First-order geometric multilevel optimization for discrete tomography

Plier, Jan; Savarino, Fabrizio; Kocvara, Michal; Petra, Stefania

DOI:
10.1007/978-3-030-75549-2_16

Document Version
Peer reviewed version

Citation for published version (Harvard):
Plier, J, Savarino, F, Kocvara, M & Petra, S 2021, First-order geometric multilevel optimization for discrete tomography. in A Elmoataz, J Fadili, Y Queau, J Rabin & L Simon (eds), Scale Space and Variational Methods in Computer Vision: 8th International Conference, SSVM 2021, Virtual Event, May 16–20, 2021, Proceedings. Lecture Notes in Computer Science, vol. 12679, Springer, pp. 191-203, SSVM 2021: Scale Space and Variational Methods in Computer Vision, 16/05/21. https://doi.org/10.1007/978-3-030-75549-2_16

First-Order Geometric Multilevel Optimization for Discrete Tomography

Jan Plier[1,3], Fabrizio Savarino[2][0000−0001−9629−8486], Michal Kočvara[4], and Stefania Petra[3][0000−0002−7189−2275]

[1]Heidelberg Institute for Theoretical Studies, Germany
[2]Image and Pattern Analysis Group, Heidelberg University, Germany
[3]Mathematical Imaging Group, Heidelberg University, Germany
[4]School of Mathematics, University of Birmingham, United Kingdom and Institute of Information Theory and Automation, Prague, Czech Republic

Abstract. Discrete tomography (DT) naturally leads to a hierarchy of models of varying discretization levels. We employ *multilevel optimization (MLO)* to take advantage of this hierarchy: while working at the fine level we compute the search direction based on a coarse model. Importing concepts from information geometry to the n-orthotope, we propose a smoothing operator that only uses first-order information and incorporates constraints smoothly. We show that the proposed algorithm is well suited to the ill-posed reconstruction problem in DT, compare it to a recent MLO method that nonsmoothly incorporates box constraints and demonstrate its efficiency on several large-scale examples.

1 Introduction

This paper introduces a *geometric* multilevel optimization approach for solving

$$\min_{x\in\mathbb{R}^n} f(x), \qquad f(x) = \mathrm{KL}(Ax, b) + \lambda\|Dx\|_{1,\tau} + \delta_C(x) \tag{1}$$

to recover a discretized function x on a spatial domain from linear projection measurements $b = Ax$ by minimizing the Kullback-Leibler (KL) divergence and a sparsity promoting prior subject to box constraints $C = [l, u] \subset \mathbb{R}^n_+$ – see Fig. 1 for an illustration. We aim at exploiting 'geometry' in a twofold way. On one hand, multiple grid sizes are used for discretizing the domain with different resolutions, which mitigates the ill-posedness of the inverse recovery problem at coarser levels. On the other hand, by turning the bounded interior of the convex feasible set into a Riemannian manifold, the geometry of the space makes first-order updates of the iterate x more efficient. Our approach combines these design aspects in a principled manner using problem (1) and discrete tomography [1]

Acknowledgments: Dr. Jan Plier gratefully acknowledges the generous and invaluable support of the Klaus Tschira Foundation.

as a scenario that is representative for a range of approaches to inverse problems using constrained convex optimization.

Related work. Seminal work on unconstrained smooth multilevel optimization includes [2–4]. These ideas were elaborated for nonsmooth convex optimization in [5, 6]. Our approach is applicable to such problems but essentially relies on smoothness induced by changing the geometry of the feasible set. Regarding discrete tomography, multiresolution approaches include [7, 8] with a focus on filtered backprojection and heuristics for acceleration, whereas our approach solves a constrained optimization problem.

