A geometric integration approach to non-smooth and non-convex optimisation

Martin Benning[1], Matthias Ehrhardt[2], GRW Quispel[3],
Erlend Skaldehaug Riis[4], Torbjørn Ringholm[5], Carola-Bibiane Schönlieb[4]

[1]Queen Mary University of London, UK
[2]University of Bath, UK
[3]La Trobe University, Australia
[4]University of Cambridge, UK
[5]Norwegian University of Science and Technology, Norway

Variational Methods and Optimization in Imaging
IHP, Paris

UNIVERSITY OF CAMBRIDGE

Outline

1. Geometric numerical integration and the discrete gradient method
2. The DG method for nonsmooth, nonconvex optimisation
3. Beyond gradient flow

Geometric numerical integration and the discrete gradient method

Optimisation and numerical integration

- $\min_{x\in\mathbb{R}^n} V(x), \qquad \dot{x}(t) = -\nabla V(x(t)) \qquad$ (gradient flow)
- Forward Euler $\quad\rightarrow\quad x^{k+1} = x^k - \tau\nabla V(x^k)$,
  Backward Euler $\quad\rightarrow\quad x^{k+1} = x^k - \tau\nabla V(x^{k+1})$.
- Numerical integration and analysis of ODEs
  - Optimisation scheme $\rightarrow$ ODE
    Accelerated gradient descent $\rightarrow \ddot{x} + \frac{3}{t}\dot{x} = -\nabla V(x)$
    [Su, Boyd, Candès (2016)]
  - Numerical integration tools to improve efficiency
    Runge-Kutta methods for stiff ODEs $\rightarrow$ Larger time steps
    [Eftekhari, Vandereycken, Vilmart, Zygalakis (2018)]
  - Discretisation methods for structure preservation of ODE
    Symmetry preservation $\rightarrow$ acceleration phenomenon
    [Betancourt, Jordan, Wilson (2018)]

Geometric integration and discrete gradients

- Geometric numerical integration
  - ODEs have structure (conservation laws, symplectic structure.)
  - Aim: Preserve structure when numerically solving ODEs
- Discrete gradient (DG) method
  - Preserves first integrals; energy conservation and dissipation laws, Lyapunov functions[1]
- Optimisation methods
  - Want to solve $\min_{x \in \mathbb{R}^n} V(x)$.
  - Apply DG method to gradient flow to preserve dissipative structure

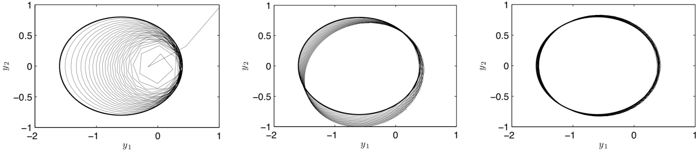


Figure: Dahlby, Owren, Yaguchi (2011).

'Why geometric numerical integration?' [Iserles, Quispel (2018)]
[1] McLachlan, Quispel, and Robidoux (1999).

Discrete gradient method

Definition

For a smooth function $V : \mathbb{R}^n \to \mathbb{R}$, a discrete gradient $\overline{\nabla} V(x, y)$ satisfies

(i) $\lim_{y \to x} \overline{\nabla} V(x, y) = \nabla V(x)$ (consistency),

(ii) $\langle \overline{\nabla} V(x, y), y - x \rangle = V(y) - V(x)$ (mean value).

$$\min V : \mathbb{R}^n \to \mathbb{R}$$

Discrete gradient method.

$$x^{k+1} = x^k - \tau \overline{\nabla} V(x^k, x^{k+1})$$

Dissipative:

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

Gradient flow.

$$\dot{x}(t) = -\nabla V(x(t)).$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} V(x(t)) &= \langle \dot{x}(t), \nabla V(x(t)) \rangle \\ &= -\|\nabla V(x(t))\|^2 \\ &= -\|\dot{x}(t)\|^2. \end{aligned}$$

Theorem[1]

Suppose V is C^1-smooth, coercive, and bounded below, $0 < c < \tau < C$, and $x^{k+1} = x^k - \tau\overline{\nabla}V(x^k, x^{k+1})$. Then

- $\nabla V(x^k) \to 0$,
- $\|x^{k+1} - x^k\| \to 0$,
- (x^k) has an accumulation point x^*, and it satisfies $\nabla V(x^*) = 0$.

Figure: Inpainting with Itoh-Abe DG method[1]

[1] Grimm, McLachlan, McLaren, Quispel and Schönlieb (2017).

