A SMOOTHING DIRECT SEARCH METHOD FOR MONTE CARLO-BASED CONSTRAINED NONSMOOTH NONCONVEX OPTIMIZATION

XIAOJUN CHEN*, C. T. KELLEY†, FENGMIN XU‡, AND ZAIKUN ZHANG§

February 14, 2017

Abstract. We propose and analyze a smoothing direct search algorithm for finding a minimizer of a nonsmooth nonconvex function over a convex set, where the objective function values cannot be computed directly, but are approximated by Monte Carlo simulation. In the algorithm, we adjust the stencil size, the sample size, and the smoothing parameter simultaneously so that the stencil size goes to zero faster than the smoothing parameter and the square root of the sample size goes to infinity faster than the reciprocal of the stencil size. We prove that with probability one any accumulation point of the sequence generated by the algorithm is a Clarke stationary point. We report on numerical results from statistics and financial applications.

Key words. Sampling methods, direct search algorithm, Monte Carlo simulation, non-smooth optimization, smoothing functions, Clarke stationarity

AMS subject classifications. 65K05, 65K10, 90C30

1. Introduction.

In this paper, we consider the nonsmooth minimization problem

$$\min_{x\in X} f(x), \tag{1.1}$$

where X is a nonempty closed convex set and $f : X \to R$ is a locally Lipschitz continuous function. We propose and analyze a smoothing direct search method in the case where the function value $f(x)$ cannot be computed directly, but is approximated by a Monte Carlo simulation $f^N(x)$, where N is the sample size of the Monte Carlo simulation.

In [9] we considered a similar problem in the more general situation where the objective function was not everywhere defined and capturing the domain of f was part of the problem. In this paper the objective function is everywhere defined and can be approximated by a smoothing approach. The results in this paper exploit the structure of that special case to simplify the analysis and improve the efficiency of the method via smoothing.

A typical example of the class of objective functions of interest here is the composite nonsmooth function

$$f(x) = G(x, F(x)) \tag{1.2}$$

where $G : R^{n+m} \to R$ is Lipschitz continuous and $F : R^n \to R^m$ is continuously differentiable and the function value $F(x)$ is approximated by a Monte Carlo simulation. For example, the expected value version of the stochastic variational inequality problem [11,36]: Given the induced probability space $(\Xi \subset R^\ell, \mathcal{A}, \mathcal{P})$ and a convex set $X \subseteq R^n$, find $x^* \in X$ such that

$$(x - x^*)^T F(x^*) \geq 0, \quad \forall x \in X, \tag{1.3}$$

where $F(x) := E[\phi(\xi, x)]$, and $\phi : \Xi \times R^n \to R^m$ is continuously differentiable with respect to x for almost all $\xi \in \Xi$ and $\mathcal{A}$-measurable with respect to ξ. The stochastic variational inequality problem (1.3) reduces to the stochastic complementarity problem:

$$x \geq 0, \quad F(x) \geq 0, \quad x^T F(x) = 0, \tag{1.4}$$

*Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China (maxjchen@polyu.edu.hk). The work of this author was partially supported by Hong Kong Research Grant Council grant PolyU153000/15p.

†North Carolina State University, Department of Mathematics, Box 8205, Raleigh, NC 27695-8205, USA (Tim_Kelley@ncsu.edu). The work of this author was partially supported by the Consortium for Advanced Simulation of Light Water Reactors (www.casl.gov), and Simulation of Nuclear Reactors under U.S. Department of Energy Contract No. DE-AC05-00OR22725, Army Research Office grant #W911NF-16-1-0504 and National Science Foundation Grant DMS-1406349.

‡ School of Economics and Finance, Xi'an Jiaotong University, Xi'an, 710049, China (fengminxu@mail.xjtu.edu.cn). The work of this author was partially supported by the Chinese Natural Science Foundation 11571271, 71331001.

§Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China (zaikun.zhang@polyu.edu.hk). The work of this author was partially supported by the start-up grant 1-ZVHT from The Hong Kong Polytechnic University.

when $X = R^n_+ = \{x \in R^n \mid x \geq 0\}$, and the system of stochastic nonsmooth equations:

$$F(x) = 0$$

when $X = R^n$. In this case, the approximation is via Monte Carlo simulation

$$F(x) = E[\phi(\xi, x)] \approx F^N(x) := \frac{1}{N}\sum_{i=1}^{N} \phi(\xi^i, x),$$

where N is the sample size and ξ^i, $i = 1, \ldots, N$ are observations of $\xi \in \Xi$.

We can express problem (1.3) as a minimization problem [16]

$$\min_{x \in R^n} \|x - \mathrm{Proj}_X(x - F(x))\| \tag{1.5}$$

where Proj_X is the projection onto the set X. In this formulation the optimal function value is zero.

Another example of problem (1.1) is the ℓ_1 regularized minimization problem

$$\min_{x \in R^n} \|F(x)\|_1 + \lambda\|x\|_1, \tag{1.6}$$

where $F(x)$ is approximated by a Monte Carlo simulation.

We will exploit the composition structure by using a smoothing function for f. We will show that if f is replaced by the outcome of Monte Carlo simulation and one has full knowledge of the nonsmoothness, we can develop a smoothing direct search method with Monte Carlo simulation, which has global convergence to a Clarke stationary point of problem (1.1) with probability one (w.p.1.).

For example, if f has the form (1.2), we can define the smoothing function for f as

$$\hat{f}(x, \mu) = \hat{G}(x, F(x), \mu) \tag{1.7}$$

and, when $F(x)$ is replaced by the Monte Carlo outcome $F^N(x)$, we set the smoothing Monte Carlo simulation as

$$\tilde{f}(x, \mu, N) = \hat{G}(x, F^N(x), \mu). \tag{1.8}$$

For any fixed smoothing parameter $\mu > 0$, the function $\hat{f}$ is continuously differentiable with respect to x. The main contribution of this paper is to propose a smoothing direct search algorithm with Monte Carlo simulation for solving problem (1.1) and prove the convergence of the algorithm when the stencil size h and smoothing parameter μ go to zero with the rate $h/\mu \to 0$, and the sample size N goes to infinity with the rate $(h\sqrt{N})^{-1} \to 0$.

Convergence analysis of direct search algorithms for smooth optimization problems where function values can be computed exactly have been well studied in [14, 15, 18, 22, 35]. Nonsmooth problems have been considered in [1–3, 9, 22]. Algorithms for problems where the function evaluations require embedded Monte Carlo simulations have been considered for nonlinear equations [40, 41] and for optimization problems [9]. The new algorithm in this paper exploits the structure of the problem and properties of smoothing methods to allow for using coordinate basis as fixed stencil search directions, simplifying the approaches of [2, 9, 22] for nonsmooth problems while preserving the convergence results.

This paper is organized as follows. In section 2, we present a smoothing direct search algorithm for problem (1.1) where the function values $f(x)$ can be computed directly, and prove the convergence of the algorithm. In section 3, we extend the algorithm and convergence analysis to a smoothing direct search algorithm for (1.1) where the function values $f(x)$ cannot be computed directly, but are approximated by Monte Carlo simulation. In section 4, we present numerical experiments which include examples from statistical learning, and portfolio selection using test problems from the OR-Library [4] and real data from the Shanghai-Shenzhen stock market.

2. A smoothing direct search method. We begin by reviewing sampling direct search methods in the context of the smooth optimization problem. Let the set of search directions be an orthonormal basis $V = \{v_1, v_2, \ldots, v_n\}$. Let h be the stencil size along those search directions. A stencil centered at x with h is the set of points $\{x \pm hv_i\}_{i=1}^n \cup \{x\}$. Stencil failure in this specific instance means

$$f(x) \leq f(x \pm hv_i) \quad \text{for } i = 1, \ldots n. \tag{2.1}$$

More general stencils can be used [2, 22, 24, 25] but are not needed for the applications in this paper.

For simplicity, in this paper we will use

$$V \;=\; \{e_1, e_2, \ldots, e_n\}\,, \tag{2.2}$$

for each iteration. Here e_i is the i-th coordinate vector. The algorithms and convergence analysis can be extended to an orthononormal basis.

It is easy to show [15, 21, 22] that if f is Lipschitz continuously differentiable, then (2.1) implies that there is $\Gamma(x) > 0$ such that

$$\|\nabla f(x)\| \leq h\Gamma(x), \tag{2.3}$$

and that for any bounded set $\Omega \subset R^n$ there is a positive constant Γ_Ω such that

$$\sup_{x\in\Omega} \Gamma(x) \leq \Gamma_\Omega. \tag{2.4}$$

Sampling methods evaluate the objective function at the points of the stencil. If the current point is the best (stencil failure at the current point), then the stencil size is reduced. If the current point is not the best on the stencil, then the new best point becomes the current point. Algorithm `direct_search` is a version of the method for minimizing a continuously differentiable objective function f within a convex set X.