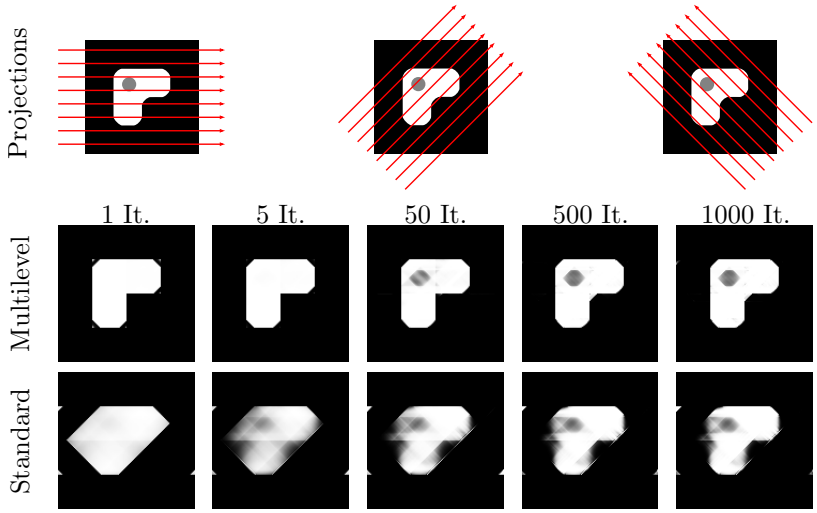

Fig. 1: **Scenario and approach. Top row:** In discrete tomography, we reconstruct *finite* range functions from a finite set of parallel projections. The incidence relation of projection rays and the discretized domain (pixels in 2D, voxels in 3D) is represented by a matrix A. All line integrals are collected in a vector b. **Bottom row:** Comparing to a standard iterative reconstruction scheme for solving (1), our novel multilevel approach recovers more efficiently the large scale structure and subsequently the fine scale structure of the unknown function.

Contribution and organization. Section 2 introduces essential concepts of multilevel optimization. Our geometric multilevel optimization approach is introduced in Section 3, building on [9]. In Section 4, we show results of single- and multilevel optimization and compare them to [5]. Our large scale experiments show that our approach is on par with the state of the art and holds potential for further elaboration.

Basic notation. $\langle \cdot, \cdot \rangle$ denotes the standard inner product on $\mathbb{R}^n$, ∇f the gradient and $\nabla^2 f$ the Hessian of a sufficiently differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. We denote componentwise multiplication of vectors by $uv = (u_1 v_1, \ldots, u_n v_n)^\top$

and, for strictly positive vectors $v \in \mathbb{R}^n_{++}$, componentwise division by $\frac{u}{v}$. Likewise, the functions e^x and $\log x$ apply componentwise to a vector x. For a smooth Riemannian manifold $(\mathcal{M}, g)$ with metric g, $T_x\mathcal{M}$ denotes the tangent space at $x \in \mathcal{M}$ and $\mathrm{d}_x f : T_x\mathcal{M} \to \mathbb{R}$ the differential (aka tangent map) of a smooth function $f : \mathcal{M} \to \mathbb{R}$. The Riemannian gradient $\nabla_{\mathcal{M}} f(x) \in T_x\mathcal{M}$ of f is uniquely defined by $\mathrm{d}_x f[\xi] = g_x\left(\nabla_{\mathcal{M}} f(x), \xi\right),\ \forall \xi \in T_x\mathcal{M}$.

2 Multilevel Optimization in Euclidean Space

Two level optimization. We next describe a two grid cycle, that is computing an update x^+ at a fine grid from the current iterate x. This is done either by a search direction obtained from a model defined on a coarse grid using a much smaller number of variables (*coarse correction*) or, whenever coarse correction is not effective, by a standard local approximation defined on the fine grid (*fine correction*). The general approach is summarized in Algorithm 2.1.

Algorithm 2.1: Two Level Optimization

initialization: Set $i = 0$ and choose initial point x, two grids, transfer operators R and P and a coarse representation $\overline{f}$ of the objective.
repeat
  if *condition to use coarse model is satisfied at* x **then**
    Define coarse model $\overline{\psi}(\overline{y}; \overline{x}, \overline{f}, R\nabla f(x))$. `/* coarse model */`
    Find a descent direction $\overline{d}$ w.r.t. the fine objective f at x using $\overline{\psi}$.
    Set $d = P\overline{d}$.
    Find $\alpha > 0$ such that $f(x + \alpha d) < f(x)$. `/* line search */`
    $x \leftarrow x + \alpha d$.
  else
    Apply one iteration of the monotone fine level algorithm to find x^+ with $f(x^+) < f(x)$ and update $x \leftarrow x^+$.
  Increment $i \leftarrow i + 1$.
until *a stopping rule is met.*

The key question is how to represent the problem on a coarser grid. The starting point is the coarse discretization of the fine grid objective $f\colon \mathbb{R}^n \to \mathbb{R}$, denoted by $\overline{f}\colon \mathbb{R}^{\overline{n}} \to \mathbb{R}$, that represents f on the coarse grid in a meaningful way. We use the following notation

$y \in \mathbb{R}^n$: fine grid variable,	$\overline{y} \in \mathbb{R}^{\overline{n}}$: coarse grid variable,
$d \in \mathbb{R}^n$: search direction on fine grid,	$\overline{d} \in \mathbb{R}^{\overline{n}}$: search direction on coarse grid.

