0 \quad \text{for all } r \in [c_2, c_3].$$

4: Set  $\underline{\mu} \leftarrow c_1^2/c_3$  and  $\bar{\mu} \leftarrow 1/c_2^2$ .

5: Set

$$\mu \leftarrow \max \left\{ \frac{\bar{\mu}}{\underline{\mu}}, \frac{1}{\underline{\mu}} \right\} \in (1, \infty).$$

6: Choose  $\underline{\eta} \in (0, 1)$  such that  $\underline{\eta}\mu \in (0, 1)$ .

7: **terminate** and **return**  $(\underline{\eta}, \mu)$

---

## A.2 Line search

Given a descent direction for  $f$  at  $x_k$ , the line search is intended to find a stepsize satisfying the weak Armijo-Wolfe conditions; see (21) and, e.g., [35]. However, as motivated by [11], the line search may terminate early if the sample set size indicator  $p_k$  is less than the prescribed integer  $p \in [n + 1, \infty)$  or may ignore the curvature condition (21b)—and switch to a “backtracking Armijo” line search—if both  $p_k \geq p$  and a certain number of iterations of the line search have already been performed without termination. This potential switch to a backtracking Armijo line search may be needed due to potential nonsmoothness of  $f$ , since under Assumption 1.1 one can only show that an Armijo-Wolfe line search can “bracket” a stepsize satisfying the Armijo-Wolfe conditions (21); see [26, Theorem 4.7]. One could ensure finite termination of an Armijo-Wolfe line search, without having to switch to a backtracking Armijo line search as a backup, with a stronger assumption on  $f$ , such as it being weakly lower semismooth; see [32].

Our line search subroutine is stated as Algorithm 5. Given  $\underline{\eta} \in (0, 1)$  and  $\bar{\eta} \in (\underline{\eta}, 1)$ , the Armijo-Wolfe conditions that we use are stated as (21). We remark that (21a) does not use the directional derivative of  $f$  at  $x_k$  along  $d_k$ , as is typical in the Armijo condition in the context of smooth optimization; rather, it uses squared norms of search direction quantities, which is common in GS algorithms when nonnormalized search directions are used; see, e.g., [24, Eq. (4.2)].

A proof of Lemma 2.5(c) can be found in that for [11, Lemma 2.3].

## A.3 Iterate perturbation

Gradient sampling algorithms require that each iterate lies in the set of points over which the objective function  $f$  is differentiable (for well-posedness of the algorithm) or even continuously differentiable (for the convergence guarantees); see [5]. For our proposed algorithms, we employ Algorithm 6 to ensure that each

---

**Algorithm 5** Armijo-Wolfe Line Search

---

**Require:**  $(\underline{\eta}, \overline{\eta}, \underline{\alpha})$  from outer algorithm;  $\overline{\alpha} \in [\underline{\alpha}, \infty)$ ;  $\alpha_{\text{init}} \in (0, \infty)$ ;  $\gamma \in (0, 1)$ .

```
1: Set  $l \leftarrow 0$ ,  $u \leftarrow \overline{\alpha}$ , and  $\alpha_k \leftarrow \alpha_{\text{init}}$ .
2: if  $d_k = 0$  then
3:   terminate and return  $\alpha_k$ .
4: end if
5: for all  $\ell \in \mathbb{N}$  do
6:   if  $p_k < p$  and  $\alpha_k < \underline{\alpha}$  then
7:     set  $\alpha_k \leftarrow 0$ , terminate, and return  $\alpha_k$ .           [truncate and take null stepsize]
8:   end if
9:   if  $\alpha_k < \underline{\alpha}$  then
10:    set  $l \leftarrow 0$ .                                           [switch to backtracking Armijo line search]
11:  end if
12:  if (21) holds or both  $\alpha_k < \underline{\alpha}$  and (21a) hold then
13:    terminate and return  $\alpha_k$ .                                   [success]
14:  end if
15:  if (21a) does not hold then
16:    set  $u \leftarrow \alpha_k$ ;
17:  else
18:    set  $l \leftarrow \alpha_k$ .
19:  end if
20:  set  $\alpha_k \leftarrow (1 - \gamma)l + \gamma u$ .
21: end for
```

