Foundations of Computational Mathematics (2022) 22:1351–1394
https://doi.org/10.1007/s10208-020-09489-2

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

A Geometric Integration Approach to Nonsmooth, Nonconvex Optimisation

Erlend S. Riis[1] · Matthias J. Ehrhardt[2] · G. R. W. Quispel[3] · Carola-Bibiane Schönlieb[1]

Received: 20 July 2018 / Revised: 22 November 2020 / Accepted: 22 November 2020 /
Published online: 29 July 2021

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 can be superior to state-of-the-art derivative-free optimisation methods in solving nonsmooth problems while still remaining competitive in terms of efficiency.

Keywords Geometric numerical integration · Discrete gradient methods · Derivative-free optimisation · Nonconvex optimisation · Nonsmooth optimisation · Clarke subdifferential · Bilevel optimisation

Communicated by Arieh Iserles.

All authors acknowledges support from the European Union Horizon 2020 research and innovation programmes under the Marie Skłodowska-Curie Grant Agreement No. 691070. ESR, MJE, and CBS acknowledges support from the Cantab Capital Institute for the Mathematics of Information. ESR acknowledges support from London Mathematical Society. MJE and CBS acknowledges support from Leverhulme Trust project "Breaking the non-convexity barrier", EPSRC Grant "EP/M00483X/1", and EPSRC centre "EP/N014588/1". GRWQ acknowledges support from the Australian Research Council and is grateful to the Mittag–Leffler Institute for a productive stay. Moreover, CBS acknowledges support from the RISE project NoMADS and the Alan Turing Institute. MJE acknowledges support from the EPSRC (EP/S026045/1) and the Faraday Institution (EP/T007745/1).

Extended author information available on the last page of the article

Mathematics Subject Classification 49M25 · 49Q15 · 65K10 · 90C15 · 90C26 · 90C56 · 94A08

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 $\{x \in \mathbb{R}^n : V(x) \leq M\}$ 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. [33] 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 in 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

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

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 while also preserving structures of the continuous system—see [35,52]

[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.

for an introduction. Discrete gradient methods can be applied to the first-order gradient systems to preserve energy conservation laws, dissipation laws, as well as Lyapunov functions [30,40,53,65]. 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:

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

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. [33] 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* [40] (also known as coordinate increment discrete gradient)[2] is defined for differentiable functions V as:

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

where $[\overline{\nabla} V(x, y)]_i := [\nabla V(y_1, \ldots, y_i, x_{i+1}, \ldots, x_n)]_i$ if $y_i = x_i$. Solving an iterate of the discrete gradient method (1.4) with the Itoh–Abe discrete gradient is equivalent

[2] There are infinitely many discrete gradients, each with a corresponding discrete gradient method. See [33] for further examples.

to successively solving n scalar equations of the form:

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

or $x_i^{k+1} = x_i^k$ in the cases where $[\nabla V(x_1^{k+1}, \dots, x_{i-1}^{k+1}, x_i^k, \dots, x_n^k)]_i = 0$, i.e. when there is directional stationarity.

To make this scheme meaningful for nondifferentiable V, we only need to adapt the criteria of directional stationarity, as the scheme is otherwise derivative-free. For this we use Clarke directional stationarity, Definition 2.7. We proceed to describe an extension of the Itoh–Abe discrete gradient method for nonsmooth functions.

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 of descent, 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

$$x^{k+1} = \begin{cases} x^k, & \text{if } V \text{ is stationary at } x^k \text{ along } d^k \text{ (see Definition 2.7), or} \\ x^k + \tau_k \beta_k d^k, & \text{where } \beta_k \neq 0 \text{ solves } \beta_k = -\frac{V(x^k + \tau_k \beta_k d^k) - V(x^k)}{\tau_k \beta_k}. \end{cases} \tag{1.6}$$

For notational brevity, we write $dV(x, y) := (V(x) - V(y)/\|x - y\|$ when $x \neq y$ and $dV(x, x) := 0$.

We formalise this method in Algorithm 1. Note that the two conditions in (1.6) are not mutually exclusive, but at least one will hold (Lemma 3.1). 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.

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

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:

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

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

Ehrhardt et al. [24] studied discrete gradient methods in the smooth setting, asserting linear convergence rates for functions that satisfy the Polyak–Łojasiewicz inequality [42]. Benning et al. [7] propose a Bregman Itoh–Abe optimisation scheme by applying the Itoh–Abe discrete gradient to the inverse scale space flow, enabling one to speed up convergence by promoting structural priors such as sparsity. Celledoni et al. [15] extend the Itoh–Abe discrete gradient method to optimisation on Riemannian manifolds. Furthermore, the application of Itoh–Abe discrete gradient methods to smooth optimisation problems is well documented and includes convex variational regularisation problems for image analysis [33], nonconvex image inpainting problems with Euler's elastica regularisation [66], for which it outperformed gradient-based schemes, and the popular Gauss–Seidel method and successive-over-relaxation (SOR) methods for solving linear systems [55]. Pathiraja and Reich [60] apply discrete gradient methods to gradient flow systems with applications to computational Bayesian inference, observing rapid convergence of the scheme applied to Fokker–Planck dynamics.

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

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

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

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

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

where f^δ is a noisy image and $R_\alpha(\cdot) \equiv R(\cdot, \alpha)$ is a regularisation function that depends on a regularisation parameter α. 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. To ensure that $\alpha \mapsto u_\alpha$ is a well-defined mapping, we assume

that the variational problem is strictly convex for all admissible parameters α. This would follow, e.g. from convexity of R_α. In Sect. 6.2, we apply Itoh–Abe methods to solve these problems.

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. [13,22,45,58]. Furthermore, model parameter optimisation problems appear in many other applications. These include optimising for the management of water resources [28], approximation of a transmembrane protein structure in computational biology [31], image registration in medical imaging [59], the building of wind farms [23], and solar energy utilisation in architectural design [41], 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. For recent reviews on derivative-free optimisation, we refer the reader to Audet and Hare [3] and Larson et al. [46].

While there is 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 evaluations, rather than the numerical differences. A significant portion of derivative-free methods is direct search methods, and the most well known of these is the Nelder–Mead method (also known as the downhill simplex method) [56].

Alternatively, model-based methods that build a local quadratic model based on evaluations are well-documented [14,63,64]. 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. [27] formulate a derivative-free line search method termed DFN and analyse its convergence properties for nonsmooth functions for the Clarke subdifferential, in the constrained setting. Building on the DFN algorithm, Liuzzi and Truemper [50] formulate a derivative-free method that is a hybrid between DFN and MADS. There are similarities between the DFN scheme in [27] and our approach to Itoh–Abe methods, both in the implementation of the derivative-free scheme and in their the-

oretical analysis. For example, both schemes implement line search schemes which are robust to nonsmoothness in similar ways, and in both cases, Clarke stationarity is proven, given sufficient density assumptions for the sequence of directions $(d^k)_{k\in\mathbb{N}}$ on the unit sphere.

With these similarities in mind, we highlight some ways in which this work differs from [27]. A central motivation for this work is the extension of discrete gradient methods, as an established concept in geometric numerical integration, to nondifferentiable systems, in particular for solving optimisation problems. The realisation and particular implementation of the discrete gradient scheme in Sect. 5 is in this sense secondary.[3] Moreover, this builds on previous research into discrete gradients methods in optimisation, including [24] where convergence rates are established for smooth problems, and [7] where the Itoh–Abe methods are generalised to Bregman methods, e.g. for faster, sparse optimisation.

In comparing the theoretical analyses, we also remark that the theorems in Sect. 3.2 are stronger, as they establish convergence to stationary points for directions $(d^k)_{k\in\mathbb{N}}$ chosen both determinstically and stochastically, through cyclical density and probabilistic arguments, respectively, while the corresponding result in [27, Proposition 2.7] assumes the density of a subsequence $(d^{k_j})_{j\in\mathbb{N}}$. However, we believe that it should be possible to extend our novel arguments in Sect. 3.2 to DFN.

We mention the random search scheme, given by

$$x^{k+1} = x^k - \tau_k \frac{V(x^k + \beta_k d^k) - V(x^k)}{\beta_k} d^k, \tag{1.8}$$

where $\beta_k > 0$, $\tau_k > 0$, and $(d^k)_{k\in\mathbb{N}}$ are independent, random draws from a probability distribution. This explicit scheme was proposed by Polyak [62] and later studied by Nesterov [57] for derivative-free optimisation of nonsmooth, convex functions. In [57], convergence rates are obtained for $\mathbb{E}(V(x^k)) - \min V$, based on analysis of the *Gaussian smoothing* function $V_\beta(x^k) := \mathbb{E}_{d^k} V(x^k + \beta d^k)$, i.e. the expectation with respect to d^k, where the directions $(d^k)_{k\in\mathbb{N}}$ are drawn from a Gaussian distribution. We remark that the randomised Itoh–Abe method (1.6) corresponds to (1.8) with an additional, implicit coupling between τ_k and β_k as described in (1.6). However, in contrast to the stochastic analysis in [57], the optimality analysis for the randomised Itoh–Abe methods in Sect. 3.2 is based on the deterministic estimates (1.7), inherited by the structure preservation of the discrete gradient methods. Whether a similar analysis could be applied to obtain convergence rates for randomised Itoh–Abe methods in the nonsmooth setting is beyond the scope of this paper, and is left as a future topic of inquiry.

Amongst the challenges that nonconvexity poses in optimisation, a central one is that V is not bounded below by its first-order approximation, and, on a related note, the concept of a subdifferential becomes significantly more complicated in this setting (see Sect. 2 for details on generalised subdifferentials). While the main focus of [57] is convex optimisation, in Sect. 7 the authors derive complexity estimates for nonconvex problems with respect to the norm of the gradient of the Gaussian smoothing V_μ. As is

[3] Indeed, alternative approaches to solving the Itoh–Abe method (1.6) can be found, e.g. in [7,15,24,33,66].

expected, the complexity estimate is significantly affected by nonconvexity, and it is not shown whether these estimates for V_μ can be extended to the nonsmooth function V or its subgradients. In contrast, while the analysis for the Itoh–Abe methods in Sect. 3.2 holds for nonsmooth, nonconvex functions, it is not clear whether the analysis could be notably simplified by assuming convexity of V.

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* [19] is required at each iterate to construct a linear approximation to the objective function—see [43] for an introduction. A close alternative to bundle methods are *gradient sampling methods* (see [12] for a recent review by Burke et al.), where the descent direction is determined by sampling gradients in a neighbourhood of the current iterate. Curtis and Que [21] formulated a hybrid method between the gradient sampling scheme of [20] and the well-known quasi-Newton method BFGS adapted for nonsmooth problems [49]. These methods have convergence guarantees in the Clarke subdifferential framework, under the assumption that the objective function is differentiable in an open, dense set. Last, we mention a derivative-free scheme based on gradient sampling methods, proposed by Kiwiel [44], 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. Sect. 2 provides a background on the Clarke subdifferential for nonsmooth, nonconvex analysis. In Sect. 3, the main theoretical results of the paper are presented, namely existence and optimality results in the stochastic and deterministic setting. In Sect. 4, we briefly discuss the Itoh–Abe

discrete gradient for general coordinate systems. In Sects. 5 and 6, the numerical implementation is described and results from example problems are presented. A conclusion is given in Sect. 7.

2 Nonconvex Optimisation

In this section, we introduce the Clarke subdifferential framework [19], the most popular framework for nonsmooth, nonconvex optimality analysis, due to its nice analytical properties. It generalises the gradient of a differentiable function, as well as the subdifferential [26] of a convex function, hence the term *Clarke subdifferential*. Francis H. Clarke introduced the framework in his doctoral thesis in 1973 [18], in which he used the term *generalised gradients*.