Algorithm `direct_search` (x, f, h)

for forever **do**
  $f_{base} = f(x)$
  $f_{min} = \min\{f(y) \,|\, y = x \pm hv,\ v \in V \text{ and } y \in X\}$
  $\hat{y} \in \{y \,|\, f(y) = f_{min},\ y = x \pm hv,\ v \in V \text{ and } y \in X\}$
  if $f_{min} \geq f_{base}$ **then**
    $h \leftarrow h/2$
  else
    $x \leftarrow \hat{y}$
  end if
end for

In Algorithm `direct_search`, we choose the initial point $x \in X$.

The convergence proof of Algorithm `direct_search` is based on the stencil directions such that if stencil failure happens at the current point, then some type of approximate necessary condition holds. This idea can be made very general with different stencils and different smoothness requirements on the objective function f [1–3, 9, 22].

We consider the following first-order stationarity measure

$$\chi(x) \;=\; \max_{x+d\in X, \|d\|\leq 1} [-d^T \nabla f(x)]. \tag{2.5}$$

It is easy to check that, if $x \in X$ is a local minimizer of (1.1), then $\chi(x) = 0$.

In our convergence analysis (Proposition 2.1, Theorem 2.4, Theorem 3.1), we assume that

$$X \;=\; \{x \in R^n : \ell \leq x \leq u\}. \tag{2.6}$$

In (2.6), $l \in \{R \cup -\infty\}^n, u \in \{R \cup \infty\}^n$ and the inequalities are componentwise.

PROPOSITION 2.1. *Assume that* (2.6) *holds and that f is Lipschitz continuously differentiable and has bounded level sets. Let $\{x_k\}$ be the sequence generated from Algorithm* `direct_search`. *Then*

$$\liminf_{k\to\infty} \chi(x_k) = 0. \tag{2.7}$$

Moreover, stencil failure happens at infinitely many iterates, and for each limit point x of the stencil failure iterates, it holds that

$$\chi(x) \;=\; 0. \tag{2.8}$$

Proof. For the initial iterate $x_0 \in X$ and the initial stencil size h_0, set

$$\Lambda_0 \;=\; \{x \in X : f(x) \le f(x_0)\}, \tag{2.9}$$

and

$$X_t \;=\; \Lambda_0 \cap \left\{ x_0 + \sum_{i=1}^{n} \frac{j_i h_0}{2^t} e_i : j_i = 0, \pm 1, \pm 2, \cdots \right\}, \quad t = 0, 1, 2, \ldots. \tag{2.10}$$

Then Λ_0 is bounded, and consequently X_t is a finite set. Stencil failure occurs at $\tilde{x}_t$, the last iterate contained in X_t and the size of the stencil at that iteration is $\tilde{h}_t = h_0/2^t$. For each $t \ge 1$, define

$$I_t^+ \;=\; \{i : \tilde{x}_t + \tilde{h}_t e_i \in X, 1 \le i \le n\}, \tag{2.11}$$

$$I_t^- \;=\; \{i : \tilde{x}_t - \tilde{h}_t e_i \in X, 1 \le i \le n\}, \tag{2.12}$$

and denote $g_t = \nabla f(\tilde{x}_t)$. Let L be the Lipschitz constant of ∇f. Then by the definition of stencil failure and Taylor's theorem,

$$0 \;\le\; f(\tilde{x}_t + \tilde{h}_t e_i) - f(\tilde{x}_t) \;\le\; \tilde{h}_t e_i^T g_t + \frac{L}{2}\tilde{h}_t^2 \quad \text{for all} \quad i \in I_t^+, \tag{2.13}$$

$$0 \;\le\; f(\tilde{x}_t - \tilde{h}_t e_i) - f(\tilde{x}_t) \;\le\; -\tilde{h}_t e_i^T g_t + \frac{L}{2}\tilde{h}_t^2 \quad \text{for all} \quad i \in I_t^-, \tag{2.14}$$

and consequently,

$$e_i^T g_t \;\ge\; -\frac{L\tilde{h}_t}{2} \quad \text{for all} \quad i \in I_t^+, \tag{2.15}$$

$$e_i^T g_t \;\le\; \frac{L\tilde{h}_t}{2} \quad \text{for all} \quad i \in I_t^-. \tag{2.16}$$

By the assumption, there exists a positive constant Υ such that

$$\|\nabla f(x)\| \;\le\; \Upsilon \quad \text{for all} \quad x \in \Lambda_0. \tag{2.17}$$

In specific, $\|g_t\| \le \Upsilon$. Therefore, for each d such that $\tilde{x}_t + d \in X$ and $\|d\| \le 1$, it holds that

$$\begin{aligned}
-d^T g_t \;&=\; -\sum_{i=1}^{n} d_i (g_t)_i \\
&=\; -\sum_{i \in I_t^+ \setminus I_t^-} d_i (g_t)_i - \sum_{i \in I_t^- \setminus I_t^+} d_i (g_t)_i - \sum_{i \in I_t^+ \cap I_t^-} d_i (g_t)_i - \sum_{i \notin I_t^+ \cup I_t^-} d_i (g_t)_i \\
&\le\; n \max\left\{ \frac{L\tilde{h}_t}{2}, \Upsilon\tilde{h}_t \right\} + n\frac{L\tilde{h}_t}{2} + n\Upsilon\tilde{h}_t \\
&\le\; 3n \max\left\{ \frac{L\tilde{h}_t}{2}, \Upsilon\tilde{h}_t \right\},
\end{aligned} \tag{2.18}$$

where the first inequality uses the fact that for each d with $\tilde{x}_t + d \in X$, $i \notin I_t^-$ implies $d_i > -\tilde{h}_t$ and $i \notin I_t^+$ implies $d_i < \tilde{h}_t$.

Hence, we obtain

$$\chi(\tilde{x}_t) \;\le\; 3n \max\left\{ \frac{L\tilde{h}_t}{2}, \Upsilon\tilde{h}_t \right\} \;\to\; 0 \quad \text{when} \quad t \to \infty. \tag{2.19}$$

Since $\{\tilde{x}_t\}$ is a subsequence of $\{x_k\}$, we conclude that

$$\liminf_{k\to\infty} \chi(x_k) \;=\; 0. \tag{2.20}$$

If x is an accumulation point of the stencil failure iterates $\{\tilde{x}_t\}$, then continuity of χ implies that $\chi(x) = 0$. □

2.1. Nonsmooth f. In this subsection, we consider problem (1.1) where f is nonsmooth and F can be computed directly.

We will use smoothing methods which approximate f by a parameterized family of smoothing functions $\hat{f}(\cdot, \mu)$ where $\mu > 0$ is the smoothing parameter.

We formally give a definition of smoothing functions used in this paper.

DEFINITION 2.2. *[8] Let $f : R^n \to R$ be a locally Lipschitz continuous function. We call $\hat{f} : R^n \times (0, \infty) \to R$ a smoothing function of f, if $\hat{f}(\cdot, \mu)$ is continuously differentiable and $\nabla \hat{f}(\cdot, \mu)$ is Lipschitz continuous in R^n for any fixed $\mu \in (0, \infty)$, and*

$$\lim_{x \to \hat{x}, \mu \downarrow 0} \hat{f}(x, \mu) = f(\hat{x}).$$

Throughout this subsection we let $\nabla \hat{f}$ denote the gradient $\hat{f}$ with respect to x.

ASSUMPTION 2.1.

(i) There are constants $c_1, c_2 \ge 0$ such that for any $x \in R^n$, $\mu \in (0, 1]$,

$$|f(x) - \hat{f}(x, \mu)| \le \mu(c_1 + c_2|f(x)|). \tag{2.21}$$

(ii) $\hat{f}$ satisfies the gradient consistency condition,

$$\partial f(x) = \mathrm{con}\{v \,|\, \nabla \hat{f}(x_k, \mu_k) \to v, \text{ for } x_k \to x, \ \mu_k \downarrow 0\}, \tag{2.22}$$

where "con" denotes the convex hull and $\partial f(x)$ is the Clarke subgradient at x.

(iii) There are $\Upsilon > 0$, $\Gamma > 0$ and $\mu_- > 0$ such that $\|\nabla \hat{f}(x, \mu)\| \le \Upsilon$ and

$$\|\nabla \hat{f}(x, \mu) - \nabla \hat{f}(y, \mu)\| \le \frac{\Gamma}{\mu} \|x - y\| \tag{2.23}$$

uniformly in x, $y \in \Omega$, and $\mu \in (0, \mu_-)$, where Ω is a convex and compact set.

In section 4, we use examples to illustrate the definition of smoothing functions and Assumption 2.1.

The proof that the stencil fails infinitely often for smooth f depends on the fact that f has bounded level sets. Assumption 2.1 implies that $\hat{f}(x, \mu)$ has bounded level sets in both variables if f does. Lemma 2.3 states this precisely.

LEMMA 2.3. *Assume that f has bounded level sets and that part (i) of Assumption 2.1 holds. Let $\hat{\mu} = 1/(1 + c_2)$. Then for any $M \in R$, the set*

$$\Omega_M = \bigcup_{\mu \le \hat{\mu}} \{x \in X : \hat{f}(x, \mu) \le M\}$$

is bounded.