We assume linear maps $R\colon \mathbb{R}^n \to \mathbb{R}^{\overline{n}}$ and $P\colon \mathbb{R}^{\overline{n}} \to \mathbb{R}^n$, called *restriction* and *prolongation* to be given, for translating various quantities between the coarse

and fine grid level. Typically, information transfer between levels is done via linear interpolation or simple injection as in classical multigrid techniques [10].

Coarse model. The central principle is defining the *coarse grid model*, first proposed in [2], given by

$$\overline{\phi}_x(\overline{y}) := \overline{f}(\overline{y}) - \langle \overline{v}_x, \overline{y} \rangle, \tag{2a}$$

$$\overline{v}_x := \nabla \overline{f}(\overline{x}) - R\nabla f(x), \quad \overline{x} := Rx, \tag{2b}$$

which is based on a linear modification of the coarse grid objective $\overline{f}$. In the following, we drop the explicit dependence of $\overline{\phi}$ and $\overline{v}$ on x for simplifying notation. The objective of the coarse grid model is to determine a gradient-like descent direction in an efficient way using a much smaller number of coarse grid variables. For the *initial* iterate of the coarse grid $\overline{x}$ defined in (2b), we have

$$\nabla \overline{\phi}(\overline{x}) = R\nabla f(x). \tag{3}$$

This property, also known as the *first order coherence condition*, ensures that a critical point of the objective function on the fine grid is also a critical point of the coarse model when transferred to the coarse grid. Note, that at this stage we have *not* imposed a relation between the intergrid transfer operators P and R. The update is defined as

$$x^+ = x + \alpha d, \tag{4a}$$

$$d = P(\overline{y_*} - \overline{x}), \qquad \overline{y_*} = \arg\min_{\overline{y}} \overline{\phi}(\overline{y}). \tag{4b}$$

Remark 1. We should underline that $\overline{y_*}$ in (4b) is typically replaced by an *inexact* solution of the coarse model (2) obtained by some iterative method.

Relation to the FAS. The coarse model $\overline{\phi}$ is closely related to the coarse grid correction equation of the FAS (full approximation scheme) in the context of multigrid methods for nonlinear equation [10, Chap. 5.3]. Applying FAS for solving the nonlinear critical point equation $\nabla f(y) = 0$ at the approximation x gives the *coarse grid correction*, see [10, Eq. 5.3.13]

$$\nabla \overline{f}(\overline{x} + \overline{y}) - \nabla \overline{f}(\overline{x}) = \overline{r}, \quad \overline{r} := 0 - R\nabla f(x), \tag{5}$$

that needs to be solved for $\overline{y}$. In FAS *both* the current approximation x and the residual, here $r := 0 - \nabla f(x)$, are transferred to the coarse grid. The coarse grid correction is defined in terms of $\overline{x}$, $\overline{r}$ and the coarse representation of the nonlinear equation. A solution of the coarse grid correction equation (5) is a critical point of $\overline{y} \mapsto \overline{\phi}(\overline{x} + \overline{y})$ in (2).

Bregman gap, coarse model-based descent direction. The following notion will be used for evaluating the coarse grid model.

Definition 1 (Coarse model, Bregman gap). *Given a differentiable function $\overline{f} \colon \mathbb{R}^{\overline{n}} \to \mathbb{R}$ and $\overline{x} \in \mathbb{R}^{\overline{n}}$ define the coarse model by*

$$\overline{\psi}_{x,\overline{x}}(\overline{y}) := B_{\overline{f}}(\overline{x} + \overline{y}, \overline{x}) + \langle \nabla f(x), P\overline{y} \rangle, \tag{6a}$$

with Bregman gap

$$B_{\overline{f}}(\overline{x}+\overline{y},\overline{x}) := \overline{f}(\overline{x}+\overline{y}) - \overline{f}(\overline{x}) - \langle \nabla \overline{f}(\overline{x}), \overline{y}\rangle. \tag{6b}$$

Again we drop the explicit dependence of $\overline{\psi}$ on x and $\overline{x}$ for simplifying notation. The rational behind this definition is that it allows to efficiently obtain a descent direction, as the gap function is always nonnegative for any convex function $\overline{f}$.

Lemma 1. *Assume that $\overline{f}$ is convex and $\overline{\psi}(\overline{d}) < 0$ holds. Then $d := P\overline{d}$ is a descent direction satisfying $\langle \nabla f(x), d\rangle < 0$.*

Proof. Since $\overline{f}$ is convex, the statement follows from $B_{\overline{f}}(\overline{x}+\overline{y},\overline{x}) \geq 0$ for all $\overline{y}$.