2.1 The Clarke Subdifferential

Throughout the rest of 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 a function V that is Lipschitz continuous 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 [19, Proposition 2.2.7], and coincides with the derivative at points of strict differentiability [19, Proposition 2.2.4]. It can equivalently be characterised as [19, Theorem 2.5.1]

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

where $\mathscr{D}(V)$ is the set of differentiable points of V, and co denotes the convex hull of the set. We additionally state two useful results, both of which can be found in Chapter 2 of [19].

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 at x: 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 to nonsmooth, nonconvex functions. For example, the Michel–Penot subdifferential [54] coincides with the Gâteaux derivative when this exists, unlike the Clarke subdifferential, which is larger and only coincides with strict derivatives [29]. However, the outer semicontinuity of the Clarke subdifferential makes it in most cases the preferred framework for analysis. See [8] 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*.[4] 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 $x^k = [\frac{1}{k}, 0]^{\mathrm{T}}$ and $y^k = [0, \frac{1}{k}]^{\mathrm{T}}$. Then, for all k, the Itoh–Abe discrete gradient is

$$\overline{\nabla} V(x^k, y^k) = [1, 1]^{\mathrm{T}}.$$

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

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 directionally stationary at x^* along d* if

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

[4] For convex functions, $p \in \mathbb{R}^n$ is an ε-approximate subgradient if, for all $y \in \mathbb{R}^n$, it is the case that $V(y) \geq V(x) + \langle p, y - x\rangle - \varepsilon$ [37].

Remark 2.8 A point x^* is Clarke stationary if and only if V is Clarke directionally stationary at x^* along d for all $d \in S^{n-1}$.

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

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

If V is pseudoconvex, then any Clarke stationary point is a global minimum [4]. Clarke [19] 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 and equals $V^o(x; d)$ for each $d \in \mathbb{R}^n$. 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, in Lemma 3.1, we prove that the discrete gradient update (1.6) 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 locally Lipschitz 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

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

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 $\beta \in (0,\delta)$,

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

Choosing $\beta_1 = \min\{\delta/2, \varepsilon/2\}$, we obtain

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

On the other hand, as V is bounded below, there is $M < 0$ such that $V(x+\tau\beta d) - V(x) > M$ for all $\beta \in \mathbb{R}$. Setting $\beta_2 = \sqrt{|M|/\tau}$, we derive

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

Note that the above inequality holds for all $\beta \geq \beta_2$, so we have $\beta_1 < \beta_2$. 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 [70, Theorem 4.23] that there is $\beta \in (\beta_1, \beta_2)$ that solves the discrete gradient equation:

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

The following lemma, which is an adaptation of [33, Theorem 1] for the nonsmooth setting, summarises some useful properties of the methods.

Lemma 3.2 *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) $\lim_{k\to\infty} dV(x^k, x^{k+1}) = 0$.
(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 non-increasing, $V(x^k) \to V^*$ for some limit V^*. Therefore, by (1.7)

$$V(x^0) - V^* = \sum_{k=0}^{\infty} V(x^k) - V(x^{k+1}) = \sum_{k=0}^{\infty} \tau_k dV(x^k, x^{k+1})^2$$
$$\geq \tau_{\min} \sum_{k=0}^{\infty} dV(x^k, x^{k+1})^2.$$

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 that

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

which shows properties (ii) and (iii).

Last, we show (iv). Since $(V(x^k))_{k\in\mathbb{N}}$ is a non-increasing 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. □

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 3.3 *With the same assumption for V as in Lemma 3.2, 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 is a subset of $\{x \in \mathbb{R}^n \ : \ V(x) \leq V(x^0)\}$. 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 S is disconnected. Since S is closed, this is equivalent to there being disjoint, closed, nonempty sets S_1, S_2 such that $S = S_1 \cup S_2$. Choose $\varepsilon > 0$ such that $\|x - y\| \geq \varepsilon$ for all $x \in S_1$, $y \in S_2$, and define the nonempty, closed set $S_3 = \{x \in \mathbb{R}^n \ : \ \|x - y\| \geq \varepsilon/3 \text{ for all } y \in S\}$. As $\|x^k - x^{k+1}\| \to 0$, there is $K \in \mathbb{N}$ such that $\|x^k - x^{k+1}\| < \varepsilon/3$ for all $k \geq K$. It follows that if $x^k \in S_1 + B_{\varepsilon/3}(0)$ and $x^j \in S_2 + B_{\varepsilon/3}(0)$ for $j \geq k \geq K$, then there is $l \in (k, j)$ such that $x^l \in S_3$. As $(x^k)_{k\in\mathbb{N}}$ has accumulation points in S_1 and S_2, it thus follows that it also has a subsequence in S_3. As $(x^k)_{k\in\mathbb{N}}$ is a bounded sequence and S_3 is closed, this subsequence admits a convergent subsequence with limit in S_3. This contradicts the assumption that all accumulation points of $(x^k)_{k\in\mathbb{N}}$ are in $S_1 \cup S_2$.

To show that V is constant on S, we simply note that $(V(x^k))_{k\in\mathbb{N}}$ is a non-increasing sequence and $V(x^*) = \lim_{k\to\infty} V(x^k)$ for all $x^* \in S$, from which the result follows.

Finally, we show by contradiction that S has empty interior. 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 $x^j \in B_\varepsilon(x)$. However, as V is constant on S, $V(x^j) = \min_{k\in\mathbb{N}} V(x^k)$. Since $(V(x^k))_{k\in\mathbb{N}}$ is a non-increasing sequence and $\|x^k - x^{k+1}\|^2 = \tau_k(V(x^k) - V(x^{k+1}))$, it follows that $x^k = x^j$ for all $k \geq j$. Therefore, $S = \{x^j\}$, which contradicts the assumption that S has nonempty interior. □

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 probability density of Ξ has support almost everywhere 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 $(d^{nk+j})_{k\in\mathbb{N}}$ are independently drawn from S^{n-1} for $j = 1, \ldots, n$, 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 3.4 *With the same assumption for V as in Lemma 3.2, 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 density of Ξ has almost everywhere support in S^{n-1}. Then, every accumulation point of $(x^k)_{k\in\mathbb{N}}$ is almost surely Clarke stationary.*

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.

We first 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 $d \in S^{n-1}$, $\varepsilon > 0$ 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 $B_\eta(x)$, 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 step.

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

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

Clearly,

$$X = \bigcup_{m\in\mathbb{N}} X_m. \tag{3.3}$$

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

$$(3.3) \quad \text{and} \quad X_m \subset \bigcup_{i\in\mathbb{N}} Y_i^{(m)} \text{ for all } m \in \mathbb{N} \quad \Longrightarrow \quad X \subset \bigcup_{m\in\mathbb{N}} \bigcup_{i\in\mathbb{N}} Y_i^{(m)}.$$

Since a countable union of countable sets is countable, we conclude with the following statement. There exists sequences $(y^i)_{i\in\mathbb{N}} \subset \mathbb{R}^n$, $(\varepsilon_i)_{i\in\mathbb{N}} \subset (0,\infty)$, $(\delta_i)_{i\in\mathbb{N}} \subset (0,\infty)$, and $(\tilde{d}^i)_{i\in\mathbb{N}} \subset S^{n-1}$, such that for all $z \in B_{\delta_i}(y^i) =: B_i$, all $e \in B_{\delta_i}(\tilde{d}^i) \cap S^{n-1} =: D_i$, and all $\lambda \in (0, \delta_i)$, we have

$$\frac{V(z-\lambda e)-V(z)}{\lambda} \leq -\varepsilon_i, \tag{3.4}$$

and such that

$$X \subset \bigcup_{i\in\mathbb{N}} B_i. \tag{3.5}$$

Finally, we want to show that for each $i \in \mathbb{N}$, $\mathbb{P}(S \cap B_i \neq \varnothing) = 0$. For this, we fix i and set $m := \min_{x\in B_i} V(x)$, $M := \max_{x\in B_i} V(x)$, and $\mu := \min\{\varepsilon_i^2 \tau_{\min}, \frac{\delta_i^2}{\tau_{\max}}\} > 0$. We first show for any $k \in \mathbb{N}$ that

$$x^k \in B_i \text{ and } d^k \in D_i \quad \Longrightarrow \quad V(x^k) - V(x^{k+1}) \geq \mu. \tag{3.6}$$

To show this, we write $x^{k+1} = x^k - \lambda d^k$ for some $\lambda \in \mathbb{R}$, and consider separately the cases $|\lambda| < \delta_i$ and $|\lambda| \geq \delta_i$. In the first case, it follows from (3.4) that $\lambda \in (0, \delta_i)$, and furthermore that

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

However, by (1.7), we also have

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

and, combining these equations, we get

$$\varepsilon_i \tau_k \leq \|x^{k+1} - x^k\|.$$

This in return implies

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

Otherwise, if $|\lambda| > \delta_i$, then by (1.7)

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

Thus, (3.6) holds.

Next, choose $K \in \mathbb{N}$ such that $K\mu > M - m$, and suppose that there are $K+1$ indices $k_1, \ldots k_{K+1}$ such that $x^{k_j} \in B_i$ and $d^{k_j} \in D_i$ for $j = 1, \ldots, K+1$. Then, using (3.6) and the fact that $(V(x^k))_{k\in\mathbb{N}}$ is a non-increasing sequence, we have

$$\begin{aligned} M - m &\geq V(x^{k_1}) - V(x^{k_{K+1}}) \geq V(x^{k_1}) - V(x^{k_K}) + \mu \geq \cdots \geq V(x^{k_1}) - V(x^{k_1}) + K\mu \\ &> M - m, \end{aligned}$$

which is a contradiction. Thus, for there to be a subsequence of $(x^k)_{k\in\mathbb{N}}$ in B_i, there can be no more than K instances where $x^k \in B_i$ and d^k is drawn in D_i. Since $(d^k)_{k\in\mathbb{N}}$ are independent draws, the probability of this is 0, as we rigorously demonstrate in the next paragraph.

For $j = 1, 2, \ldots$, denote by $E_j \in \mathbb{N} \cup \{+\infty\}$ the index k for which x^k is the jth iterate of $(x^k)_{k\in\mathbb{N}}$ in B_i, where we set $E_j = +\infty$ if there are fewer than j elements of $(x^k)_{k\in\mathbb{N}}$ in B_i. Next, we construct a sequence of random variables $(f^j)_{j\in\mathbb{N}}$, where if $E_j \in \mathbb{N}$ we set $f^j := d^{E_j}$, and otherwise f^j is an independent draw from Ξ. It follows that $(f^j)_{j\in\mathbb{N}}$ is a sequence of independent draws from Ξ. Finally, for $j = 1, 2, \ldots$, set $F_j := 1$ if $f^j \in D_i$ and 0 otherwise, and define $G_j := \sum_{k=1}^{j} F_j$. We observe that

$$\mathbb{P}((x^k)_{k\in\mathbb{N}} \text{ has subsequence in } B_i) \leq \mathbb{P}(\limsup_{j\to\infty} G_j \leq K) = 0,$$

where the latter equality holds because $(F_j)_{j=1}^{\infty}$ is a sequence of independent events with $\mathbb{P}(F_j = 1) = \mathbb{P}_{\Xi}(d \in D_i) > 0$. This concludes the proof. □

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 3.5 A sequence $(d^k)_{k\in\mathbb{N}} \subset U$ is *cyclically dense* in U if for each $\varepsilon > 0$ there is $N \in \mathbb{N}$ so that for all $k \in \mathbb{N}$, the set $\left\{d^k, \ldots, d^{k+N-1}\right\}$ forms an ε-cover of U,

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

In Appendix A, we show that randomly drawn sequences are almost surely not cyclically dense, hence the separate treatment of stochastic and deterministic schemes.