Proof. First note that $\mu \leq \hat{\mu}$ implies

$$1 - c_2\mu \geq 1 - c_2\hat{\mu} = \hat{\mu} > 0.$$

Now let $x \in \Omega_M$. Then there is $\mu \leq \hat{\mu}$ such that $\hat{f}(x,\mu) \leq M$. According to Part (i) of Assumption 2.1,

$$f(x) - \hat{f}(x,\mu) \leq c_1\mu + c_2\mu|f(x)|. \tag{2.24}$$

Hence either $f(x) \leq 0$, or

$$f(x) \leq \frac{\hat{f}(x,\mu) + c_1\mu}{1 - c_2\mu} \leq \frac{M + c_1\hat{\mu}}{1 - c_2\hat{\mu}}.$$

Therefore,

$$\Omega_M \ \subset \ \left\{ x \in X : f(x) \leq \max\left\{0, \frac{M + c_1\hat{\mu}}{1 - c_2\hat{\mu}}\right\}\right\}, \tag{2.25}$$

which is a bounded set. □

In the case where f is known exactly and there is no embedded Monte Carlo simulation, we propose a smoothing direct search algorithm, Algorithm `smoothing_search` that decreases μ and h simultaneously, but in a way that ensures $h = o(\mu)$ as $\mu \to 0$, which will be important in the convergence analysis.

Algorithm `smoothing_search` $(x, \hat{f}, h, \mu, \tau)$

for forever **do**
  $\hat{f}_{base} = \hat{f}(x,\mu)$
  $\hat{f}_{min} = \min\{\hat{f}(y,\mu) \,|\, y = x \pm hv,\ v \in V \text{ and } y \in X\}$
  $\hat{y} \in \{y \,|\, \hat{f}(y,\mu) = \hat{f}_{min},\ y = x \pm hv,\ v \in V \text{ and } y \in X\}$
  if $\hat{f}_{min} \geq \hat{f}_{base}$ **then**
    $h \leftarrow h/2; \quad \mu \leftarrow \mu/2^{\tau}$
  else
    $x \leftarrow \hat{y}$
  end if
end for

In Algorithm `smoothing_search`, $\tau \in (0,1)$ is an input parameter.

As an extension of (2.5), we define

$$\tilde{\chi}(x) \ = \ \min_{v \in \partial f(x)} \ \max_{x+d \in X, \|d\| \leq 1} -d^T v \tag{2.26}$$

to measure the first-order sationarity of x with respect to problem (1.1) when f is locally Lipschitz continuous but not necessarily differentiable, where $\partial f(x)$ is the Clarke subdifferential of f at x [13]. If f is smooth, then $\tilde{\chi}(\cdot)$ is the same as $\chi(\cdot)$. Moreover, if x is a local minimizer of problem (1.1), then there exists a $v \in \partial f(x)$ such that

$$\max_{x+d \in X, \|d\| \leq 1} -d^T v = 0,$$

that is, $\tilde{\chi}(x) = 0$.

The convergence result follows the same argument as in the proof of Proposition 2.1 and Assumption 2.1 on the smoothing function of f.

THEOREM 2.4. *Assume that Assumption 2.1 holds with $0 < \mu \leq 1/(1 + c_2)$ for the initial μ and that f has bounded level sets. Let $\{x_k, \mu_k\}$ be the iterates generated by Algorithm* `smoothing_search`*, and*

$$\chi_k(x) \ = \ \max_{x+d \in X, \|d\| \leq 1} [-d^T \nabla \hat{f}(x, \mu_k)]. \tag{2.27}$$

Then

$$\liminf_{k\to\infty} \chi_k(x_k) \;=\; 0. \tag{2.28}$$

Moreover, stencil failure happens at infinitely many iterates, and for each limit point x of the stencil failure iterates, it holds that

$$\tilde{\chi}(x) \;=\; 0. \tag{2.29}$$

Proof. For the initial iterate x_0, the initial smoothing parameter μ_0, and the initial stencil size h_0, set

$$\Lambda_0 \;=\; \bigcup_{\mu\le\mu_0} \{x \in X : f(x,\mu) \le f(x_0,\mu_0)\}, \tag{2.30}$$

and

$$X_t \;=\; \Lambda_0 \cap \left\{ x_0 + \sum_{i=1}^{n} \frac{j_i h_0}{2^t} e_i : j_i = 0, \pm 1, \pm 2, \cdots \right\}, \quad t = 0, 1, 2, \ldots. \tag{2.31}$$

Lemma 2.3 ensures that Λ_0 is bounded. X_t is a finite set that contains at least one iterate. As in the proof of Proposition 2.1, we denote the last iterate in X_t, where stencil failur occurs by $\tilde{x}_t$, the corresponding stecil size by $\tilde{h}_t = h_0/2^t$, and the corresponding smoothing parameter by $\tilde{\mu}_t = \tilde{h}_t^\tau$. For each $t \ge 1$, define I_t^+ and I_t^- in the same way as in the proof of Proposition 2.1, and denote $\hat{g}_t = \nabla \hat{f}(\tilde{x}_t, \tilde{\mu}_t)$. According to the definition of stencil failure and Taylor expansion, noticing part (iii) of Assumption 2.1, we have

$$0 \;\le\; \hat{f}(\tilde{x}_t + \tilde{h}_t e_i, \tilde{\mu}_t) - \hat{f}(\tilde{x}_t, \tilde{\mu}_t) \;\le\; \tilde{h}_t e_i^T \hat{g}_t + \frac{\Gamma}{2\tilde{\mu}_t}\tilde{h}_t^2 \quad \text{for all} \quad i \in I_t^+, \tag{2.32}$$

$$0 \;\le\; \hat{f}(\tilde{x}_t - \tilde{h}_t e_i, \tilde{\mu}_t) - \hat{f}(\tilde{x}_t, \tilde{\mu}_t) \;\le\; -\tilde{h}_t e_i^T \hat{g}_t + \frac{\Gamma}{2\tilde{\mu}_t}\tilde{h}_t^2 \quad \text{for all} \quad i \in I_t^-, \tag{2.33}$$

and consequently,

$$e_i^T \hat{g}_t \;\ge\; -\frac{\Gamma\tilde{h}_t}{2\tilde{\mu}_t} \quad \text{for all} \quad i \in I_t^+, \tag{2.34}$$

$$e_i^T \hat{g}_t \;\le\; \frac{\Gamma\tilde{h}_t}{2\tilde{\mu}_t} \quad \text{for all} \quad i \in I_t^-. \tag{2.35}$$

By Assumption 2.1, there exists a positive constant Υ such that $\|\hat{g}_t\| \le \Upsilon$. Using similar argument to those for (2.18) and (2.19), we have

$$\max_{\tilde{x}_t + d \in X, \|d\| \le 1} -d^T \hat{g}_t \;\le\; 3n \max\left\{ \frac{\Gamma\tilde{h}_t}{2\tilde{\mu}_t},\; \Upsilon\tilde{h}_t \right\}. \tag{2.36}$$

Noticing the fact that $\tilde{h}_t/\tilde{\mu}_t \to 0$, we have

$$\max_{\tilde{x}_t + d \in X, \|d\| \le 1} -d^T \hat{g}_t \;\to\; 0 \quad \text{when} \quad t \to \infty, \tag{2.37}$$

which implies (2.28).

Let x be a limit point of $\{\tilde{x}_t\}$, and $\{\tilde{x}_{t_i}\}$ be a subsequence that converges to x. Since $\{\hat{g}_t\}$ is bounded, we may as well suppose that $\{\hat{g}_{t_i}\}$ converges to a point v (if not, replace $\{t_i\}$ by an appropriately chosen subsequence). Let

$$d^* \;\in\; \operatorname*{argmax}_{x + d \in X, \|d\| \le 1} -d^T v, \tag{2.38}$$

and

$$y_i \;=\; \frac{x - \tilde{x}_{t_i} + d^*}{\max\{\|x - \tilde{x}_{t_i} + d^*\|,\, 1\}}. \tag{2.39}$$

Then $\|y_i\| \le 1$, and $\tilde{x}_{t_i} + y_i \in X$ due to the convexity of X ($\tilde{x}_{t_i} + y_i$ lies on the line segment between $\tilde{x}_{t_i}$ and $x + d^*$). Hence

$$\begin{aligned} 0 \;\le\; \max_{x+d\in X, \|d\|\le 1} -d^T v \;&=\; -\,(d^*)^T v \\ &= \lim_{i\to\infty} -y_i^T \hat{g}_{t_i} \\ &\le \lim_{i\to\infty} \max_{\tilde{x}_{t_i}+d\in X, \|d\|\le 1} -d^T \hat{g}_{t_i} \\ &= 0. \end{aligned} \tag{2.40}$$

Notice that $v \in \partial f(x)$ according to the gradient consistency of $\hat{f}$. By the definition (2.26) of $\tilde{\chi}(\cdot)$, we have

$$0 \;\le\; \tilde{\chi}(x) \;\le\; \max_{x+d\in X, \|d\|\le 1} -d^T v \;=\; 0, \tag{2.41}$$

which completes the proof. □

3. Smoothing direct search method with Monte Carlo simulations. In this section, we propose a smoothing direct search method based on Monte Carlo simulations to problem (1.1), where f is not necessarily differentiable and cannot be computed directly. Deterministic direct search methods for nonsmooth optimization problems have been studied in [1–3, 9, 18, 22, 31].

