Link to publication

University of Bath

A geometric integration approach to nonsmooth, nonconvex optimisation

Erlend S. Riis · **Matthias J. Ehrhardt** · **G. R. W. Quispel** ·
Carola-Bibiane Schönlieb

Abstract The optimisation of nonsmooth, nonconvex functions without access to gradients is a particularly challenging problem that is frequently encountered, for example in model parameter optimisation problems. Bilevel optimisation of parameters is a standard setting in areas such as variational regularisation problems and supervised machine learning. We present efficient and robust derivative-free methods called randomised Itoh–Abe methods. These are generalisations of the Itoh–Abe discrete gradient method, a well-known scheme from geometric integration, which has previously only been considered in the smooth setting. We demonstrate that the method and its favourable energy dissipation properties are well-defined in the nonsmooth setting. Furthermore, we prove that whenever the objective function is locally Lipschitz continuous, the iterates almost surely converge to a connected set of Clarke stationary points. We present an implementation of the methods, and apply it to various test problems. The numerical results indicate that the randomised Itoh–Abe methods are superior to state-of-the-art derivative-free optimisation methods in solving nonsmooth problems while remaining competitive in terms of efficiency.

Erlend S. Riis
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
E-mail: e.s.riis@damtp.cam.ac.uk

Matthias J. Ehrhardt
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
E-mail: m.j.ehrhardt@damtp.cam.ac.uk

G. R. W. Quispel
Department of Mathematics and Statistics, La Trobe University, Victoria 3086, Australia
E-mail: r.quispel@latrobe.edu.au

Carola-Bibiane Schönlieb
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
E-mail: cbs31@cam.ac.uk

1 Introduction

We consider the unconstrained optimisation problem

$$\min_{x \in \mathbb{R}^n} V(x), \tag{1.1}$$

where the objective function V is locally Lipschitz continuous, bounded below and coercive—the latter meaning that $\{x \in \mathbb{R}^n : V(x) \le M\}$ is compact for all $M \in \mathbb{R}$. The function may be nonconvex and nonsmooth, and we assume no knowledge besides point evaluations $x \mapsto V(x)$. To solve (1.1), we present *randomised Itoh–Abe methods*, a generalisation of the *Itoh–Abe discrete gradient method*. The latter is a derivative-free optimisation scheme,[1] that has previously only been considered for differentiable functions.

Discrete gradient methods, a tool from geometric numerical integration, are optimisation schemes that inherit the energy dissipation of continuous gradient flow. The iterates of the methods monotonically decrease the objective function, for all time steps, and Grimm et al. [28] recently provided a convergence theory for solving (1.1) in the continuously differentiable setting. We extend the concepts and results of their work and show that the Itoh–Abe discrete gradient method can be applied in the nonsmooth case, and, furthermore, that the favourable dissipativity property of the methods extends to this setting. Furthermore, we prove that for locally Lipschitz continuous functions the iterates converge to a set of stationary points, defined by the Clarke subdifferential framework.

1.1 Gradient flow and the discrete gradient method

For a differentiable function $V : \mathbb{R}^n \to \mathbb{R}$, gradient flow is the ODE system defined by

$$\dot{x} = -\nabla V(x), \qquad x(0) = x_0 \in \mathbb{R}^n, \tag{1.2}$$

where the dot represents differentiation with respect to time. By applying the chain rule, we compute

$$\frac{\mathrm{d}}{\mathrm{d}t} V(x(t)) = \langle \nabla V(x(t)), \dot{x}(t) \rangle = -\|\nabla V(x(t))\|^2 = -\|\dot{x}(t)\|^2 \le 0, \tag{1.3}$$

where $\|x\|$ denotes the 2-norm $\sqrt{\langle x, x \rangle}$. This implies that gradient flow is inherently an energy dissipative system.

In the field of geometric numerical integration, one studies methods for numerically solving ODEs that also preserve structures of the continuous system—see [30, 45] for an introduction. Discrete gradient methods can be applied to first-order gradient systems to preserve energy conservation laws, dissipation laws, as well as Lyapunov functions [25, 34, 46, 57]. They are defined as follows.

Definition 1.1 *Let* $V : \mathbb{R}^n \to \mathbb{R}$ *be continuously differentiable. A* discrete gradient *is a continuous mapping* $\overline{\nabla} V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ *that satisfies the two following properties.*

$$\begin{cases} \langle \overline{\nabla} V(x, y), y - x \rangle = V(y) - V(x) & \text{(consistency)} \\ \lim_{y \to x} \overline{\nabla} V(x, y) = \nabla V(x) & \text{(mean value property)} \end{cases} \quad \text{for all } x, y \in \mathbb{R}^n.$$

---

[1] Not to be confused with another derivative-free method with the same name proposed by Bagirov et al. [5], which uses a different concept of a discrete gradient.

We now introduce the discrete gradient method for optimisation. For $x^0 \in \mathbb{R}^n$ and time steps $\tau_k > 0$, $k \in \mathbb{N}$, we solve

$$x^{k+1} = x^k - \tau_k \overline{\nabla} V(x^k, x^{k+1}). \tag{1.4}$$

We apply the above mean value property to derive that the iterates decrease V.

$$
\begin{aligned}
V(x^{k+1}) - V(x^k) &= \langle \overline{\nabla} V(x^k, x^{k+1}), x^{k+1} - x^k \rangle \\
&= -\tau_k \| \overline{\nabla} V(x^k, x^{k+1}) \|^2 = -\frac{1}{\tau_k} \| x^{k+1} - x^k \|^2.
\end{aligned} \tag{1.5}
$$

Note that the decrease holds for all time steps $\tau_k > 0$, and that (1.5) can be seen as a discrete analogue of the dissipative structure of gradient flow (1.3), replacing derivatives by finite differences.

Grimm et al. [28] proved that for coercive, continuously differentiable functions, the iterates of (1.4) converge to a set of stationary points, provided that there are strictly positive constants $\tau_{\min}, \tau_{\max}$ such that $\tau_k \in [\tau_{\min}, \tau_{\max}]$ for all $k \in \mathbb{N}$.

1.1.1 Itoh–Abe methods

The *Itoh–Abe discrete gradient* [34] (also known as coordinate increment discrete gradient)[2] is defined as

$$
\overline{\nabla} V(x, y) = \begin{pmatrix} \frac{V(y_1, x_2, \ldots, x_n) - V(x)}{y_1 - x_1} \\ \frac{V(y_1, y_2, x_3, \ldots, x_n) - V(y_1, x_2, \ldots, x_n)}{y_2 - x_2} \\ \vdots \\ \frac{V(y) - V(y_1, \ldots, y_{n-1}, x_n)}{y_n - x_n} \end{pmatrix}.
$$

Solving an iterate of the discrete gradient method (1.4) with the Itoh–Abe discrete gradient is equivalent to successively solving n scalar equations of the form

$$
\begin{aligned}
x_1^{k+1} &= x_1^k - \tau_k \frac{V(x_1^{k+1}, x_2^k, \ldots, x_n^k) - V(x^k)}{x_1^{k+1} - x_1^k} \\
x_2^{k+1} &= x_2^k - \tau_k \frac{V(x_1^{k+1}, x_2^{k+1}, x_3^k, \ldots, x_n^k) - V(x_1^{k+1}, x_2^k, \ldots, x_n^k)}{x_2^{k+1} - x_2^k} \\
&\vdots \\
x_n^{k+1} &= x_n^k - \tau_k \frac{V(x^{k+1}) - V(x_1^{k+1}, x_2^{k+1}, \ldots, x_{n-1}^{k+1}, x_n^k)}{x_n^{k+1} - x_n^k}.
\end{aligned}
$$

We generalise the Itoh–Abe discrete gradient method to *randomised Itoh–Abe methods* accordingly. Let $(d^k)_{k \in \mathbb{N}} \subset S^{n-1}$ be a sequence of directions, where S^{n-1} denotes the unit sphere $\{x \in \mathbb{R}^n : \|x\| = 1\}$. The directions can be drawn from a random distribution or chosen deterministically. At the kth step, we update

$$x^{k+1} \mapsto x^k - \tau_k \beta_k d^k, \text{ where } \beta_k \neq 0 \text{ solves } \beta_k = -\frac{V(x^k - \tau_k \beta_k d^k) - V(x^k)}{\tau_k \beta_k}, \tag{1.6}$$

Algorithm 1 Randomised Itoh–Abe method

Input: starting point x^0, directions $(d^k)_{k\in\mathbb{N}}$, time steps $(\tau_k)_{k\in\mathbb{N}}$.

---

for $k = 0, 1, 2, \dots$ **do**
    Update $x^{k+1} = x^k - \tau_k \beta_k d^k$ via (1.6)
end for

If no such β_k exists, we set $x^{k+1} = x^k$. We formalise this method in Algorithm 1. We assume throughout the paper that the time steps $(\tau_k)_{k\in\mathbb{N}}$ are bounded between two strictly positive constants $\tau_{\min}, \tau_{\max}$, which can take arbitrary values.

Observe that if $(d^k)_{k\in\mathbb{N}}$ cycle through the standard coordinates $(e^i)_{i=1}^n$ with the rule $d^k = e^{[k \bmod n]+1}$, then computing n steps of (1.6) corresponds to one step of (1.4) with the Itoh–Abe discrete gradient. Furthermore, the dissipation properties (1.5) can be rewritten as

$$V(x^{k+1}) - V(x^k) = -\tau_k \left(\frac{V(x^{k+1}) - V(x^k)}{\|x^{k+1} - x^k\|} \right)^2 = -\frac{1}{\tau_k} \|x^{k+1} - x^k\|^2. \tag{1.7}$$

Consequently, the dissipative structure of the Itoh–Abe methods is well-defined in a derivative-free setting.

Ehrhardt et al. [20] studied Itoh–Abe methods in the smooth setting, asserting linear convergence rates for functions that satisfy the Polyak–Łojasiewicz inequality [36]. Furthermore, the application of Itoh–Abe discrete gradient methods to smooth optimisation problems is well-documented. Applications include convex variational regularisation problems for image analysis [28], nonconvex image inpainting problems with Euler's elastica regularisation [59], for which it outperformed gradient-based schemes, and the popular Gauss–Seidel method and successive-over-relaxation (SOR) methods for solving linear systems [48].

1.2 Bilevel optimisation and blackbox problems

An important application for developing derivative-free solvers is *model parameter optimisation problems*. The setting for this class of problems is as follows. A model depends on some tunable parameters $\alpha \in \mathbb{R}^n$, so that for a given parameter choice α, the model returns an output u_α. There is a cost function Φ, which assigns to output u_α a numerical score $\Phi(u_\alpha) \in \mathbb{R}$, which we want to minimise. The associated model parameter optimisation problem becomes

$$\alpha^* \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \Phi(u_\alpha).$$

A well-known example of parameter optimisation problems is supervised machine learning.

In this paper, we consider one instance of such problems in image analysis, called *bilevel optimisation* of variational regularisation problems. Here, the model is given by a variational regularisation problem for image denoising,

$$u_\alpha \in \operatorname*{argmin}_{u} \frac{1}{2} \|u - f^\delta\|^2 + R_\alpha(u),$$

where f^δ is a noisy image and α is the regularisation parameter. For training data with desired reconstruction $u^\dagger$, we consider a scoring function Φ that estimates the discrepancy between $u^\dagger$ and the reconstruction u_α. In Section 6.2, we apply Itoh-Abe methods to solve these problems.

---

[2] There are infinitely many discrete gradients, each with a corresponding discrete gradient method. See [28] for further examples.

Bilevel optimisation problems, and model parameter optimisation problems in general, pose several challenges. They are often nonconvex and nonsmooth, due to the nonsmoothness and nonlinearity of $\alpha \mapsto u_\alpha$. Furthermore, the model simulation $\alpha \mapsto u_\alpha$ is an algorithmic process for which gradients or subgradients cannot easily be estimated. Such problems are termed *blackbox optimisation problems*, as one only has access to point evaluations of the function. It is therefore of great interest to develop efficient and robust derivative-free methods for such optimisation problems.

There is a rich literature on bilevel optimisation for variational regularisation problems in image analysis, c.f. e.g. [12, 58, 39, 51]. Furthermore, model parameter optimisation problems appear in many other applications. These include optimising for the management of water resources [23], approximation of a transmembrane protein structure in computational biology [26], image registration in medical imaging [52], the building of wind farms [19], and solar energy utilisation in architectural design [35], to name a few.

1.3 Related literature on nonsmooth, nonconvex optimisation

Although nonsmooth, nonconvex problems are known for their difficulty compared to convex problems, a rich optimisation theory has grown since the 1970s. As the focus of this paper is derivative-free optimisation, we will compare the methods' convergence properties and performance to other derivative-free solvers. Audet and Hare [3] recently provided a reference work for this field.

While there a myriad of derivative-free solvers, few provide convergence guarantees for nonsmooth, nonconvex functions. Audet and Dennis Jr [2] introduced the *mesh adaptive direct search* (MADS) method for constrained optimisation, with provable convergence guarantees to stationary points for nonsmooth, nonconvex functions. Direct search methods evaluate the function at a finite polling set, compare the evaluations, and update the polling set accordingly. Such methods only consider the ordering of the evaluations, rather than the numerical differences. A significant portion of derivative-free methods are direct search methods, and the most well known of these is the Nelder–Mead method (also known as the downhill simplex method) [49].

Alternatively, model-based methods that build a local quadratic model based on evaluations are well-documented [13, 55, 56]. While such methods tend to work well in practice, they are normally designed only for smooth functions, so their performance on nonsmooth functions is not guaranteed.

Fasano et al. [22] formulated a derivative-free line search method termed DFN and analyse its convergence properties for nonsmooth functions for the Clarke subdifferential, in the constrained setting. The Itoh–Abe methods share many similarities with DFN, such as performing line searches along dense directions, and they employ a similar convergence analysis. However, the line searches of these methods differ, and the Itoh–Abe methods are in particular motivated by discrete gradient methods and preserving the dissipative structure of gradient flow. Furthermore, our convergence analysis is more comprehensive, considering both stochastic and deterministic methods, as well as obtaining stronger convergence results in the deterministic case. Building on the aforementioned algorithm, Liuzzi and Truemper [43] formulated a derivative-free method that is a hybrid between DFN and MADS.

Furthermore, we note the resemblance of randomised Itoh–Abe methods (1.6), when $(d^k)_{k \in \mathbb{N}}$ is randomly, independently drawn from S^{n-1}, to the random search method proposed by Polyak in [54] and studied for nonsmooth, convex functions by Nesterov in [50], given by

$$x^{k+1} = x^k - \tau_k \frac{V(x^k + \beta_k d^k) - V(x^k)}{\beta_k} d^k,$$

where d^k is randomly, independently drawn from S^{n-1}. The implicit equation (1.6) can be treated as a line search rule for the above method, with constraints imposed by $\tau_{\min}, \tau_{\max}$.