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

Example 3.6 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 $\delta = \lambda_{k+r} - \lambda_k$, where we know that $\delta \neq 0$, 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 $\delta \in (-\varepsilon, \varepsilon)$ on [0, 1]. Setting $N = r\left\lceil \frac{1}{|\delta|} \right\rceil + k$, where $\lceil 1/|\delta| \rceil$ denotes the smallest integer greater than or equal to $1/|\delta|$, it is clear that for any $j \in \mathbb{N}$, the set $\{\lambda_j, \lambda_{j+1}, \ldots, \lambda_{j+N-1}\}$ 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 3.7 *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 3.4, where X is the set of nonstationary points (3.1) and is covered by a countable collection of open balls (3.5),

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

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 3.2 (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

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

Then, by cyclical density, we can choose N such that the corresponding directions $\{d^{k_j}, d^{k_j+1}, \ldots, d^{k_j+N-1}\}$ form a δ_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 3.4, 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. □

3.3 Necessity of Search Density and Lipschitz Continuity

In what follows, we examine the necessity of some of the assumptions made in the convergence theorems.

It is well known that 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 $\{e^1, \ldots, e^n\}$. For example, consider the function $V(x, y) = \max\{x, y\}$ and the starting point $x^0 = [1, 1]^{\mathrm{T}}$. With the standard Itoh–Abe discrete gradient method, the iterates would remain at x^0, even though this point is nonstationary.

Furthermore, the following example demonstrates that it is necessary for the set of directions $(\pm d^k)_{k\in\mathbb{N}}$ to be dense on S^{n-1}.

Example 3.8 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 $x^0 = [-1, 0]^{\mathrm{T}}$. For $\theta \in [-\pi/2, \pi/2]$, let $d = [\cos\theta, \sin\theta]^{\mathrm{T}}$. 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, either the directions $(d^k)_{k\in\mathbb{N}}$ (or $(-d^k)_{k\in\mathbb{N}}$) need to include a convergent subsequence to the direction $[1, 0]^{\mathrm{T}}$. As this direction is arbitrary, we deduce that $(\pm d^k)_{k\in\mathbb{N}}$ must be dense.

Theorem 3.4 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 Clarke [19, Proposition 2.1.1. (b)], the mapping $(y, d) \mapsto V^o(y; d)$ is upper semicontinuous for y near x, due to 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 that for a subsequence $(x^{k_j})_{j\in\mathbb{N}}$ such that $x^{k_j} \to x^*$ and $d^{k_j} \to d^*$, we have

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

Without local Lipschitz continuity, it is possible to have

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

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

Example 3.9 We first fix the iterates $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^2$ and $(d^k)_{k\in\mathbb{N}} \subset S^1$, and construct the function $V : \mathbb{R}^2 \to \mathbb{R}$ subsequently. Let $(d^k)_{k\in\mathbb{N}}$ be a cyclically dense sequence in S^1 and assume without loss of generality that $[0, 1]^{\mathrm{T}} \notin (d^k)_{k\in\mathbb{N}}$. 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 $x^0 = [0, 0]^{\mathrm{T}}$ and define $(x^k)_{k\in\mathbb{N}}$ to be

$$x^{k+1} = x^k - \frac{1}{(k+1)^2} d^k.$$

Note that since $d_1^k < 0$ for all k, $x_1^{k+1} > x_1^k$ for all k, so none of the iterates coincide. Furthermore, the sequence $(x^k)_{k\in\mathbb{N}}$ has a unique limit, which we denote by x^*.

Next, we define a piecewise affine path $\rho : \mathbb{R} \to \mathbb{R}$ accordingly. If $r < x_1^0$, set $\rho(r) := x_2^0$, and if $r > x_1^*$, set $\rho(r) := x_2^*$. Otherwise, $r \in [x_1^0, x_1^*]$ and there are unique $\lambda \in [0, 1]$ and $k \in \mathbb{N}$ such that $r = \lambda x_1^k + (1-\lambda)x_1^{k+1}$. In this case, we set $\rho(r) := \lambda x_2^k + (1-\lambda)x_2^{k+1} = x_2^k + (r - x_1^k)d_2^k/d_1^k$.

We then define $V : \mathbb{R}^2 \to \mathbb{R}$ accordingly. On $(x^k)_{k\in\mathbb{N}}$, we set

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

If $r = \lambda x_1^k + (1-\lambda)x_1^{k+1}$ for some $\lambda \in [0, 1]$ and $k \in \mathbb{N}$, set

$$V(r, \rho(r)) := \lambda V(x^k) + (1-\lambda)V(x^{k+1}) = V(x^k) + \frac{r - x^k}{d_1^k (k+1)^2}.$$

Otherwise, if $r < x_1^0$, set $V(r, \rho(r)) := V(x^0)$, and if $r > x_1^*$, then set $V(r, \rho(r)) := V(x^*)$. Finally, for each $[r, x_2]^{\mathrm{T}} \in \mathbb{R}^2$, define $V(r, x_2) := V(r, \rho(r)) - (x_2 - \rho(r))$.

One can verify that this constitutes a well-defined, continuous function on $\mathbb{R}^2$. Furthermore, the iterates $(x^k)_{k\in\mathbb{N}}$ satisfy (1.6) with $\tau_k = 1$. Finally, observe that for all $x \in \mathbb{R}^2$, we have $V^o(x; [0, 1]^{\mathrm{T}}) = -1$. Therefore, x^* is not a stationary point of V.

One may observe directly for this function that $V^o(\cdot; \cdot)$ is not upper semicontinuous, and hence that V is not locally Lipschitz continuous at x^*. Namely, passing

to a subsequence $(d^{k_j})_{j\in\mathbb{N}}$ such that $d^{k_j} \to [0,1]^{\mathrm{T}}$, we conclude from the above construction that $\limsup_{j\to\infty} V^o(x^{k_j}; d^{k_j}) \geq 0$, while $V^o(x^*; [0,1]^{\mathrm{T}}) = -1$.

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 $V^o(x^*; \pm e^i) \geq 0$ for $i = 1, \ldots, n$. One may, for example, consider functions of the form:

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

where E is a continuously differentiable function that may be nonconvex and $\|x\|_1$ denotes $|x_1| + \cdots + |x_n|$, and $\lambda > 0$. See, for example, Proposition 2.3.3 and the subsequent corollary in [19], 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

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

where $\times$ denotes the Cartesian product and

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

Since this paper is chiefly concerned with the blackbox setting where no particular structure of V is assumed, we do not include an 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 problems where Clarke stationarity is equivalent to Clarke directional stationarity along the standard coordinates, one can easily adapt Theorem 3.4 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 $\|x\|_p^p = |x_1|^p + \cdots + |x_n|^p$. 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 for all k. For this, we consider each block of n directions to be indepen-

dently drawn from a random distribution on the set of orthogonal transformations on $\mathbb{R}^n$, denoted by $O(n)$.

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 $R^{-1} = R^{\mathrm{T}}$. Equivalently, R maps one orthonormal basis of $\mathbb{R}^n$ to another.

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

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

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

where

$$\left(\hat{\nabla}_R V(x, y)\right)_i := \begin{cases} \frac{V\left(x+\sum_{j=1}^{i}\langle y-x, f^j\rangle f^j\right) - V\left(x+\sum_{j=1}^{i-1}\langle y-x, f^j\rangle f^j\right)}{\langle y-x, f^i\rangle}, & \text{if } \langle y-x,\, f^i\rangle \neq 0, \\ \langle \nabla V\left(x + \sum_{j=1}^{i-1}\langle y-x,\, f^j\rangle f^j\right),\, f^i\rangle, & \text{otherwise.} \end{cases}$$

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

Proposition 4.3 *If $V : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, then $\overline{\nabla}_R V$ is a discrete gradient.*

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

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

The convergence property $\lim_{y\to x} \overline{\nabla}_R V(x, y) = \nabla V(x)$ follows directly from continuous differentiability of V. □

Thus, we can implement schemes that are formally discrete gradient methods and also fulfil the convergence theorems in Sect. 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 $d^k = e^{[k \bmod n]+1}$. Performing n steps of this method is equivalent to one step with the standard Itoh–Abe discrete gradient method.
2. *Random pursuit Itoh–Abe method* The directions are independently drawn from a random distribution Ξ on S^{n-1}. We assume that the density of Ξ has support almost everywhere.
3. *Rotated Itoh–Abe method* For each $k \in \mathbb{N}$, the block of n consecutive directions $(d^{kn+1}, d^{kn+2}, \dots, d^{(k+1)n})$ is drawn from a random distribution on $O(n)$, so that the directions form an orthonormal basis. We assume that each draw from $O(n)$ is independent.

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 returns 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. The exception to this is when 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 tailor-made scalar solver for (1.6) that balances the trade-off between decreasing $\beta \mapsto V(x^k + \beta d^k)$ as much as possible within the time step constraints $\tau_{\min}, \tau_{\max}$ and using minimal function evaluations. Rather than solving for a given τ_k, it ensures that there exists some $\tau_k \in [\tau_{\min}, \tau_{\max}]$ that matches the output x^{k+1}. The algorithm uses a preliminary time step $\overline{\tau} \in [\tau_{\min}, \tau_{\max}]$, which we set to $\overline{\tau} = \sqrt{\tau_{\min}\tau_{\max}}$. This method is particularly suitable when $\tau_{\min} \ll \tau_{\max}$, and it can be replaced by any other scalar root solver. The algorithm calls the functions interpolationStep and backtrackingFunction, whose algorithms can be found in Appendix.

As the algorithm does not use a fixed time step, it is notationally convenient to make β_k include τ_k, so that the update becomes $x^{k+1} = x^k + \beta_k d^k$ where β_k/τ_k solves (1.6). With this in mind, we define an admissible output $x^{k+1} = x^k + \beta d^k$ of Algorithm 3 as follows, for a specified tolerance $\varepsilon > 0$ and time step bounds $\tau_{\min} < \tau_{\max}$:

$$\exists \tau \in [\tau_{\min}, \tau_{\max}],\ \tilde{\beta} \in [\beta - \varepsilon, \beta + \varepsilon] \text{ s.t. } \tilde{\beta}/\tau \text{ solves (1.6) for time step } \tau. \quad (5.1)$$

The Python code is available at https://github.com/esriis/itohabe_optimisation.

6 Examples

In this section, we use the randomised Itoh–Abe methods to solve several nonsmooth, nonconvex problems. In Sect. 6.1, we consider some well-known optimisation

Algorithm 2 Randomised Itoh–Abe method with solver and stopping criterion. The function innerSolver is described in Algorithm 3.

Input:

x^0	starting point
$(d^k)_{k\in\mathbb{N}}$	directions
$(\tau_{\min}, \tau_{\max})$	time step bounds
$\overline{\tau} \in (\tau_{\min}, \tau_{\max})$	proposed time step
$\varepsilon > 0$	tolerance
$\eta > 0$	tolerance for decrease
$\sigma \in (0, 1)$	search parameter
K	maximal number of iterations
M	maximal number of consecutive directions without descent before stopping
$m = 0$	initialise counter

for $k = 0, \dots, K-1$ **do**
  $x^{k+1} \leftarrow \text{innerSolver}(x^k, d^k, \tau_{\min}, \tau_{\max}, \overline{\tau}, \varepsilon, \sigma)$
  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

Algorithm 3 Solver for Itoh–Abe step (1.6). The function interpolationStep, defined in Algorithm 4, returns a point of descent β, which is chosen in such a way that the subsequent backtrackingFunction procedure, defined in Algorithm 6, is guaranteed to find a solution to (5.1). For full details, see Appendix B.