Algorithm `mc_smoothing_search` for the embedded Monte Carlo case is a simple extension of Algorithm `smoothing_search`.

Algorithm `mc_smoothing_search` $(x, \tilde{f}, h, \mu, N, \tau, \gamma)$
  for forever **do**
    $\tilde{f}_{base} = \tilde{f}(x, \mu, N)$
    $\tilde{f}_{min} = \min\{\tilde{f}(y, \mu, N) \mid y = x \pm hv, v \in V \text{ and } y \in X\}$
    $\hat{y} \in \{y \,|\, \tilde{f}(y, \mu, N) = \tilde{f}_{min},\ y = x \pm hv,\ v \in V \text{ and } y \in X\}$
    if $\tilde{f}_{min} \ge \tilde{f}_{base}$ **then**
      $h \leftarrow h/2; \quad \mu \leftarrow \mu/2^\tau; \quad N \leftarrow 4^\gamma N$
    else
      $x \leftarrow \hat{y}$
    end if
  end for

In Algorithm `mc_smoothing_search`, $\tau \in (0, 1)$ and $\gamma > 1$ are input parameters. The objective function f is evaluated through $\tilde{f}(x, \mu, N)$, the Monte Carlo simulation of $\hat{f}(x, \mu)$ with sample size N, where $\hat{f}(x, \mu)$ is a smoothing function of f that is defined in Definition 2.2 and satisfies Assumption 2.1.

As in Proposition 2.1 and Theorem 2.4, we need level boundedness to establish the convergence of the algorithm. Here impose the level boundedness by Assumption 3.1. Let x_0 denote the initial point of Algorithm `mc_smoothing_search`, μ_0 the initial smoothing parameter, and N_0 the initial sample size.

ASSUMPTION 3.1. *There are $\Gamma > 0$ and a bounded convex set D such that set*

$$\Omega \;=\; \bigcup_{\mu\le\mu_0, N\ge N_0} \Big\{x \in X : \tilde{f}(x, \mu, N) \le \Gamma\Big\} \subseteq D \tag{3.1}$$

for each realization of Algorithm `mc_smoothing_search`.

It is possible to weaken Assumption 3.1 by requiring that the boundedness holds with probability 1, and our results will still hold with minor modifications of the argument.

The following is an assumption on the effectiveness of $\tilde{f}(\cdot,\cdot,N)$ as an approximation of $\hat{f}(\cdot,\cdot)$.

ASSUMPTION 3.2. *There exist constants $\delta \in (0,1)$ and $c_f > 0$ such that*

$$\text{Prob}\left(\sup_{x\in\Omega} |\hat{f}(x,\mu) - \tilde{f}(x,\mu,N)| > \frac{c_f}{\sqrt{N}}\right) < \delta \tag{3.2}$$

for each $\mu \in (0,\mu_0)$.

Consider the composite nonsmooth function in the form (1.2). Suppose there exist constants $\delta \in (0,1)$ and $c_F > 0$ such that

$$\text{Prob}\left(\sup_{x\in\Omega} \|F(x) - F^N(x)\| > \frac{c_F}{\sqrt{N}}\right) < \delta, \tag{3.3}$$

where $F^N(x)$ is the Monte Carlo outcome of $F(x)$ with the sample size N. If G is Lipschitz continuous at $F(x)$ and $\hat{G}$ satisfies Assumption 2.1, then there is a constant L_F which is independent of N and μ, such that

$$\|\hat{G}(x,F(x),\mu) - \hat{G}(x,F^N(x),\mu)\| \le L_F \|F(x) - F^N(x)\|.$$

This, together with (3.3), implies that Assumption 3.2 holds with $c_f = L_F c_F$.

THEOREM 3.1. *Suppose that the Monte Carlo simulations in Algorithm* `mc_smoothing_search` *are mutually independent for different N. Assume that* (2.6) *and Assumptions 2.1, 3.1, 3.2 hold. Let $\{x_k, \mu_k, N_k\}$ be the sequence generated by Algorithm* `mc_smoothing_search`*. Then*

$$\text{Prob}\left(\liminf_{k\to\infty} \chi_k(x_k) = 0\right) \;=\; 1, \tag{3.4}$$

where χ_k is defined in (2.27)*, and*

$$\text{Prob}(\{x_k\} \textit{ has an accumulation point } x \textit{ such that } \tilde{\chi}(x) = 0) \;=\; 1. \tag{3.5}$$

Proof. Let Ω be the set defined in Assumption 3.1, and

$$X_t \;=\; \Omega \cap \left\{ x_0 + \sum_{i=1}^{n} \frac{j_i h_0}{2^t} e_i : j_i = 0, \pm 1, \pm 2, \cdots \right\}, \quad t = 0,1,2,\ldots. \tag{3.6}$$

Then Ω is bounded. As in the proof of Proposition 2.1, X_t is a finite set that contains at least one iterate. As before we denote the last iterate in X_t, where stencil failure occurs by $\tilde{x}_t$, the corresponding stencil size by $\tilde{h}_t = h_0/2^t$, and the corresponding smoothing parameter by $\tilde{\mu}_t = \mu_0/2^{t\tau}$. The sample size at this point in the iteration is $\tilde{N}_t = 4^{t\gamma} N_0$. For each integer $t \ge 1$, define I_t^+ and I_t^- in the same way as in the proof of Proposition 2.1, denote $\hat{g}_t = \nabla \hat{f}(\tilde{x}_t, \tilde{\mu}_t)$, and consider the event

$$E_t \;=\; \left\{ \sup_{x\in\Omega} |\hat{f}(x,\mu) - \tilde{f}(x,\mu,\tilde{N}_t)| \le \frac{c_f}{\sqrt{\tilde{N}_t}} \right\}. \tag{3.7}$$

By assumption, $\{E_t\}$ are mutually independent, and

$$\text{Prob}(E_t) \;\ge\; 1 - \delta \;>\; 0$$

for each t. Therefore,

$$\text{Prob}(E_t \text{ happens for infinitely many } t) \;=\; 1. \tag{3.8}$$

When E_t happens, similar to (2.34) and (2.35), from

$$0 \;\le\; \hat{f}(\tilde{x}_t + \tilde{h}_t e_i, \tilde{\mu}_t) - \hat{f}(\tilde{x}_t, \tilde{\mu}_t) \;\le\; \tilde{h}_t e_i^T \hat{g}_t + \frac{\Gamma}{2\tilde{\mu}_t}\tilde{h}_t^2 + \frac{2c_f}{\tilde{h}_t\sqrt{\tilde{N}_t}} \quad \text{for all} \quad i \in I_t^+, \tag{3.9}$$

$$0 \;\le\; \hat{f}(\tilde{x}_t - \tilde{h}_t e_i, \tilde{\mu}_t) - \hat{f}(\tilde{x}_t, \tilde{\mu}_t) \;\le\; -\tilde{h}_t e_i^T \hat{g}_t + \frac{\Gamma}{2\tilde{\mu}_t}\tilde{h}_t^2 + \frac{2c_f}{\tilde{h}_t\sqrt{\tilde{N}_t}} \quad \text{for all} \quad i \in I_t^-, \tag{3.10}$$

we have

$$e_i^T \hat{g}_t \;\ge\; -\frac{\Gamma\tilde{h}_t}{2\tilde{\mu}_t} - \frac{2c_f}{\tilde{h}_t\sqrt{\tilde{N}_t}} \quad \text{for all} \quad i \in I_t^+, \tag{3.11}$$

$$e_i^T \hat{g}_t \;\le\; \frac{\Gamma\tilde{h}_t}{2\tilde{\mu}_t} + \frac{2c_f}{\tilde{h}_t\sqrt{\tilde{N}_t}} \quad \text{for all} \quad i \in I_t^-. \tag{3.12}$$

By Assumption 2.1, there exists a positive constant Υ such that $\|\hat{g}_k\| \le \Upsilon$. Using similar argument as (2.18) and (2.19), we obtain from (3.11) and (3.12) that

$$\max_{\tilde{x}_t + d \in X, \|d\| \le 1} -d^T \hat{g}_t \;\le\; 3n \max\left\{ \frac{\Gamma\tilde{h}_t}{2\tilde{\mu}_t} + \frac{2c_f}{\tilde{h}_t\sqrt{\tilde{N}_t}},\; \Upsilon\tilde{h}_t \right\}. \tag{3.13}$$

Hence, by (3.8) and the fact that $\tilde{h}_t \to 0$, $\tilde{h}_t/\tilde{\mu}_t \to 0$, and $\tilde{h}_t\sqrt{\tilde{N}_t} \to \infty$, we have

$$\text{Prob}(\{\chi_k(x_k)\} \text{ has a subsequence that converges to zero}) \;=\; 1, \tag{3.14}$$

which implies (3.4).

