

Goldstein Stationarity in Lipschitz Constrained Optimization

Benjamin Grimmer^{1†} and Zhichao Jia^{2†}

¹Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA.

²H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

Contributing authors: grimmer@jhu.edu; zjia75@gatech.edu;

[†]These authors contributed equally to this work.

Abstract

We prove the first convergence guarantees for a subgradient method minimizing a generic Lipschitz function over generic Lipschitz inequality constraints. No smoothness or convexity (or weak convexity) assumptions are made. Instead, we utilize a sequence of recent advances in Lipschitz unconstrained minimization, which showed convergence rates of $\mathcal{O}(1/\delta\epsilon^3)$ towards reaching a “Goldstein” stationary point, that is, a point where an average of gradients sampled at most distance δ away has size at most ϵ . We generalize these prior techniques to handle functional constraints, proposing a subgradient-type method with similar $\mathcal{O}(1/\delta\epsilon^3)$ guarantees on reaching a Goldstein Fritz-John or Goldstein KKT stationary point, depending on whether a certain Goldstein-style generalization of constraint qualification holds.

Keywords: Constrained Optimization, Nonsmooth Nonconvex Optimization, Goldstein Stationarity, Subgradient Methods, Constraint Qualification

1 Introduction

This work considers a quite general family of constrained optimization problems where both the objective and constraint functions may be nonconvex and nonsmooth. The only structure assumed is Lipschitz continuity of these functions. Specifically, we

consider problems of the following form

$$p_\star = \begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & g_i(x) \leq 0, \quad i = 1, \dots, m \end{cases} \quad (1)$$

with objective $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and constraints $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$. We assume these functions are M -Lipschitz continuous on a neighborhood of the feasible region $\{x \mid g_i(x) \leq 0\}$, but do not assume convexity or differentiability. We propose an iterative subgradient-type method that only computes one gradient (or subgradient-like vector) of one of these functions at each iteration. For now, we leave general the exact notion of our first-order oracle $\mathcal{G}_h(x)$ and associated subdifferential by $\partial h(x)$. Throughout, this will either be the Clarke subdifferential or a certain nonstandard directional subdifferential.

For unconstrained minimization (i.e., $m = 0$), in 1977, Goldstein [1] proposed an idealized algorithm achieving descent by repeatedly moving in minimal norm subgradient directions, eventually reaching a related notion of approximate stationarity. Mahdavi-Amiri and Yousefpou [2] revisited this idea, giving a more practical minimization algorithm approximating these descent directions. The groundbreaking work of Zhang et al. [3] showed assuming only Lipschitz continuity is sufficient for provable convergence guarantees towards stationarity. Subsequently, Davis et al. [4] and Kong and Lewis [5] built on this using the more standard Clarke subgradient oracle and avoiding the use of randomness, respectively. A key insight enabling these works to prove guarantees for such generic problems is considering a weakened notion of stationarity, seeking points where a convex combination of gradients computed nearby is small. Formally, we say x is an (ϵ, δ) -Goldstein stationary point of h if $\text{dist}(0, \partial_\delta h(x)) \leq \epsilon$ where $\partial_\delta h(x)$ denotes the Goldstein subdifferential defined as

$$\partial_\delta h(x) = \text{conv} \left\{ \bigcup_{y \in B(x, \delta)} \partial h(y) \right\}$$

where $B(x, \delta)$ is the closed ball of radius δ around x and $\text{conv}(\cdot)$ denotes the convex hull. This condition is a natural relaxation of the optimality condition $0 \in \partial h(x)$.

Davis et al. [4] showed for any Lipschitz h , using $O(1/\delta\epsilon^3)$ gradient evaluations, an (ϵ, δ) -Goldstein stationary point can be found in expectation (where the randomness arises from sampling a random point for each gradient evaluation). Kong and Lewis [5] showed an entirely deterministic method could find such a point in $O(1/\delta\epsilon^4)$ subgradient-like evaluations (although depending on a certain nonconvexity modulus). Cutkosky et al. [6] derived a $O(1/\delta\epsilon^3)$ rate using only stochastic gradient evaluations (beyond the scope of this work) and matching lower bounds when $M = \Omega(\sqrt{\frac{\epsilon}{\delta}})$. Both Davis et al. and Kong and Lewis use a double-loop method, only differing in their inner loop's implementation, which approximates Goldstein's descent direction.