Input:

$x \in \mathbb{R}^n$	current point
$d \in S^{n-1}$	direction
$\tau_{\min} > 0$	time step lower bound
$\tau_{\max} \in (\tau_{\min}, +\infty)$	time step upper bound
$\overline{\tau} \in (\tau_{\min}, \tau_{\max})$	proposed time step
$\varepsilon > 0$	tolerance for β
$\sigma \in (0, 1)$	search parameter

Output:

$y \in \mathbb{R}^n$	solution to (5.1)

function INNERSOLVER(x,d,$\tau_{\min}$,$\tau_{\max}$,$\overline{\tau}$,ε,σ)
  if $(V(x) - V(x + \varepsilon d))/\varepsilon^2 \leq 1/\tau_{\min}$ **then** ▷ check for direction of descent
    $d \leftarrow -d$
    if $(V(x) - V(x + \varepsilon d))/\varepsilon^2 \leq 1/\tau_{\min}$ **then** ▷ if V is stationary at x along d up to tolerance
      return x ▷ then $\beta = 0$ solves (5.1)
    end if
  end if
  $f : \beta \mapsto V(x + \beta d) - V(x)$ ▷ define scalar function
  $\beta \leftarrow \text{interpolationStep}(f, \tau_{\min}, \tau_{\max}, \overline{\tau}, \varepsilon, \sigma)$ ▷ do parabolic interpolation step
  $\beta \leftarrow \text{backtrackingFunction}(f, \beta, \tau_{\min}, \tau_{\max}, \varepsilon, \sigma)$ ▷ do backtracking procedure
  return $x + \beta d$ ▷ output solution to (5.1)
end function

challenges developed by Rosenbrock and Nesterov. In Sect. 6.2, we solve bilevel optimisation of parameters in variational regularisation problems.[5]

We compare our method to state-of-the-art derivative-free optimisation methods Py-BOBYQA [14,64] and the LT-MADS solver provided by NOMAD [2,47,48]. For purposes of comparing results across solvers for these problems, we do not measure objective function value against iterates, but against function evaluations.

6.1 Rosenbrock Functions

We consider the well-known Rosenbrock function [68]

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

Its global minimiser $[1, 1]^{\mathrm{T}}$ 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 Fig. 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 additionally consider a nonsmooth variant of (6.1), termed Nesterov's (second) nonsmooth Chebyshev–Rosenbrock function [34],

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

In this case too, the global minimiser $[1, 1]^{\mathrm{T}}$ is located along a narrow path. Furthermore, there is a nonminimising, stationary point at $[0, -1]^{\mathrm{T}}$, 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 Fig. 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 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 $= 2$,

[5] Test images are taken from the Berkeley database [51]. Available online: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/BSDS300/html/dataset/images.html.

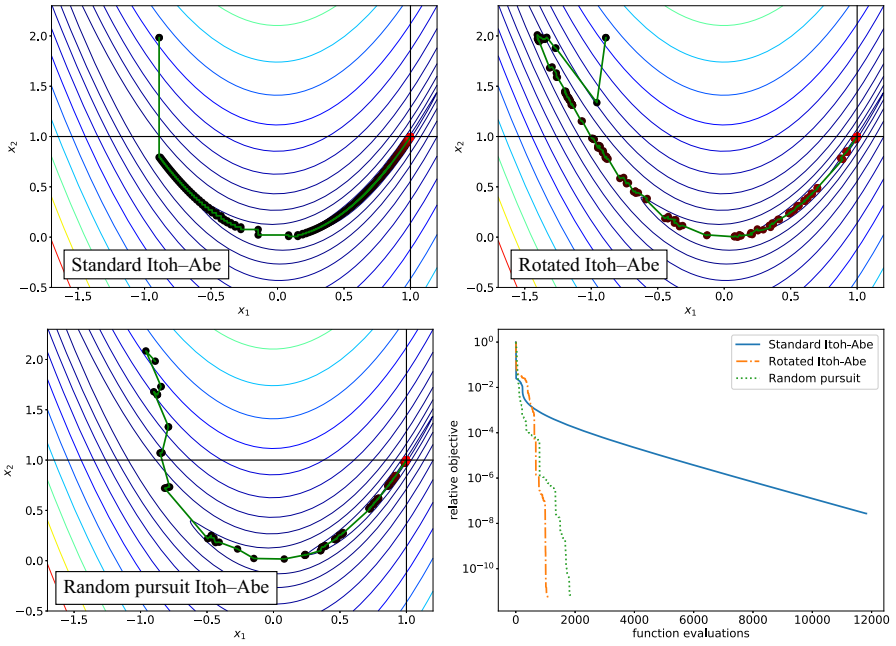


Fig. 1 Comparison of three variants of the Itoh–Abe method applied to the Rosenbrock function. Top left: Itoh–Abe method with standard frame. Top right: rotated Itoh–Abe method. Bottom left: Itoh–Abe method with random pursuit. Bottom right: convergence rates of the relative objective $\frac{V(x^k)-V^*}{V(x^0)-V^*}$ for the three variants

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 Figs. 3 and 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 LT-MADS method locates the nonminimising stationary point at $[0, -1]^{\mathrm{T}}$, 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 $n = 1, 2$.

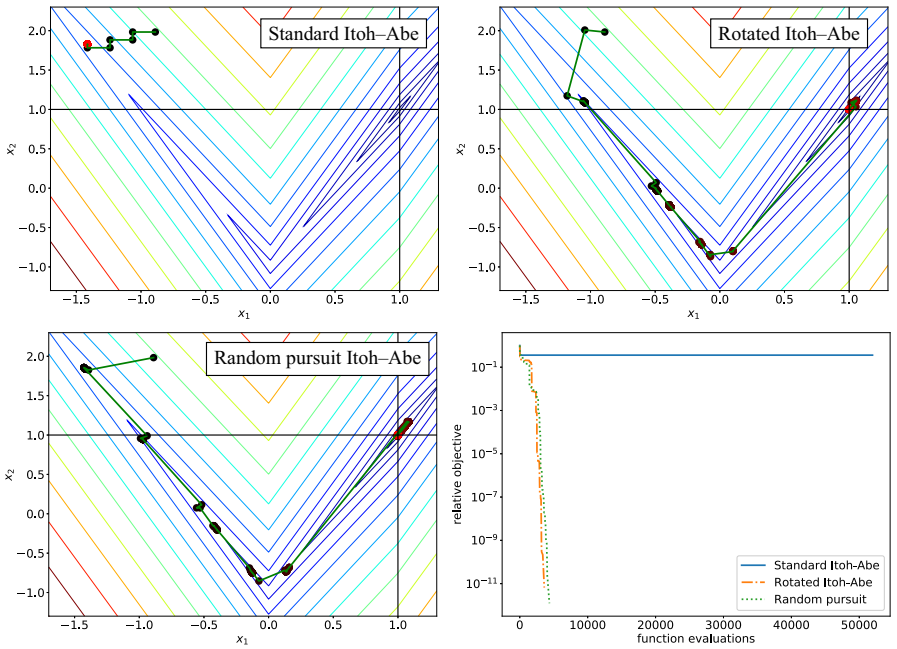


Fig. 2 Comparison of three variants of the Itoh–Abe method applied to Nesterov's nonsmooth Chebyshev–Rosenbrock function. Top left: Itoh–Abe method with standard frame. Top right: rotated Itoh–Abe. Bottom left: Itoh–Abe with random pursuit. Bottom right: convergence rates of the relative objective $\frac{V(x^k)-V^*}{V(x^0)-V^*}$ for the three variants

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

$$u_\alpha \in \arg\min_u \frac{1}{2}\|u - f^\delta\|^2 + R_\alpha(u). \tag{6.3}$$

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 deter-

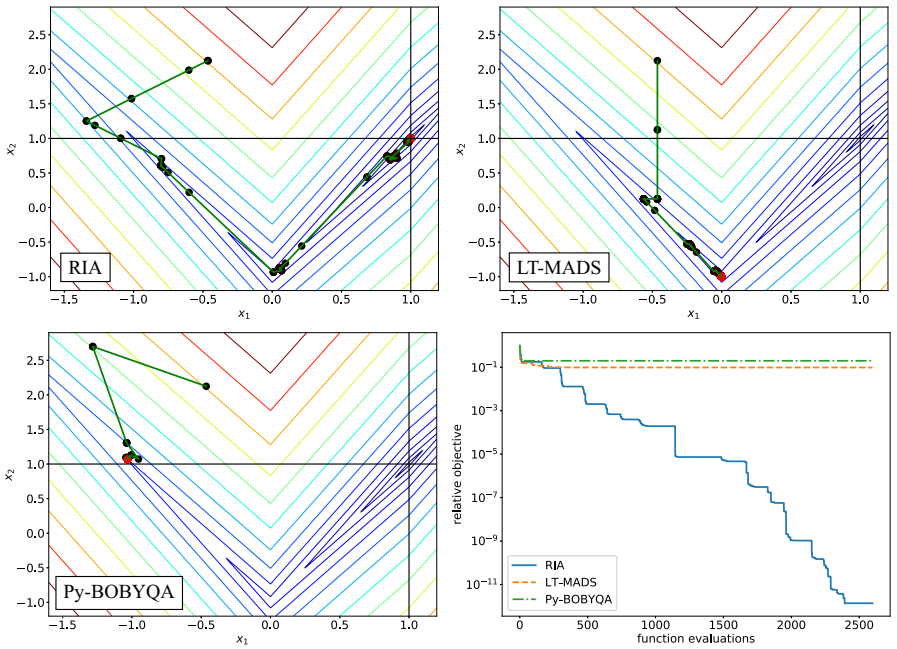


Fig. 3 Comparison of rotated Itoh–Abe method, LT-MADS and Py-BOBYQA applied to Nesterov's nonsmooth Chebyshev–Rosenbrock function. Top left: the iterates from the Itoh–Abe method locate the unique minimiser to an order of accuracy of about 10^{-11}. Top right: the iterates from the LT-MADS method locate 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

mine how heavily to regularise, and sometimes adjust other features of the regulariser. See [6,39,71] for an overview of variational regularisation methods.

We list some common regularisers in image analysis. *Total variation* (TV) [11,69] is given by the function $R_\alpha(u) := \alpha\,\mathrm{TV}(u)$, where $\alpha \in (0, \infty)$, and

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

This is one of the most common regularisers for image denoising. See Fig. 5 for an example of denoising with TV regularisation. We also consider its second-order generalisation, *total generalised variation* [9,10], $R_\alpha(u) = \mathrm{TGV}_\alpha^2(u)$, where $\alpha = [\alpha_1, \alpha_2]^\mathrm{T} \in (0,\infty)^2$ and

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

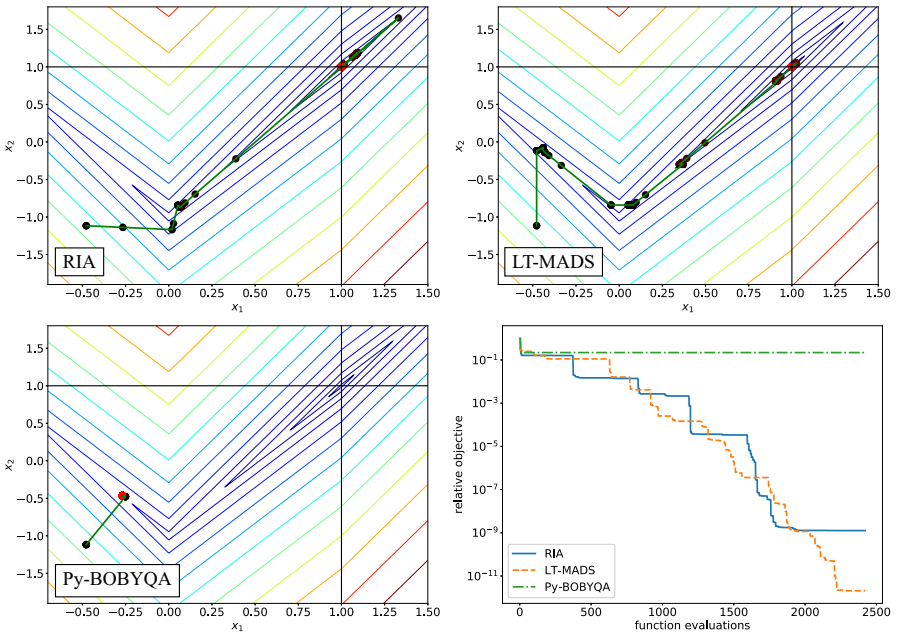


Fig. 4 Comparison of rotated Itoh–Abe method, LT-MADS and Py-BOBYQA applied to Nesterov's non-smooth Chebyshev–Rosenbrock function with a different starting point. Top left: the iterates from the Itoh–Abe method locate the unique minimiser to an order of accuracy of about 10^{-11}. Top right: the iterates 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