While our focus is on derivative-free methods, we also mention some popular methods for nonsmooth, nonconvex optimisation that use gradient or subgradient information. Central in nonsmooth optimisation are *bundle methods*, where a *subgradient* [16] is required at each iterate to construct a linear approximation to the objective function—see [37] for an introduction. A close alternative to bundle methods are *gradient sampling methods* (see [11] for a recent review by Burke et al.), where the descent direction is determined by sampling gradients in a neighborhood of the current iterate. Curtis and Que [18] formulated a hybrid method between the gradient sampling scheme of [17] and the well known quasi-Newton method BFGS adapted to nonsmooth problems [42]. These methods have convergence guarantees in the Clarke subdifferential framework, under the assumption that the objective function is differentiable on an open, dense set. Last, we mention a derivative-free scheme based on gradient sampling methods, proposed by Kiwiel [38], where gradients are replaced by Gupal's estimates of gradients of the Steklov averages of the objective function. This method has convergence guarantees in the Clarke subdifferential framework, but has a high computational cost in terms of function evaluations per iterate.

1.4 Contributions

In this paper, we formulate randomised Itoh–Abe methods for nonsmooth functions. We prove that the method always admits a solution, and that the iterates converge to a set of Clarke stationary points, for any locally Lipschitz continuous function, and both for deterministic and randomly chosen search directions. Consequently, the scope of discrete gradient methods for optimisation is significantly broadened, and we conclude that the dissipativity properties of gradient flow can be preserved even beyond differentiability. Ultimately, this provides a new, robust, and versatile optimisation scheme for nonsmooth, nonconvex functions.

The theoretical convergence analysis for the Itoh–Abe methods is thorough and foundational, and we provide examples that demonstrate that the conditions of the convergence theorem are not just sufficient, but necessary. Furthermore, the statements and proofs are sufficiently general so that they can be adapted to other schemes, such as the aforementioned DFO method, thus enhancing the theory of these methods as well.

We show that the method works well in practice, by solving bilevel optimisation problems for variational regularisation problems, as well as solving benchmark problems such as Rosenbrock functions.

The rest of the paper is structured as follows. Section 2 provides a background on the Clarke subdifferential for nonsmooth, nonconvex analysis. In Section 3, the main theoretical results of the paper are presented, namely existence and optimality results in the stochastic and deterministic setting. In Section 4, we briefly discuss the Itoh–Abe discrete gradient for general coordinate systems. In Section 5 and Section 6, the numerical implementation is described and results from example problems are presented. Finally, a conclusion is given in Section 7.

2 Nonconvex optimisation

In this section, we introduce the Clarke subdifferential framework [16] for nonsmooth, nonconvex optimisation. This is the most popular framework for this setting, due to its nice analytical properties. It generalises the gradient of a differentiable function, as well as the subdifferential [21] of a convex function, hence the term *Clarke subdifferential*. Francis H. Clarke introduced the framework in his doctoral thesis in 1973 [15], in which he termed it *generalised gradient*.

2.1 The Clarke subdifferential

Throughout the paper, for $\varepsilon > 0$ and $x \in \mathbb{R}^n$, we denote by $B_\varepsilon(x)$ the open ball $\{y \in \mathbb{R}^n : \|y - x\| < \varepsilon\}$.

Definition 2.1 *V is* Lipschitz of rank L near x *if there exists $\varepsilon > 0$ such that for all $y, z \in B_\varepsilon(x)$, one has*

$$|V(y) - V(z)| \leq L\|y - z\|.$$

V is locally Lipschitz continuous if the above property holds for all $x \in \mathbb{R}^n$.

Definition 2.2 *For V Lipschitz near x and for a vector $d \in \mathbb{R}^n$, the* Clarke directional derivative *is given by*

$$V^o(x; d) = \limsup_{y \to x, \, \lambda \downarrow 0} \frac{V(y + \lambda d) - V(y)}{\lambda}.$$

Definition 2.3 *Let V be locally Lipschitz and $x \in \mathbb{R}^n$. The* Clarke subdifferential *of V at x is given by*

$$\partial V(x) = \left\{ p \in \mathbb{R}^n : V^o(x; d) \geq \langle d, p \rangle \text{ for all } d \in \mathbb{R}^n \right\}.$$

An element of $\partial V(x)$ is called a Clarke subgradient.

The subdifferential ∂V is well-defined for locally Lipschitz functions, coincides with the standard subdifferential for convex functions [16, Proposition 2.2.7], and coincides with the derivative at points of strict differentiability [16, Proposition 2.2.4]. It can equivalently be characterised as

$$\partial V(x) = \operatorname{co}\left\{ d \in \mathbb{R}^n : \exists (x^k)_{k \in \mathbb{N}} \subset \mathscr{D}(V) \text{ s.t. } x^k \to x \text{ and } \nabla V(x^k) \to d \right\},$$

where $\mathscr{D}(V)$ is the set of differentiable points, and co denotes the convex hull of the set [16, Theorem 2.5.1]. We additionally state two useful results, both of which can be found in Chapter 2 of [16].

Proposition 2.4 *Suppose V is locally Lipschitz continuous. Then*

  (i) *$\partial V(x)$ is nonempty, convex and compact, and if V is Lipschitz of rank L near x, then $\partial V(x) \subseteq B_L(0)$.*
 (ii) *$\partial V(x)$ is outer semicontinuous. That is, for all $\varepsilon > 0$, there exists $\delta > 0$ such that*

$$\partial V(y) \subset \partial V(x) + B_\varepsilon(0), \quad \text{for all } y \in x + B_\delta(0).$$

There are alternative frameworks for generalising differentiability of nonsmooth, nonconvex functions. For example, the Michel–Penot subdifferential [47] coincides with the Gâteaux derivative when this exists, unlike the Clarke subdifferential, which is larger and only coincides with strict derivatives [24]. However, the Clarke subdifferential is outer semicontinuous, making it in most cases the preferred framework for analysis. See [7] by Borwein and Zhu for a survey of various subdifferentials, published on the 25th birthday of the Clarke subdifferential.

2.1.1 Discrete gradients versus subgradients

By definition, when V is continuously differentiable, any discrete gradient $\overline{\nabla} V(x, y)$ converges to the gradient $\nabla V(x)$ as $y \to x$. However, for nondifferentiable V, discrete gradients do not necessarily approximate a subgradient or even an ε-*approximate subgradient*.[3] This is demonstrated by the following example.

Example 2.5 *Let $V(x_1, x_2) := \sqrt{x_1^2 + x_2^2}$, and set $x^k = [\frac{1}{k}, 0]^T$ and $x_2^k = [0, \frac{1}{k}]^T$. Then, for all k, the Itoh–Abe discrete gradient is*

$$\overline{\nabla} V(x_1^k, x_2^k) = [1, 1]^T.$$

Thus, $x_1^k \to [0, 0]^T$, $x_2^k \to [0, 0]^T$ and $\overline{\nabla} V(x_1^k, x_2^k) \to [1, 1]^T$. However, $[1, 1]^T$ is not in $\partial V(0, 0) = B_1(0, 0)$. In fact, for all $\varepsilon > 0$, we have $[1, 1]^T \notin \partial_\varepsilon V(0, 0)$.

---

[3] For convex functions, $p \in \mathbb{R}^n$ is an ε-approximate subgradient if, for all $y \in \mathbb{R}^n$, we have $V(y) \geq V(x) + \langle p, y - x \rangle - \varepsilon$ [32].

2.1.2 Clarke stationary points

Definition 2.6 $x^* \in \mathbb{R}^n$ *is a* Clarke stationary point *of V if $0 \in \partial V(x^*)$.*

For our purposes, we also define Clarke directional stationarity.

Definition 2.7 (Directional Clarke stationarity) *For a direction $d \in \mathbb{R}^n \setminus \{0\}$, we say that V is* Clarke direc-tionally stationary at x^* along d if

$$\min\{V^o(x;d), V^o(x;-d)\} \geq 0.$$

Remark 2.8 *A point x^* is Clarke stationary if and only if V is Clarke directionally stationary at x^* along d for all $d \in \mathbb{R}^n$ such that $\|d\| = 1$.*

Any local maxima and minima are stationary. If V is convex, then stationary points coincide with the global minima. For more general classes of functions, the concept of Clarke stationary points also reduces to convex first-order optimality conditions.

Definition 2.9 ([53]) *A locally Lipschitz continuous function V is* pseudoconvex *if, for all $x, y \in \mathbb{R}^n$,*

$$f(y) < f(x) \implies \forall p \in \partial V(x), \quad \langle p, y - x \rangle < 0.$$

If V is pseudoconvex, then any Clarke stationary point is a global minimum [4]. Clarke [16] also introduced the notion of *regularity*.

Definition 2.10 *A function V is* regular *at x if the directional derivative*

$$V'(x;d) := \lim_{\lambda \downarrow 0} \frac{V(x + \lambda d) - V(x)}{\lambda}$$

exists for all $d \in \mathbb{R}^n$ and $V^o(x;d) = V'(x;d)$. If this holds for all x, we say that V is regular.

For a regular function, a point is Clarke stationary if and only if the directional derivative is nonnegative in all directions. For example, convex functions are regular, and strict differentiability at a point implies regularity at a point. However, for nonregular functions, x^* can simultaneously be Clarke stationary and have negative directional derivatives in a neighbourhood of directions.

3 The discrete gradient method for nonsmooth optimisation

In this section, we present the main theoretical results for the randomised Itoh–Abe methods. In particular, we prove that the update (1.6),

$$x^{k+1} \mapsto x^k - \tau_k \beta_k d^k, \quad \text{where } \beta_k \neq 0 \text{ solves} \quad \beta_k = -\frac{V(x^k - \tau_k \beta_k d^k) - V(x^k)}{\tau_k \beta_k},$$

admits a solution for all $\tau_k > 0$. We also prove under minimal assumptions on V and $(d^k)_{k \in \mathbb{N}}$ that the iterates converge to a connected set of Clarke stationary points, both in a stochastic and deterministic setting.

3.1 Existence result

Lemma 3.1 *Suppose V is a continuous function bounded below, and that $x \in \mathbb{R}^n$, $d \in S^{n-1}$ and $\tau > 0$. Then one of the following statements hold.*

(i) There is a $\beta \neq 0$ that solves (1.6), i.e. that satisfies $\frac{V(x-\tau\beta d)-V(x)}{\tau\beta} = -\beta$.
(ii) V is Clarke directionally stationary at x along d.

Proof. Suppose the second statement does not hold. Then there is $\varepsilon > 0$ such that

$$\min\{V^o(x;-d), V^o(x;d)\} < -\varepsilon,$$

so assume without loss of generality that $V^o(x;-d) < -\varepsilon$. By definition of V^o, there is $\delta > 0$ such that for all $\beta \in (0,\delta)$,

$$\frac{V(x-\tau\beta d)-V(x)}{\tau\beta} \leq -\varepsilon/2.$$

Taking $\beta \to 0$, we get that

$$\lim_{\beta\to 0^+} \frac{V(x-\tau\beta d)-V(x)}{\tau\beta^2} \leq \lim_{\beta\to 0^+} -\frac{\varepsilon}{2\beta} = -\infty,$$

so there is a $\beta_1 \in (0,\delta)$ such that

$$\frac{V(x-\tau\beta_1 d)-V(x)}{\tau\beta_1^2} < -1.$$

On the other hand, as V is bounded below, we have

$$\liminf_{\beta\to\infty} \frac{V(x-\tau\beta_2 d)-V(x)}{\tau\beta_2^2} \geq \lim_{\beta\to\infty} \frac{\min\{0, V(x-\tau\beta_2 d)\}-V(x)}{\tau\beta_2^2} = 0.$$

Thus there is β_2 such that

$$\frac{V(x-\tau\beta_2 d)-V(x)}{\tau\beta_2^2} > -1.$$

Since the mapping $\beta \mapsto \frac{V(x-\tau\beta d)-V(x)}{\tau\beta^2}$ is continuous for $\beta \in (0,\infty)$, we conclude by the intermediate value theorem [62, Theorem 4.23] that there is $\beta \in (\beta_1,\beta_2)$ that solves the discrete gradient equation

$$\frac{V(x-\tau\beta d)-V(x)}{\tau\beta^2} = -1.$$

$\square$

Remark 3.2 *Note that by the above proof it is straightforward to identify an interval in which a solution to (1.6) exists, allowing for the use of standard root solver algorithms.*

The following lemma, which is an adaptation of [28, Theorem 1] for the nonsmooth setting, summarises some useful properties of the methods.

Lemma 3.3 *Suppose that V is continuous, bounded from below and coercive, and let $(x^k)_{k\in\mathbb{N}}$ be the iterates produced by (1.6). Then, the following properties hold.*

(i) $V(x^{k+1}) \leq V(x^k)$.
(ii) $\lim_{k\to\infty} \frac{V(x^{k+1})-V(x^k)}{\|x^{k+1}-x^k\|} = 0$.

(iii) $\lim_{k \to \infty} \|x^k - x^{k+1}\| = 0$.

(iv) $(x^k)_{k \in \mathbb{N}}$ has an accumulation point x^*.

Proof. Property *(i)* follows from the equation $V(x^{k+1}) - V(x^k) = -\tau_k \beta_k^2$.

Next we show properties *(ii)* and *(iii)*. Since V is bounded below and $(V(x^k))_{k \in \mathbb{N}}$ is decreasing, $V(x^k) \to V^*$ for some limit V^*. Therefore, by (1.7)

$$V(x^0) - V^* = \sum_{k=0}^{\infty} V(x^k) - V(x^{k+1}) = \sum_{k=0}^{\infty} \tau_k \left(\frac{V(x^k) - V(x^{k+1})}{\|x^{k+1} - x^k\|} \right)^2$$

$$\geq \tau_{\min} \sum_{k=0}^{\infty} \left(\frac{V(x^k) - V(x^{k+1})}{\|x^{k+1} - x^k\|} \right)^2.$$

Similarly, by (1.7)

$$V(x^0) - V^* = \sum_{k=0}^{\infty} V(x^k) - V(x^{k+1}) = \sum_{k=0}^{\infty} \frac{1}{\tau_k} \|x^k - x^{k+1}\|^2 \geq \frac{1}{\tau_{\max}} \sum_{k=0}^{\infty} \|x^k - x^{k+1}\|^2.$$

We conclude

$$\lim_{k \to \infty} \frac{V(x^k) - V(x^{k+1})}{\|x^{k+1} - x^k\|} = \lim_{k \to \infty} \|x^{k+1} - x^k\| = 0,$$

which shows properties *(ii)* and *(iii)*.