This work generalizes these prior works to the functionally constrained setting (1). We show that a simple modification of previous double-loop methods can solve generic constrained problems of the form (1). To analyze this, we propose new Goldstein-type generalizations of classic Fritz-John, KKT, and Constraint Qualification conditions

(see Definitions 1-3). Our Theorems 1 and 2 show that using the same inner-loop as Davis et al. [4], an approximate Goldstein Fritz John or KKT point is reached at the same rate, depending on whether constraint qualification holds, in expectation. Similarly, using the inner loop of Kong and Lewis [5] extends their deterministic rate to functional constraints. Associated dual multipliers can be extracted from our method, so stopping criteria and certificates of approximate stationarity follow immediately.

1.1 Related Works and a Sketch of our Algorithm

We first review classic results on Lagrangian optimality conditions and provably good methods for weakly convex constrained problems and for Lipschitz unconstrained problems. Subsequently, Section 2 formalizes all the necessary definitions and our algorithm. Section 3 then states and proves our convergence theorems.

Lagrangians, KKT, and Constraint Qualification. The Lagrangian of (1) for any nonnegative Lagrange multipliers $\lambda \in \mathbb{R}^m$ is denoted by $L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i g_i(x)$. We say a feasible point $x^* \in \mathbb{R}^n$ is KKT stationary if there exist such multipliers where

$$\begin{cases} \partial f(x^*) + \sum_{i=1}^m \lambda_i \partial g_i(x^*) \ni 0, \\ \lambda_i g_i(x^*) = 0, \quad \forall i = 1 \dots m. \end{cases} \quad (2)$$

Generally, this condition is not necessary for x^* to be a minimizer of (1). Instead, one can only ensure every minimizer x^* is Fritz-John (FJ) stationary, meaning that there exist nonnegative multipliers $\gamma_0, \dots, \gamma_m$, not all zero, such that

$$\begin{cases} \gamma_0 \partial f(x^*) + \sum_{i=1}^m \gamma_i \partial g_i(x^*) \ni 0, \\ \gamma_i g_i(x^*) = 0, \quad \forall i = 1 \dots m. \end{cases} \quad (3)$$

The above FJ condition implies the KKT condition whenever the ‘‘constraint qualification’’ (CQ) condition $0 \notin \partial \max\{g_i(x) \mid g_i(x^*) = 0\}(x^*)$ holds at x^* ¹. Hence, when CQ holds, KKT is a necessary optimality condition. Our theory gives guarantees on the approximate attainment of FJ or KKT conditions, depending on whether a slightly strengthened constraint qualification condition holds.

Prior Methods for Weakly Convex Constrained Problems. Previous works [7–9] have addressed nonsmooth nonconvex constrained problems under the additional assumption that f, g_0, \dots, g_m are all ρ -weakly convex. A function h is ρ -weakly convex if $h + \frac{\rho}{2} \|\cdot\|^2$ is convex. Under this additional condition, one can achieve an objective decrease on (1) by solving the following strongly convex proximal subproblem

$$\begin{cases} \min_{x \in \mathbb{R}^n} & f(x) + \frac{\hat{\rho}}{2} \|x - x_k\|^2 \\ \text{s.t.} & g_i(x) + \frac{\hat{\rho}}{2} \|x - x_k\|^2 \leq 0, \quad i = 1, \dots, m \end{cases} \quad (4)$$

for any $\hat{\rho} > \rho$. The works [7–9] all considered double-loop methods where the inner loop approximately solves this subproblem to produce the next iterate of the outer

¹In particular, constraint qualification ensures γ_0 is positive. Hence $\lambda_i = \gamma_i/\gamma_0$ satisfies KKT.

loop. In all these works, convergence rates on the order of $O(1/\epsilon^4)$ were proven towards approximate KKT stationarity holding near the iterates. Additionally, [9] showed that without constraint qualification, approximate FJ stationarity is still reached at the same rate. A direct, single-loop approach was recently developed by [10], attaining the same rates via a much simpler iteration.

Prior Methods for Lipschitz Unconstrained Problems. Inspired by Goldstein [1], previous works [2–6] have proposed their algorithms for solving nonsmooth nonconvex unconstrained problems under Lipschitz continuity. These works considered double-loop methods, where the inner loop computes an approximate minimal norm element of the Goldstein subdifferential, either being close to zero or producing at least some descent on the objective function. These descent directions are then followed in an outer loop until a small Goldstein subgradient is found.

Sketch of our Method for Lipschitz, Constrained Problems. Without weak convexity, the proximal subproblem (4) becomes intractable. Instead, at each outer iteration k , we consider the unconstrained Lipschitz subproblem of minimizing $h_{x_k}(x) := \max\{f(x) - f(x_k), g_i(x)\}$. Using a subgradient norm minimizing inner loop, we either identify a descent on h_{x_k} from x_k , yielding a feasible descent for (1), or certify that x_k is approximately Goldstein stationary to h_{x_k} , yielding approximate FJ or KKT stationarity for (1).