When $\liminf_{k\to\infty} \chi_k(x_k) = 0$, let $\{k_i\}$ be the index sequence such that $\chi_{k_i}(x_{k_i}) \to 0$. Since $\{x_{k_i}\}$ is bounded (guaranteed the boundedness of Ω), it has an accumulation point x. By the same argument that leads to the second part of Theorem 2.4, we have that $\tilde{\chi}(x) = 0$. Thus (3.5) holds. □

4. Numerical experiments. In this section, we test Algorithm `mc_smoothing_search` on two problems: a stochastic optimization problem arising from censored regression and a two-stage optimization problem arising from portfolio management. The problems in § 4.1 and 4.2.1 are derived from applications, but use synthetic data to enable us to control the sample size.

4.1. Censored regression. We consider the following regularized censored regression problem [5, 26, 27, 38, 39]

$$\begin{aligned} &\min_{x\in R^n} \; f(x) \\ &\text{s.t.} \; -e \le x \le e, \end{aligned} \tag{4.1}$$

where

$$f(x) = E_{c,y}[(\max(c^T x, 0) - y)^2] + \lambda \sum_{i=1}^{n} \log(1 + |x_i|). \tag{4.2}$$

Here the random variable pair (c, y) represents a data set of interest ($c \in R^n$, $y \in R$), $\lambda > 0$ is a regularization parameter and e is the vector with all its entries being one. The regularization term

$$\lambda \sum_{i=1}^{n} \log(1 + |x_i|)$$

is in the objective function to enforce sparsity.

We assume that $c \sim N(0, I)$, and $y = \max(c^T x^* + \epsilon, 0)$ for some underlying ground truth feature x^* and unobservable noise $\epsilon \sim N(0, \sigma^2)$. Moreover, we assume x^* is sparse, that is, x^* has few nonzero entries. Using

the concave regularized model (4.1), we want to recover a sparse feature x to approximate x^* as accurate as possible given that $x^* \in \{x \mid \|x\|_\infty \leq 1\} = [-e, e]$.

The functions $(\max(c^T x, 0) - y)^2$ and $\log(1 + |x_i|)$ are not differentiable. See Appendix A for smoothing functions for $\max(., 0)$ and $|\cdot|$. Using the smoothing functions, we can define a smoothing function $\hat{f}_{c,y}(x, \mu)$ for $(\max(c^T x, 0) - y)^2$ which satisfies Assumption 2.1. From the convexity of $(\max(c^T x, 0) - y)^2$, the Clarke subdifferential and the expectation can be exchanged, that is,

$$\partial E_{c,y}[(\max(c^T x, 0) - y)^2] = E_{c,y}[\partial(\max(c^T x, 0) - y)^2]$$

(see [13]). Moreover, $E_{c,y}[\hat{f}_{c,y}(x, \mu)]$ is a smoothing function for $E_{c,y}[(\max(c^T x, 0) - y)^2]$, and satisfies Assumption 2.1 (see [7]). Problem (4.1) is a constrained nonsmooth nonconvex optimization problem where the objective function values cannot be computed directly.

It is easy to see that Assumption 3.1 holds, since X is bounded. Moreover, the objective function is level bounded. Since $(\max(c^T x, 0) - y)^2 \geq 0$ for any $x \in R^n$, we have that for any $N > 0$,

$$\{x \mid \frac{1}{N} \sum_{i=1}^{N} [(\max(c_i^T x, 0) - y_i)^2] + \lambda \sum_{i=1}^{n} \log(1 + |x_i|) \leq M\} \subseteq \{x \mid \lambda \sum_{i=1}^{n} \log(1 + |x_i|) \leq M\},$$

where $M > 0$ is a large positive number. Hence Assumption 3.1 holds even if X is not bounded, because of the boundedness of the set $\{x \mid \lambda \sum_{i=1}^{n} \log(1 + |x_i|) \leq M\}$.

In practice the data in these problems are limited. To mimic the finite size of the data we will pose an approximation to problem (4.1) that replaces the expectation with the sample average of a finite, but large, data set. We will manage the sampling in the algorithm by randomly sampling from that data set. To this end, we randomly generate a true feature $x^* \in R^{20}$ whose 5 nonzero entries are from uniform distribution on $[-1, 1]$. Independently, we generate samples c_i from $c \sim N(0, I)$ and ϵ_i from $\epsilon \sim N(0, 0.01)$ with a sample size 10^7. Let $y_i = \max(c_i^T x^* + \epsilon_i, 0)$. The new problem is an approximation to (4.1) with a finite data set. We have

$$f(x) = \frac{1}{10^7} \sum_{i=1}^{10^7} [(\max(c_i^T x, 0) - y_i)^2] + \lambda \sum_{i=1}^{n} \log(1 + |x_i|). \tag{4.3}$$

We used the regularization parameter $\lambda = 10^{-2}$. This is large enough to capture the sparsity exactly and small enough to allow us to observe several iterations before the iteration stagnates.

We configure the optimization as follows:

- The algorithmic parameters are $c = 2$, $\gamma = 1.5$, and $\tau = 0.5$.
- We begin with $h = 0.5$ and terminate when $h \leq 10^{-3}$.
- $N = 100$ and $\mu = 0.1$ at the beginning of the iteration.

Given N, for each evaluation of f, we independently and randomly choose vectors $(c_i, y_i), i = 1, \ldots, N$ from the data set $(c_i, y_i), i = 1, \ldots, 10^7$ generated above. Then we compute smoothing approximation $\tilde{f}(x, \mu, N)$ of the following function

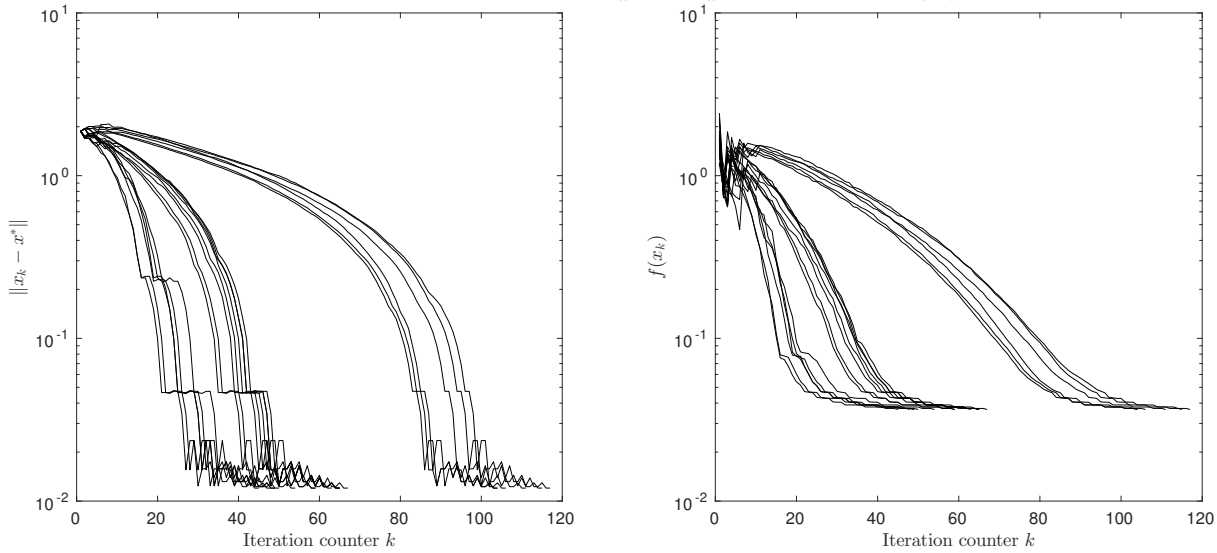
$$f(x, N) = \frac{1}{N} \sum_{i=1}^{N} [(\max(c_i^T x, 0) - y_i)^2] + \lambda \sum_{i=1}^{n} \log(1 + |x_i|), \tag{4.4}$$

by using smoothing functions of $\max(., 0)$ and $|\cdot|$ in Appendix A.

For the initial point $x^0 = (0, \ldots, 0)^T$ we performed 20 runs of Algorithm `mc_smoothing_search`. In Figure 4.1, we show histories of the difference $\|x^k - x^*\|$ and the value $f_{base}(x^k)$, where one search through the stencil is regarded one iteration.

Figure 4.1 illustrates several properties of the algorithm and the problem. At the end of the iteration, all of the iteration histories are very similar. The theory would lead one to expect similar histories if N is large. On the other hand, the initial value of N is large enough to cause considerable variation in f early in the

FIG. 4.1. *Histories of the difference* $\|x^k - x^*\|$ *and the value* $f_{base}(x^k)$



iteration. This variation accounts for the differences in the histories. Finally, the iteration stagnates in the terminal phase when the differences from the iterates and x^* are roughly at the level of the regularization parameter. The reason for this is that the regularization term would dominate the error term when x is near x^*. While a smaller regularization would defer the stagnation, it would make it harder to capture the sparsity. Our choice of $\lambda = 10^{-2}$ captures the sparsity exactly. At each final iterate x^k, we have $x_i^k = x_i^* = 0$ for all $i \in S^c$, where $S = \{i \,|\, x_i^* \neq 0, i = 1, \ldots, n\}$, the support set of x^*.