Last, we show *(iv)*. Since $(V(x^k))_{k \in \mathbb{N}}$ is a decreasing sequence, the iterates $(x^k)_{k \in \mathbb{N}}$ belong to the set $\left\{ x \in \mathbb{R}^n : V(x) \leq V(x^0) \right\}$. Therefore, by coercivity of V, the iterates $(x^k)_{k \in \mathbb{N}}$ are bounded, and admit an accumulation point. $\qquad\square$

We denote by S the limit set of $(x^k)_{k \in \mathbb{N}}$, which is the set of accumulation points,

$$S = \left\{ x^* \in \mathbb{R}^n : \exists (x^{k_j})_{j \in \mathbb{N}} \text{ s.t. } x^{k_j} \to x^* \right\}.$$

By the above lemma, S is nonempty. We now prove further properties of the limit set.

Lemma 3.4 *The limit set S is compact, connected and has empty interior. Furthermore, V is constant on S.*

Proof. Boundedness of S follows from coercivity of V combined with the fact that S belongs to the level set $\{x \in \mathbb{R}^n : V(x) \leq V(x^0)\}$. Since any accumulation point of S is also an accumulation point of $(x^k)_{k \in \mathbb{N}}$, S is closed. Hence S is compact.

We prove connectedness by contradiction. Suppose there are two disjoint, nonempty open sets A and B such that $S \subset A \cup B$. The sequence $(x^k)_{k \in \mathbb{N}}$ jumps between A and B infinitely many times and $\|x^{k+1} - x^k\| \to 0$, which implies that there is a subsequence of $(x^{k_j})_{j \in \mathbb{N}}$ in $\mathbb{R}^n \setminus (A \cup B)$. However, $(x^{k_j})_{j \in \mathbb{N}}$ is a bounded sequence and has an accumulation point, which must belong in $\mathbb{R}^n \setminus (A \cup B)$. This contradicts the assumption that all accumulation points of $(x^k)_{k \in \mathbb{N}}$ are in $A \cup B$.

We show that S has empty interior by contradiction. Suppose S contains an open ball $B_\varepsilon(x)$ in $\mathbb{R}^n$. Then as $\|x^{k+1} - x^k\| \to 0$, there is a $j \in \mathbb{N}$ such that $x^j \in B_\varepsilon(x) \subset S$. However, as V takes the same value on all of S, we deduce that $V(x^j) = \lim_{k \to \infty} V(x^x)$. Since $(V(x^k))_{k \in \mathbb{N}}$ is a decreasing sequence, $V(x^k) = V(x^j)$ for all $k > j$. It follows from (1.7) that $x^k = x^j$ for all $k > j$. Therefore, $S = \{x^j\}$, which contradicts the assumption that S has nonempty interior.

Last, since $(V(x^k))_{k \in \mathbb{N}}$ is a decreasing sequence and $V(x^*) = \lim_{k \to \infty} V(x^k)$ for all $x^* \in S$, it follows that V is constant on S. $\qquad\square$

3.2 Optimality result

We now proceed to the main result of this paper, namely that all points in the limit set S are Clarke stationary. We consider the stochastic case and the deterministic case separately.

In the stochastic case, we assume that the directions $(d^k)_{k \in \mathbb{N}}$ are randomly, independently drawn, and that the support of the probability density of Ξ is dense in S^{n-1}. It is straightforward to extend the proof to the case where $(d^{nk+1}, \ldots, d^{n(k+1)})$ are drawn as an orthonormal system under the assumptions that the directions $(d^{nk+1})_{k \in \mathbb{N}}$ are independently drawn from S^{n-1} and that the support of the density of the corresponding marginal distribution is dense in S^{n-1}.

We define X to be the set of nonstationary points,

$$X = \{x \in \mathbb{R}^n : 0 \notin \partial V(x)\}. \tag{3.1}$$

Theorem 3.5 *Let $(x^k)_{k \in \mathbb{N}}$ solve (1.6) where $(d^k)_{k \in \mathbb{N}}$ are independently drawn from the random distribution Ξ, and suppose that the support of the density of Ξ is dense in S^{n-1}. Then $\mathbb{P}(S \cap X \neq \varnothing) = 0$, i.e. the limit set S is almost surely in the set of stationary points.*

Proof. We will construct a countable collection of open sets $(B_j)_{j \in \mathbb{N}}$, such that $X \subset \bigcup_{j \in \mathbb{N}} B_j$ and so that for all $j \in \mathbb{N}$ we have $\mathbb{P}(S \cap B_j \neq \varnothing) = 0$. Then the result follows from countable additivity of probability measures.

First, we show that for every $x \in X$, there is $d \in S^{n-1}$, $\varepsilon > 0$, and $\delta > 0$ such that

$$\frac{V(y - \lambda e) - V(y)}{\lambda} \leq -\varepsilon, \quad \forall y \in B_\delta(x),\ e \in B_\delta(d) \cap S^{n-1},\ \lambda \in (0, \delta). \tag{3.2}$$

To show this, note that if $x \in X$, then by definition there is $d \in S^{n-1}$ and $\varepsilon > 0$ such that

$$V^o(x; -d) = \limsup_{\substack{y \to x \\ \lambda \downarrow 0}} \frac{V(y - \lambda d) - V(y)}{\lambda} \leq -\varepsilon.$$

Therefore, there is $\eta > 0$ such that for all $\lambda \in (0, \eta)$ and all $y \in B_\eta(x)$, we have

$$\frac{V(y - \lambda d) - V(y)}{\lambda} \leq -\varepsilon/2.$$

As V is Lipschitz continuous around $B_\eta(x)$, it is clear that the mapping

$$e \mapsto \frac{V(y - \lambda e) - V(y)}{\lambda},$$

is also locally Lipschitz continuous (of the same rank). It follows that there exists $\delta \in (0, \eta)$ such that for all $y \in B_\delta(x)$, all $e \in B_\delta(d) \cap S^{n-1}$, and all $\lambda \in (0, \delta)$, we have

$$\frac{V(y - \lambda e) - V(y)}{\lambda} \leq -\varepsilon/3.$$

This concludes the first part.

Next, for $m \in \mathbb{N}$, we define the set

$$X_m = \left\{ x \in X : \text{(3.2) holds for some } d \in S^{n-1}, \varepsilon > 0 \text{ and all } \delta < 1/m \right\}.$$

Clearly

$$X = \bigcup_{m \in \mathbb{N}} X_m.$$

Let $(y^i)_{i \in \mathbb{N}}$ be a dense sequence in X_m, which exists because $\mathbb{Q}^n$ is both countable and dense in $\mathbb{R}^n$. We define $Y_i^{(m)} = B_\delta(y^i)$, where $\delta = \frac{1}{m+1}$. Therefore,

$$X_m \subset \bigcup_{i \in \mathbb{N}} Y_i^{(m)} \quad \Longrightarrow \quad X \subset \bigcup_{m \in \mathbb{N}} \bigcup_{i \in \mathbb{N}} Y_i^{(m)}.$$

Since a countable union of countable sets is countable, we conclude with the following statement. For each $i \in \mathbb{N}$ there is $y^i \in \mathbb{R}^n$, $\varepsilon_i, \delta_i > 0$, and $d^i \in S^{n-1}$, such that for all $z \in B_{\delta_i}(y^i)$, all $e \in B_{\delta_i}(d^i) \cap S^{n-1}$, and all $\lambda \in (0, \delta_i)$, we have

$$\frac{V(z - \lambda e) - V(z)}{\lambda} \leq -\varepsilon_i,$$

and such that

$$X \subset \bigcup_{i \in \mathbb{N}} B_{\delta_i}(y^i). \tag{3.3}$$

Finally, we show that for each $i \in \mathbb{N}$, almost surely, $S \cap B_{\delta_i}(y^i) = \varnothing$. For a given i, write $B_i := B_{\delta_i}(y^i)$, and define $m := \min_{x \in B_j} V(x)$, $M := \max_{x \in B_j} V(x)$. We argue accordingly: The existence of an accumulation point of $(x^k)_{k \in \mathbb{N}}$ in B_i would imply that there is a subsequence $(x^{k_j})_{j \in \mathbb{N}} \subset B_i$. Suppose $x^{k_j} \in B_i$ and $d^{k_j} \in B_{\delta_i}(d^i)$, so that $x^{k_j+1} = x^{k_j} - \lambda d^{k_j}$ for some $\lambda > 0$. If $\lambda < \delta_i$, then

$$V(x^{k_j} - \lambda d^{k_j}) - V(x^{k_j}) \leq -\varepsilon_i \lambda = -\varepsilon_i \| x^{k_j+1} - x^{k_j} \|.$$

However,

$$V(x^{k_j+1}) - V(x^{k_j}) = -\frac{1}{\tau_{k_j}} \| x^{k_j+1} - x^{k_j} \|^2,$$

so, combining these equations, we get

$$\varepsilon_i \tau_{k_j} \leq \| x^{k_j+1} - x^{k_j} \|.$$

This in return implies

$$V(x^{k_j}) - V(x^{k_j+1}) \geq \varepsilon_i^2 \tau_{\min}.$$

On the other hand, if $\lambda \geq \delta_i$, then

$$V(x^{k_j}) - V(x^{k_j+1}) \geq \frac{\delta_i^2}{\tau_{\max}}.$$

Setting $\mu = \min \left\{ \varepsilon_i^2 \tau_{\min}, \frac{\delta_i^2}{\tau_{\max}} \right\}$, it follows that whenever $x^{k_j} \in B_i$ and $d^{k_j} \in B_{\delta_i}(d^i)$, then

$$V(x^{k_j}) - V(x^{k_j+1}) \geq \mu.$$

Choosing $K \in \mathbb{N}$ such that $K\mu > M - m$, we know that this event only has to occur K times for $(x^{k_j})_{j \in \mathbb{N}}$ to leave B_i. In other words, almost surely, there is no subsequence $(x^{k_j})_{j \in \mathbb{N}} \subset B_i$. This concludes the proof. $\quad\square$

3.2.1 Deterministic case

We now cover the deterministic case, in which $(d^k)_{k \in \mathbb{N}}$ is required to be *cyclically dense*.

Definition 3.6 *A sequence* $(d^k)_{k \in \mathbb{N}} \subset S^{n-1}$ *is* cyclically dense *in* S^{n-1} *if, for all* $\varepsilon > 0$*, there is* $N \in \mathbb{N}$ *such that for any* $k \in \mathbb{N}$*, the set* $\{d^k, \dots, d^{k+N-1}\}$ *forms an* ε*-cover of* S^{n-1},

$$S^{n-1} \subset \bigcup_{i=k+1}^{k+N-1} B_\varepsilon(d^i).$$

Remark 3.7 *Randomly drawn sequences are almost surely not cyclically dense, hence the separate treatment of the stochastic and deterministic methods.*

Many constructions of dense sequences are also cyclically dense. We provide an example of such a sequence on the unit interval $[0, 1]$.

Example 3.8 *Let* $\sigma \in (0, 1)$ *be an irrational number and define the sequence* $(\lambda_k)_{k \in \mathbb{N}}$ *in* $[0, 1]$ *by*

$$\lambda_k = (\sigma k) \pmod{1} = \sigma k - \lfloor \sigma k \rfloor,$$

where $\lfloor \sigma k \rfloor$ *denotes the largest integer less than or equal to* σk.

 To see that $(\lambda_k)_{k \in \mathbb{N}}$ *is cyclically dense in* $[0, 1]$*, set* $\varepsilon > 0$ *and note by sequential compactness of* $[0, 1]$ *that there is* $k, r \in \mathbb{N}$ *such that* $|\lambda_k - \lambda_{k+r}| < \varepsilon$*. We can write* $\delta = |\lambda_k - \lambda_{k+r}| > 0$*, where we know that* δ *is strictly positive, as no value can be repeated in the sequence due to* σ *being irrational. By modular arithmetic, we have for any* $l \in \mathbb{N}$,

$$\lambda_{k+rl} = \lambda_k + l\delta \pmod{1}.$$

In other words, the subsequence $(\lambda_{k+rl})_{l \in \mathbb{N}}$ *moves in increments of* $\delta < \varepsilon$ *on* $[0, 1]$*. Setting* $N = r \lceil \frac{1}{\delta} \rceil + k$*, where* $\lceil \delta \rceil$ *denotes the smallest integer greater than or equal to* δ*, it is clear that for any* $j \in \mathbb{N}$*, the set* $\{\lambda_j, \lambda_{j+1}, \dots, \lambda_{j+N-1}\}$ *forms an* ε*-cover of* $[0, 1]$.

 One could naturally extend this construction to higher dimensions $[0, 1]^n$*, by choosing* n *irrational numbers such that the ratio of any two of these numbers is also irrational.*

Theorem 3.9 *Let* $(x^k)_{k \in \mathbb{N}}$ *solve* (1.6)*, where* $(d^k)_{k \in \mathbb{N}}$ *are cyclically dense. Then all accumulation points* $x^* \in S$ *satisfy* $0 \in \partial V(x^*)$.

Proof. We consider the setup in the proof to Theorem 3.5, where X is the set of nonstationary points (3.1) and is covered by a countable collection of open balls (3.3),

$$X \subset \bigcup_{i \in \mathbb{N}} B_{\delta_i}(y^i).$$

We will show that an accumulation point $x^* \in S$ cannot belong to the ball $B_{\delta_i}(y^i)$, from which it follows that S is a subset of the set of stationary points. For contradiction, suppose that there is a subsequence $(x^{k_j})_{j \in \mathbb{N}} \to x^* \in B_{\delta_i}(y^i)$. By Lemma 3.3 *(iii)*, since $\|x^k - x^{k+1}\| \to 0$ as $k \to \infty$, we deduce that for any $N \in \mathbb{N}$, there is $j \in \mathbb{N}$ such that

$$\{x^{k_j}, x^{k_j+1}, \dots, x^{k_j+N-1}\} \subset B_{\delta_i}(y^i).$$

By cyclical density, we can choose N such that the corresponding directions $\{d^{k_j}, d^{k_j+1}, \dots, d^{k_j+N-1}\}$ form an ε_i-cover of S^{n-1}. Therefore, there exists $x^k \in B_{\delta_i}(y^i)$ and $d^k \in B_{\delta_i}(e^i)$, so we can argue as in Theorem 3.5, that

$$V(x^k) - V(x^{k+1}) \geq \mu,$$

where $\mu = \min\left\{\varepsilon_i^2 \tau_{\min}, \frac{\delta_i^2}{\tau_{\max}}\right\}$. If $(x^{k_j})_{j \in \mathbb{N}}$ had a limit in $B_{\delta_i}(y^i)$, this would happen arbitrarily many times, which is a contradiction. This concludes the proof. $\qquad\square$

3.3 Necessity of search density and Lipschitz continuity

For nonsmooth problems, it is necessary to employ a set of directions $(d^k)_{k\in\mathbb{N}}$ larger than the set of basis coordinates $\{e^1,\ldots,e^n\}$. To see this, consider the function $V(x,y) = \max\{x,y\}$ and the starting point $x^0 = [1,1]^T$. With the standard Itoh–Abe discrete gradient method, the iterates would remain at x^0, even though this point is nonstationary.

We show with a simple example that the assumption of density of $(d^k)_{k\in\mathbb{N}}$ in Theorem 3.5 is not only sufficient, but also necessary.