2 Goldstein KKT Stationarity and our Algorithm

For a fixed notion of subdifferential $\partial h(x)$ (two such models are formalized in Sections 2.2 and 2.3), we define Goldstein-type measures of approximate notions of Fritz-John (3) and KKT stationarities (2) for the constrained problem (1).

Definition 1. We say a feasible point x is (δ, ϵ, η) -Goldstein Fritz-John (GFJ) stationary for problem (1) if there exists $(\gamma_0, \gamma) \geq 0$ with $\gamma_0 + \sum_{i=1}^m \gamma_i = 1$, such that $\text{dist}(0, \gamma_0 \partial_\delta f(x) + \sum_{i=1}^m \gamma_i \partial_\delta g_i(x)) \leq \epsilon$ and $\max_{y \in B(x, \delta)} |\gamma_i g_i(y)| \leq \eta, \forall i = 1, \dots, m$.

Definition 2. We say a feasible point x is (δ, ϵ, η) -Goldstein KKT (GKKT) stationary for problem (1) if there exists $\lambda \geq 0$ such that $\text{dist}(0, \partial_\delta f(x) + \sum_{i=1}^m \lambda_i \partial_\delta g_i(x)) \leq \epsilon$ and $\max_{y \in B(x, \delta)} |\lambda_i g_i(y)| \leq \eta, \forall i = 1, \dots, m$.

GFJ stationarity implies GKKT stationarity when $\gamma_0 > 0$ in Definition 1. A larger value of γ_0 leads to stronger guarantees on approximate GKKT stationarity. In particular, GFJ stationarity is equivalent to GKKT stationarity if $\gamma_0 = 1$. Constraint Qualification (CQ) classically provides a means to bound γ_0 away from zero. Below, we generalize CQ to use Goldstein subdifferentials.

Definition 3. We say a feasible point x satisfies (a, b, c) -Goldstein Constraint Qualification (GCQ) if all $\gamma \geq 0$ with $\sum_{i=1}^n \gamma_i = 1$ and $\gamma_i > 0$ only if $g_i(x) \geq -c$ have $\text{dist}(0, \sum_{i=1}^n \gamma_i \partial_a g_i(x)) \geq b$.

2.1 The Proposed Constrained Goldstein Subgradient Method

We make the following pair of modeling assumptions throughout.

Assumption 1. The optimal objective value p_* is finite and for some $\Delta > 0$, f and each g_i are M -Lipschitz continuous within distance Δ of $\{x \mid g_i(x) \leq 0\}$.

Assumption 2. An initial feasible point x_0 is known (i.e. $g(x_0) \leq 0$).

Algorithm 1 Constrained Goldstein Subgradient Method

Require: $\delta, \tilde{\epsilon}, x_0$ with $g(x_0) \leq 0$

for $k = 0, 1, \dots$, **do**

 Compute $\zeta_k = \mathcal{A}(x_k, h_{x_k}, \delta, \tilde{\epsilon})$, an approximate minimal element of $\partial_\delta h_{x_k}(x_k)$

if $\|\zeta_k\| \leq \tilde{\epsilon}$ **then return** x_k

$x_{k+1} = x_k - \delta \zeta_k / \|\zeta_k\|$, $k = k + 1$

end for

Without loss of generality, the m constraints of (5) can be combined into one:

$$p_\star = \begin{cases} \min_{x \in \mathbb{R}^n} & f(x) \\ \text{s.t.} & g(x) := \max_{i=1, \dots, m} g_i(x) \leq 0 . \end{cases} \quad (5)$$

Our proposed subgradient method then maintains a feasible solution x_k and proceeds by, in an inner loop, computing in a descent direction of

$$\min h_{x_k}(z) := \max\{f(z) - f(x_k), g(z)\} , \quad (6)$$

and then moving in this direction a fixed distance $\delta < \Delta$ in an outer loop. To facilitate this, we require some subgradient-type oracle \mathcal{G} be provided for f, g , and h_x . The two particular computational models we consider here are formalized in Assumptions 3a and 3b, with \mathcal{G} as a gradient almost everywhere or as a specialized directional derivative matching subgradient, respectively. In general, all we require is the following.