4.2. Portfolio management. Consider ν assets. Let $u \in R^\nu$ denote the random returns of them, and

$$r = E[u], \qquad C = E[(u - r)(u - r)^T]. \tag{4.5}$$

Here r is the vector of expected returns of the different assets, and C is the covariance matrix of the return on the assets in the portfolio. When r and C are known, as discussed in [34], the Markowitz mean-variance model [28, 29] for portfolio selection can be formulated as

$$\begin{aligned} \min \ & \frac{1}{2} w^T C w - \eta r^T w \\ \text{s.t.} \ & e^T w \ = \ 1 \\ & a \ \leq \ w \ \leq b, \end{aligned} \tag{4.6}$$

where w denotes the weights of the assets in the portfolio, $a \in R^\nu$ and $b \in R^\nu$ $(a \leq b)$ are lower and upper bounds enforced on w, and η is a nonnegative parameter (called the risk aversion factor) to balance the conflicting aspects of minimizing the risk measured by $w^T C w$ and maximizing the expected return measured by $r^T w$. The Markowitz mean-variance model [28, 29] was first proposed and solved when the total return is known. The model captures the essence of two conflicting aspects in portfolio management; namely, the risk and the return.

The use of mean variance analysis in portfolio selection normally requires the knowledge of means, variances, and covariances of returns of all securities under consideration. However, in general, these data are not known exactly. Treating their estimates as if they were the exact parameters can lead to suboptimal portfolio choices.

The experiments reported in [17, 20, 23] show that, influenced by the sampling error, portfolios selected with the mean-variance model by Markowitz are not as efficient as an equally weighted portfolio. Other results [12, 30] show that the mean-variance model tends to magnify the errors associated with the estimates.

In this section, we consider an optimal parameter selection model based on the Markowitz mean-variance model to find optimal parameters for portfolio selection.

For simplicity, here we only consider the case where C is positive definite and the feasible set $\{w \,|\, e^T w = 1, a \le w \le b\}$ is nonempty. Given a, b, and η, problem (4.6) has a unique solution w. In other words, w is uniquely defined by a, b, and η, the values of which will determine the quality of the portfolio selected by solving problem (4.6). A common measure for the quality is the Sharpe ratio [37]

$$\mathrm{SR} = \frac{r^T w}{\sqrt{w^T C w}}.$$

The Sharpe ratio characterizes how well the return of an asset compensates the investor for the risk taken. In general, a strategy is better than others if its Sharpe ratio is higher.

In practice, a, b, and η are usually set by investors empirically according to their preferences. We consider selecting them by solving the two-stage optimization problem

$$\begin{aligned} \max_{(a,b,\eta)\in\Omega} \ & \frac{r^T w(a,b,\eta)}{\sqrt{w(a,b,\eta)^T C w(a,b,\eta)}} \\ \text{where } w(a,b,\eta) = \operatorname{argmin} \ & \frac{1}{2} w^T C w - \eta r^T w \\ \text{s.t.} \ & e^T w \ = \ 1 \\ & a \ \le \ w \ \le b, \end{aligned} \tag{4.7}$$

where the feasible set

$$\Omega = [\underline{a}, \overline{a}] \times [\underline{b}, \overline{b}] \times [\underline{\eta}, \overline{\eta}] \quad \text{with} \quad \underline{\eta} < \overline{\eta}$$

is given. The number of variables of the first level problem is

$$\#\{i | \underline{a}_i < \overline{a}_i, i = 1, \ldots, \nu\} + \#\{i | \underline{b}_i < \overline{b}_i, i = 1, \ldots, \nu\} + 1.$$

For example, if we choose

$$\underline{a}_i = \overline{a}_i = 0, \text{ for } i \ne 1, \quad \underline{a}_1 = 0, \ \overline{a}_1 = 1 \quad \text{and} \quad \underline{b}_i = \overline{b}_i = 1, \text{ for } i \ne 2, \quad \underline{b}_2 = 0, \ \overline{b}_2 = 1 \tag{4.8}$$

then the number of variables of the first level problem is 3.

Finding an optimal parameters a, b, η is a challenging problem, since the solution $w(a,b,\eta)$ of the second stage optimization problem is not differentiable and the covariance matrix C and the vector of expected return r cannot be computed directly in general. We will use the interior point method ($w - a = s > 0, b - w = t > 0$) to solve the second stage problem of (4.7) and define a smoothing function $w_\mu(a,b,\eta)$ [32]. In particular, we use Algorithm `mc_smoothing_search` to solve

$$\begin{aligned} \max_{(a,b,\eta)\in\Omega} \ & \frac{r^T w_\mu(a,b,\eta)}{\sqrt{w_\mu(a,b,\eta)^T C w_\mu(a,b,\eta)}} \\ \text{where } w_\mu(a,b,\eta) = \operatorname{argmin} \ & \frac{1}{2} w^T C w - \eta r^T w - \mu \sum_{i=1}^{\nu} \log(s_i) - \mu \sum_{i=1}^{\nu} \log(t_i) \\ \text{s.t.} \ & e^T w \ = \ 1 \\ & w - a - s = 0 \\ & b - w - t = 0. \end{aligned} \tag{4.9}$$

In this section, we report numerical results that we got with Algorithm `mc_smoothing_search` for five standard data sets from the OR-Library [4] and the CSI 300 index from Shanghai-Shenzhen stock market. The data are the weekly or daily prices of the component stocks for the six stock market indices drawn from different countries. See Table 4.1 for the description of the data sets. The ν and T columns are the number of the component assets included in the index and the number of the observations for the assets, respectively.

TABLE 4.1
Description of the six real data sets

Data set	ν	Location	Index	T	Description
data 1	31	Hong Kong	Hang Seng	291	weekly prices from 1992 to 1997
data 2	85	Germany	DAX 100	291	weekly prices from 1992 to 1997
data 3	89	UK	FTSE 100	291	weekly prices from 1992 to 1997
data 4	98	USA	S&P 100	291	weekly prices from 1992 to 1997
data 5	225	Japan	Nikkei 225	291	weekly prices from 1992 to 1997
CSI300	300	China	CSI 300	401	daily prices from 2011 to 2013

We report on two experiments: randomly generated problems in §4.2.1, which use the mean and the covariance matrix generated from the real data in Table 4.1 and rolling window procedures for out-of-sample comparison in §4.2.2, which use the stock prices to generate the returns of assets and the covariance matrix by Monte Carlo simulation.

4.2.1. Randomly generated problems. We choose the following parameters as input data of Algorithm `mc_smoothing_search`:

$$h = 0.5, \quad \mu = 0.1, \quad N = 100, \quad \tau = 0.5, \quad \gamma = 1.5.$$

We choose the feasible set Ω as in (4.8).

For all tests, we terminated the algorithm if the stencil size is less than 10^{-2}.

For each data set in Table 4.1, we first calculate the average $\hat{r} \in R^\nu$ and the covariance matrix $\hat{C} \in R^{\nu\times\nu}$ for the returns of the assets. Given a sample size N, we generate i.i.d. random vectors $u_i \in R^\nu$, $i = 1, \ldots, N$ normally distributed with mean $\hat{r}$ and covariance matrix $\hat{C}$, that is

$$u_i = \hat{r} + \hat{C}^{\frac{1}{2}}\text{randn}(\nu, 1), \qquad i = 1, 2, ..., N,$$

and then take r^N to be the sample average of $u_i, i = 1, 2, ..., N$, and C^N to be the sample covariance matrix

$$r^N = \frac{1}{N}\sum_{i=1}^{N} u_i \quad \text{and} \quad C^N = \frac{1}{N}\sum_{i=1}^{N}(u_i - r^N)(u_i - r^N)^T.$$

Then we compute the smoothing approximation for the problem of minimizing the negative Sharpe ratio

$$\begin{aligned} \tilde{f}(x, \mu, N) = & -\frac{(r^N)^T w_\mu(x)}{\sqrt{w_\mu(x)^T C^N w_\mu(x)}} \\ & \text{where } w_\mu(x) = \text{argmin } \frac{1}{2}w^T C^N w - \eta(r^N)^T w - \mu\sum_{i=1}^{\nu}\log(s_i) - \mu\sum_{i=1}^{\nu}\log(t_i) \\ & \qquad \text{s.t. } e^T w = 1 \\ & \qquad\qquad w - a - s = 0 \\ & \qquad\qquad b - w - t = 0, \end{aligned} \tag{4.10}$$

where $x = (a_1, b_2, \eta)$.

For each data set in Table 4.1, we use Algorithm `mc_smoothing_search` to solve problem (4.7) with the starting point $(a_1, b_2, \eta) = (0, 1, 0.5)$. Table 4.2 presents the results. From Table 4.2, we can see the value of objective function (Sharpe ratio) at the final iteration is bigger than the value of objective function at the point $w = e/\nu$, which is a feasible point of problem (4.6) with $a = 0$ and $b = e$.