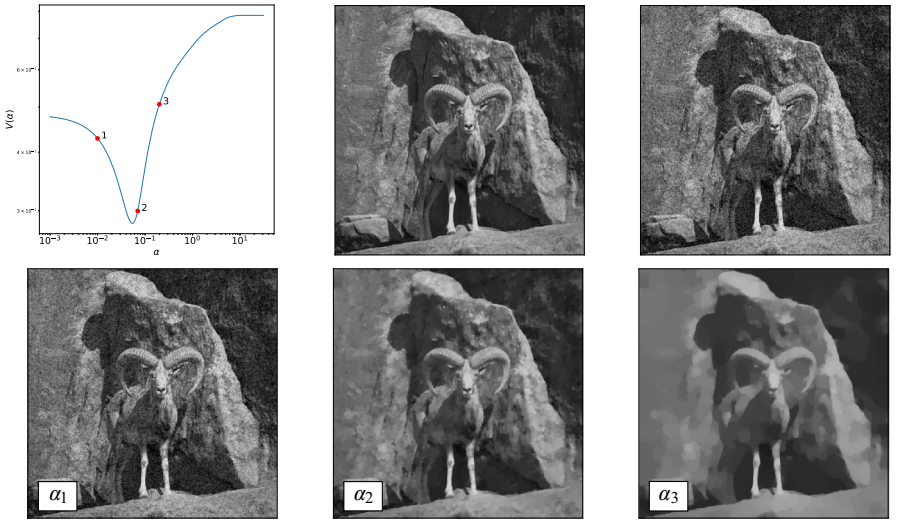


Fig. 5 TV denoising reconstructions for different regularisation parameters, with the SSIM scoring function (see (6.6)). Top: graph of V in (6.5), ground truth image, and noisy image, respectively. Bottom: reconstructions for parameter choices, $\alpha_1 = 10^{-2}$, $\alpha_2 = 7 \times 10^{-2}$, $\alpha_3 = 2 \times 10^{-1}$, respectively

For a linear, bounded operator $W : L^2(\Omega) \to \ell^2$, the basis pursuit regulariser

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

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

As illustrated in Fig. 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

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

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

$$V(\alpha) := \Phi(u_\alpha). \tag{6.5}$$

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[6]. Therefore, one typically does not have access to gradient or subgradient information[7] 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 us 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.

For the numerical implementation, we discretise the domain Ω and represent the image u as a vector in some Euclidean space $\mathbb{R}^m$.

[6] Accounting for inexact functions evaluations in the algorithm is beyond the scope of this paper. However, for bilevel learning, this issue was recently addressed by Ehrhardt and Roberts [25].

[7] While automatic differentiation [32] 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 [57] for further discussion of why derivative-free optimisation schemes are still needed.

6.2.2 Wavelet Denoising

We consider the wavelet denoising problem

$$u_\alpha = \operatorname*{arg\,min}_{u\in\mathbb{R}^m} \frac{1}{2}\|u - f^\delta\|^2 + \alpha\|Wu\|_1,$$

where W is the Haar wavelet transform. In particular, W is an orthogonal matrix, which implies that the regularisation problem has the unique solution

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

where T_α is the shrinkage operator defined by

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

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 Fig. 6 for the numerical results.

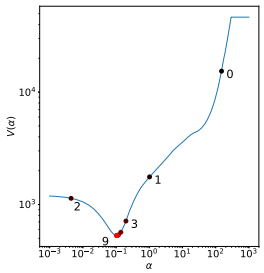
We also optimise α with respect to the scoring function $\Phi(u) := 1 - \mathrm{SSIM}(u, u^\dagger)$, where $\mathrm{SSIM} : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is the *structural similarity* function [72] defined for two images u and v as

$$\mathrm{SSIM}(u, v) := \frac{(2\mu_u\mu_v + c)(2\sigma_{uv} + C)}{(\mu_u^2 + \mu_v^2 + c)(\sigma_u^2 + \sigma_v^2 + C)}. \tag{6.6}$$

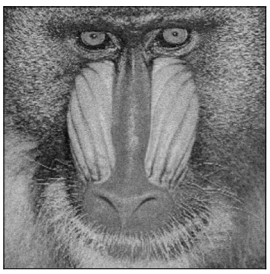
Here, μ_u is the mean intensity of u, σ_u is the unbiased estimate of the standard deviation of u, and σ_{uv} is the correlation coefficient between u and v:

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

The constants $c > 0$ and $C > 0$ are small parameters which provide stability to the measure and are set to 0.01 and 0.03, respectively. 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 Fig. 7 for the numerical results.

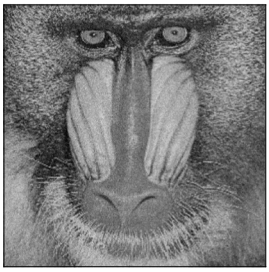


(a) Plot with labels.

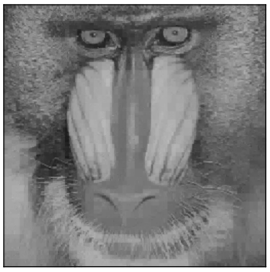


(b) Noisy image

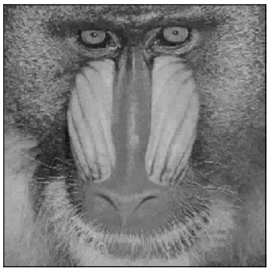
(c) $k = 1$. $\alpha = 1.02$.



(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 Haar wavelet denoising with $\|\cdot\|^2$ scoring function and the Itoh–Abe method. Top left: plot of iterates of the Itoh–Abe method. Top middle: the noisy image. The rest: image denoising results at different iterates k

6.2.3 Total Variation Denoising

We consider the TV denoising problem

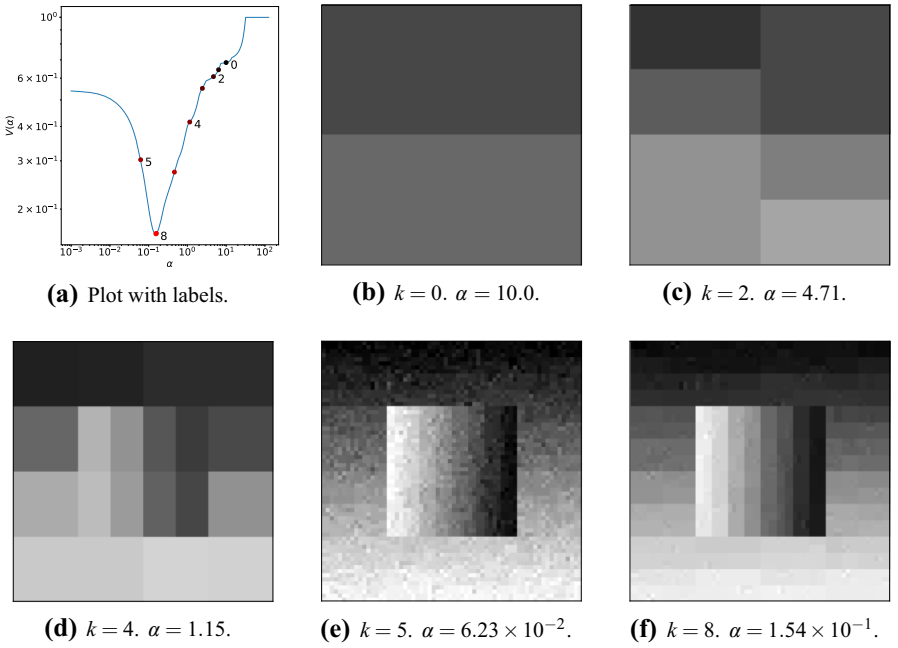
$$u_\alpha = \operatorname*{arg\,min}_{u \in \mathbb{R}^m} \frac{1}{2}\|u - f^\delta\|^2 + \alpha\,\mathrm{TV}(u),$$

with the SSIM scoring function. We solve the above denoising problem using 300 iterations of the PDHG method [16]. 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 Fig. 8 for the numerical results.

6.2.4 Total Generalised Variation Regularisation

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



(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. 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

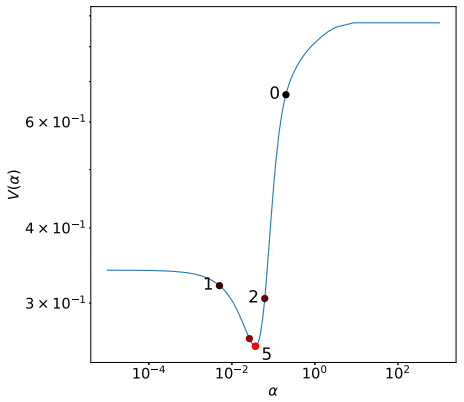
npt $= 2(n + 1)$ and the parameters of LT-MADS to DIRECTION_TYPE $=$ LT $2N$. See the results for two different starting points in Figs. 10 and 11. We note that the objective function is approximately stationary across a range of values, which leads to the different points of convergence, and different 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 neighbourhood of the minimiser.

6.3 Continuity and Boundedness Properties of Bilevel Problems

For the theoretical analysis in this paper, we assume that the objective function V is locally Lipschitz continuous, bounded below, and coercive. In what follows, we briefly discuss whether bilevel problems of the form (6.4) will satisfy these properties.

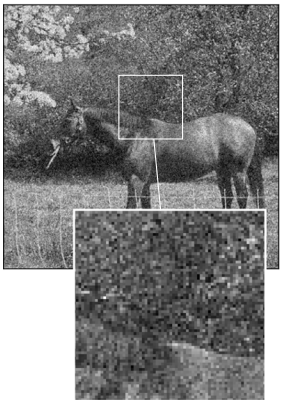
We first consider the assumption that V is bounded below. Since $V(\alpha) = \Phi(u_\alpha)$, this holds if Φ is bounded below, as is the case for $\Phi(u) = \|u - u^\dagger\|^2/2$ and $1 - \mathrm{SSIM}(u, u^\dagger)$ [72].

Coercivity does not hold for bilevel problems in general. This is because while the parameter domain $\mathbb{R}^n$ or $(0, \infty)^n$ is unbounded, the image in $\mathbb{R}^m$, $\{u_\alpha \in \mathbb{R}^m : \alpha \in \mathbb{R}^n\}$, is often bounded (one may verify this, for example, for $u_\alpha = \arg\min_{u \in \mathbb{R}} (u-1)^2/2 + \alpha|u|, \alpha > 0$), leading to $V(\alpha)$ flattening out as $\|\alpha\|$ grows. However, for non-coercive objective functions, if the iterates of the Itoh–Abe method $(x^k)_{k \in \mathbb{N}}$ admit an accumu-

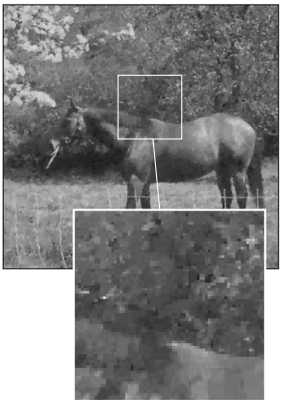


(a) Plot with labels.