Assumption 3. For any $x, z \in \mathbb{R}^n$, subgradient-type oracles are known with $\mathcal{G}_f(z) \in \partial f(z), \mathcal{G}_g(z) \in \partial g(z), \mathcal{G}_{h_x}(z) \in \partial h_x(z)$ satisfying

$$\mathcal{G}_{h_x}(z) \in \begin{cases} \{\mathcal{G}_f(z)\} & \text{if } f(z) - f(x) > g(z) \\ \{\mathcal{G}_f(z), \mathcal{G}_g(z)\} & \text{if } f(z) - f(x) = g(z) \\ \{\mathcal{G}_g(z)\} & \text{if } f(z) - f(x) < g(z) . \end{cases}$$

We require a subroutine outputting an approximate minimal norm element of the Goldstein subdifferential set of h_{x_k} at x_k in each outer iteration k . We consider any subroutine $\mathcal{A}(x, h, \delta, \epsilon)$ producing a direction $\zeta \in \mathbb{R}^n$ defined by $t \leq T_{\mathcal{A}}(M, \epsilon, \delta, \tau)$ subgradient oracle calls with probability at least $(1 - \tau)$ satisfying

$$\|\zeta\| \leq \epsilon \quad \text{or} \quad h(x) - h(x - \delta \zeta / \|\zeta\|) \geq C \delta \epsilon ,$$

with $\zeta = \sum_{i=1}^t w_i \mathcal{G}_{h_x}(z_i)$ for some points $z_i \in B(x, \delta)$ and weights $w \geq 0, \sum_{i=1}^t w_i = 1$ for some fixed $C \in (0, 1)$ and $\tau \in [0, 1)$.

The evaluation of such a subroutine forms the inner loop of our method. As two examples, Davis et al. [4] gave a randomized searching subroutine, which we denote as *RandSearch*(x, h, δ, ϵ), and Kong and Lewis [5] gave a deterministic linesearching subroutine defined below as *BisecSearch*(x, h, δ, ϵ). Our proposed algorithm can use any such subroutine as a black-box inner loop, formalized in Algorithm 1. The following

two lemmas show finite termination and a relationship between Goldstein stationarity on h_{x_k} and GFJ and GKKT stationarity (proofs deferred to Section 3).

Lemma 1. *Algorithm 1 always terminates with a feasible x_k with $k \leq \left\lceil \frac{f(x_0) - p_*}{C\delta\epsilon} \right\rceil$.*

Lemma 2. *If a subroutine \mathcal{A} reports a feasible x is (δ, ϵ) -Goldstein stationary to h_x , then x is $(\delta, \epsilon, 3M\delta)$ -GFJ stationary for problem (5). If $(\delta, \hat{\epsilon}, M\delta)$ -GCQ holds at x with $\hat{\epsilon} > \epsilon$, then x is $(\delta, \epsilon(\hat{\epsilon} + M)/(\hat{\epsilon} - \epsilon), 3M\delta(\hat{\epsilon} + M)/(\hat{\epsilon} - \epsilon))$ -GKKT stationary.*

The Lagrange multiplier λ_k associated with each iteration x_k can be easily computed: Let z_i and w_i denote the points and weights associated with the construction of ζ_k . By Assumption 3, letting $I_f = \{i \mid \mathcal{G}_f(z_i) = \mathcal{G}_{h_{x_k}}(z_i)\}$, we have $\zeta_k = \sum_{i \in I_f} w_i \mathcal{G}_f(z_i) + \sum_{i \notin I_f} w_i \mathcal{G}_g(z_i)$. Provided $\sum_{i \in I_f} w_i > 0$, which GCQ ensures, $\lambda_k = \sum_{i \notin I_f} w_i / \sum_{i \in I_f} w_i$ certifies approximate GKKT stationarity.

2.2 Oracles and Minimum Norm Subroutine of Davis et al. [4]

Note that the assumed uniform Lipschitz continuity ensures gradients exist almost everywhere for f , g , and h_{x_k} (by Rademacher). Davis et al. [4] leverage this by only requiring subgradients at randomly sampled points that are differentiable with probability one. Such an oracle can often be produced via automatic differentiation.

Assumption 3a. *Oracles computing $\mathcal{G}_f(x) = \nabla f(x)$ and $\mathcal{G}_g(x) = \nabla g(x)$ almost everywhere in \mathbb{R}^n are known.*

Given these oracles, a gradient oracle for h_{x_k} almost everywhere can be easily constructed, returning $\nabla f(x)$ if $f(x) - f(x_k) \geq g(x)$ and $\nabla g(x)$ otherwise. Using this oracle, the randomized search subroutine of [4] finds an approximate minimum norm Goldstein subgradient as defined in Algorithm 2. The natural corresponding subdifferential is the set of all Clarke subgradients, given by $\partial h(x) = \text{conv} \{\lim_{i \rightarrow \infty} \nabla h(x_i) \mid x_i \rightarrow x, x_i \in \text{dom}(\nabla h)\}$.