Remark 2. Whenever $R = P^\top$ holds (a standard assumption[1] in multigrid literature [10]), the coarse model $\overline{\psi}$ and the 'shifted' coarse model $\overline{y} \mapsto \overline{\phi}(\overline{x}+\overline{y})$ only differ by a constant that depends on x and $\overline{x} = Rx$. Indeed, using $\overline{v}_x$ from (2b) we rewrite

$$\begin{aligned}
\overline{\phi}(\overline{x}+\overline{y}) &= \overline{f}(\overline{x}+\overline{y}) - \langle \nabla \overline{f}(\overline{x}) - R\nabla f(x), \overline{x}+\overline{y}\rangle && \text{(7a)}\\
&= B_{\overline{f}}(\overline{x}+\overline{y},\overline{x}) + \langle R\nabla f(x), \overline{y}\rangle + const && \text{(7b)}\\
&\overset{R=P^\top}{=} \overline{\psi}(\overline{y}) + const. && \text{(7c)}
\end{aligned}$$

In the following we disregard constant terms in (7) and consider (the simplified) coarse model $\overline{\psi}$ from (6a).

Remark 3. The first-order coherence applied to $\overline{\psi}$ now reads $\nabla\overline{\psi}(0) = P^\top \nabla f(x)$.

Remark 4. The coarse model $\overline{\psi}$ incorporates both first order information of the fine objective and second order information of the coarse objective. Indeed, for $\overline{f} \in C^2$ we can write

$$\overline{\psi}(\overline{y}) = \langle \nabla f(x), P\overline{y}\rangle + B_{\overline{f}}(\overline{x}+\overline{y},\overline{x}) = \langle \nabla f(x), P\overline{y}\rangle + \frac{1}{2}\langle \overline{y}, \nabla^2 \overline{f}(\overline{z}), \overline{y}\rangle$$

for some $\overline{z} \in \{(1-t)(\overline{x}+\overline{y}) + t\overline{x}\}_{t\in[0,1]}$. We interpret the first term in $\overline{\psi}$ as the first-order Taylor expansion of $f(x+P\overline{y})$ at the current iterate x on the fine grid and ignore $f(x)$ as it is a constant with respect to $\overline{y}$. Hence, coarse model $\overline{\psi}$ resembles the quadratic approximation model in *single level* optimization

$$q_x(y) := f(x) + \langle \nabla f(x), y\rangle + \frac{1}{2}\langle y, H_x y\rangle, \tag{8}$$

where H_x is a symmetric positive definite approximation of $\nabla^2 f(x)$.

Coarse correction condition. We adopt the following criteria from [4]

$$\|P^\top \nabla f(x)\| \geq \kappa \|\nabla f(x)\| \quad \text{and} \quad \|P^\top \nabla f(x)\| > \varepsilon, \tag{9}$$

[1] Our model does not require this assumption.

where $\kappa \in (0, \min(1, \|P\|))$ and $\varepsilon \in (0, 1)$. The above criteria prevent us from using the coarse model for computing a descent direction when $\overline{x} \approx \overline{x} + \overline{d}$, i.e. the coarse correction direction $\overline{d}$ is close to 0.

Box constrained coarse model. We now extend the coarse model $\overline{\psi}$ to box constraints in order to approach (1) by MLO. We introduce

$$\min_{\overline{y}} \overline{\psi}(\overline{y}) \qquad \text{subject to} \quad \bar{l}_{x,\overline{x},P} \leq \overline{y} \leq \overline{u}_{x,\overline{x},P}, \tag{10}$$

where the bounds at the coarse level are defined as

$$(\bar{l}_{x,\overline{x},P})_j = \overline{x}_j + \frac{1}{\|P\|_\infty} \max_{i=1,\dots,n} \begin{cases} (l-x)_i, & \text{if } P_{ij} > 0, \\ (x-u)_i, & \text{if } P_{ij} < 0, \end{cases} \tag{11a}$$

$$(\overline{u}_{x,\overline{x},P})_j = \overline{x}_j + \frac{1}{\|P\|_\infty} \min_{i=1,\dots,n} \begin{cases} (u-x)_i, & \text{if } P_{ij} > 0, \\ (x-l)_i, & \text{if } P_{ij} < 0, \end{cases} \tag{11b}$$

and adopted from [11]. A closely related coarse model was considered in [5]. In our notation, we drop the dependency of these bounds on x, $\overline{x}$ and P. The above definitions also handle negative elements in P (as in e.g. cubic interpolation). The next result states that box constraints are preserved by prolongation.