Examples of discrete gradients

- Gonzalez (midpoint) DG[1]:

$$\overline{\nabla} V(x,y) = \nabla V\left(\tfrac{x+y}{2}\right) + \frac{V(y)-V(x)-\langle \nabla V(\frac{x+y}{2}), y-x\rangle}{\|x-y\|^2}(y-x).$$

- Mean value DG[2]:

$$\overline{\nabla} V(x,y) = \int_0^1 \nabla V\big((1-s)x + sy\big)\,\mathrm{d}s.$$

- Itoh-Abe (coordinate increment) DG[3]:

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

[1] Gonzalez (1996). [2] Celledoni, Grimm, et al. (2012). [3] Itoh and Abe (1988).

Applications

- Image inpainting with Euler's elastica (nonconvex).[1]
- Successive-over-relaxation (SOR) and the Gauss-Seidel method, for linear systems $Ax = b$.[2]
  - Kaczmarz methods (by extension)

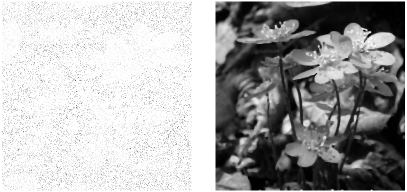


Figure: Inpainting with Itoh-Abe DG method[1].

[1] Ringholm, Lasić and Schönlieb (2017). [2] Miyatake, Sogabe, Zhang (2017).

Itoh–Abe method for nonsmooth, nonconvex optimisation

Rewrite IADG method as

$$x^{k+1} = x^k - \alpha d^k, \quad \text{s.t.} \quad \alpha = \tau_k \frac{V(x^k - \alpha d^k) - V(x^k)}{\alpha}, \qquad (1)$$
$$d^k \in S^{n-1} := \{d \in \mathbb{R}^n \ : \ \|d\| = 1\}.$$

- Derivative-free gradient flow dissipative structure

$$V(x^{k+1}) - V(x^k) = -\frac{1}{\tau^k}\|x^{k+1} - x^k\|^2 = -\tau^k \left(\frac{V(x^{k+1} - V(x^k)}{\|x^{k+1} - x^k\|}\right)^2$$

- Well-defined for nonsmooth functions; computationally tractable
- Descends along directions $(d^k)_{k\in\mathbb{N}}$
    - Standard IADG: Let (d^k) cycle through coordinates e^i
    - Can also randomise: Draw d^k randomly from $(e^i)_{i=1}^n$ or from S^{n-1}

Motivations

- When the function is nonsmooth, nonconvex, and black-box
  - Bilevel optimisation of variational regularisation problems
  - Parameter optimisation of model simulations

    'Optimal camera placement to measure distances regarding static and dynamic obstacles' [Hänel et al. (2012)]
- When gradients are computationally expensive
- When problem is poorly conditioned or stiff.

The DG method for nonsmooth, nonconvex optimisation

Clarke subdifferential for nonsmooth, nonconvex analysis

- The Clarke subdifferential[1] $\partial V(x)$ introduced by F. Clarke (1973).
- $\partial V(x) = \text{co}\left\{p \quad \text{s.t.} \quad x^k \to x,\ \nabla V(x^k) \to p\right\}$ (Convex hull of limiting gradients).
- Generalises gradient, and classical subdifferential for convex functions.
- Nice analytical properties for locally Lipschitz continuous functions. (outer semicontinuity; mean value theorem; convex, compact, non-empty sets)
- x is Clarke stationary when $0 \in \partial V(x)$.
- Clarke directional derivative:

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

- $0 \in \partial V(x) \quad \Longleftrightarrow \quad V^o(x; d) \geq 0$ for all $d \in S^{n-1}$

[1] Clarke (1990).

Ingredients of proof

Want to prove accumulation points of $(x^k)_{k\in\mathbb{N}}$ are Clarke stationary.

- Local Lipschitz continuity of $V \Rightarrow$ upper semicontinuity of $V^o(\,\cdot\,;\cdot\,)$.
  (closely related to outer semicontinuity of ∂V)
- $\|x^{k+1} - x^k\| \to 0$, $\qquad \frac{V(x^{k+1})-V(x^k)}{\|x^{k+1}-x^k\|} \to 0$.
- $(x^{k_j}, d^{k_j}) \to (x^*, d^*)$ and $\liminf_{j\to\infty} V^o(x^{k_j}; d^{k_j}) \geq 0 \Rightarrow V^o(x^*; d^*) \geq 0$.
- Need to ensure $\exists$ dense subsequence d^{k_j} corresponding to $x^{k_j} \to x^*$.
  - For random $(d^k)_{k\in\mathbb{N}}$, $\iff$ full support of distribution of d^k on S^{n-1}.
  - For determinstic (d^k), $\iff$ "cyclical" density.