2.3 Oracles and Minimum Norm Subroutine of Kong et al. [5]

Alternatively, Kong and Lewis [5] showed an entirely deterministic approach could be used at the cost of requiring additional mild assumptions and a stronger subgradient-type oracle. The directional derivative of any function $h : \mathbb{R}^n \rightarrow \mathbb{R}$ is defined by

$$\nabla_v h(x) = \lim_{t \rightarrow 0^+} \frac{h(x + tv) - h(x)}{t},$$

for any point $x \in \mathbb{R}^n$ and any direction $v \in \mathbb{R}^n$. Kong and Lewis [5]’s subroutine requires an oracle able to compute subgradient-like vectors in agreement with this directional derivative and bounds on the following quantities. For any one-dimensional function $\ell : [0, \delta] \rightarrow \mathbb{R}$, we denote the “concave deviation” by

$$c_\delta(\ell) = \inf\{M \geq 0 \mid \ell + s \text{ is convex for some convex } M\text{-Lipschitz } s : [0, \delta] \rightarrow \mathbb{R}\}.$$

This notion extends to multivariate functions $h : \mathbb{R}^n \rightarrow \mathbb{R}$ by considering each restriction of h to a line segment $\ell_{h,x,y}(r) = h(x + r(y - x)/\delta)$ for $\|x - y\| \leq \delta$. We denote

Algorithm 2 *RandSearch*(z, h, δ, ϵ)

Require: a feasible z, h, δ, ϵ

Let $t = 0, \zeta_0 = \nabla h_z(y_0)$, where $y_0 \sim B(z, \delta)$

while $\|\zeta_t\| > \epsilon$ and $\delta\|\zeta_t\|/4 \geq h_z(z) - h_z(z - \delta\zeta_t/\|\zeta_t\|)$ **do**

 Choose any r satisfying $0 < r < \|\zeta_t\| \cdot \sqrt{1 - (1 - \|\zeta_t\|^2/128M^2)^2}$

 Sample y_t uniformly from $B(\zeta_t, r)$

 Choose s_t uniformly at random from the segment $[z, z - \delta y_t/\|y_t\|]$

$\zeta_{t+1} = \operatorname{argmin}_{\zeta \in [\zeta_t, \nabla h_z(s_t)]} \|\zeta\|, \quad t = t + 1$

end while

Algorithm 3 *BisecSearch*(z, h, δ, ϵ)

Require: a feasible z, h, δ, ϵ

Let $t = 0, \zeta_0 = \nabla h_z(z)$

while $\|\zeta_t\| > \epsilon$ and $\delta\epsilon/3 > h_z(z) - h_z(z - \delta\hat{\zeta}_t)$ **do**

 Use bisection [5, Algorithm 3.1] to find r_t satisfying $\langle H_z(z_t(r_t), \hat{\zeta}_t), \hat{\zeta}_t \rangle - \epsilon/2 < 0$

$\zeta_{t+1} = \operatorname{argmin}_{\zeta \in [\zeta_t, H_z(z_t(r_t), \hat{\zeta}_t)]} \|\zeta\|, \quad t = t + 1$

end while

the nonconvexity modulus of h by the largest concave deviation among these

$$\Lambda_h(\delta) = \sup_{\|x-y\| \leq \delta} \{c_\delta(\ell_{h,x,y})\}.$$

Assumption 3b. *The functions f and g are directionally differentiable, have oracles producing directional subgradient maps $F(x, v), G(x, v)$ satisfying $\langle F(x, v), v \rangle = \nabla_v f(x)$, $\langle G(x, v), v \rangle = \nabla_v g(x)$, and have $\Lambda_f(\delta)$ and $\Lambda_g(\delta)$ both finite.*

For any $x, z, v \in \mathbb{R}^n$, we define the associated directional subgradient map for h_x as

$$H_x(z, v) = \begin{cases} F(z, v) & \text{if } f(z) - f(x) > g(z), \text{ or } f(z) - f(x) = g(z) \text{ and } \nabla_v f(z) \geq \nabla_v g(z) \\ G(z, v) & \text{if } f(z) - f(x) < g(z), \text{ or } f(z) - f(x) = g(z) \text{ and } \nabla_v f(z) < \nabla_v g(z). \end{cases}$$

This is justified by noting the directional derivative of h_x at z is given by

$$\nabla_v h_x(z) = \begin{cases} \nabla_v f(z), & f(z) - f(x) > g(z) \\ \max\{\nabla_v f(z), \nabla_v g(z)\}, & f(z) - f(x) = g(z) \\ \nabla_v g(z), & f(z) - f(x) < g(z). \end{cases}$$

Note $\Lambda_{h_x}(\delta) \leq \Lambda_f(\delta) + \Lambda_g(\delta) < \infty$ since any convex Lipschitz functions s_f, s_g such that $\ell_{f,x,y} + s_f$ and $\ell_{g,x,y} + s_g$ are convex have $\ell_{h_z,x,y} + s_f + s_g$ convex.