Lemma 2 ([11, Lemma 4.3]). *Let $x, l, u \in \mathbb{R}^n$ with $l < u$, $P : \mathbb{R}^{\overline{n}} \to \mathbb{R}^n$ and $\bar{l}$ and $\overline{u}$ be defined as in* (11)*. Consider any $\overline{d} \in [\bar{l}, \overline{u}]$. Then $l \leq x + P\overline{d} \leq u$ holds.*

In the unconstrained case, it suffices to test whether $\|P^\top \nabla f(x)\|$ is large enough compared to $\|\nabla f(x)\|$, see (9). However, this criterion is inadequate for the box-constrained problem. Instead, we use the scaled gradient [12],

$$G(x) = S(x)\nabla f(x), \qquad S(x) = \operatorname{diag}(s_1(x), \dots, s_n(x)), \tag{12a}$$

$$s_i(x) = \begin{cases} \min\{1, x_i - l_i\}, & \text{if} \quad (\nabla f(x))_i > 0, \\ \min\{1, u_i - x_i\}, & \text{if} \quad (\nabla f(x))_i < 0, \\ \min\{1, x_i - l_i, u_i - x_i\}, & \text{if} \quad (\nabla f(x))_i = 0, \end{cases} \tag{12b}$$

and replace $\nabla f(x)$ by $G(x)$ in (9). This gives

$$\|P^\top G(x)\| \geq \kappa \|G(x)\| \quad \text{and} \quad \|P^\top G(x)\| > \varepsilon, \tag{13}$$

where $\kappa \in (0, \min(1, \|P\|))$ and $\varepsilon \in (0, 1)$. One can show that

$$s_i(x) \begin{cases} = 0, & \text{if} \quad x_i = l_i \quad \text{and} \quad (\nabla f(x))_i > 0, \\ = 0, & \text{if} \quad x_i = u_i \quad \text{and} \quad (\nabla f(x))_i < 0, \\ \geq 0, & \text{if} \quad x_i \in \{l_i, u_i\} \quad \text{and} \quad (\nabla f(x))_i = 0, \\ > 0, & \text{otherwise.} \end{cases}$$

Thus any x with $G(x) = 0$ is a stationary point of the box-constrained fine level problem.

Application to discrete tomography. We now represent problem (1) on a coarser grid and evaluate the coarse grid model $\overline{\psi}$ from (6a). We assume that images are discretized on $n = N \times N$ grid points in a two dimensional domain in $\mathbb{R}^2$. Using the one-dimensional discrete derivative operator $\partial_d \colon \mathbb{R}^d \to \mathbb{R}^d$, $(\partial_d)_{ij} = -1$, if $i = j < d$, $(\partial_d)_{ij} = +1$ if $j = i + 1 \leq d$ and $(\partial_d)_{ij} = 0$ otherwise, along each spatial direction, we define the discrete gradient matrix of an $N \times N$ discrete image by $D := \binom{D_1}{D_2} = \binom{\partial_N \otimes I_N}{I_N \otimes \partial_N}$, where $\otimes$ stands for the Kronecker product and I_N is the identity matrix of dimension N. Analogously, we define the discrete gradient on a coarse $\overline{n} = \overline{N} \times \overline{N}$ grid and denote it by $\overline{D}$.

We denote the projection matrix at the coarser level by $\overline{A}$. Next, we show that the specific ray geometry corresponding to the coarse grid can be selected independently of the ray geometry at the fine grid, as we do not need to transfer the projection information between levels. To this end, we evaluate the Bregman gap in (6b), as only this term involves the coarse objective $\overline{f}$.

Lemma 3. *Denote the data term in f from (1) by $p(y) := \mathrm{KL}(Ay, b)$ and the regularizer with $q(y) := \|Dy\|_{1,\tau} := \langle \rho_\tau(Dy), \mathbb{1} \rangle$, where ρ_τ is the Huber function applied component-wise. Assume $\overline{A}$ and $\overline{D}$ are given on the coarse grid. Then*

$$B_{\overline{f}}(\overline{y}, \overline{x}) = KL(\overline{A}\overline{y}, \overline{A}\overline{x}) + \lambda B_{\overline{q}}(\overline{y}, \overline{x}). \tag{14}$$

Proof. A simple calculation shows $B_{\overline{p}}(\overline{y}, \overline{x}) = KL(\overline{A}\overline{y}, \overline{A}\overline{x})$. Then the result follows from linearity of the Bregman gap $B_{p+\lambda q}(y, x) = B_p(y, x) + \lambda B_q(y, x)$.