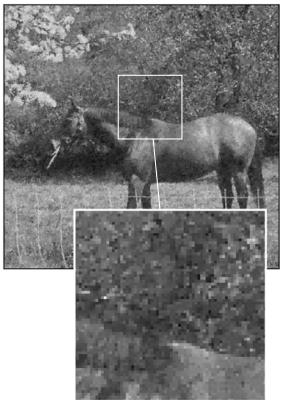
(b) $k = 0.\ \alpha = 2.00 \times 10^{-1}$.



(c) $k = 1.\ \alpha = 5.74 \times 10^{-3}$.



(d) $k = 2.\ \alpha = 6.18 \times 10^{-2}$.

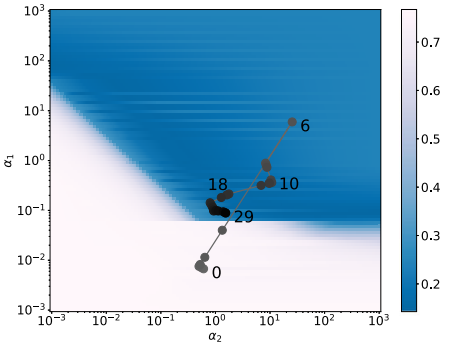


(e) $k = 5.\ \alpha = 3.63 \times 10^{-2}$.

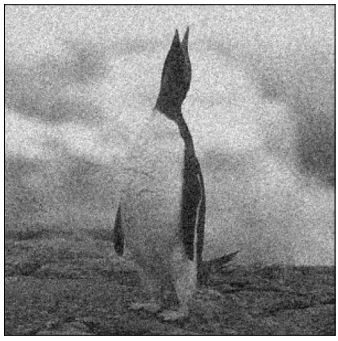
Fig. 8 TV denoising with SSIM scoring function and the Itoh–Abe method. Top left: plot of iterates of the Itoh–Abe method. The rest: image denoising result at different iterates k, with a zoom to show the difference

lation point, then it is straightforward to verify that the results from Theorems 3.4 and 3.7 still hold. Alternatively, one can impose coercivity by including regularisation of the parameters in the scoring function.

Third, we consider local Lipschitz continuity of bilevel problems. A sufficient condition for this is that both the scoring function Φ and the solution mapping $\alpha \mapsto u_\alpha$ are locally Lipschitz continuous. Local Lipschitz continuity of the scoring functions follows from continuous differentiability of $\|\cdot\|^2/2$ and SSIM, which can be verified directly. Local Lipschitz continuity of the solution mapping $\alpha \mapsto u_\alpha$ for general bilevel problems is beyond the scope of this paper, but we demonstrate it for a class of bilevel problems and afterwards mention some previous works that address this issue.



(a)



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

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

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

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

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

Fig. 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 iterations k

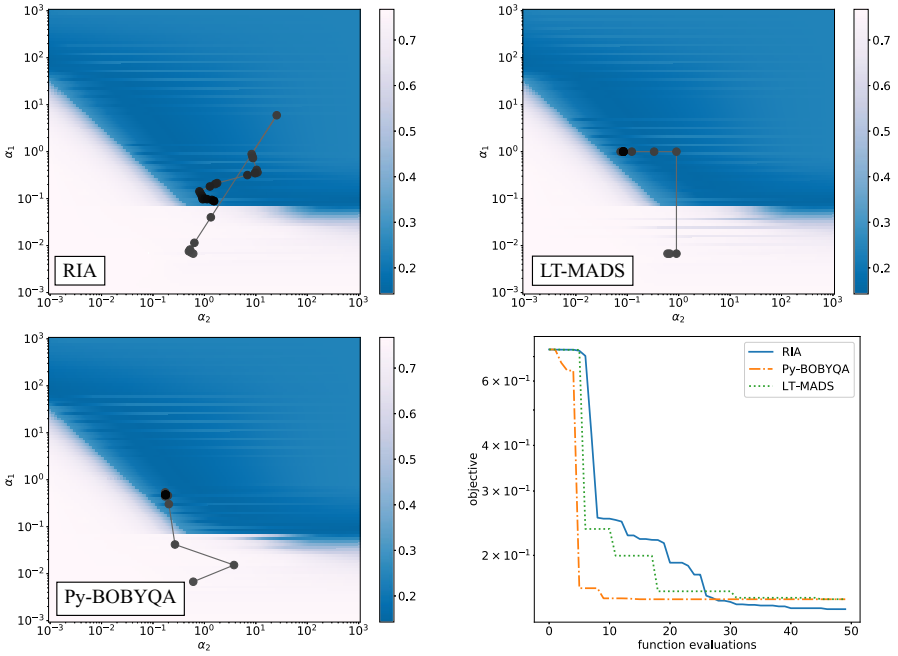


Fig. 10 Comparison of optimisation methods for TGV denoising with SSIM scoring function. Top left: plot of iterates of the Itoh–Abe method. Top right: plot of iterates of the LT-MADS method. Bottom left: plot of iterates of the Py-BOBYQA method. Bottom right: comparison of convergence rates for the methods with respect to function evaluations

We consider bilevel problems where for $\alpha > 0$

$$u_\alpha := \arg\min_u \left\{ \frac{1}{2}\|u - f^\delta\|^2 + \alpha R(u) \right\},$$

where R is a convex, proper function that is Lipschitz continuous of rank $L > 0$. For $\beta > \alpha > 0$, there exist unique subgradients $p_\alpha \in \partial R(u_\alpha)$ and $p_\beta \in \partial R(u_\beta)$ such that

$$u_\alpha = f^\delta - \alpha p_\alpha, \qquad u_\beta = f^\delta - \beta p_\beta.$$

Then, we compute

$$\begin{aligned} \|u_\alpha - u_\beta\|^2 &= \langle u_\beta - u_\alpha, \alpha p_\alpha - \beta p_\beta \rangle \\ &= (\beta - \alpha)\langle u_\alpha - u_\beta, p_\beta \rangle - \alpha \langle u_\alpha - u_\beta, p_\alpha - p_\beta \rangle \\ &\leq |\beta - \alpha| \|u_\alpha - u_\beta\| \|p_\beta\| \leq L|\beta - \alpha| \|u_\alpha - u_\beta\|, \end{aligned}$$

where the first and second inequalities follow from convexity and Lipschitz continuity of R, respectively, see [19, Propositions 2.2.9, 2.1.2]. Thus, we obtain $\|u_\alpha - u_\beta\| \leq$

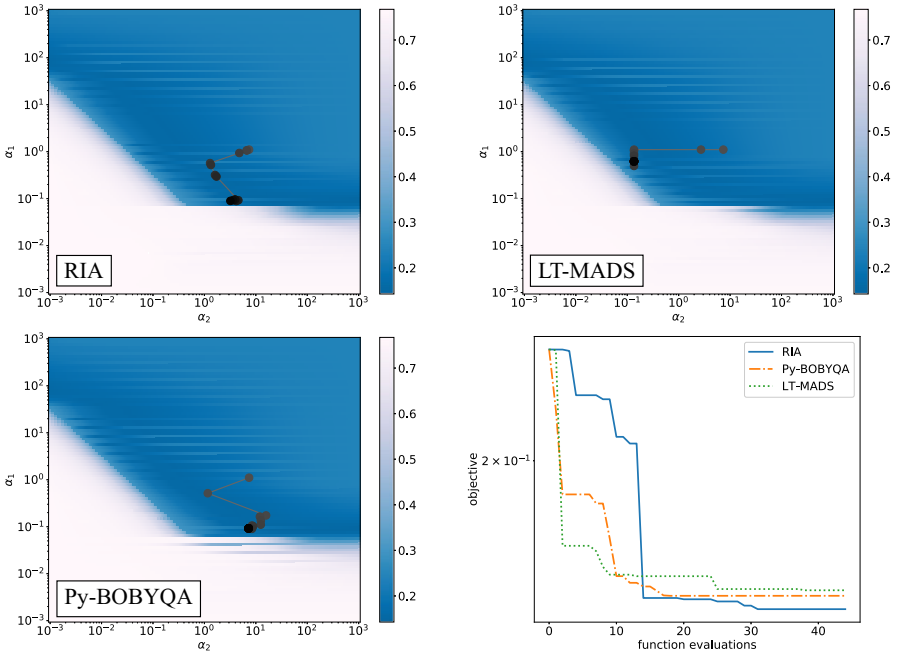


Fig. 11 Comparison of optimisation methods for TGV denoising with SSIM scoring function for a different starting point. Top left: plot of iterates of the Itoh–Abe method. Top right: plot of iterates of the LT-MADS method. Bottom left: plot of iterates of the Py-BOBYQA method. Bottom right: comparison of convergence rates for the methods with respect to function evaluations

$L|\beta - \alpha|$. This demonstrates local Lipschitz continuity for several of the bilevel problems we consider, including TV and wavelet denoising in Sects. 6.2.3 and 6.2.2.

We did not consider continuity properties of $\alpha \mapsto u_\alpha$ for more general regularisation problems. However, we refer the reader to the *strong regularity condition* formulated by Robinson [67], which can be used to show local Lipschitz continuity of solution mappings, for example, as was done by Hintermüller and Wu [38] for blind deconvolution problems.

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 extend 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 non-smooth Chebyshev–Rosenbrock functions. The efficiency of the Itoh–Abe discrete gradient method has already been demonstrated elsewhere for smooth problems [33,55,66] and sparse optimisation [7]. We also consider its application to bilevel learning problems and compare its performance to the derivative-free Py-BOBYQA and LT-MADS methods.

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 further analysing the Lipschitz continuity properties of bilevel optimisation for variational regularisation problems.

Acknowledgements The authors give thanks to Lindon Roberts for helpful discussions and for providing code for Py-BOBYQA, to Antonin Chambolle for helpful discussions, and to the reviewers for useful feedback.

Appendix A Probability Theory

In what follows, we show that a sequence $(d^k)_{k\in\mathbb{N}} \subset S^{n-1}$ of independent draws from a probability distribution Ξ is almost surely not cyclically dense. This is a simple consequence of the second Borel–Cantelli lemma [17, Theorem 4.2.4].

Proposition A.1 *Let $(d^k)_{k\in\mathbb{N}} \subset S^{n-1}$ be a sequence of independent draws from Ξ on S^{n-1}. Then, $(d^k)_{k\in\mathbb{N}}$ is almost surely not cyclically dense.*

Proof For $\varepsilon > 0$, $d^* \in S^{n-1}$, choose a set $U \subset S^{n-1}$ such that $p := \mathbb{P}_\Xi(d \in U) \in (0, 1)$ and $B_\varepsilon(d^*) \cap U = \varnothing$. For $k \in \mathbb{N}$, set $A_k = 1$ if $d^k \in U$ and $A_k = 0$ otherwise. Then, $(A_k)_{k\in\mathbb{N}}$ is a sequence of independent Bernoulli trials with success probability p.

We set $k_0 = 0$ and define recursively k_{i+1} as the largest integer such that $k_{i+1} - k_i \leq \log_{1/p}(i)$ for $i \in \mathbb{N}$. Denote by E_i the event that $A_j = 1$ for all $k_i \leq j < k_{i+1}$. Then, $(E_i)_{i\in\mathbb{N}}$ is a sequence of independent events, and $\mathbb{P}(E_i) = p^{k_{i+1}-k_i} \geq 1/i$. Thus, $\sum_{i\in\mathbb{N}} \mathbb{P}(E_i) = +\infty$, so by the second Borel–Cantelli lemma [17, Theorem 4.2.4], almost surely infinitely many of the events E_i occur.