Main theorem

Theorem (Ehrhardt, Riis, Quispel, Schönlieb (2018))

Let $V : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz continuous, coercive function that is bounded below. Suppose (x^k) are the iterates from the generalised Itoh-Abe DG method with appropriate sequence of directions $(d_k)_{k\in\infty}$. Then

- *The iterates converge to a nonempty, connected, compact set of accumulation points.*
- *All accumulation points are Clarke stationary.*
- *All accumulation points take the same value on V.*

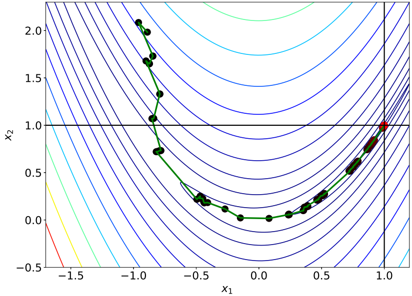
Other properties of DG method

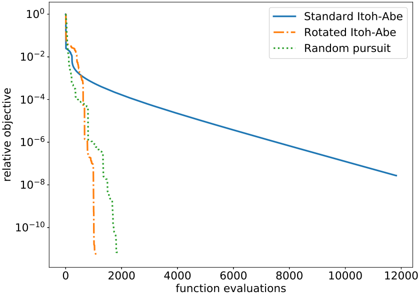
When V is C^1-smooth, the DG methods inherit properties from gradient descent/flow:

- Convergence rates:[1] $V(x^k) - V^* \to 0$.
  - $\mathcal{O}(1/k)$ if V is convex.
  - Linearly if V is Polyak–Łojasiewicz function/ strongly convex.
  - Itoh–Abe method has marginally better convergence rate than coordinate descent.
- Kurdyka–Łojasiewicz inequality $\implies$ $(x^k)_{k\in\mathbb{N}}$ converges.
- Properties hold for all time steps $c < (\tau^k)_{k\in\mathbb{N}} < C, \quad c, C > 0$.

[1] Ehrhardt, Riis, Ringholm, Schönlieb (2018).

Rosenbrock function





Beyond gradient flow

Bregman distance and inverse scale space flow

- Let $J : \mathbb{R}^n \to \mathbb{R}$ be convex, e.g.

$$J(x) = \gamma\|x\|_1 \quad \text{or} \quad \gamma\|x\|_1 + \|x\|^2/2 \quad \text{or} \quad \gamma\,\mathrm{TV}(x).$$

- Define Bregman distance (notion of distance induced by J)

$$D_J^p(x, y) = J(y) - J(x) - \langle p, y - x\rangle \geq 0, \quad p \in \partial J(x).$$

- For inverse problem $b = Ax$, want to solve

$$\min_x J(x) \text{ s.t. } b = Ax \quad \text{or} \quad \min_x V(x) + \lambda J(x),$$

where $V(x) = \|Ax - b\|^2/2$.

- Consider inverse scale space (ISS) flow[1]

$$\partial_t p(t) = -\partial V(x(t)), \quad p(t) \in \partial J(x(t)).$$

[1] Burger, Gilboa, Osher, Xu (2006).

Discretisations of inverse scale space

$$\partial_t p(t) = -\partial V(x(t)), \quad p(t) \in \partial J(x(t)).$$

- Backward Euler $\rightarrow$ Bregman iterations[1]:

$$p^{k+1} = p^k - \tau^k \nabla V(x^{k+1}), \quad p^{k+1} \in \partial J(x^{k+1})$$
$$\iff x^{k+1} = \arg\min_x \tau^k V(x) + D_J^{p^k}(x^k, x).$$

- Forward Euler $\rightarrow$ Linearised Bregman iterations[1]:

$$p^{k+1} = p^k - \tau^k \nabla V(x^k), \quad p^{k+1} \in \partial J(x^{k+1})$$
$$\iff x^{k+1} = \arg\min_x \tau^k \left(V(x^k) + \langle \nabla V(x^k), x - x^k \rangle \right) + D_J^{p^k}(x^k, x).$$

- DG method $\rightarrow$ Bregman DG method:

$$p^{k+1} = p^k - \tau^k \overline{\nabla} V(x^k, x^{k+1}), \quad p^{k+1} = \partial J(x^{k+1}).$$

[1] Benning, Burger (2018).