Remark 5. One can show that the observation above applies to all variational models that involve a data term formulated by means of a Bregman divergence.

Final algorithm. We now particularize the steps of the general framework in Algorithm 2.1.

- Line 3: We choose the coarse correction condition as in (13).
- Line 4: We use the box constrained coarse model in (10).
- Line 5: We obtain $\overline{d}$ with a few iterations of the projected gradient method with inexact line search [13] until $\overline{\psi}(\overline{d}) < 0$ holds.
- Line 6: We employ a full weighting operator [10].
- Line 7: This line search may be omitted due to our choice of the restricted box (11), see Lemma 2.
- Line 10: As f is not gradient Lipschitz continuous, we do fine corrections via the projected gradient with inexact (rather than fixed) line search.

The algorithm can be implemented also recursively using multiple levels.

3 Geometric Approach

We focus on the minimization of f subject to box constraints in a smooth setting.

Riemannian geometry of the box. Following [14] we turn the open box into a manifold

$$(\mathcal{M}, g), \qquad \mathcal{M} := (l, u), \quad l, u \in \mathbb{R}^n, \quad l < u \tag{15}$$

with the Hessian Riemannian metric $g_x(v,w) = \langle v, H_x w\rangle,\ v,w \in T_x\mathcal{M} = \mathbb{R}^n$, induced by $h(x) = \langle \mathbb{1}, (x-l)\log(x-l) + (u-x)\log(u-x)\rangle$ (a convex Legendre function [15, Chapter 26]) and its Hessian given by $H_x := \nabla^2 h(x) = \operatorname{Diag}\left(\frac{u-l}{(x-l)(u-x)}\right)$. The Riemannian gradient is now given by

$$\nabla_{\mathcal{M}} f(x) = H_x^{-1}\nabla f(x) = \frac{(x-l)(u-x)}{u-l}\nabla f(x). \tag{16}$$

Though the choice of h may appear arbitrary at this point, it will prove beneficial in connection with the constructed retraction below.

Retraction. Conceptually, any reasonable numerical first-order update for the minimization of f has to map the Riemannian gradient $\nabla_{\mathcal{M}} f(x)$ from the tangent space $T_x\mathcal{M}$ at a current point $x \in \mathcal{M}$ onto the manifold in a meaningful way in order to produce an update $x^+ \in \mathcal{M}$. On a Riemannian manifold, the natural candidate for this purpose is given by the exponential map with respect to the Levi-Civita connection. However, in our case it can be shown that this map is only defined around a small neighborhood of $0 \in T_x\mathcal{M}$ and does not extend onto all of $T_x\mathcal{M}$. To overcome this limitation, we consider a *retraction map* $\mathscr{R}_x \colon T_x\mathcal{M} \to \mathcal{M}$, smoothly varying in $x \in \mathcal{M}$, which is required to fulfill

$$\text{(i)}\quad \mathscr{R}_x(0) = x, \quad 0 \in T_x\mathcal{M} \quad \text{and} \quad \text{(ii)}\quad d\mathscr{R}_x(0) = \mathrm{id}_{T_x\mathcal{M}}, \quad \forall x \in \mathcal{M}. \tag{17}$$

These conditions ensure that the curve $\gamma(t) := \mathscr{R}_x(tv)$ realizes the tangent vector $v \in T_x\mathcal{M}$ at $x \in \mathcal{M}$ by satisfying $\gamma(0) = x$ and $\dot\gamma(0) = v$. See [16, Section 4] for more background and details of retraction maps.

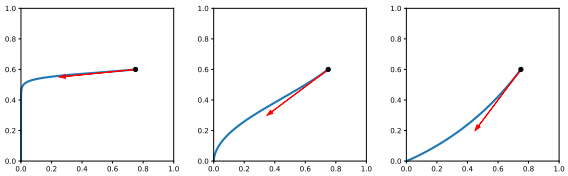

Fig. 2: Behavior of the retraction $\widetilde{\exp}_x$ from (18) on $\mathcal{M} = (0,1)^2$ in terms of the curve $\gamma(t) = \widetilde{\exp}_x(tv)$ for various choices of $v \in T_x\mathcal{M} = \mathbb{R}^2$. This illustrates that $\gamma(t)$ realizes the tangent vector v near $t = 0$ and never leaves the manifold $\mathcal{M}$.