Example 3.10 *We suppose $V : \mathbb{R}^2 \to \mathbb{R}$ is defined by $V(x_1, x_2) = |x_1| + N|y_2|$ for some $N \in \mathbb{N}$, and set $x^0 = [-1,0]^T$. For $\theta \in [-\pi/2, \pi/2]$, let $d = [\cos\theta, \sin\theta]^T$. Then $-d$ is a direction of descent if and only if $\theta \in (-\arctan(1/N), \arctan(1/N))$. This interval can be made arbitrarily small by choosing N to be sufficiently large. Therefore, for an Itoh–Abe method to descend from x^0 for arbitrary functions, the directions $(d^k)_{k\in\mathbb{N}}$ need to include a convergent subsequence to the direction $[1,0]^T$. As this direction is arbitrary, we deduce that $(d^k)_{k\in\mathbb{N}}$ must be dense.*

Theorem 3.5 also assumes that V is locally Lipschitz continuous. We briefly discuss why this assumption is necessary, and provide an example to show that for functions that are merely continuous, the theorem no longer holds.

By Proposition 2.1.1. (b) in [16], the mapping $(y, d) \mapsto V^o(y; d)$ is upper semicontinuous for y in a neighbourhood of x, due to the local Lipschitz continuity of V near x. That is,

$$V^o(y^*; d^*) \geq \limsup_{y \to y^*, d \to d^*} V^o(y; d).$$

This property is crucial for the convergence analysis of Itoh–Abe methods, as it implies

$$V^o(x^*, d^*) \geq \limsup_{k \in \mathbb{N}} V^o(x^k; d^k) = 0.$$

Without local Lipschitz continuity, it is possible to have

$$x^k \to x^*, \quad d^k \to d^*, \text{ and } V^o(x^k; d^k) \to 0, \qquad \text{but } V^o(x^*; d^*) < 0.$$

In this case, there is no guarantee that the limit x^* is Clarke stationary. We demonstrate this with an example.

Example 3.11 *We will first state the iterates $(x^k)_{k\in\mathbb{N}}$ and then construct a function $V : \mathbb{R}^2 \to \mathbb{R}$ that fits these iterates. Let $(d^k)_{k\in\mathbb{N}}$ be a cyclically dense sequence in S^1 and assume without loss of generality that $[0,1]^T \notin (d^k)_{k\in\mathbb{N}}$. Replacing d^k with $-d^k$ does not change the step in (1.6), so we assume that $d_1^k < 0$ for all k. We set $x^0 = [0,0]^T$ and define $(x^k)_{k\in\mathbb{N}}$ and $(V(x^k))_{k\in\mathbb{N}}$ to be*

$$x^{k+1} = x^k - \frac{1}{(k+1)^2} d^k, \qquad V(x^{k+1}) = V(x^k) - \frac{1}{(k+1)^4}, \qquad V(x^0) = 0.$$

Since $\sum_{k\in\mathbb{N}} \|x^k - x^{k+1}\| < \infty$, it follows that x^k converges to some limit x^*, and $V(x^k)$ clearly decreases to a limit $V^* \in \mathbb{R}$. Furthermore, these steps satisfy (1.6) with $\tau_k = 1$. We then define V on the line segments $[x^k, x^{k+1}] = \{\lambda x^k + (1 - \lambda)x^{k+1} : \lambda \in [0,1]\}$ by interpolating linearly from $(x^k, V(x^k))$ to $(x^{k+1}, V(x^{k+1}))$.

Next, we define V on $\mathbb{R}^2$ as a function that linearly decreases everywhere in the direction $[0,1]^T$, and so that its value is consistent with the values given on the predefined line segments $[x^k, x^{k+1}]$. Note that this is a well-defined and continuous function, since each line in the direction $[0,1]^T$ crosses at most one point on at most one line segment, due to our assumptions on $(d^k)_{k\in\mathbb{N}}$.

We conclude the example by noting that the limit x^* is not Clarke stationary—in fact, no point is Clarke stationary—since $V^o(x; [0,1]^T) = -1$ for all x.

3.4 Nonsmooth, nonconvex functions with further regularity

For a large class of nonsmooth optimisation problems (convex and nonconvex), the objective function is sufficiently regular so that the standard Itoh–Abe discrete gradient method is also guaranteed to converge to Clarke stationary points. These are functions V for which $x^* \in \mathbb{R}^n$ is Clarke stationary if and only if $V^o(x^*; e^i) \geq 0$ for $i = 1, \ldots, n$. One example is functions of the form

$$V(x) = E(x) + \lambda \|x\|_1,$$

where E is a continuously differentiable function that may be nonconvex, $\|x\|_1$ denotes $|x_1| + \ldots + |x_n|$, and $\lambda > 0$. See for example Proposition 2.3.3 and the subsequent corollary in [16], combined with the fact that the nonsmooth component of V, i.e. $\|\cdot\|_1$, separates into n coordinate-wise scalar functions. This implies that the Clarke subdifferential is given by

$$\partial V(x) = \{\nabla E(x)\} + \lambda \prod_{i=1}^{n} \operatorname{sgn}(x_i),$$

where $\prod$ denotes the Cartesian product and

$$\operatorname{sgn}(x_i) := \begin{cases} \{1\}, & \text{if } x_i > 0, \\ \{-1\}, & \text{if } x_i < 0, \\ [-1, 1], & \text{if } x_i = 0. \end{cases}$$

Since this paper is chiefly concerned with the blackbox setting where no particular structure of V is assumed, we do not include a rigorous analysis of the convergence properties of the standard Itoh–Abe discrete gradient method for functions of the above form. However, we point out that for nonsmooth, nonconvex optimisation problems where Clarke stationarity is equivalent to Clarke directional stationarity along the standard coordinates, one can adapt Theorem 3.5 in a straightforward manner to prove that the iterates converge to a set of Clarke stationary points when the directions $(d^k)_{k \in \mathbb{N}}$ are drawn from the standard coordinates $(e^i)_{i=1}^{n}$.

Furthermore, one could drop the requirement that V is locally Lipschitz continuous, and replace $\|x\|_1$ with $\|x\|_p^p$, where $p \in (0, 1)$, and $\|x\|_p^p = |x_1|^p + \ldots + |x_n|^p$. This too is beyond the scope of this paper.

4 Rotated Itoh–Abe discrete gradients

We briefly discuss a randomised Itoh–Abe method that retains the Itoh–Abe discrete gradient structure, by ensuring that the directions $(d^{kn+1}, d^{kn+2}, \ldots, d^{k(n+1)})$ are orthonormal. Equivalently, we consider each block of n directions to be independently drawn from a random distribution on the set of orthogonal transformations on $\mathbb{R}^n$, denoted by $\mathrm{O}(n)$.

Definition 4.1 *The* orthogonal group of dimension n, $O(n)$, *is the set of orthogonal matrices in $\mathbb{R}^n$, i.e. matrices R which satisfy $R^{-1} = R^T$. Equivalently, R maps one orthonormal basis of $\mathbb{R}^n$ to another.*

Each element of $\mathrm{O}(n)$ corresponds to a *rotated* Itoh–Abe discrete gradient.

Definition 4.2 (Rotated Itoh–Abe discrete gradient) *Suppose $R \in O(n)$ maps the basis $(e^i)_{i=1}^{n}$ to another orthonormal basis $(f^i)_{i=1}^{n}$, i.e. $Rf^i = e^i$. For continuously differentiable functions V, the* rotated Itoh–Abe discrete gradient, *denoted by $\overline{\nabla}_R V$, is given by*

$$\overline{\nabla}_R V(x, y) = R^T \hat{\nabla}_R V(x, y),$$

where

$$\left(\hat{\nabla}_R V(x,y)\right)_i := \frac{V\left(x+\sum_{j=1}^{i}\langle y-x,f^j\rangle f^j\right) - V\left(x+\sum_{j=1}^{i-1}\langle y-x,f^j\rangle f^j\right)}{\langle y-x,f^i\rangle}.$$

It is straightforward to check that it is a discrete gradient, as defined for continuously differentiable functions V.

Proposition 4.3 $\overline{\nabla}_R V$ *is a discrete gradient.*

Proof. For any $x,y \in \mathbb{R}^n$, $x \neq y$,

$$\begin{aligned}
\langle \overline{\nabla}_R V(x,y), y-x \rangle & \\
&= \langle R^T \hat{\nabla}_R V(x,y), y-x \rangle \\
&= \langle \hat{\nabla}_R V(x,y), R(y-x) \rangle \\
&= \sum_{i=1}^{n} \frac{V\left(x+\sum_{j=1}^{i}\langle y-x,f^j\rangle f^j\right) - V\left(x+\sum_{j=1}^{i-1}\langle y-x,f^j\rangle f^j\right)}{\langle y-x,f^i\rangle} \cdot \langle y-x,f^i\rangle \\
&= \sum_{i=1}^{n} V\left(x+\sum_{j=1}^{i}\langle y-x,f^j\rangle f^j\right) - V\left(x+\sum_{j=1}^{i-1}\langle y-x,f^j\rangle f^j\right) \\
&= V(y) - V(x).
\end{aligned}$$

The convergence property $\lim_{y \to x} \overline{\nabla}_R V(x,y) = \nabla V(x)$ is immediate, providing V is continously differentiable. $\qquad\square$

Thus, we can implement schemes that are formally discrete gradient methods, and also fulfill the convergence theorem in Section 3.

5 Numerical implementation

We consider three ways of choosing $(d^k)_{k\in\mathbb{N}}$.

1. *Standard Itoh–Abe method.* The directions cycle through the standard coordinates, with the rule $d^k = e^{[(k-1) \bmod n]+1}$. Performing n steps of this method is equivalent to one step with the standard Itoh–Abe discrete gradient method.
2. *Random pursuit.* The directions are independently drawn from a random distribution Ξ on S^{n-1}. We assume that the support of the density of Ξ is dense in S^{n-1}.
3. *Rotated Itoh–Abe method.* For each $k \in \mathbb{N}$, the block of n directions $(d^{kn+1}, d^{kn+2}, \dots, d^{(k+1)n})$ is drawn from a random distribution on $O(n)$, the orthogonal group of dimension n. In other words, the directions form an orthonormal basis. This retains the discrete gradient structure of the standard Itoh–Abe discrete gradient method. We assume that each draw from $O(n)$ is independent, and, for notational continuity, we denote by Ξ the marginal distribution of d^{kn+1} on S^{n-1}, and again assume that the support of the density is dense in S^{n-1}.

We formalise an implementation of randomised Itoh–Abe methods with two algorithms, an inner and an outer one. Algorithm 3 is the inner algorithm and solves (1.6) for x^{k+1}, given x^k, d^k and time step bounds $\tau_{\min}, \tau_{\max}$. Algorithm 2 is the outer algorithm, which calls the inner algorithm for each iterate x^k, and provides a stopping rule for the methods. The stopping rule in Algorithm 2 takes two positive integers

K and M as parameters, such that the algorithm stops either after K iterations, or when the iterates have not sufficiently decreased V in the last M iterations. We typically set $M \approx n$, n being the dimension of the domain, unless the function V is expected to be highly irregular or nonsmooth, in which case we choose a larger M, as directions are generally prone to yield insufficient decrease. This stopping rule can be replaced by any other heuristic.

Algorithm 3 is a tailormade scalar solver for (1.6) that balances the tradeoff between optimally decreasing V given constraints $\tau_{\min}, \tau_{\max}$ and using minimal function evaluations. Rather than solving for a given τ_k, it ensures that there exists $\tau_k \in [\tau_{\min}, \tau_{\max}]$ that matches the output x^{k+1}. It requires a preliminary $\tau \in [\tau_{\min}, \tau_{\max}]$, which we heuristically chose as $\tau = \sqrt{\tau_{\min}\tau_{\max}}$. This method is particularly suitable when $\tau_{\min} \ll \tau_{\max}$, and can be replaced by any other scalar root finder algorithm.

The randomised Itoh–Abe methods have been implemented on Python and will be made available on GITHUB upon acceptance of this manuscript.

Algorithm 2 Randomised Itoh–Abe method with solver and stopping criterion

---

Input: starting point x^0, directions $(d^k)_{k\in\mathbb{N}}$, time step bounds $(\tau_{\min}, \tau_{\max})$, tolerance for function reduction η, maximal number of iterations K, maximal number of consecutive directions without descent before stopping M, internal solver described by Algorithm 3.
Initialise: counter $m = 0$.

---

for $k = 0, \ldots, K-1$ **do**
    Update $x^{k+1} \leftarrow (x^k, d^k, \tau_{\min}, \tau_{\max})$ via Algorithm 3
    if $V(x^k) - V(x^{k+1}) \leq \eta$ **then**
        $m = m + 1$
    else
        $m = 0$
    end if
    if $m \geq M$ **then**
        Terminate
    end if
end for

---

6 Examples

In this section, we use the randomised Itoh–Abe methods to solve several nonsmooth, nonconvex problems. In Section 6.1, we consider some well known optimisation challenges developed by Rosenbrock and Nesterov. In Section 6.2, we solve bilevel optimisation of parameters in variational regularisation problems.[4]

We compare our method to state-of-the-art derivative-free optimisation methods Py-BOBYQA [13, 56] and the LT-MADS solver provided by NOMAD [2, 41, 40]. For purposes of comparing results across solvers for these problems, we do not measure objective function value against iterates, but objective function value against function evaluations.

---

[4] Test images are taken from the Berkeley database [44]. Available online: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html.