Suppressing the dependence on the given direction v , we denote $\mathcal{G}_f(z) = F(z, v)$, $\mathcal{G}_g(z) = G(z, v)$, and $\mathcal{G}_{h_x}(z) = H_x(z, v)$. One corresponding subdifferential is the set of all directional subgradients, $\partial h(x) = \{\zeta \mid \exists v \in \mathbb{R}^n \text{ s.t. } \langle \zeta, v \rangle = \nabla_v h(x)\}$.

Using such oracles, the bisection search subroutine of [5] finds an approximate minimum norm Goldstein subgradient as defined in Algorithm 3. For a given $z \in \mathbb{R}^n$,

this subroutine constructs a sequence of Goldstein subgradients ζ_t of h_z at z , eventually satisfying one of the needed conditions. Defining $\hat{\zeta}_t = \zeta_t/\|\zeta_t\|$, $z_t(r) = z + (r - \delta)\hat{\zeta}_t$ and $l_t(r) = h_z(z_t(r)) - \epsilon r/2$ on $[0, \delta]$, knowing $l_t(0) > l_t(\delta)$ (an average rate of decrease), the critical step in this construction is the bisection [5, Algorithm 3.1], finding a location where the right derivative $l'_{t+}(r) = \lim_{\Delta r \rightarrow 0^+} (l_t(r + \Delta r) - l_t(r))/\Delta r$ of $l_t(r)$ is negative by bisection. Finite nonconvexity modulus guarantees termination.

3 Convergence Guarantees and Analysis

We provide two main theorems on FJ and KKT guarantees, using Algorithm 2 or 3, respectively, as the subroutines. These follow primarily from our Lemmas 1 and 2.

Theorem 1. *Given Assumptions 1 and 2, for any $\epsilon > 0$ and $0 < \delta < \Delta$, Algorithm 1 with $\tilde{\epsilon} = \epsilon$ will terminate with an $(\delta, \epsilon, 3M\delta)$ -GFJ stationary point. Under Assumption 3a and using Algorithm 2 as the subroutine, with probability $1 - \tau$, this requires at most*

$$\left\lceil \frac{4(f(x_0) - p_\star)}{\delta\epsilon} \right\rceil \left\lceil \frac{64M^2}{\epsilon^2} \right\rceil \left\lceil 2 \log \left(\frac{4(f(x_0) - p_\star)}{\tau\delta\epsilon} \right) \right\rceil$$

first-order oracle calls. Under Assumption 3b and using Algorithm 3, this becomes

$$\left\lceil \frac{3(f(x_0) - p_\star)}{\delta\epsilon} \right\rceil \left\lceil \frac{16M^2}{\epsilon^2} \right\rceil \left(1 + \left\lceil \frac{12(\Lambda_f(\delta) + \Lambda_g(\delta))}{\epsilon} \right\rceil \right).$$

Proof. According to Lemma 1, we attain a (δ, ϵ) -Goldstein stationary point x_k to h_{x_k} with $g(x_k) < 0$ with $k \leq \left\lceil \frac{f(x_0) - p_\star}{C\delta\epsilon} \right\rceil$. According to Lemma 2, x is a $(\delta, \epsilon, 3M\delta)$ -GFJ stationary solution for problem (5). By Corollary 5 and Theorem 6 in [4], each call to Algorithm 2 uses at most $\lceil 64M^2/\epsilon^2 \rceil \lceil 2 \log(1/\tau) \rceil$ gradient evaluations with probability at least $1 - \tau$ and $C = 1/4$. Likewise, by Theorem 4.3, Corollary 5.6 and Theorem 6.6 in [5], Algorithm 3 requires no more than $\lceil 16M^2/\epsilon^2 \rceil (1 + \lceil 12(\Lambda_f(\delta) + \Lambda_g(\delta))/\epsilon \rceil)$ subgradient evaluations with $C = 1/3$. \square