Proposition 1. *Let $(\mathcal{M}, g)$ be given by* (15)*. Then the map*

$$\widetilde{\exp}\colon T\mathcal{M} \to \mathcal{M}, \qquad \widetilde{\exp}_x(v) := l + \frac{(u-l)(x-l)e^{\frac{u-l}{(x-l)(u-x)}v}}{u - x + (x-l)e^{\frac{u-l}{(x-l)(u-x)}v}} \tag{18}$$

is a proper retraction map.

Proof. The relative interior of the probability 2-simplex is given by $\operatorname{relint}(\Delta_2) = \{p \in \mathbb{R}^2 | p > 0 \text{ and } p_1 + p_2 = 1\} =: \mathcal{S}_2$. For any index $i \in [n]$ we can identify the interval (l_i, u_i) with $\mathcal{S}_2$ via $F_i \colon (l_i, u_i) \to \mathcal{S}_2$, by sending a point $x_i \in (l_i, u_i)$ to $F_i(x_i) := \frac{1}{u_i - l_i} \begin{pmatrix} u_i - x_i \\ x_i - l_i \end{pmatrix}$. The manifold S_2 possesses an exponential map with respect to the so called *e-connection* from information geometry [17, 18], which is defined on all of $T_p\mathcal{S}_2$ and given by $\exp_p(v) = \langle p, e^{\frac{v}{p}} \rangle^{-1} p e^{\frac{v}{p}}$ at any $p \in \mathcal{S}_2$ and $v \in T_p\mathcal{S}_2$. Since exponential maps always fulfill condition (ii) of (17), the map $\exp_p \colon T_p\mathcal{S}_2 \to \mathcal{S}_2$ and therefore also the pullback onto (l_i, u_i) under F_i

$$F_i^{-1}\big(\exp_{F_i(x_i)}(dF_{i,x}(v))\big) = \big(\widetilde{\exp}_x(v)\big)_i$$

are both proper retractions. Applying this argument to each coordinate $i \in [n]$ proves the statement. □

The retraction in (18) allows us to compute updates on the manifold based on numerical operation in the tangent space. Due to the simple structure of the constraints, this can be done separately for each coordinate. Furthermore, as a consequence of the choice for h, the corresponding Hessian H_x defined before equation (16) exactly matches the exponent in the expression for $\widetilde{\exp}_x$ in (18). Thus, applying $\widetilde{\exp}_x$ to the Riemannian gradient (16) simplifies to

$$\widetilde{\exp}_x\big(-\alpha \nabla_{\mathcal{M}} f(x)\big) = l + \frac{(u-l)(x-l)e^{-\alpha \nabla f(x)}}{u - x + (x-l)e^{-\alpha \nabla f(x)}}. \tag{19}$$

Coarse grid model, coarse grid correction. For $x \in \mathcal{M}$ and $\overline{x} = Rx$ define $\overline{\mathcal{M}} := (\overline{l}, \overline{u}) \subset \mathbb{R}^{\overline{n}}$ endowed with the Riemannian geometry from (15). Further consider $\overline{\psi}(\overline{y}) = B_{\overline{f}}(\overline{y}, \overline{x}) + g_x(\nabla_{\mathcal{M}} f(x), P(\overline{y} - \overline{x}))$. To find a $\overline{y} \in \overline{\mathcal{M}}$ such that $\overline{\psi}(\overline{y}) < 0$ we employ the Riemannian gradient method, see [16, Alg. 1].

Final algorithm. Algorithm 3.1 summarizes a multilevel implementation of the two grid general framework of Algorithm 2.1 specified to our geometric setting. Note that in Algorithm 3.1, just as $\overline{f}$ represents f on the first coarse level, $\overline{\overline{f}}$ represents f on the second coarse level.

Remark 6. Strictly speaking, tangent vectors from different vector spaces $T_x\mathcal{M}$ and $T_{x'}\mathcal{M}$ are incompatible unless parallel transport is used. This issue arises in Algorithm 3.1, line 5, when the tangent vector field $\nabla_{\mathcal{M}} f$ is evaluated at a fine grid point x, as part of the coarse grid model. Since $\mathcal{M}$ is an open subset of an ambient Euclidean space with a trivial tangent bundle, however, this problem is merely a *formal* one, and we deliberately ignore it throughout this paper.