Note that $k_{i+1} - k_i \to \infty$ as $i \to \infty$, since $\log_{1/p}(i) \to \infty$. Hence, if infinitely many of the events E_i occur, then the sequence $(d^k)_{k\in\mathbb{N}}$ contains strings of consecutive draws of arbitrary length which do not intersect with $S^{n-1}\backslash U \supset B_\varepsilon(d^*) \cap S^{n-1}$. This violates the conditions for cyclical density. Hence, $(d^k)_{k\in\mathbb{N}}$ is almost surely not cyclically dense. □

Appendix B Algorithms for Inner Solver

In this section, we provide full details on the approach to solving the Itoh–Abe discrete gradient equation (1.6). We furthermore ascertain that this algorithm terminates in a finite number of steps, and outputs a solution to (5.1).

Algorithm 4 Parabolic interpolation step. The algorithm seeks to achieve a large function decrease through a parabolic interpolation step. Initial steps $\beta_0, \beta_1, \beta_2$ are incrementally increased until parabolic interpolation criteria holds. If the function slope becomes too flat, i.e. $-f(\beta_2)/\beta_2^2 < 1/\tau_{\max}$, before this happens, then the algorithm returns β_2.

Input:

$f : \mathbb{R} \to \mathbb{R}$	scalar function
$\tau_{\min} > 0$	time step lower bound
$\tau_{\max} \in (\tau_{\min}, +\infty)$	time step upper bound
$\overline{\tau} \in (\tau_{\min}, \tau_{\max})$	proposed time step
$\varepsilon > 0$	tolerance for β
$\sigma \in (0, 1)$	search parameter

Output:

$\beta > 0$	best step after parabolic interpolation step

1: **function** INTERPOLATIONSTEP($f, \tau_{\min}, \tau_{\max}, \overline{\tau}, \varepsilon, \sigma$)
2: $\beta_0, \beta_1, \beta_2 \leftarrow 0, \varepsilon, -\dfrac{f(\varepsilon)\overline{\tau}}{\varepsilon}$ ▷ $\beta_2/\overline{\tau}$ solves (1.6) for linearisation $f(\beta) \approx \beta f(\varepsilon)/\varepsilon$
3: $\beta = \text{ParabolicDescent}(f, \beta_0, \beta_1, \beta_2)$ ▷ do parabolic descent step
4: **while** $\beta = 0$ **do** ▷ while parabolic criteria not fulfilled, increase β_i
5: **if** $-f(\beta_2)/\beta_2^2 < 1/\tau_{\max}$ **then**
6: **return** β_2 ▷ if slope is too flat, return β_2
7: **end if**
8: $\beta_0, \beta_1, \beta_2 \leftarrow \beta_1, \beta_2, \beta_2/\sigma$
9: $\beta = \text{ParabolicDescent}(f, \beta_0, \beta_1, \beta_2)$
10: **end while**
11: $\beta \leftarrow \arg\min\{f(\beta) \ : \ \beta \in \{\beta, \beta_1, \beta_2\}\}$
12: **return** β
13: **end function**

In what follows, we prove several statements about Algorithms 4 and 6.

Proposition B.1 *Algorithms 4 and 6 terminate after a finite number of steps.*

Proof We first prove this for Algorithm 4. The only loop statement in this algorithm is the line segment 4–10. For this loop to go on indefinitely, the difference quotients

$$\frac{f(c/\sigma^{j+1}) - f(c/\sigma^j)}{c/\sigma^{j+1} - c/\sigma^j} < 0$$

with $c = -f(\varepsilon)\overline{\tau}/\varepsilon > 0$, would need to be non-increasing with respect to $j \in \mathbb{N}$, due to the condition in Algorithm 5, line 2. This and the fact that $c/\sigma^j \to \infty$ as $j \to \infty$ imply that $f(c/\sigma^j) \to -\infty$, hence violating the assumption that V and f are bounded below. Thus, Algorithm 4 terminates in a finite number of steps.

We prove the same for Algorithm 6. In this algorithm, there are three loop statements, starting at lines 6, 11, and 21, respectively. The first and third loops are

Algorithm 5 Parabolic descent. If the slope of f from (β_0, β_1) to (β_1, β_2) increases, then do parabolic interpolation step. Otherwise, return 0.

Input:
$f : \mathbb{R} \to \mathbb{R}$ scalar function
$\beta_0 > 0$ first evaluation point
$\beta_1 > \beta_0$ second evaluation point
$\beta_2 > \beta_1$ third evaluation point
Output:
$\beta \geq 0$ parabolic descent step

function PARABOLICDESCENT(f,β_0,β_1,β_2)
  if $\frac{f(\beta_1)-f(\beta_0)}{\beta_1-\beta_0} \geq \frac{f(\beta_2)-f(\beta_1)}{\beta_2-\beta_1}$ **then** ▷ if parabolic criteria not fulfilled, return 0
    return 0
  else ▷ otherwise, do parabolic descent (see [36, Section 6.2.2])
    return $\beta_1 - \frac{1}{2}\frac{(\beta_1-\beta_0)^2(f(\beta_1)-f(\beta_2)) - (\beta_1-\beta_2)^2(f(\beta_1)-f(\beta_0))}{(\beta_1-\beta_0)(f(\beta_1)-f(\beta_2)) - (\beta_1-\beta_2)(f(\beta_1)-f(\beta_0))}$
  end if
end function

guaranteed to terminate, as otherwise we would have $\beta^* - \beta_* \to 0$, violating the criteria that $\beta^* - \beta_* > \varepsilon$. The second loop will terminate as a consequence of V (and thus f) being bounded below.

Finally, we verify that Algorithm 3 indeed solves (5.1).

Proposition B.2 *For input x and d, the output of Algorithm 3, $x + \beta d$, is a solution to the Itoh–Abe scalar equation as defined in (5.1).*

Proof We prove this on a case-by-case basis.

The first case is that Algorithm 3 finds in line 4 that V is directionally stationary at x along d up to ε-tolerance, and sets $\beta = 0$. If V is indeed directionally stationary at x along d, then clearly output x satisfies (5.1). Otherwise, then without loss of generality, we can assume that $V^o(x; d) < 0$, and by arguing as in the proof to Lemma 3.1, there exists $\beta \in (0, \varepsilon)$ such that $V(x + \beta d) - V(x) < -\beta^2/\tau_{\min}$. This together with the inequality $V(x+\varepsilon d) - V(x) \geq -\varepsilon^2/\tau_{\min}$ and the intermediate value theorem implies that there exists $\tilde{\beta} \in (\beta, \varepsilon)$ such that $\tilde{\beta}/\tau_{\min}$ solves (1.6) for the time step $\tau_{\min}$. It would follow that x satisfies (5.1).

The next cases are in Algorithm 6, lines 3, 16, and 18. In each of these cases, it follows by definition that the output $x + \beta d$ solves (5.1).

The final case is Algorithm 6, line 30, via line 27, as the outcome of the while loop starting at line 21. If, at the termination of the loop, the first condition fails, meaning that $-f(\beta)/\beta^2 \in [1/\tau_{\max}, 1/\tau_{\min}]$, then trivially the output $x + \beta d$ solves (5.1). Otherwise, if the second condition fails, then $|\beta^* - \beta_*| \leq \varepsilon$, and by the bounds $-f(\beta^*)/\beta^{*2} < 1/\tau_{\max}$ and $-f(\beta_*)/\beta_*^2 > 1/\tau_{\min}$, the continuity of $\beta \mapsto -f(\beta)/\beta^2$, and the intermediate value theorem [70, Theorem 4.23], for an arbitrary $\tau \in [\tau_{\min}, \tau_{\max}]$, there is $\tilde{\beta} \in [\beta_*, \beta^*]$ such that $\tilde{\beta}/\tau$ solves (1.6) for τ. Furthermore, since $\beta \in (\beta_*, \beta^*)$, we have $|\beta - \tilde{\beta}| < \varepsilon$ and thus $x + \beta d$ solves (5.1). This concludes the proof. □

Algorithm 6 Backtracking procedure. This algorithm finds a solution to (5.1), given input β from the interpolationStep in Algorithm 4. If $-f(\beta)/\beta^2 < 1/\tau_{\max}$, a solution is found in the interval $(0, \beta)$, and if $-f(\beta)/\beta_2 > 1/\tau_{\min}$, then a solution is found in $(\beta, +\infty)$, through standard search procedures.

Input:

$\beta > 0$	initial step
$f : \mathbb{R} \to \mathbb{R}$	scalar function
$\tau_{\min} > 0$	time step lower bound
$\tau_{\max} \in (\tau_{\min}, +\infty)$	time step upper bound
$\varepsilon > 0$	tolerance for β
$\sigma \in (0, 1)$	search parameter

Output:

$\beta > 0$	step corresponding to solution $x + \beta d$ to (5.1)

function BACKTRACKINGFUNCTION($f, \beta, \tau_{\min}, \tau_{\max}, \varepsilon, \sigma$)
  if $-f(\beta)/\beta^2 \in [1/\tau_{\max}, 1/\tau_{\min}]$ **then** ▷ backtracking part
    return β ▷ if time step constraints hold, return solution to (5.1)
  else if $-f(\beta)/\beta^2 < 1/\tau_{\max}$ **then** ▷ slope too flat $\implies$ there exists a solution $< \beta$ to (5.1)
    $\beta^* \leftarrow \beta$, $\beta_* \leftarrow \sigma\beta^*$
    while $-f(\beta_*)/\beta_*^2 < 1/\tau_{\max}$, $\beta^* - \beta_* > \varepsilon$ and $\beta_* > \varepsilon$ **do**
      $\beta^* \leftarrow \beta_*$, $\beta_* \leftarrow \sigma\beta_*$ ▷ decrease β_*, β^* until β_* is lower bound
    end while
  else ▷ slope too steep $\implies$ there exists a solution $> \beta$ to (5.1)
    $\beta_* \leftarrow \beta$, $\beta^* \leftarrow \beta^*/\sigma$
    while $-f(\beta^*)/\beta^{*2} > 1/\tau_{\min}$ **do**
      $\beta_* \leftarrow \beta^*$, $\beta^* \leftarrow \beta^*/\sigma$ ▷ increase β_*, β^* until β^* is upper bound
    end while
  end if
  if $-f(\beta^*)/\beta^{*2} \in [1/\tau_{\max}, 1/\tau_{\min}]$ **then** ▷ if β^* or β_* solve (5.1), return this
    return β^*
  else if $-f(\beta_*)/\beta_*^2 \in [1/\tau_{\max}, 1/\tau_{\min}]$ **then**
    return β_*
  else ▷ otherwise, find solution to (5.1) β in (β_*, β^*)
    $\beta \leftarrow \beta_* + \sigma(\beta^* - \beta_*)$
    while $-f(\beta)/\beta^2 \notin [1/\tau_{\max}, 1/\tau_{\min}]$ and $\beta^* - \beta_* > \varepsilon$ **do**
      if $-f(\beta)/\beta^2 < 1/\tau_{\max}$ **then**
        $\beta^* \leftarrow \beta$ ▷ if slope is too flat, reduce steps
      else
        $\beta_* \leftarrow \beta$ ▷ if slope is too steep, increase steps
      end if
      $\beta \leftarrow \beta_* + \sigma(\beta^* - \beta_*)$
    end while
  end if
  return β
end function

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-Łsojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
2. Audet, C., Dennis Jr, J.E.: Mesh adaptive direct search algorithms for constrained optimization. SIAM J. Optim. **17**(1), 188–217 (2006)
3. Audet, C., Hare, W.: Derivative-Free and Blackbox Optimization, 1st edn. Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, Switzerland (2017)