Algorithm 3 Solver for Itoh–Abe step (1.6)

Input: current point x, direction d, time step upper bound $\tau_{\max}$, time step lower bound $\tau_{\min}$, predicted time step $\tau = \sqrt{\tau_{\min}\tau_{\max}}$, tolerance for x, ε, scalar $\sigma \in (0,1)$.

if $V(x + \varepsilon d) \geq V(x)$ **then**
    $d = -d$
    if $V(x + \varepsilon d) \geq V(x)$ **then**
        return x (stationary along d)
    end if
end if
Solve for β assuming linear extrapolation of V and with predicted τ (assume for simplicity $\beta > \varepsilon$):

$$\beta = -\frac{V(x + \varepsilon d) - V(x)}{\varepsilon\tau}$$

$$x^0 = x, \quad x^1 = x + \varepsilon d, \quad x^2 = x + \beta d$$

while V is concave between x^0, x^1 and x^2 (meaning $\frac{V(x^2) - V(x^1)}{x^2 - x^1} \leq \frac{V(x^1) - V(x^0)}{x^1 - x^0}$) **do**
    $\beta = \beta/\sigma$, $x^2 = x + \beta d$.
end while
Do step of parabolic interpolation (see [31, Section 6.2.2]) between x^0, x^1 and x^2, i.e.

$$y = x^1 - \frac{1}{2}\frac{(x^1 - x^0)^2(V(x^1) - V(x^2)) - (x^1 - x^2)^2(V(x^1) - V(x^0))}{(x^1 - x^0)(V(x^1) - V(x^2)) - (x^1 - x^2)(V(x^1) - V(x^0))}$$

while Parabolic step has not decreased V **do**
    Update parabolic interpolation points x^i, $i = 0, 1, 2$.
end while
$y = x^i$ is optimal point from parabolic interpolation step
while $\frac{|V(y) - V(x)|}{\|y - x\|^2} \notin \left[1/\tau_{\max}, 1/\tau_{\min}\right]$ **do**
    if $\frac{|V(y) - V(x)|}{\|y - x\|^2} > 1/\tau_{\min}$ **then**
        $y = y/\sigma$
    else
        $y = \sigma y$
    end if
end while
return y

6.1 Rosenbrock functions

We consider the well-known Rosenbrock function [60]

$$V(x, y) = (1 - x)^2 + 100(y - x^2)^2. \tag{6.1}$$

Its global minimiser $[1, 1]^T$ is located in a narrow, curved valley, which is challenging for the iterates to navigate. We compare the three variants of the Itoh–Abe method, for which we set the algorithm parameters $\varepsilon = 10^{-5}$, $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^2$, $\eta = 10^{-9}$, and $M = 30$. See Figure 6.1 for the numerical results. All three methods converge to the global minimiser, which shows that the Itoh–Abe methods are robust. Unsurprisingly, the random pursuit method and the rotated Itoh–Abe method, which descend in varying directions, perform significantly better than the standard Itoh–Abe method.

We also consider a nonsmooth variant of (6.1), termed Nesterov's (second) nonsmooth Chebyshev–Rosenbrock function [29],

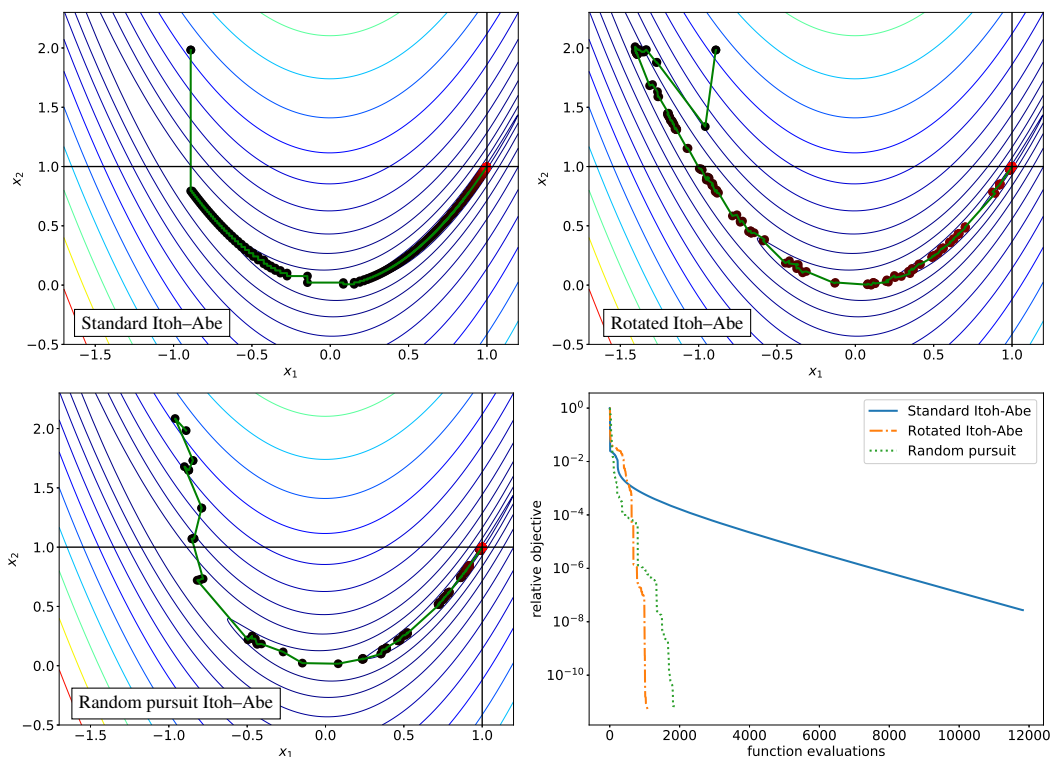$$V(x, y) = \frac{1}{4}|x - 1| + |y - 2|x| + 1|. \tag{6.2}$$

Fig. 6.1: Comparison of three variants of the Itoh–Abe method applied to the Rosenbrock function. Top left: Itoh–Abe method with standard frame. Top right: Rotated Itoh–Abe method. Bottom left: Itoh–Abe method with random pursuit. Bottom right: Convergence rates of the relative objective $\frac{V(x^k)-V^*}{V(x^0)-V^*}$ for the three variants, displayed with a log–log plot.

In this case too, the global minimiser $[1,1]^T$ is located along a narrow path. Furthermore, there is a nonminimising, stationary point at $[0,-1]^T$, which is nonregular—i.e. it has negative directional derivatives).

We also compare the three Itoh–Abe methods for this example, and set the algorithm parameters $\varepsilon = 10^{-10}$, $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^2$, $\eta = 10^{-16}$, and $M = 100$. See Figure 6.2 for the results from this. As can be seen, the standard Itoh–Abe discrete gradient method is not suitable for the irregular paths and nonsmooth kinks of the objective function, and stagnates early on. The two randomised Itoh–Abe methods perform better, as they descend in varying directions. For the remaining 2D problems in this paper, we will consider the rotated Itoh–Abe method, although we could just as well have used the random pursuit method. For higher-dimensional problems, we recommend the random pursuit method.

We also compare the performance of the randomised Itoh–Abe (RIA) method to Py-BOBYQA and LT-MADS for Nesterov's nonsmooth Chebyshev–Rosenbrock function. We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-10}$, $\tau_{\min} = 10^{-4}$, $\tau_{\max} = 10^2$, $\eta = 10^{-16}$, and $M = 100$, the parameters of Py-BOBYQA to rhobeg = 2, rhoend = 10^{-16} and npt = $(n+1)(n+2)/2$, and the parameters of LT-MADS to DIRECTION_TYPE = LT $2N$ and MIN_MESH_SIZE = 10^{-13}. See Figure 6.3 and 6.4 for the numerical results for two different starting points. In the first case, the Itoh–Abe method successfully converges to the global minimiser, the
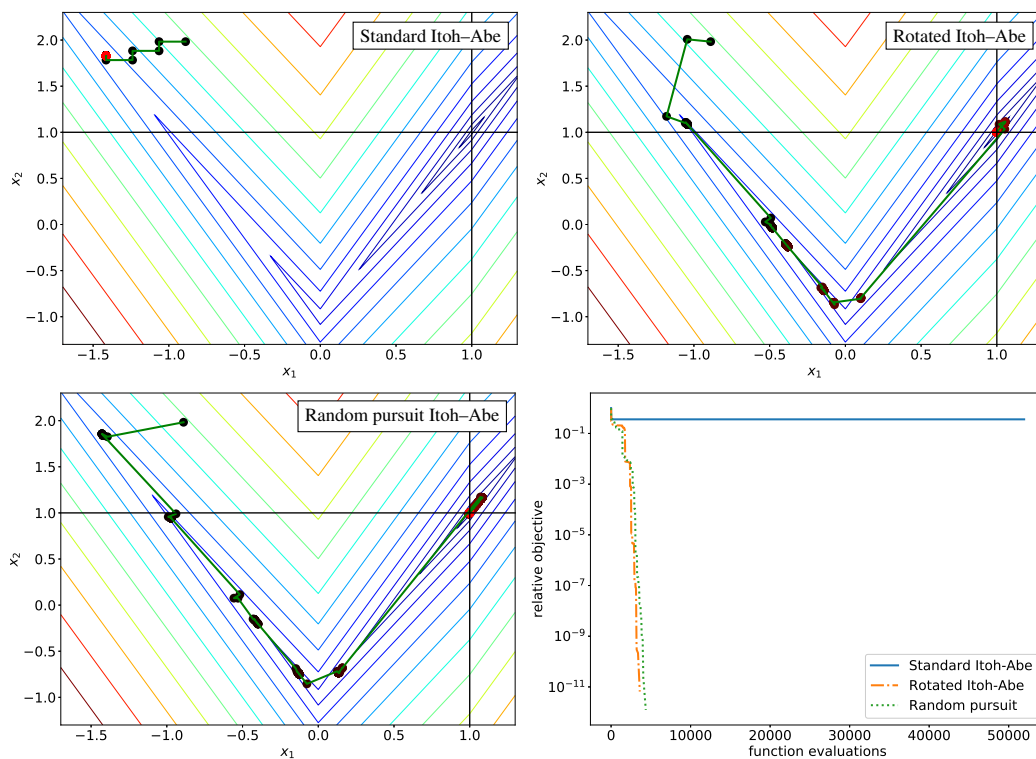
Fig. 6.2: Comparison of three variants of the Itoh–Abe method applied to Nesterov's nonsmooth Chebyshev–Rosenbrock function. Top left: Itoh–Abe method with standard frame. Top right: Rotated Itoh–Abe. Bottom left: Itoh–Abe with random pursuit. Bottom right: Convergence rates of the relative objective $\frac{V(x^k) - V^*}{V(x^0) - V^*}$ for the three variants, displayed with a log–log plot.

LT-MADS method locates the nonminimising stationary point at $[0, -1]^T$, while the Py-BOBYQA iterates stagnate at a kink, reflecting the fact that the method is not designed for nonsmooth functions. In the second case, both the Itoh–Abe method and LT-MADS locate the minimiser, while the Py-BOBYQA iterates stagnate at a kink.

6.2 Bilevel parameter learning in image analysis

In this subsection, we consider the Itoh–Abe method for solving bilevel optimisation problems for the learning of parameters of variational imaging problems. We restrict our focus to denoising problems, although the same method could be applied to any inverse problem. We first consider one-dimensional bilevel problems with wavelet and TV denoising, and two-dimensional problems with TGV denoising. In the TGV case, we compare the randomised Itoh–Abe method to the Py-BOBYQA and LT-MADS methods. Throughout this section, we set $M = n$, where $n = 1, 2$.
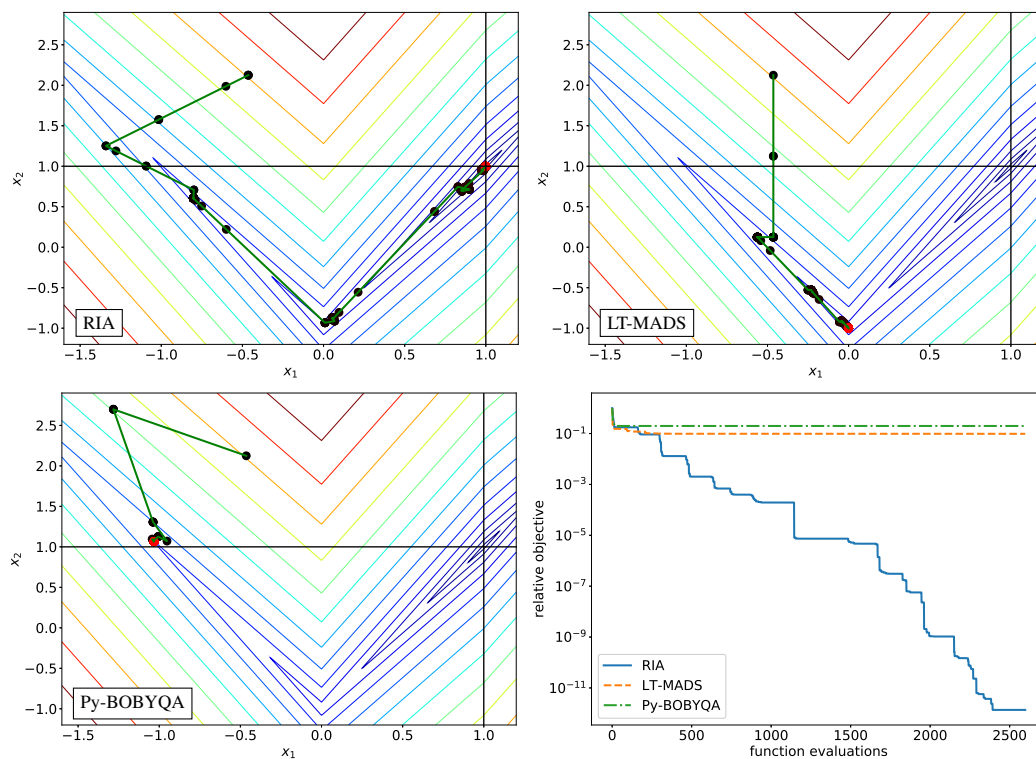
Fig. 6.3: Comparison of rotated Itoh–Abe method, LT-MADS and Py-BOBYQA applied to Nesterov's non-smooth Chebyshev–Rosenbrock function. Top left: The iterates from the Itoh–Abe method locate the unique minimiser to an order of accuracy of about 10^{-11}. Top right: The iterates from the LT-MADS method locate the nonminimising stationary point. Bottom left: The iterates from the Py-BOBYQA method stagnate due to nonsmoothness. Bottom right: A plot of the relative objective $\frac{V(x^k)-V^*}{V(x^0)-V^*}$ with respect to function evaluations, for each method.