Theorem 2. *Given Assumptions 1, 2, for any $\epsilon > 0$ and $0 < \delta < \Delta$ such that $(\delta, \sigma, M\delta)$ -GCQ holds for all x , Algorithm 1 with $\tilde{\epsilon} = \sigma\epsilon/(\epsilon + \sigma + M)$ will terminate with an $(\delta, \epsilon, 3M\delta(\epsilon + \sigma + M)/\sigma)$ -Goldstein KKT point. Under Assumption 3a and using Algorithm 2 as the subroutine, with probability $1 - \tau$, this requires at most*

$$\left\lceil \frac{4(f(x_0) - p_\star)}{\delta\tilde{\epsilon}} \right\rceil \left\lceil \frac{64M^2}{\tilde{\epsilon}^2} \right\rceil \left\lceil 2 \log \left(\frac{4(f(x_0) - p_\star)}{\tau\delta\tilde{\epsilon}} \right) \right\rceil$$

first-order oracle calls. Under Assumption 3b and using Algorithm 3, this becomes

$$\left\lceil \frac{3(f(x_0) - p_\star)}{\delta\tilde{\epsilon}} \right\rceil \left\lceil \frac{16M^2}{\tilde{\epsilon}^2} \right\rceil \left(1 + \left\lceil \frac{12(\Lambda_f(\delta) + \Lambda_g(\delta))}{\tilde{\epsilon}} \right\rceil \right).$$

Proof. By Lemmas 1 and 2, our choice $\tilde{\epsilon} = \frac{\sigma\epsilon}{\epsilon+\sigma+M} < \sigma$ ensures x_k is a feasible $(\delta, \tilde{\epsilon}, 3M\tilde{\epsilon})$ -GFJ point. Then $(\delta, \sigma, M\delta)$ -GCQ implies x_k is $(\delta, \epsilon, 3M\delta(\epsilon + \sigma + M)/\sigma)$ -GKKT stationary, where ϵ is derived by $\epsilon = \tilde{\epsilon}(\sigma + M)/(\sigma - \tilde{\epsilon})$, and $3M\delta(\epsilon + \sigma + M)/\sigma = 3M\delta(\sigma + M)/(\sigma - \tilde{\epsilon})$. Using the same subroutine bounds of [4, 5] used in the prior proof of Theorem 1 finishes the proof. \square

3.1 Proof of Lemma 1

We prove this by inductively showing feasibility is maintained and the objective value decreases by at least $C\delta\epsilon$ each iteration until $\|\zeta_k\| \leq \epsilon$. Supposing $\|\zeta_k\| > \epsilon$, our black-box subroutine finds $h_{x_k}(x_k) - h_{x_k}(x_{k+1}) \geq C\delta\epsilon$. Since $h_x(x) = \max\{f(x) - f(x), g(x)\} = 0$ whenever $g(x) \leq 0$, we know $\max\{f(x_{k+1}) - f(x_k), g(x_{k+1})\} \leq -C\delta\epsilon$. Thus we have descent $f(x_k) - f(x_{k+1}) \geq C\delta\epsilon$ and maintain feasibility $g(x_{k+1}) \leq -C\delta\epsilon$. Given an initial objective gap of $f(x_0) - p_*$, this occurs at most $\left\lceil \frac{f(x_0) - p_*}{C\delta\epsilon} \right\rceil$ times.

3.2 Proof of Lemma 2

If x is (δ, ϵ) -Goldstein stationary to h_x , there exists t points $z_i \in B(x, \delta)$ with (sub)gradients $\mathcal{G}_{h_x}(x_i) \in \partial h_x(x_i)$ and weights $w > 0$ with $\sum_{i=1}^t w_i = 1$ such that $\zeta = \sum_{i=1}^t w_i \mathcal{G}_{h_x}(x_i)$ has $\|\zeta\| \leq \epsilon$. By Assumption 3, letting $I_f = \{i \mid \mathcal{G}_f(z_i) = \mathcal{G}_{h_x}(z_i)\}$, we have $\zeta = \sum_{i \in I_f} w_i \mathcal{G}_f(z_i) + \sum_{i \notin I_f} w_i \mathcal{G}_g(z_i)$. This motivates selecting $\gamma_0 = \sum_{i \in I_f} w_i \geq 0$ and $\gamma = \sum_{i \notin I_f} w_i \geq 0$. Indeed this selection has $\gamma_0 + \gamma = 1$ and $\text{dist}(0, \gamma_0 \partial_\delta f(x) + \gamma \partial_\delta g(x)) \leq \epsilon$. Lastly, for establishing approximate Goldstein Fritz-John stationarity, we verify complementary slackness. This trivially holds if $\gamma = 0$. Otherwise let $i \notin I_f$ and observe that $|g(z)| \leq 3M\delta$ for any $z \in B(x, \delta)$ since we have