4. Aussel, D.: Subdifferential properties of quasiconvex and pseudoconvex functions: unified approach. J. Optim. Theory. Appl. **97**(1), 29–45 (1998)
5. Bagirov, A.M., Karasözen, B., Sezer, M.: Discrete gradient method: derivative-free method for nonsmooth optimization. J. Optim. Theory. Appl. **137**(2), 317–334 (2008)
6. Benning, M., Burger, M.: Modern regularization methods for inverse problems. Acta Numer. **27**, 1–111 (2018)
7. Benning, M., Riis, E.S., Schönlieb, C.B.: Bregman Itoh–Abe methods for sparse optimisation. J. Math. Imaging Vision (2020). https://doi.org/10.1007/s10851-020-00944-x
8. Borwein, J.M., Zhu, Q.J.: A survey of subdifferential calculus with applications. Nonlinear Ana. Theory Methods Appl. **38**(6), 687–773 (1999)
9. Bredies, K., Holler, M.: A TGV-based framework for variational image decompression, zooming, and reconstruction. part I: Analytics. SIAM J. Imag. Sci. **8**(4), 2814–2850 (2015)
10. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imag. Sci. **3**(3), 492–526 (2010)
11. Burger, M., Osher, S.: A guide to the TV zoo. In: M. Burger, S. Osher (eds.) Level Set and PDE Based Reconstruction Methods in Imaging: Cetraro, Italy 2008, pp. 1–70. Springer International Publishing, Cham (2013)
12. Burke, J.V., Curtis, F.E., Lewis, A.S., Overton, M.L., Simões, L.E.: Gradient sampling methods for nonsmooth optimization. In: A.M. Bagirov, M. Gaudioso, N. Karmitsa, M.M. Mäkelä, S. Taheri (eds.) Numerical Nonsmooth Optimization: State of the Art Algorithms, pp. 201–225. Springer International Publishing, Cham (2020)
13. Calatroni, L., Chung, C., De Los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: Bilevel approaches for learning of variational imaging models arXiv:1505.02120 (2015)
14. Cartis, C., Fiala, J., Marteau, B., Roberts, L.: Improving the flexibility and robustness of model-based derivative-free optimization solvers. ACM Trans. Math. Software **45**(3), 1–41 (2019)
15. Celledoni, E., Eidnes, S., Owren, B., Ringholm, T.: Dissipative numerical schemes on riemannian manifolds with applications to gradient flows. SIAM J. Sci. Comput. **40**(6), A3789–A3806 (2018)
16. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vision **40**(1), 120–145 (2011)
17. Chung, K.: A Course in Probability Theory, 3rd edn. Academic Press, San Diego (2001)
18. Clarke, F.H.: Necessary conditions for nonsmooth problems in optimal control and the calculus of variations. Ph.D. thesis, University of Washington (1973)
19. Clarke, F.H.: Optimization and Nonsmooth Analysis, 1st edn. Classics in Applied Mathematics. SIAM, Philadelphia (1990)
20. Curtis, F.E., Que, X.: An adaptive gradient sampling algorithm for non-smooth optimization. Optim. Methods Softw. **28**(6), 1302–1324 (2013)
21. Curtis, F.E., Que, X.: A quasi-Newton algorithm for nonconvex, nonsmooth optimization with global convergence guarantees. Math. Program. Comput. **7**(4), 399–428 (2015)
22. De los Reyes, J.C., Schönlieb, C.B., Valkonen, T.: Bilevel parameter learning for higher-order total variation regularisation models. J. Math. Imaging Vision **57**(1), 1–25 (2017)
23. DuPont, B., Cagan, J.: A hybrid extended pattern search/genetic algorithm for multi-stage wind farm optimization. Optim. Eng. **17**(1), 77–103 (2016)
24. 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:1805.06444 (2018)
25. Ehrhardt, M.J., Roberts, L.: Inexact derivative-free optimization for bilevel learning. J. Math. Imaging. Vis. **63**, 580–600 (2021)
26. Ekeland, I., Téman, R.: Convex Analysis and Variational Problems, 1st edn. SIAM, Philadelphia, PA, USA (1999)
27. Fasano, G., Liuzzi, G., Lucidi, S., Rinaldi, F.: A linesearch-based derivative-free approach for nonsmooth constrained optimization. SIAM J. Optim. **24**(3), 959–992 (2014)
28. 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. Adv. Water Resour. **31**(5), 743–757 (2008)
29. Giles, J.R.: A survey of Clarke's subdifferential and the differentiability of locally Lipschitz functions. In: A. Eberhard, R. Hill, D. Ralph, B.M. Glover (eds.) Progress in Optimization: Contributions from Australasia, pp. 3–26. Springer US, Boston, MA (1999)

30. Gonzalez, O.: Time integration and discrete Hamiltonian systems. J. Nonlinear Sci. **6**(5), 449–467 (1996)
31. Gray, G.A., Kolda, T.G., Sale, K., Young, M.M.: Optimizing an empirical scoring function for transmembrane protein structure determination. INFORMS J. Comput. **16**(4), 406–418 (2004)
32. Griewank, A., Walther, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia (2008)
33. Grimm, V., McLachlan, R.I., McLaren, D.I., Quispel, G.R.W., Schönlieb, C.B.: Discrete gradient methods for solving variational image regularisation models. J. Phys. A: Math. Theor. **50**(29) (2017). https://doi.org/10.1088/1751-8121/aa747c
34. Gürbüzbalaban, M., Overton, M.L.: On Nesterov's nonsmooth Chebyshev–Rosenbrock functions. Nonlinear Anal. Theory Methods Appl. **75**(3), 1282–1289 (2012)
35. 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)
36. Heath, M.T.: Scientific Computing: An Introductory Survey, 1st edn. McGraw-Hill, New York (2002)
37. Hintermüller, M.: A proximal bundle method based on approximate subgradients. Comput. Optim. Appl. **20**(3), 245–266 (2001)
38. Hintermüller, M., Wu, T.: Bilevel optimization for calibrating point spread functions in blind deconvolution. Inverse Prob. Imaging **9**(4), 1139–1169 (2015)
39. Ito, K., Jin, B.: Inverse Problems: Tikhonov Theory And Algorithms, 1st edn. Series On Applied Mathematics. World Scientific Publishing Company, Singapore (2014)
40. Itoh, T., Abe, K.: Hamiltonian-conserving discrete canonical equations based on variational difference quotients. J. Comput. Phys. **76**(1), 85–102 (1988)
41. Kämpf, J.H., Robinson, D.: Optimisation of building form for solar energy utilisation using constrained evolutionary algorithms. Energy Build. **42**(6), 807–814 (2010)
42. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In: P. Frasconi, N. Landwehr, G. Manco, J. Vreeken (eds.) Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811. Springer, Springer International Publishing, Cham (2016)
43. Kiwiel, K.C.: Methods of descent for nondifferentiable optimization, vol. 1133, 1st edn. Springer, Berlin (1985)
44. Kiwiel, K.C.: A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. SIAM J. Optim. **20**(4), 1983–1994 (2010)
45. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM J. Imag. Sci. **6**(2), 938–983 (2013)
46. Larson, J., Menickelly, M., Wild, S.M.: Derivative-free optimization methods. Acta Numer. **28**, 287–404 (2019)
47. Le Digabel, S.: Algorithm 909: NOMAD: Nonlinear optimization with the MADS algorithm. ACM Trans. Math. Software **37**(4), 1–15 (2011)
48. Le Digabel, S., Tribes, C., Audet, C.: NOMAD user guide. technical report g-2009-37. Tech. rep., Les cahiers du GERAD (2009)
49. Lewis, A.S., Overton, M.L.: Nonsmooth optimization via quasi-Newton methods. Math. Program. **141**, 135–163 (2013)
50. Liuzzi, G., Truemper, K.: Parallelized hybrid optimization methods for nonsmooth problems using NOMAD and linesearch. Comput. Appl. Math. **37**, 3172–3207 (2018)
51. 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)
52. 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)
53. McLachlan, R.I., Quispel, G.R.W., Robidoux, N.: Geometric integration using discrete gradients. Philos. Trans. A Math. Phys. Eng. Sci. **357**(1754), 1021–1045 (1999)
54. Michel, P., Penot, J.P.: Calcul sous-différentiel pour des fonctions lipschitziennes et non lipschitziennes. C. R. Acad. Sci. Paris **298**, 269–272 (1984)
55. Miyatake, Y., Sogabe, T., Zhang, S.L.: On the equivalence between SOR-type methods for linear systems and the discrete gradient methods for gradient systems. J. Comput. Appl. Math. **342**, 58–69 (2018)
56. Nelder, J.A., Mead, R.: A simplex method for function minimization. Comput. J. **7**(4), 308–313 (1965)

57. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017)
58. Ochs, P., Ranftl, R., Brox, T., Pock, T.: Bilevel optimization with nonsmooth lower level problems. In: J.F. Aujol, M. Nikolova, N. Papadakis (eds.) Scale Space and Variational Methods in Computer Vision, pp. 654–665. Springer International Publishing, Cham (2015)
59. Oeuvray, R., Bierlaire, M.: A new derivative-free algorithm for the medical image registration problem. Int. J. Model. Simul. **27**(2), 115–124 (2007)
60. Pathiraja, S., Reich, S.: Discrete gradients for computational Bayesian inference. J. Comput. Dyn. **6**(2), 385–400 (2019)
61. Penot, J.P., Quang, P.H.: Generalized convexity of functions and generalized monotonicity of set-valued maps. J. Optim. Theory. Appl. **92**(2), 343–356 (1997)
62. Polyak, B.T.: Introduction to Optimization, 1st edn. Optimization Software, Inc., New York (1987)
63. Powell, M.J.D.: The NEWUOA software for unconstrained optimization without derivatives. In: G. Di Pillo, M. Roma (eds.) Large-Scale Nonlinear Optimization, 1st edn., pp. 255–297. Springer US, Boston, MA (2006)
64. Powell, M.J.D.: The BOBYQA algorithm for bound constrained optimization without derivatives. Tech. rep., University of Cambridge (2009)
65. Quispel, G.R.W., Turner, G.S.: Discrete gradient methods for solving ODEs numerically while preserving a first integral. J. Phys. A: Math. Gen. **29**(13), 341–349 (1996)
66. Ringholm, T., Lazic, J., Schonlieb, C.B.: Variational image regularization with Euler's elastica using a discrete gradient scheme. SIAM J. Imaging Sci. **11**(4), 2665–2691 (2018)
67. Robinson, S.M.: Strongly regular generalized equations. Math. Oper. Res. **5**(1), 43–62 (1980)
68. Rosenbrock, H.H.: An automatic method for finding the greatest or least value of a function. Comput. J. **3**(3), 175–184 (1960)
69. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1-4), 259–268 (1992)
70. Rudin, W.: Principles of Mathematical Analysis, 3rd edn. International series in pure and applied mathematics. McGraw-Hill, New York (1976)
71. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging, 1st edn. Applied Mathematical Sciences. Springer, New York (2008)
72. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)

Affiliations

Erlend S. Riis[1] · Matthias J. Ehrhardt[2] · G. R. W. Quispel[3] · Carola-Bibiane Schönlieb[1]

✉ Erlend S. Riis
erlend.s.riis@gmail.com

Matthias J. Ehrhardt
m.ehrhardt@bath.ac.uk

G. R. W. Quispel
r.quispel@latrobe.edu.au

Carola-Bibiane Schönlieb
cbs31@cam.ac.uk

[1] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

2 Department of Mathematical Sciences, University of Bath, Bath, UK

3 Department of Mathematics and Statistics, La Trobe University, Victoria 3086, Australia