6.2.1 Setup for variational regularisation problem

Consider an image $u^\dagger \in L^2(\Omega)$, for some domain $\Omega \subset \mathbb{R}^2$, and a noisy image

$$f^\delta = u^\dagger + \text{ noise}.$$

To recover a clean image from the noisy one, we consider a parametrised family of regularisers,

$$\left\{ R_\alpha : L^2(\Omega) \to [0,\infty] \; : \; \alpha \in [0,\infty)^n \right\},$$

and solve the variational regularisation problem

$$u_\alpha \in \underset{u}{\arg\min} \frac{1}{2} \| u - f^\delta \|^2 + R_\alpha(u). \tag{6.3}$$
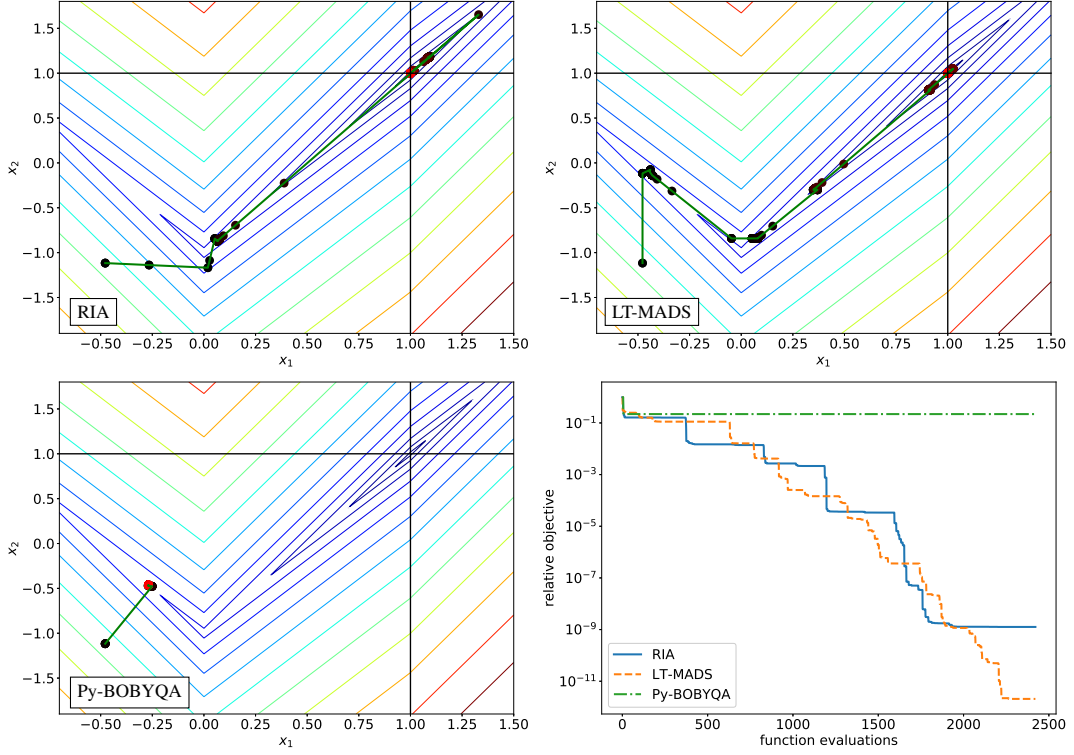
Fig. 6.4: Comparison of rotated Itoh–Abe method, LT-MADS and Py-BOBYQA applied to Nesterov's non-smooth Chebyshev–Rosenbrock function with a different starting point. Top left: The iterates from the Itoh–Abe method locate the unique minimiser to an order of accuracy of about 10^{-11}. Top right: The iterates from the Py-BOBYQA method stagnate due to nonsmoothness. Bottom left: The iterates from the LT-MADS method locate the nonminimising stationary point. Bottom right: A plot of the relative objective $\frac{V(x^k)-V^*}{V(x^0)-V^*}$ with respect to function evaluations, for each method.

The first term in (6.3), the *data fidelity* term, ensures that the reconstruction approximates f^δ. The regulariser term serves to denoise the reconstruction, by promoting favourable features such as smooth regions and sharp edges. The parameters α determine how heavily to regularise, and sometimes adjust other features of the regulariser. See [6, 33, 63] for an overview of variational regularisation methods.

We list some common regularisers in image analysis. *Total variation* (TV) [10, 61] is given by the function $R_\alpha(u) := \alpha \, \mathrm{TV}(u)$, where $\alpha \in [0,\infty)$, and

$$\mathrm{TV}(u) := \sup \left\{ \int_\Omega u(x) \, \mathrm{div}\,\phi(x) \, \mathrm{d}x \, : \, \phi \in C^1_c(\Omega;\mathbb{R}^d), \|\phi\|_\infty \le 1 \right\}.$$

This is one of the most common regularisers for image denoising. See Figure 6.5 for an example of denoising with TV regularisation. We also consider its second-order generalisation, *total generalised variation* [9, 8],
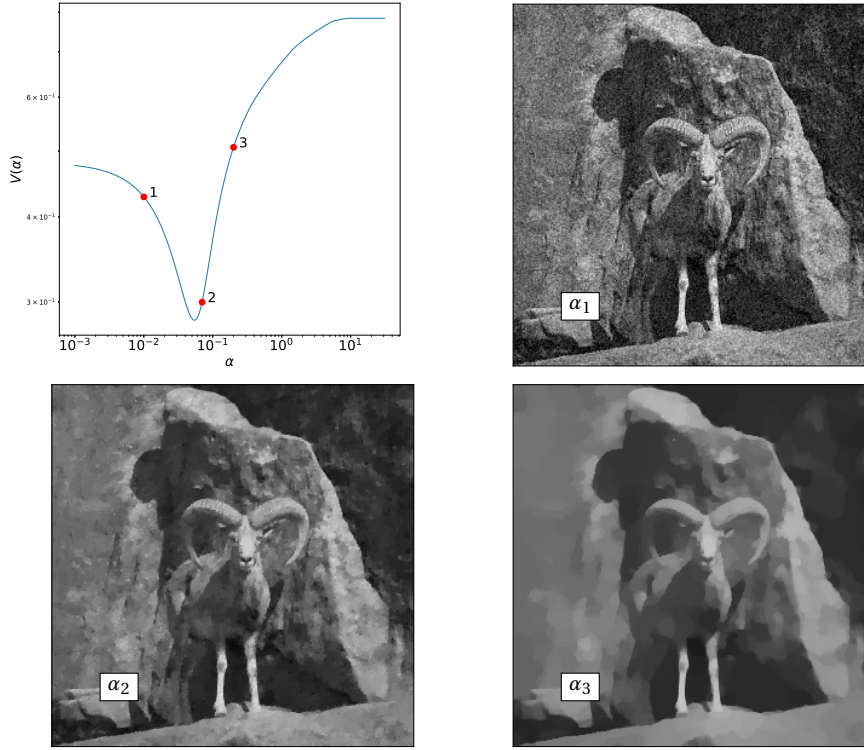
Fig. 6.5: TV denoising reconstructions for different regularisation parameters. Top left: Graph of V in (6.5). Top right: First parameter choice, $\alpha_1 = 10^{-2}$. Bottom left: Second parameter choice, $\alpha_2 = 7 \times 10^{-2}$. Bottom right: Third parameter choice, $\alpha_3 = 2 \times 10^{-1}$.

$R_\alpha(u) = \mathrm{TGV}_\alpha^2(u)$, where $\alpha = [\alpha_1, \alpha_2]^T \in [0, \infty)^2$ and

$$
\begin{aligned}
&\mathrm{TGV}_\alpha^2(u) \\
&:= \sup\left\{ \int_\Omega u(x)\,\mathrm{div}^2\,\phi(x)\,\mathrm{d}x \,:\, \phi \in C_c^2(\Omega; \mathrm{Sym}^2(\mathbb{R}^d)), \|\mathrm{div}^l\,\phi\|_\infty \leq \alpha_{l+1},\, l = 0, 1 \right\}.
\end{aligned}
$$

For a linear operator W on $L^2(\Omega)$, the basis pursuit regulariser

$$
R_\alpha(u) := \alpha \|Wu\|_1
$$

promotes sparsity of the image u in the dictionary of W.

As illustrated in Figure 6.5, the quality of the reconstruction is sensitive to α. If α is too low, the reconstruction is too noisy, while if α is too high, too much detail is removed. As it is generally not possible to ascertain the optimal choice of α a priori, a significant amount of time and effort is spent on parameter tuning. It is therefore of interest to improve our understanding of optimal parameter choices. One approach is to learn suitable parameters from training data. This requires a desired reconstruction $u^\dagger$, noisy data f^δ,

and a scoring function $\Phi : L^2(\Omega) \to \mathbb{R}$ that measures the error between $u^\dagger$ and the reconstruction u_α. The bilevel optimisation problem is given by

$$\alpha^* \in \underset{\alpha \in [0,\infty)^n}{\arg\min} \Phi(u_\alpha), \qquad \text{s.t. } u_\alpha \text{ solves (6.3).} \qquad (6.4)$$

In our case, we have strong convexity in the data fidelity term, which implies that u_α is unique for each $\alpha \in [0,\infty)^n$. We can therefore define a mapping

$$V(\alpha) := \Phi(u_\alpha). \qquad (6.5)$$

The bilevel problem (6.4) is difficult to tackle, both analytically as well as numerically. In most cases, the lower level problem (6.3) does not have a closed form formulation. Instead, a reconstruction u_α is approximated numerically with an algorithm. Therefore, one typically does not have access to gradient or subgradient information[5] for the mappings $\alpha \mapsto u_\alpha$. Furthermore, the bilevel mapping $\alpha \mapsto \Phi(u_\alpha)$ is often nonsmooth and nonconvex. Therefore, numerically solving (6.4) amounts to solving a nonsmooth, nonconvex function in a blackbox setting. We consider the application of the Itoh–Abe method for these problems.

For the numerical experiments in this paper, we reparametrise $V(\alpha)$ as $V(\exp(\alpha))$, where the exponential operator is applied elementwise on the parameters. There are two reasons for doing so. The first reason is that this paper is concerned with unconstrained optimisation, and this parametrisation allows is to optimise on $\mathbb{R}^n$ instead of $[0,\infty)^n$. The second reason is that $\exp(\alpha)$ has been found to be a preferable scaling for purposes of numerical optimisation.

6.2.2 Wavelet denoising

We consider the wavelet denoising problem

$$u_\alpha = \underset{u \in L^2(\Omega)}{\arg\min} \frac{1}{2}\|u - f^\delta\|^2 + \alpha\|Wu\|_1,$$

where W is a wavelet transform. In particular, W is an orthonormal basis, which implies that the regularisation problem has the unique solution

$$u_\alpha = W^{-1} T_\alpha(W f^\delta),$$

where T_α is the shrinkage operator defined by

$$[T_\alpha(v)]_i := \text{sgn}(v_i) \max\big(|v_i| - \alpha, 0\big).$$

We first optimise α for the scoring function

$$\Phi(u) := \frac{1}{2}\|u - u^\dagger\|^2.$$

We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-4}$, $\tau_{\min} = 10^{-1}$, $\tau_{\max} = 10$, and $\eta = 10^{-1}$. See Figure 6.6 for the numerical results.

---

[5] While automatic differentiation [27] can be useful for these purposes, it is in many cases not applicable to iterative algorithms that involve nonsmooth terms or for which the number of iterations cannot be predetermined. See [50] for further discussion of why derivative-free optimisation schemes are still needed.
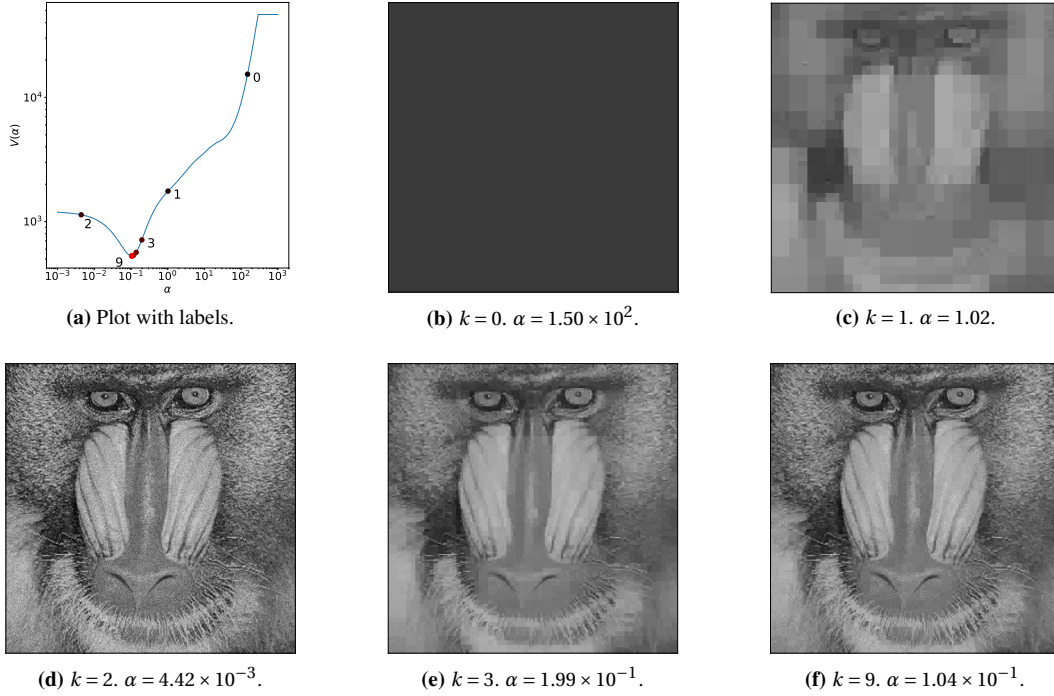
(a) Plot with labels.

(b) $k = 0$. $\alpha = 1.50 \times 10^2$.

(c) $k = 1$. $\alpha = 1.02$.

(d) $k = 2$. $\alpha = 4.42 \times 10^{-3}$.

(e) $k = 3$. $\alpha = 1.99 \times 10^{-1}$.

(f) $k = 9$. $\alpha = 1.04 \times 10^{-1}$.

Fig. 6.6: Wavelet denoising with L^2 scoring function and the Itoh–Abe method. Top left: Plot of iterates of the Itoh–Abe method. The rest: Image denoising results at different iterates k.

We also optimise α with respect to the scoring function $\Phi(u) := 1 - \text{SSIM}(u, u^\dagger)$, where SSIM is the *structural similarity* function [64]

$$\text{SSIM}(u, v) := \frac{(2\mu_u \mu_v + c)(2\sigma_{uv} + C)}{(\mu_u^2 + \mu_v^2 + c)(\sigma_u^2 + \sigma_v^2 + C)}.$$

Here μ_u is the mean intensity of u, σ_u is the unbiased estimate of the standard deviation of u, and σ_{uv} is the correlation coefficient between u and v:

$$\mu_u := \frac{1}{m} \sum_{i=1}^m u_i, \, \sigma_u := \left(\frac{1}{m-1} \sum_{i=1}^m (u_i - \mu_u)^2 \right)^{\frac{1}{2}}, \, \sigma_{uv} := \frac{1}{m-1} \sum_{i=1}^m (u_i - \mu_u)(v_i - \mu_v).$$

We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-4}$, $\tau_{\min} = 10^{-3}$, $\tau_{\max} = 10^3$, and $\eta = 10^{-2}$. See Figure 6.7 for the numerical results.

6.2.3 Total variation denoising

We consider the TV denoising problem

$$u_\alpha = \operatorname*{arg\,min}_{u \in L^2(\Omega)} \frac{1}{2} \| u - f^\delta \|^2 + \alpha \, \text{TV}(u),$$

(a) Plot with labels.

(b) $k = 0$. $\alpha = 10.0$.

(c) $k = 2$. $\alpha = 4.71$.

(d) $k = 4$. $\alpha = 1.15$.

(e) $k = 5$. $\alpha = 6.23 \times 10^{-2}$.