$$g(z) = g(z) - g(x) + g(x) \leq M\delta$$

using M -Lipschitz continuity of g on $\|z - x\| \leq \delta$ and the feasibility of x , and we have

$$g(z) = g(z) - g(z_i) + g(z_i) \geq -3M\delta$$

using M -Lipschitz continuity of g on $\|z_i - z\| \leq 2\delta$ and $g(z_i) \geq f(z_i) - f(x) \geq -M\delta$. Since $\gamma \leq 1$, $\max_{z \in B(x, \delta)} |\gamma g(z)| \leq 3M\delta$ and so x is a $(\delta, \epsilon, 3M\delta)$ -GFJ point.

Assuming $(\delta, \hat{\epsilon}, M\delta)$ -GCQ holds, we first claim $\gamma_0 \geq \frac{\hat{\epsilon} - \epsilon}{\hat{\epsilon} + M} > 0$ since

$$(1 - \gamma_0)\hat{\epsilon} = \gamma\hat{\epsilon} \leq \left\| \sum_{i \notin I_f} w_i \mathcal{G}_g(z_i) \right\| \leq \left\| \sum_{i=1}^t w_i \mathcal{G}_{h_x}(z_i) \right\| + \left\| \sum_{i \in I_f} w_i \mathcal{G}_f(z_i) \right\| \leq \epsilon + \gamma_0 M.$$

As a result, consider the Lagrangian multiplier $0 \leq \lambda := \gamma/\gamma_0 \leq \frac{\hat{\epsilon} + M}{\hat{\epsilon} - \epsilon} - 1$. Then $(\delta, \epsilon(\hat{\epsilon} + M)/(\hat{\epsilon} - \epsilon), 3M\delta(\hat{\epsilon} + M)/(\hat{\epsilon} - \epsilon))$ -GKKT stationary for problem (5) follows as

$$\text{dist}(0, \partial_\delta f(x) + \lambda \partial_\delta g(x)) \leq \frac{\epsilon}{\gamma_0} \leq \frac{\epsilon(\hat{\epsilon} + M)}{\hat{\epsilon} - \epsilon},$$

$$\max_{z \in B(x, \delta)} |\lambda g(z)| \leq \frac{\max_{z \in B(x, \delta)} |\gamma g(z)|}{\gamma_0} \leq \frac{3M\delta(\hat{\epsilon} + M)}{\hat{\epsilon} - \epsilon}.$$

Statements and Declarations. Benjamin Grimmer was supported by the Air Force Office of Scientific Research under award number FA9550-23-1-0531. No data sets were used in or generated by this work. The authors have no competing interests to declare.

References

- [1] Goldstein, A.: Optimization of lipschitz continuous functions. *Mathematical Programming* **13**, 14–22 (1977)
- [2] Mahdavi-Amiri, N., Yousefpour, R.: An effective nonsmooth optimization algorithm for locally lipschitz functions. *Journal of Optimization Theory and Applications* **155**, 180–195 (2012)
- [3] Zhang, J., Lin, H., Jegelka, S., Sra, S., Jadbabaie, A.: Complexity of finding stationary points of nonconvex nonsmooth functions. In: *International Conference on Machine Learning*, pp. 11173–11182 (2020). PMLR
- [4] Davis, D., Drusvyatskiy, D., Lee, Y.T., Padmanabhan, S., Ye, G.: A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *Advances in Neural Information Processing Systems* **35**, 6692–6703 (2022)
- [5] Kong, S., Lewis, A.: The cost of nonconvexity in deterministic nonsmooth optimization. *arXiv:2210.00652* (2022)
- [6] Cutkosky, A., Mehta, H., Orabona, F.: Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In: *International Conference on Machine Learning, ICML. Proceedings of Machine Learning Research*, vol. 202, pp. 6643–6670 (2023)
- [7] Ma, R., Lin, Q., Yang, T.: Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In: *International Conference on Machine Learning*, pp. 6554–6564 (2020). PMLR
- [8] Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and non-convex functional constrained optimization. *Mathematical Programming* **197**(1), 215–279 (2023)
- [9] Jia, Z., Grimmer, B.: First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points. *arXiv:2212.00927* (2022)
- [10] Huang, Y., Lin, Q.: Oracle Complexity of Single-Loop Switching Subgradient Methods for Non-Smooth Weakly Convex Functional Constrained Optimization (2023)