TABLE 4.2
Numerical results for the portfolio management problem with randomly generated data

Data set	data1	data2	data3	data4	data5	CSI300
Lower bound a_1	4.69E-1	4.22E-1	6.56E-1	1.25E-1	1.00E00	7.50E-1
Upper bound b_2	8.13E-1	8.75E-1	9.69E-1	9.69E-1	7.12E-1	7.50E-1
Risk aversion η	9.69E-1	5.31E-1	3.43E-1	4.38E-1	8.59E-1	6.41E-1
Optimized Sharpe ratio	1.57E-1	2.85E-1	2.51E-1	2.47E-1	9.76E-2	7.98E-2
Sharpe ratio of e/ν	1.04E-1	9.15E-2	1.53E-1	1.99E-1	-4.90E-2	-9.05E-2

4.2.2. Problems with rolling window procedures. For a given data set, assuming that the observations of the stock prices are $\{P_{i,t} : 1 \le i \le \nu, 1 \le t \le T\}$, we can compute the (logarithmic) returns of the stocks:

$$r_{i,t} = \log \frac{P_{i,t+1}}{P_{i,t}}, \ \ i = 1, \dots, \nu, \ t = 1, \dots, T-1.$$

For the purpose of numerical comparisons, we partition the data set into two subsets: a training set and a testing set. The training set, called in-sample set, consists the first half of the data set and is used to compute an optimal parameter x^* and the corresponding optimal portfolio selection $w(x^*)$. The testing set, called out-of-sample, consists the second half of the data set and is used to test how well the optimal parameter x^* and the corresponding optimal portfolio selection $w(x^*)$.

More exactly, for stock i with $i = 1, \dots, \nu$, we can use the training set to compute the in-sample expectation and the standard deviation by

$$\bar{\mu}_i = \frac{1}{M} \sum_{t=1}^{M} r_{i,t} \quad \text{and} \quad \bar{\sigma}_i = \sqrt{\frac{1}{M} \sum_{t=1}^{M} (r_{i,t} - \bar{\mu}_i)^2},$$

respectively, where $M = (T-1)/2$. As is standard in finance [19], we simulate the out-of-sample prices as follows. Let N be the sample size. Then at the j-th simulation ($1 \le j \le N$), for $M+1 \le t \le T-1$, if the price $S^{(j)}_{i-1,t}$ of stock i at an out-of-sample time $t-1$ is known, the price $S^{(j)}_{i,t}$ of this stock at time t is generated by

$$S^{(j)}_{i,t} = S^{(j)}_{i,t-1} \exp(\bar{\mu}_i + \bar{\sigma}_i Z),$$

where $S^{(j)}_{i,M} = P_{i,M}$ for all $1 \le j \le N$ and Z is randomly produced by the standard normal distribution $N(0,1)$. In a similar way, we can calculate the (logarithmic) returns by this simulation

$$r^{(j)}_{i,t} = \log \frac{S^{(j)}_{i,t+1}}{S^{(j)}_{i,t}}, \quad t = M+1, \dots, T-1.$$

For $t = M+1, \dots, T-1$, denote the column vector $r^{(j)}_t$ with its i-th component being $r^{(j)}_{i,t}$ and its average vector $\bar{r}_t = \frac{1}{N} \sum_{j=1}^{N} r^{(j)}_t$, the sample mean r^N and the sample variance C^N of the out-of-sample can be computed by

$$r^N = \frac{1}{M} \sum_{t=M+1}^{T-1} \bar{r}_t \quad \text{and} \quad C^N = \frac{1}{M} \sum_{t=M+1}^{T-1} (\bar{r}_t - r^N)(\bar{r}_t - r^N)^T.$$

Then we solve problem (4.9) with the sample mean r^N and the sample variance C^N as described in (4.10) to obtain the optimal parameter x^* and the corresponding optimal portfolio selection $w(x^*)$ by Algorithm `mc_smoothing_search`.

We choose the following parameters as input data of Algorithm `mc_smoothing_search`:

$$h = 0.5, \quad \mu = 0.1, \quad N = 10, \quad \tau = 0.5, \quad \gamma = 1.5.$$

We choose the feasible set Ω as in (4.8). For all tests, we choose the starting point $(a_1, b_2, \eta) = (0, 1, 0.5)$ and terminated the algorithm when the sample size N gets larger than 10^5.

To evaluate the quality of the optimal portfolio selection $w(x^*)$, we shall make use of the real out-of-sample data. We denote by r^{out} and C^{out} the mean and variance of the real returns of the out-of-sample set; namely,

$$r^{out} = \frac{1}{M} \sum_{t=M+1}^{T-1} r_t \quad \text{and} \quad C^{out} = \frac{1}{M} \sum_{t=M+1}^{T-1} (r_t - r^{out})(r_t - r^{out})^T,$$

where r_t is the vector formed by the stock prices $r_{i,t} (i = 1, \ldots, n)$ for $t = M + 1, \ldots, T - 1$. Then we can calculate the Sharpe ratio of the optimal solution $w(x^*)$ by using r^{out} and C^{out} as follows.

$$\mathrm{SR}^* = \frac{(r^{out})^T w(x^*)}{\sqrt{w(x^*)^T C^{out} w(x^*)}}.$$

In Table 4.3, for all the six data sets, we list the optimal values of a_1, b_2, η achieved by Algorithm `mc_smoothing_search`, and the corresponding SR^* For comparison, we also list the Sharpe ratio of the average strategy (namely, taking $1/\nu$ portion of each portfolio) using r^{out} and C^{out}. From Table 4.3, we can see that using Algorithm `mc_smoothing_search` to solve problem (4.7) can provide a portfolio strategy with higher Sharpe ratio than the average strategy for all data sets.

TABLE 4.3
Numerical results for the portfolio management problem with rolling window procedures

Data set	data1	data2	data3	data4	data5	CSI300
Lower bound a_1	1.00E-3	1.18E-2	1.12E-2	1.02E-1	4.40E-2	3.33E-3
Upper bound b_2	3.14E-1	2.31E-1	6.36E-1	4.15E-2	1.29E-2	2.53E-1
Risk aversion η	2.81E-1	9.38E-1	6.25E-1	6.25E-2	1.00E00	7.50E-1
Sharpe ratio SR^*	3.35E-1	2.36E-1	3.72E-1	5.12E-1	2.19E-1	2.19E-1
Sharpe ratio of e/ν	1.57E-1	2.10E-1	2.79E-1	3.44E-1	-3.85E-2	3.18E-3

5. Conclusions. In this paper we propose a smoothing direct search algorithm with Monte Carlo simulation **Algorithm** `mc_smoothing_search` for the constrained nonsmooth nonconvex optimization problem (1.1), where the objective function value $f(x)$ cannot be computed directly, but are approximated by Monte Carlo simulation. This algorithm updates the stencil size h, smoothing parameter μ and the sample size N simultaneously with the rate $h/\mu \to 0$, and $(h\sqrt{N})^{-1} \to 0$. We prove that any accumulation point of the sequence generated by the algorithm satisfies the first order optimality condition $\tilde{\chi}(x) = 0$ with probability one, where $\tilde{\chi}(x)$ is defined by (2.26). We report on a set of numerical experiments which illustrate the analysis and show that **Algorithm** `mc_smoothing_search` is an effective method for minimizing nonsmooth functions whose function values cannot be computed directly but are approximated by Monte Carlo simulation.

Appendix A. Smoothing functions.

We give an example of smoothing functions to explain Assumption 2.1. Let $f(x) = 2\max(0, p(x))$, where $p : R^n \to R$ is twice continuously differentiable with

$$\|\nabla p(x) \nabla p(x)^T\| \le \Gamma.$$

We use the smoothing function

$$\hat{f}(x, \mu) = p(x) + \sqrt{p(x)^2 + 4\mu^2}, \tag{A.1}$$

and $V = \{e_1, \ldots, e_n\}$, the unit coordinate directions in R^n.

Clearly part (i) of Assumption 2.1 holds with $c_1 = 2$ and $c_2 = 0$, since

$$|f(x) - \hat{f}(x, \mu)| \leq 2\mu.$$

Now we consider part (ii) of Assumption 2.1. The Clarke subgradient has the form

$$\partial f(x) = 2 \begin{cases} \nabla p(x) & \text{if } p(x) > 0 \\ \mathbf{0} & \text{if } p(x) < 0 \\ [0,1]\nabla p(x) & \text{if } p(x) = 0 \end{cases} \tag{A.2}$$

and the gradient of the smoothing function is

$$\nabla \hat{f}(x, \mu) = \left(1 + \frac{p(x)}{\sqrt{p(x)^2 + 4\mu^2}}\right) \nabla p(x).$$

Hence, we have

$$\|\nabla \hat{f}(x, \mu)\| \leq 2\|\nabla p(x)\|.$$

If p is Lipschitz continuously differentiable on a convex and compact set Ω, then there is an Υ such that $\|\nabla \hat{f}(x, \mu)\| \leq \Upsilon$ on Ω.