---

iterate lies in the set  $\mathcal{D}$  defined in Assumption 1.1. If, after the line search, the resulting trial point is contained in  $\mathcal{D}$ , i.e.,  $x_k + \alpha_k d_k \in \mathcal{D}$ , then  $x_{k+1}$  is set to be this trial point; otherwise, the iterate perturbation strategy in Algorithm 6 aims to compute  $x_{k+1} \in \mathcal{D}$  satisfying (22).

Algorithm 6 can fail if its **for** loop iterates infinitely. However, under Assumption 1.1, this is a probability zero event. In other words, it terminates finitely—meaning the subroutine runs successfully—with probability one.

A proof of Lemma 2.5(d) can be found in that for [11, Lemma 2.3].

## A.4 Hessian and inverse Hessian approximation strategy

The Hessian approximation strategy employed in [11] is conservative in the sense that it might replace a BFGS approximation with an L-BFGS approximation in order to ensure that, in certain cases, the eigenvalues of the Hessian approximation are bounded above and below away from zero. For our purposes, we employ the less conservative strategy advocated in [12], which exploits the *self-correcting* properties of BFGS updating. The subroutine we use is stated in Algorithm 7.

**Lemma A.1.** *The properties pertaining to  $\{H_k\}$  and  $\{W_k\}$  stated in Lemma 2.5 hold.*

*Proof.* Proof. Positive definiteness of  $H_{k+1}$  and  $W_{k+1}$  follows by induction and the fact that  $H_0 \succ 0$  and  $W_0 \succ 0$ . In particular, if  $s_0 = 0$  or (23) does not hold in iteration  $k = 0$ , then  $H_1 \leftarrow H_0 \succ 0$  and  $W_1 \leftarrow W_0 \succ 0$ ; otherwise, positive definiteness of  $H_1$  and  $W_1$  follows the fact that (38) implies  $s_0^T v_0$  and from well-known properties of BFGS updating. Inductively, positive definiteness of  $H_{k+1}$  and  $W_{k+1}$  for any  $k \in \mathbb{N}$  follows by the same arguments. Finally, with respect to the properties of  $\{W_k\}$ , the proof follows in a similar manner as that for [12, Corollary 3.2].  $\square$

---

**Algorithm 6** Iterate Perturbation

---

**Require:**  $\bar{\ell} \in \mathbb{N}$ .

- 1: Set  $x_{k+1} \leftarrow x_k + \alpha_k d_k$ .
- 2: **if**  $\alpha_k = 0$  or  $d_k = 0$  **then**
- 3:     **terminate** and **return**  $x_{k+1}$ .
- 4: **end if**
- 5: **for all**  $\ell \in \mathbb{N}$  **do**
- 6:     **if**  $x_{k+1} \in \mathcal{D}$  and either (22) holds or each of (22a), (22c), and  $\ell > \bar{\ell}$  hold **then**
- 7:         **terminate** and **return**  $x_{k+1}$
- 8:     **end if**
- 9:     Sample  $x_{k+1}$  from a uniform distribution over

$$\mathbb{B} \left( x_k + \alpha_k d_k, \frac{\min\{\alpha_k, \epsilon_k\} \min\{\|d_k\|_2, \|G_k y_k\|_2\}}{\ell \max\{\|d_k\|_2, \|G_k y_k\|_2\}} \right).$$

10: **end for**

---

## A.5 Sample point generation

With  $x_{k+1} \in \mathcal{D}$  in hand, a GS algorithm turns to setting the set of sample points  $\mathcal{X}_{k+1}$  and corresponding size indicator  $p_{k+1}$ . To limit the size of the sample set, which has the benefit of reducing the costs of subsequent QP solves, we follow the lead of [11], which sets  $\mathcal{X}_{k+1} \leftarrow \{x_{k+1}\}$  if, for  $(\xi, \underline{\alpha}) \in (0, \infty)^2$ , one finds that (23) holds. Otherwise, the sample set preserves points near  $x_{k+1}$  and augments it with randomly generated points, the hallmark of GS methods; see Algorithm 8.