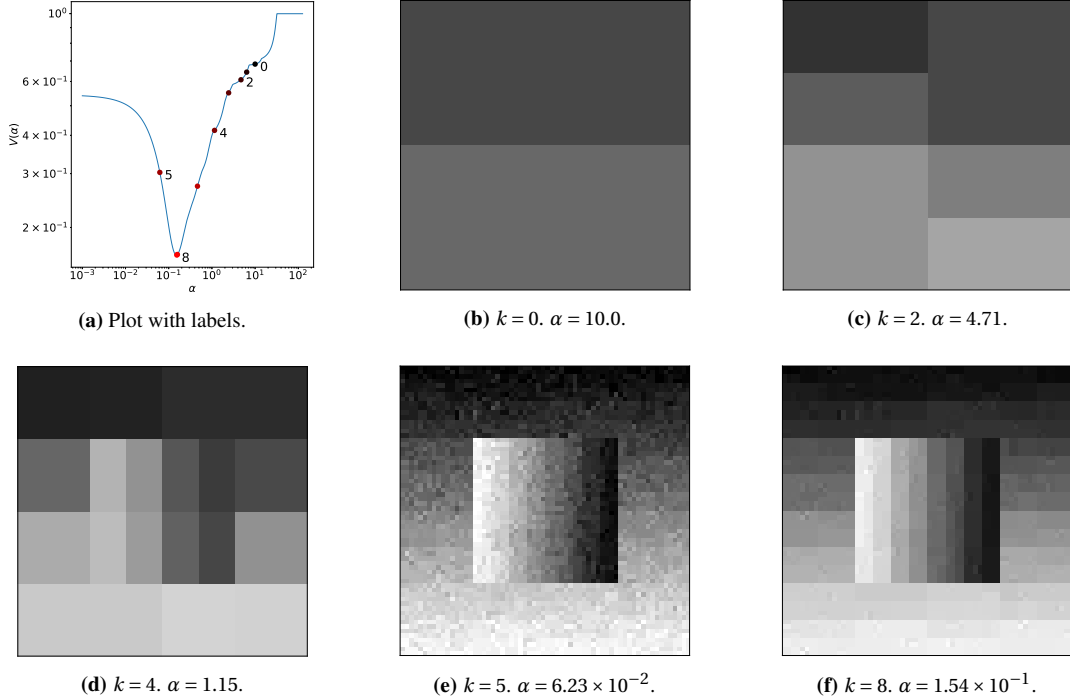**(f)** $k = 8$. $\alpha = 1.54 \times 10^{-1}$.

Fig. 6.7: Wavelet denoising with SSIM scoring function and the Itoh–Abe method. Top left: Plot of iterates of the Itoh–Abe method. The rest: Image denoising result at different iterates k.

with the SSIM scoring function. We solve the above denoising problem using the PDHG method [14]. We set the parameters of the Itoh–Abe method to $\varepsilon = 10^{-4}$, $\tau_{\min} = 10^{-5}$, $\tau_{\max} = 9 \times 10^{-4}$, and $\eta = 10^{-5}$. See Figure 6.8 for the numerical results.

6.2.4 Total generalised variation regularisation

We now consider the second-order total generalised variation (TGV) regulariser $R_{\alpha_1, \alpha_2}(u) = \mathrm{TGV}^2_{\alpha_1, \alpha_2}(u)$ for image denoising, with the scoring function

$$\Phi(u) := 1 - \mathrm{SSIM}(u, u^\dagger).$$

Like for TV denoising, we solve the denoising problem using the PDHG method. We set the parameters of the randomised Itoh–Abe (RIA) method to $\varepsilon = 10^{-1}$, $\tau_{\min} = 10^{-3}$, $\tau_{\max} = 10^5$, and $\eta = 10^{-20}$. See Figure 6.9 for the numerical results.

We compare these results to the results from the Py-BOBYQA and LT-MADS solvers. We set the parameters of Py-BOBYQA to rhobeg = 2, rhoend = 10^{-10} and npt = $2(n+1)$ and the parameters of LT-MADS to DIRECTION_TYPE = LT $2N$. See the results for two different starting points in Figure 6.10 and 6.11. We note that the objective function is approximately stationary across a range of values, which leads to the different points of convergence, and different limiting values of the objective function for different methods. We see that the methods are all of comparable efficiency, although the Itoh–Abe method is slower initially. The Itoh–Abe method seems to be the most efficient, once it is within a neighborhood of the minimiser.
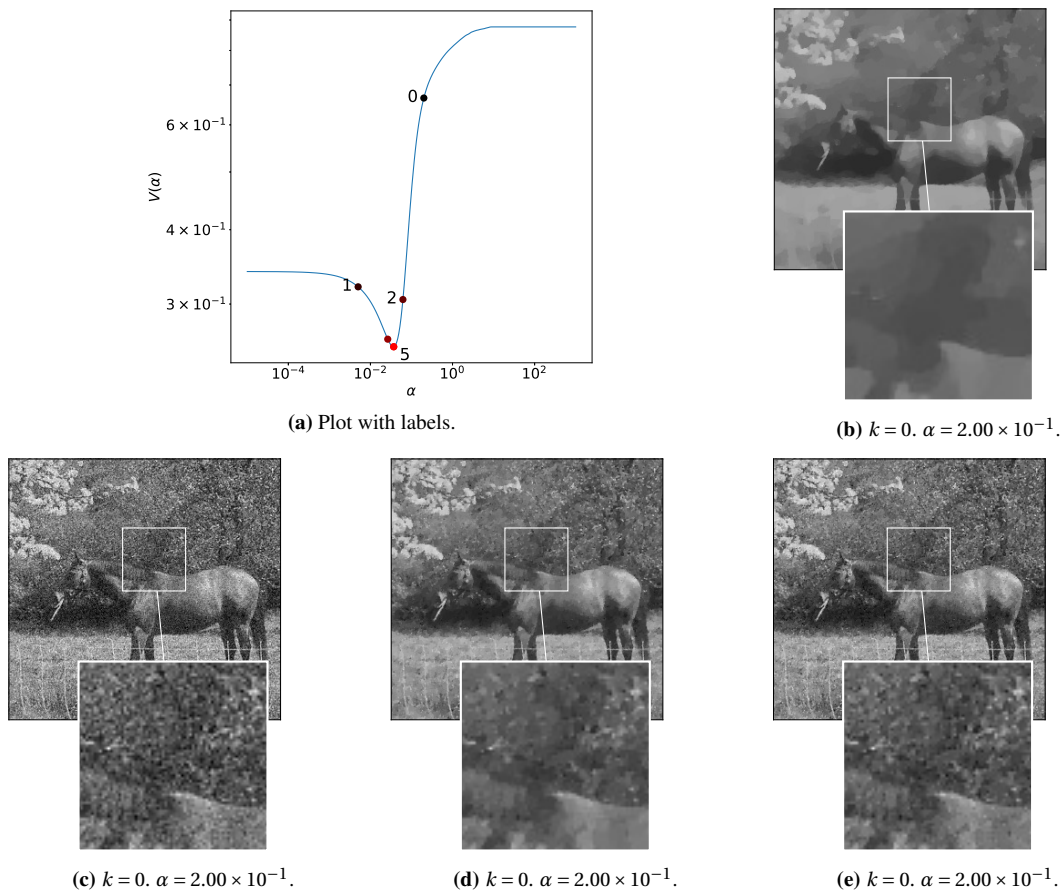
(a) Plot with labels.

(b) $k = 0$. $\alpha = 2.00 \times 10^{-1}$.

(c) $k = 0$. $\alpha = 2.00 \times 10^{-1}$.

(d) $k = 0$. $\alpha = 2.00 \times 10^{-1}$.

(e) $k = 0$. $\alpha = 2.00 \times 10^{-1}$.

Fig. 6.8: TV denoising with SSIM scoring function and the Itoh–Abe method. Top left: Plot of iterates of the Itoh–Abe method. The rest: Image denoising result at different iterates k, with a zoom to show the difference.

7 Conclusion

In this paper, we have shown that the randomised Itoh–Abe methods are efficient and robust schemes for solving unconstrained nonsmooth, nonconvex problems without the use of gradients or subgradients. Furthermore, the favourable rates of dissipativity that the discrete gradient method inherits from the gradient flow system extends to the nonsmooth case. We show, under minimal assumptions on the objective function, that the methods admit a solution that is computationally tractable, and the iterates converge to a connected set of Clarke stationary points. Through examples, the assumptions are also shown to be necessary.

The methods are shown to be robust and versatile optimisation schemes. It locates the global minimisers of the Rosenbrock function and a variant of Nesterov's nonsmooth Chebyshev–Rosenbrock functions. The efficiency of the Itoh–Abe discrete gradient method for smooth problems has already been demonstrated elsewhere [28, 48, 59]. We also consider its application to bilevel learning problems and compare its performance to the derivative-free Py-BOBYQA and LT-MADS methods.
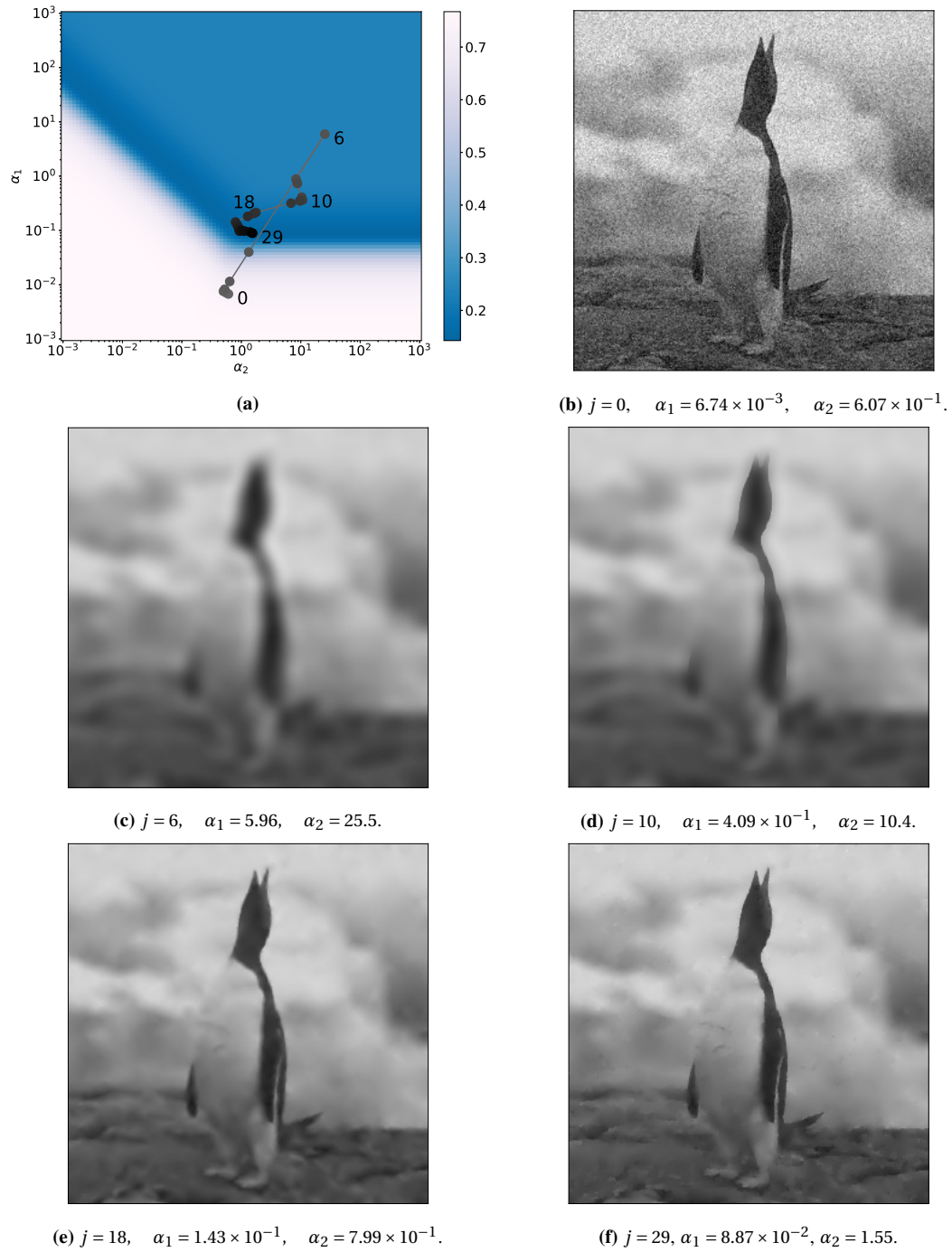
(a)

(b) $j = 0$, $\quad \alpha_1 = 6.74 \times 10^{-3}$, $\quad \alpha_2 = 6.07 \times 10^{-1}$.

(c) $j = 6$, $\quad \alpha_1 = 5.96$, $\quad \alpha_2 = 25.5$.

(d) $j = 10$, $\quad \alpha_1 = 4.09 \times 10^{-1}$, $\quad \alpha_2 = 10.4$.

(e) $j = 18$, $\quad \alpha_1 = 1.43 \times 10^{-1}$, $\quad \alpha_2 = 7.99 \times 10^{-1}$.

(f) $j = 29$, $\alpha_1 = 8.87 \times 10^{-2}$, $\alpha_2 = 1.55$.

Fig. 6.9: TGV denoising with SSIM scoring function and the Itoh–Abe method. Top left: Plot of iterates of the method. The rest: Image denoising result at different function evaluations j.
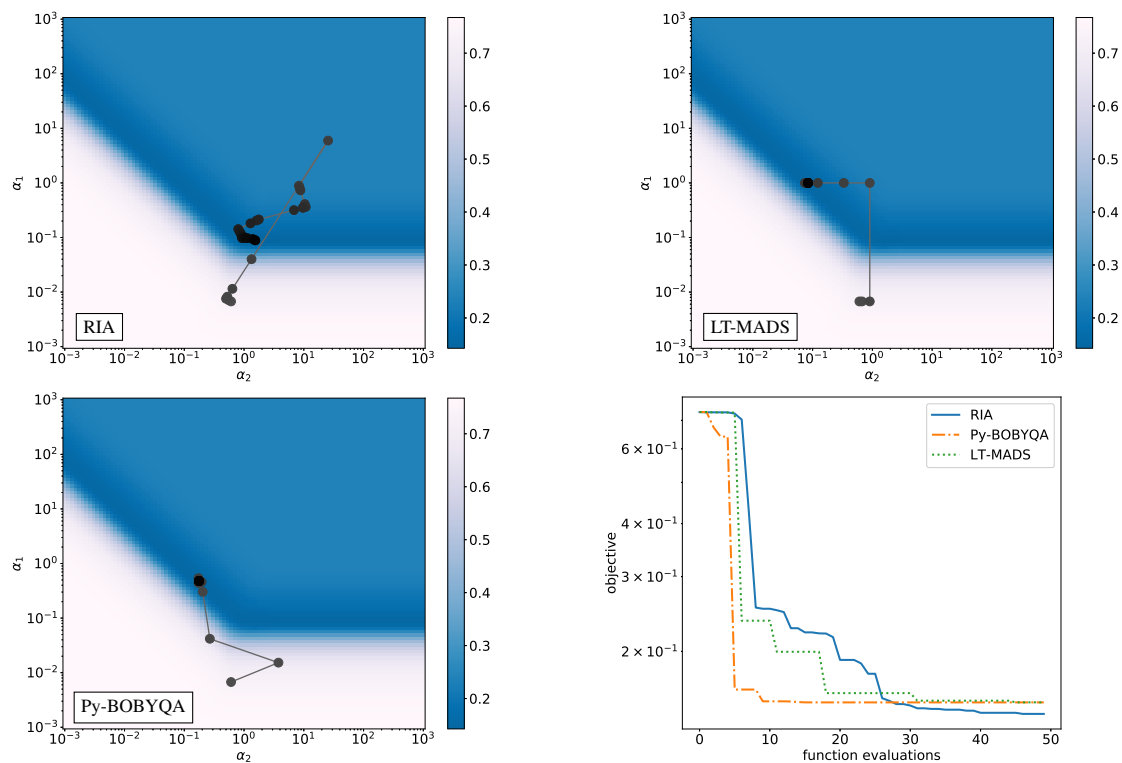
Fig. 6.10: Comparison of optimisation methods for TGV denoising with SSIM scoring function. Top left: Plot of iterates of the Itoh–Abe method. Top right: Plot of iterates of the LT-MADS method. Bottom left: Plot of iterates of the Py-BOBYQA method. Bottom right: Comparison of convergence rates for the methods with respect to function evaluations.