Bregman Itoh–Abe method

$$p_i^{k+1} = p_i^k - \tau_i^k \frac{V(x_1^{k+1}, \dots, x_i^{k+1}, x_{i+1}^k, \dots, x_n^k) - V(x_1^{k+1}, \dots, x_{i-1}^{k+1}, x_i^k, \dots, x_n^k)}{x_i^{k+1} - x_i^k}.$$

- If V is continuous, and J is continuous and strongly convex, then updates are well-defined.
- If V is convex, updates are unique.
- Iterates $(x^k)_{k\in\mathbb{N}}$ converge to Clarke stationary points (under regularity assumption).
- Implications:
  - Can adapt pre-existing methods to incorporate bias (e.g. sparsity, variational regularisation problems)
  - Handles nonsmoothness better (e.g. $\|\cdot\|_1$ kinks)

Example: Bregman Itoh–Abe for linear systems (SOR) (1/2)

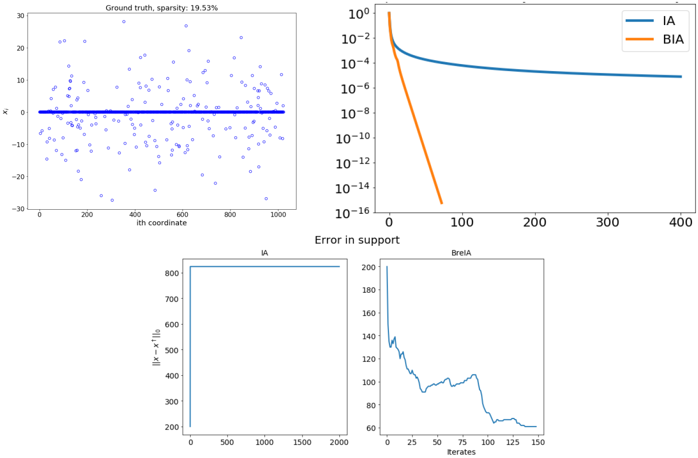


Figure: Top left: Ground truth. Top right: Normalised residual of data fidelity. Bottom: Error in support of iterate, $\text{supp}\,(x^k)$, to support of ground truth, $\text{supp}\,(x^*)$.

Example: Bregman Itoh–Abe for linear systems (SOR) (2/2)

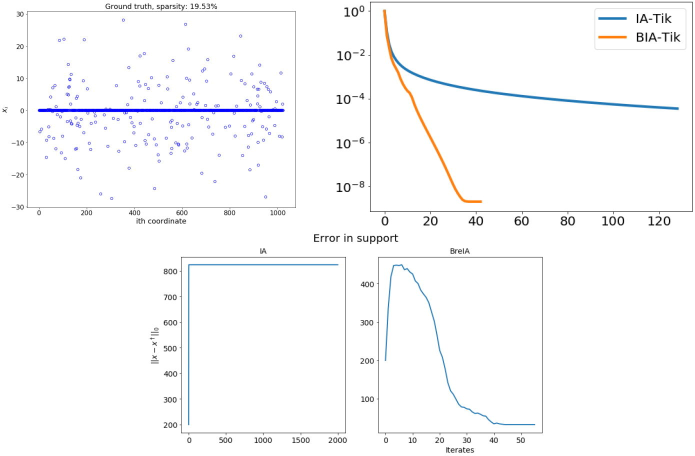


Figure: Top left: Ground truth. Top right: Normalised residual of data fidelity. Bottom: Error in support of iterate, $\operatorname{supp}(x^k)$, to support of ground truth, $\operatorname{supp}(x^*)$.

Outlook

- Accelerate Itoh–Abe DG method.
- Apply discrete gradient methods to gradient flow under different metrics(e.g. optimal transport)

Thank you for your attention!

Relevant papers

1. Riis, Ehrhardt, Quispel, Schönlieb. *A geometric integration approach to nonsmooth, nonconvex optimisation.* (2018, *preprint*)
2. Ehrhardt, Riis, Ringholm, Schönlieb. *A geometric integration approach to smooth optimisation: Foundations of the discrete gradient method.* (2018, *preprint*)
3. Grimm, McLachlan, McLaren, Quispel, Schönlieb. *Discrete gradient methods for solving variational image regularisation models.* (2017)
4. Ringholm, Lazić, Schönlieb. *Variational image regularization with Euler's elastica using a discrete gradient scheme.* (2017)