Like for Algorithm 6, one finds that Algorithm 8 can fail if its **do-while** loop iterates infinitely. However, under Assumption 1.1, this occurs with probability zero. The subroutine runs successfully with probability one.

A proof of Lemma 2.5(e) can be found in that for [11, Lemma 2.5].

---

**Algorithm 7** Hessian and Inverse Hessian Approximation Updates

---

**Require:**  $\underline{\phi} \in (0, 1)$  and  $\bar{\phi} \in (1, \infty)$  from outer algorithm.

- 1: Set  $s_k \leftarrow \alpha_k d_k$  and  $\hat{v}_k \leftarrow x_{k+1} - x_k$ .
- 2: **if**  $s_k = 0$  **then**
- 3:   set  $(H_{k+1}, W_{k+1}) \leftarrow (H_k, W_k)$ .
- 4: **else**
- 5:   Compute  $\vartheta_k$  as the smallest value in  $[0, 1]$  such that

$$v_k \leftarrow \vartheta_k s_k + (1 - \vartheta_k) y_k$$

- 6:   yields

$$\underline{\phi} \leq \frac{s_k^T v_k}{\|s_k\|_2^2} \quad \text{and} \quad \frac{\|v_k\|_2^2}{s_k^T v_k} \leq \bar{\phi}, \quad (38)$$

- 7:   then set

$$H_{k+1} \leftarrow \left( I - \frac{s_k s_k^T H_k}{s_k^T H_k s_k} \right)^T H_k \left( I - \frac{s_k s_k^T H_k}{s_k^T H_k s_k} \right) + \frac{v_k v_k^T}{s_k^T v_k}$$

and  $W_{k+1} \leftarrow \left( I - \frac{v_k s_k^T}{s_k^T v_k} \right)^T W_k \left( I - \frac{v_k s_k^T}{s_k^T v_k} \right) + \frac{s_k s_k^T}{s_k^T v_k}.$

- 8: **end if**
  - 9: **terminate** and **return**  $(H_{k+1}, W_{k+1})$ .
- 

---

**Algorithm 8** Sample Set Update

---

**Require:**  $\underline{\alpha}$  from outer algorithm;  $\xi \in (0, \infty)$ ;  $\bar{p} \in \mathbb{N}$  with  $\bar{p} \geq 1$ .

- 1: **if** (23) holds **then**
  - 2:   set  $\mathcal{X}_{k+1} \leftarrow \{x_{k+1}\}$  and  $p_{k+1} \leftarrow 0$ , **terminate**, and **return**  $(\mathcal{X}_{k+1}, p_{k+1})$ .
  - 3: **end if**
  - 4: **do**
  - 5:   set  $\mathcal{S}_{k+1}$  as a set of  $\bar{p}$  points from a uniform distribution over  $\mathbb{B}(x_{k+1}, \epsilon_{k+1})$ .
  - 6: **while**  $\mathcal{S}_{k+1} \not\subset \mathcal{D}$
  - 7:   Set  $\mathcal{X}_{k+1} \leftarrow \{x_{k+1}\} \cup (\mathcal{X}_k \cap \mathbb{B}(x_{k+1}, \epsilon_{k+1})) \cup \mathcal{S}_{k+1}$  and  $p_{k+1} \leftarrow |\mathcal{X}_{k+1}| - 1$ .
  - 8: **if**  $p_{k+1} > p$  **then**
  - 9:   remove the  $p_{k+1} - p$  eldest members of  $\mathcal{X}_{k+1}$  (except  $\{x_{k+1}\}$ ) and set  $p_{k+1} \leftarrow p$ .
  - 10: **end if**
  - 11: **terminate** and **return**  $(\mathcal{X}_{k+1}, p_{k+1})$ .
-