It is easy to see that for $p(x) \neq 0$, f is differentiable at x and

$$\partial f(x) = \nabla f(x) = \text{con}\{v \,|\, \nabla \hat{f}(x_k, \mu_k) \to v, \text{ for } x_k \to x, \ \mu_k \downarrow 0\}.$$

For $p(x) = 0$, since $0 \leq 1 + \frac{p(x)}{\sqrt{p(x)^2 + 4\mu^2}} \leq 2$, we have

$$\text{con}\{v \,|\, \nabla \hat{f}(x_k, \mu_k) \to v, \text{ for } x_k \to x, \ \mu_k \downarrow 0\} \subseteq \partial f(x).$$

Now, let $\mu_k^2 = (1 - h^2)p(x_k)^2/(4h^2)$ for some $h \in (0, 1]$. Then for $x_k \to x$ with $p(x_k) \downarrow 0$, we have $\mu_k \downarrow 0$ and

$$\nabla \hat{f}(x_k, \mu_k) = (1 + h)\nabla p(x_k) \to (1 + h)\nabla p(x),$$

and for $x_k \to x$ with $p(x_k) \uparrow 0$, we have $\mu_k \downarrow 0$ and

$$\nabla \hat{f}(x_k, \mu_k) = (1 - h)\nabla p(x_k) \to (1 - h)\nabla p(x).$$

Moreover, if we take $\mu_k = \sqrt{|p(x_k)|}$, then for $x_k \to x$, we have $p(x_k) \to 0$, $\mu_k \downarrow 0$ and

$$\nabla \hat{f}(x_k, \mu_k) = \left(1 + \frac{p(x_k)}{\sqrt{p(x_k)^2 + 4|p(x_k)|}}\right) \nabla p(x_k) \to \nabla p(x).$$

Hence we find that for $p(x) = 0$,

$$\partial f(x) = [0, 2]\nabla p(x) = \text{con}\{v \,|\, \nabla \hat{f}(x_k, \mu_k) \to v, \text{ for } x_k \to x, \ \mu_k \downarrow 0\}.$$

Finally, we consider part (iii) of Assumption 2.1. Since

$$\nabla^2 \hat{f}(x, \mu) = \left(1 + \frac{p(x)}{\sqrt{p(x)^2 + 4\mu^2}}\right) \nabla^2 p(x) + \frac{4\mu^2}{(p(x)^2 + 4\mu^2)^{\frac{3}{2}}} \nabla p(x) \nabla p(x)^T,$$

we have

$$\|\nabla^2 \hat{f}(x, \mu)\| \leq \frac{1}{2\mu} \|\nabla p(x) \nabla p(x)^T\| + 2\|\nabla^2 p(x)\| \leq \frac{\Gamma}{2\mu} + 2\|\nabla^2 p(x)\|,$$

which implies that part (iii) of Assumption 2.1 holds.

A smoothing function of $|p(x)|$ can be defined by using the relation $|p(x)| = \max(0, p(x)) + \max(0, -p(x))$ and a smoothing function of $\max(0, p(x))$. For example, using (A.1), we can have a smoothing function $\sqrt{(p(x)^2 + 4\mu^2)}$ for $|p(x)|$.

There is a detailed discussion of smoothing functions in [8].

REFERENCES

[1] C. Audet and J. E. Dennis. Analysis of generalized pattern searches. *SIAM J. Optim.*, 13(2003), pp. 889–903.

[2] C. Audet and J. E. Dennis. Mesh adaptive direct search algorithms for constrained optimization. *SIAM J. Optim.*, 17(2006), pp. 188–217.

[3] C. Audet and J. E. Dennis. A progressive barrier for derivative-free nonlinear programming. *SIAM J. Optim.*, 20(2009), pp. 445–472.

[4] J.E. Beasley, OR-Library: Distributing test problems by electronic mail, *J. Oper. Res. Soc.*, 41(1990), pp. 1069-1072. (last updated April 2016)

[5] R. Blundell and J.L. Powell, Censored regression quantiles with endogenous regression, *J. Econometrics*, 141(2007), pp. 65-83.

[6] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris, Sparse and stable Markowitz portfolios, *Proc. Natl. Acad. Sci. USA*, 106(2009), pp. 12267–12272

[7] J.V. Burke, X. Chen and H. Sun, Compute Clarke subgradients of expectation of measurable composite affine max function via smoothing, Manuscript, 2017.

[8] X. Chen, Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.*, 134(2012), pp. 71–99.

[9] X. Chen and C. T. Kelley, Optimization with hidden constraints and embedded Monte Carlo computations, *Optim. Engin.*, 17(2016), pp. 157–175.

[10] X. Chen, Z. Lu and T.K. Pong, Penalty methods for a class of non-Lipschitz optimization problems, SIAM J. Optim., 26(2016), pp. 1465-1492.

[11] X. Chen, T.K. Pong and R.J.-B. Wets, Two-stage stochastic variational inequalities: an ERM-solution procedure, *Math. Program.*, under minor revision.

[12] V. K. Chopra, Mean-variance revisited: near-optimal portfolios and sensitivity to input variations, *J. Investing*, 2(1993), pp. 51-59.

[13] F. H. Clarke, Optimization and Nonsmooth Analysis, SIAM Publisher, Philadelphia, 1990.

[14] A. R. Conn, K. Scheinberg and L. N. Vicente, Introduction to Derivative-Free Optimization, SIAM Philadelphia, 2009.

[15] J. E. Dennis and V. Torczon, Direct search methods on parallel machines, SIAM J. Optim., 1(1991), pp. 448 – 474.

[16] F. Facchinei and J.-S. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer-Verlag, New York, Inc, 2003.

[17] G. M. Frankfurter, H. E. Phillips and J. P. Seagle, Portfolio selection: the effects of uncertain means, variances and covariances, *J. Financ. Quantit. Anal.*, 6(1971), pp. 1251-1262.

[18] R. Garmanjani and L.N. Vicente, Smoothing and worst-case complexity for direct-search methods in nonsmooth optimization, *IMA J. Numer. Anal.*, 33(2012), pp. 1008–1028.

[19] J. C. Hull, Options, Futures, and Other Derivatives, Pearson, New York, USA: first edition 1988.

[20] P. Jorion, Bayes-Stein estimation for portfolio analysis, *J. Financ. Quantit. Anal.*, 21(1986), pp. 279-292.

[21] C. T. Kelley, Iterative Methods for Optimization, SIAM Publisher, Philadelphia, 1999.

[22] C. T. Kelley, Implicit Filtering, SIAM Publisher, Philadelphia, 2011.

[23] R. W. Klein and V. S. Bawa, The effect of estimation risk on optimal portfolio choice, *J. Financ. Econom,* 3(1976), pp. 215-231.

[24] T. G. Kolda, R. M. Lewis, and V. J. Torczon, Optimization by direct search: New perspectives on some classical and modern methods, SIAM Review, 45 (2003), pp. 385–482.

[25] T. G. Kolda, R. M. Lewis, and V. Torczon. Stationarity results for generating set search for linearly constrained optimization. SIAM J. Optim., 17 (2006), 943–968.

[26] G. Lan, An optimal method for stochastic composite optimization, *Math. Program.*, 133(2012), pp. 365-397.

[27] Y-F. Liu, S. Ma, Y-H. Dai, S. Zhang, A smoothing SQP framework for a class of composite L_q minimization over polyhedron, *Math. Program.*, 158(2016), pp. 467-500.

[28] H. M. Markowitz, Portfolio selection, *J. Finance*, 7(1952), pp. 77-91.

[29] H. M. Markowitz,. The optimization of a quadratic function subject to linear constraints, Naval Research Logistics Quarterly, 3(1956), pp. 111-133.

[30] R. O. Michaud, The Markowitz optimization enigma: is 'optimized' optimal?, *Financial Analysts J.*, 45(1989), pp. 31-42.

[31] Y. Nesterov, Random gradient-free minimization of convex functions, CORE, 16(2011).

[32] J. Nocedal and S. J. Wright, Numerical Optimization, 2nd ed., Springer, New York, 2006.

[33] J.M. Ortega and W.C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.

[34] J.-S. Pang, A new and efficient algorithm for a class of portfolio selection problems, *Oper. Res.*, 28(1980), pp. 754–767

[35] M. Porcelli and Ph. L. Toint, BFO: a trainable derivative-free Brute Force Optimizer, 2015.

[36] A. Shapiro, D. Dentcheva and A. Ruszczyński, Lectures on Stochastic Programming, MPS-SIAM Series on Optimization,

SIAM Philadelphia, 2009.
[37] W.F. Sharpe, The Sharpe Ratio, *J. Portfolio Management*, 21(1994), pp. 4958.
[38] R.J. Smith and R.W. Blundell, An exogeneity test for a simultaneous equation Tobit model with an application to labor supply, *Econometrica*, 54(1986), pp. 679-685.
[39] L. Taylor and T. Otsu, Estimation of nonseparable models with censored dependent variables and endogenous regressors, to appear in Econometric Reviews.
[40] A. Toth, J. A. Ellis, T. Evans, S. Hamilton, C. T. Kelley, R. Pawlowski, and S. Slattery, Local improvement results for Anderson acceleration with inaccurate function evaluations, 2016. to appear in SIAM J. Sci. Comp.
[41] J. Willert, X. Chen and C. T. Kelley, Newton's method for Monte Carlo-based residuals, *SIAM J. Numer. Anal.*, 53(2015), pp. 1738-1757