Future work will be dedicated to adapting the randomised Itoh–Abe methods for constrained optimisation problems, establishing convergence of the iterates of the method for Kurdyka-Łojasiewicz functions [1], and analysing the Lipschitz continuity properties of bilevel optimisation for variational regularisation problems.

References

1. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. Mathematics of Operations Research **35**(2), 438–457 (2010)
2. Audet, C., Dennis Jr, J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM Journal on optimization **17**(1), 188–217 (2006)
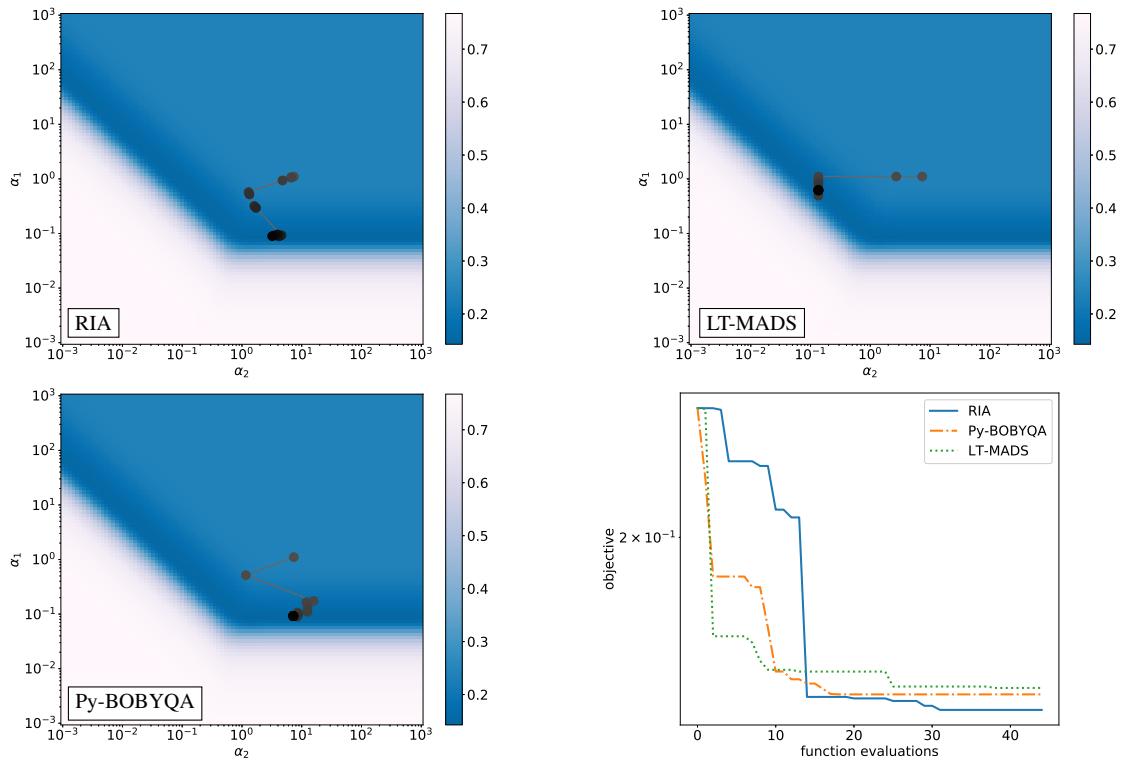
Fig. 6.11: Comparison of optimisation methods for TGV denoising with SSIM scoring function for a different starting point. Top left: Plot of iterates of the Itoh–Abe method. Top right: Plot of iterates of the LT-MADS method. Bottom left: Plot of iterates of the Py-BOBYQA method. Bottom right: Comparison of convergence rates for the methods with respect to function evaluations.

3. Audet, C., Hare, W.: Derivative-Free and Blackbox Optimization, 1st edn. Springer Series in Operations Research and Financial Engineering. Springer International Publishing (2017)

4. Aussel, D.: Subdifferential properties of quasiconvex and pseudoconvex functions: unified approach. Journal of optimization theory and applications **97**(1), 29–45 (1998)

5. Bagirov, A.M., Karasözen, B., Sezer, M.: Discrete gradient method: derivative-free method for nonsmooth optimization. Journal of Optimization Theory and Applications **137**(2), 317–334 (2008)

6. Benning, M., Burger, M.: Modern Regularization Methods for Inverse Problems. ArXiv e-prints (2018). URL http://arxiv.org/abs/1801.09922

7. Borwein, J.M., Zhu, Q.J.: A survey of subdifferential calculus with applications. Nonlinear Analysis: Theory, Methods & Applications **38**(6), 687–773 (1999)

8. Bredies, K., Holler, M.: A TGV-based framework for variational image decompression, zooming, and reconstruction. part I: Analytics. SIAM Journal on Imaging Sciences **8**(4), 2814–2850 (2015)

9. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM Journal on Imaging Sciences **3**(3), 492–526 (2010)

10. Burger, M., Osher, S.: A guide to the TV zoo. In: Level set and PDE based reconstruction methods in imaging, pp. 1–70. Springer, Berlin (2013)

11. Burke, J.V., Curtis, F.E., Lewis, A.S., Overton, M.L., Simões, L.E.: Gradient sampling methods for nonsmooth optimization. arXiv e-prints (2018). URL http://arxiv.org/abs/1804.11003
12. Calatroni, L., Chung, C., De Los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: Bilevel approaches for learning of variational imaging models. ArXiv e-prints (2015). URL http://arxiv.org/abs/1505.02120
13. Cartis, C., Fiala, J., Marteau, B., Roberts, L.: Improving the Flexibility and Robustness of Model-Based Derivative-Free Optimization Solvers. ArXiv e-prints (2018). URL http://arxiv.org/abs/1804.00154
14. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. Journal of mathematical imaging and vision **40**(1), 120–145 (2011)
15. Clarke, F.H.: Necessary conditions for nonsmooth problems in optimal control and the calculus of variations. Ph.D. thesis, University of Washington (1973)
16. Clarke, F.H.: Optimization and Nonsmooth Analysis, 1st edn. Classics in Applied Mathematics. SIAM, Philadelphia (1990)
17. Curtis, F.E., Que, X.: An adaptive gradient sampling algorithm for non-smooth optimization. Optimization Methods and Software **28**(6), 1302–1324 (2013)
18. Curtis, F.E., Que, X.: A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. Mathematical Programming Computation **7**(4), 399–428 (2015)
19. DuPont, B., Cagan, J.: A hybrid extended pattern search/genetic algorithm for multi-stage wind farm optimization. Optimization and Engineering **17**(1), 77–103 (2016)
20. Ehrhardt, M.J., Riis, E.S., Ringholm, T., Schönlieb, C.B.: A geometric integration approach to smooth optimisation: Foundations of the discrete gradient method. ArXiv e-prints (2018). URL http://arxiv.org/abs/1805.06444
21. Ekeland, I., Téman, R.: Convex Analysis and Variational Problems, 1st edn. SIAM, Philadelphia, PA, USA (1999)
22. Fasano, G., Liuzzi, G., Lucidi, S., Rinaldi, F.: A linesearch-based derivative-free approach for nonsmooth constrained optimization. SIAM Journal on Optimization **24**(3), 959–992 (2014)
23. Fowler, K.R., Reese, J.P., Kees, C.E., Dennis Jr, J., Kelley, C.T., Miller, C.T., Audet, C., Booker, A.J., Couture, G., Darwin, R.W., et al.: Comparison of derivative-free optimization methods for groundwater supply and hydraulic capture community problems. Advances in Water Resources **31**(5), 743–757 (2008)
24. Giles, J.R.: A survey of Clarke's subdifferential and the differentiability of locally Lipschitz functions. In: Progress in Optimization, pp. 3–26. Springer, Boston (1999)
25. Gonzalez, O.: Time integration and discrete Hamiltonian systems. Journal of Nonlinear Science **6**(5), 449–467 (1996)
26. Gray, G.A., Kolda, T.G., Sale, K., Young, M.M.: Optimizing an empirical scoring function for transmembrane protein structure determination. INFORMS Journal on Computing **16**(4), 406–418 (2004)
27. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2008)
28. Grimm, V., McLachlan, R.I., McLaren, D.I., Quispel, G.R.W., Schönlieb, C.B.: Discrete gradient methods for solving variational image regularisation models. Journal of Physics A: Mathematical and Theoretical **50**(29), 295201 (2017)
29. Gürbüzbalaban, M., Overton, M.L.: On Nesterov's nonsmooth Chebyshev–Rosenbrock functions. Nonlinear Analysis: Theory, Methods & Applications **75**(3), 1282–1289 (2012)
30. Hairer, E., Lubich, C., Wanner, G.: Geometric numerical integration: structure-preserving algorithms for ordinary differential equations, vol. 31, 2nd edn. Springer Science & Business Media, Berlin (2006)

31. Heath, M.T.: Scientific Computing: An Introductory Survey, 1st edn. McGraw-Hill, New York (2002)
32. Hintermüller, M.: A proximal bundle method based on approximate subgradients. Computational Optimization and Applications **20**(3), 245–266 (2001)
33. Ito, K., Jin, B.: Inverse Problems: Tikhonov Theory And Algorithms, 1st edn. Series On Applied Mathematics. World Scientific Publishing Company, Singapore (2014)
34. Itoh, T., Abe, K.: Hamiltonian-conserving discrete canonical equations based on variational difference quotients. Journal of Computational Physics **76**(1), 85–102 (1988)
35. Kämpf, J.H., Robinson, D.: Optimisation of building form for solar energy utilisation using constrained evolutionary algorithms. Energy and Buildings **42**(6), 807–814 (2010)
36. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer (2016)
37. Kiwiel, K.C.: Methods of descent for nondifferentiable optimization, vol. 1133, 1st edn. Springer, Berlin (1985)
38. Kiwiel, K.C.: A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM Journal on Optimization **20**(4), 1983–1994 (2010)
39. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM Journal on Imaging Sciences **6**(2), 938–983 (2013)
40. Le Digabel, S.: Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. ACM Transactions on Mathematical Software **37**(4), 1–15 (2011). URL https://www.gerad.ca/nomad/Project/Home.html
41. Le Digabel, S., Tribes, C., Audet, C.: NOMAD user guide. technical report g-2009-37. Tech. rep., Les cahiers du GERAD (2009)
42. Lewis, A.S., Overton, M.L.: Nonsmooth optimization via quasi-Newton methods. Mathematical Programming **141**, 135–163 (2013)
43. Liuzzi, G., Truemper, K.: Parallelized hybrid optimization methods for nonsmooth problems using NOMAD and linesearch. Computational and Applied Mathematics **37**, 3172–3207 (2018)
44. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proceedings of the 8th International Conference on Computer Vision, vol. 2, pp. 416–423. IEEE (2001)
45. McLachlan, R.I., Quispel, G.R.W.: Six lectures on the geometric integration of ODEs, p. 155–210. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge (2001)
46. McLachlan, R.I., Quispel, G.R.W., Robidoux, N.: Geometric integration using discrete gradients. Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **357**(1754), 1021–1045 (1999)
47. Michel, P., Penot, J.P.: Calcul sous-différentiel pour des fonctions lipschitziennes et non lipschitziennes. C. R. Acad. Sci. Paris **1**, 269–272 (1984)
48. Miyatake, Y., Sogabe, T., Zhang, S.L.: On the equivalence between SOR-type methods for linear systems and discrete gradient methods for gradient systems. ArXiv e-prints (2017). URL http://arxiv.org/abs/1711.02277
49. Nelder, J.A., Mead, R.: A simplex method for function minimization. The computer journal **7**(4), 308–313 (1965)
50. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Foundations of Computational Mathematics **17**(2), 527–566 (2017)
51. Ochs, P., Ranftl, R., Brox, T., Pock, T.: Bilevel optimization with nonsmooth lower level problems. In: International Conference on Scale Space and Variational Methods in Computer Vision, pp. 654–665.

Springer (2015)

52. Oeuvray, R., Bierlaire, M.: A new derivative-free algorithm for the medical image registration problem. International Journal of Modelling and Simulation **27**(2), 115–124 (2007)

53. Penot, J.P., Quang, P.H.: Generalized convexity of functions and generalized monotonicity of set-valued maps. Journal of Optimization Theory and Applications **92**(2), 343–356 (1997)

54. Polyak, B.T.: Introduction to Optimization, 1st edn. Optimization Software, Inc., New York (1987)

55. Powell, M.J.D.: The NEWUOA software for unconstrained optimization without derivatives. In: Large-scale nonlinear optimization, 1st edn., pp. 255–297. Springer, Boston (2006)

56. Powell, M.J.D.: The BOBYQA algorithm for bound constrained optimization without derivatives. Tech. rep., University of Cambridge (2009)

57. Quispel, G.R.W., Turner, G.S.: Discrete gradient methods for solving ODEs numerically while preserving a first integral. Journal of Physics A: Mathematical and General **29**(13), 341–349 (1996)

58. De los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: Bilevel parameter learning for higher-order total variation regularisation models. Journal of Mathematical Imaging and Vision **57**(1), 1–25 (2017)

59. Ringholm, T., Lazić, J., Schönlieb, C.B.: Variational image regularization with Euler's elastica using a discrete gradient scheme. ArXiv e-prints (2017). URL http://arxiv.org/abs/1712.07386

60. Rosenbrock, H.H.: An automatic method for finding the greatest or least value of a function. The Computer Journal **3**(3), 175–184 (1960)

61. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)

62. Rudin, W.: Principles of Mathematical Analysis, 3rd edn. International series in pure and applied mathematics. McGraw-Hill, New York (1976)

63. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging, 1st edn. Applied Mathematical Sciences. Springer, New York (2008)

64. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)