4 Experiments

To illustrate our approach, we compare it to a state-of-the-art first-order multilevel approach [5] (capable of handling box constraints in a Euclidean setting) which we adapt as described in Section 2 and denote it as *multilevel projected gradient (ML PG).* We denote its single-level counterpart as *projected gradient*

Algorithm 3.1: Multilevel Optimization (ML RG)

Function `MLO`$(f, \overline{f}, x, P)$
  $i \leftarrow 0$
  while x *is not optimal and* $i < i_{\max}$ **do**
    if $\|RG(x)\| \geq \kappa \|G(x)\|$ *and* $\|RG(x)\| \geq \varepsilon$ *and* $\overline{f}$ *defined* **then**
      $\overline{\psi}(\overline{y}) = D_{\overline{f}}(\overline{y}, \overline{x}) + g_x(\nabla_{\mathcal{M}} f(x), P(\overline{y} - \overline{x}))$ `/* coarse model */`
      if $\overline{\overline{f}}$ *defined* **then**
        $\overline{y} \leftarrow$ `MLO`$(\overline{\psi}, \overline{\overline{f}}, \overline{x}, \overline{P})$ `/* recursive call */`
      else
        find $\overline{y}$ with $\overline{\psi}(\overline{y}) < 0$
      $d \leftarrow P\overline{d}, \quad \overline{d} = \overline{y} - \overline{x}$ `/* descent direction */`
      $x \leftarrow \widetilde{exp}_x(\alpha d)$ `/*` $\alpha > 0$ `such that` $f(\widetilde{exp}_x(\alpha d)) < f(x)$ `*/`
    $x \leftarrow$ `RiemannianGradientDescent`(f, x)
    $i \leftarrow i + 1$
  return x

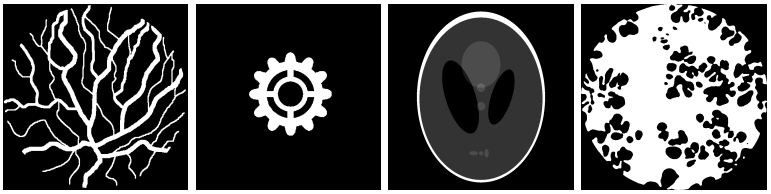

Fig. 3: Phantoms that exhibit both fine and large scale structures.

(PG). Similarly, we denote our proposed geometric multilevel approach in Section 3 with *multilevel Riemmanian gradient (ML RG)* and its single-level version as *Riemannian gradient (RG).* We summarize our results in Figure 4.

Data setup. We consider the phantoms ($n = 1024 \times 1024$) in Figure 3. We generated the projection matrices using the ASTRA-toolbox[2]. We used parallel beam projections along equidistant angles between 0 and π. The undersampling rate at the fine grid is 20%. Entry a_{ij} of projection matrix A holds the length of the line segment of the i-th projection ray passing through the j-th pixel. At every level the width of the sensor-array was set to the grid size, so that at each scale every pixel intersects with at least one projection ray. For the information transfer between levels we used as restriction the full weighting operator [10, Eq. (2.3.3.)] and set $P = R^\top$.

Implementation details. We consider 5 levels with the coarsest grid 64×64. In each coarse level, we limit the number of iterations to 10. At the finest level, we set $\lambda = 10^{-3}$ and increase it at coarser levels to $\overline{\lambda} = \lambda \cdot 2^2$. Parameter τ of

[2] https://www.astra-toolbox.com/

the Huber function is 10^{-4}. The parameters in the coarse correction condition are $\kappa = 0.49$ and $\varepsilon = 10^{-3}$.

5 Conclusion

This work is a first glimpse at ongoing research that aims at a systematic analysis and evaluation of a geometric approach to multilevel optimization. Using only first-order concepts enabled us to achieve state of the art performance. In further work, we will elaborate various ingredients of the approach like, e.g., using differential geometry for deriving optimal restriction and prolongation mappings.

References

1. Herman, G., Kuba, A.: Discrete Tomography: Foundations, Algorithms and Applications. Birkhäuser (1999)
2. Nash, S.: A multigrid approach to discretized optimization problems. Optim. Method. Softw. **14** (2000) 99–119
3. Gratton, S., Sartenaer, A., Toint, P.L.: Recursive trust-region methods for multiscale nonlinear optimization. SIAM J. Optim. **19** (2008) 414–444
4. Wen, Z., Goldfarb, D.: A line search multigrid method for large-scale nonlinear optimization. SIAM J. Optim. **20**(3) (2009) 1478–1503
5. Kočvara, M., Mohammed, S.: A first-order multigrid method for bound-constrained convex optimization. Optim. Method. Softw. **31**(3) (2016) 622–644
6. Parpas, P.: A multilevel proximal gradient algorithm for a class of composite optimization problems. SIAM J. Sci. Comput. **39**(5) (2017) 681–701
7. Roux, S., Leclerc, H., Hild, F.: Efficient binary tomographic reconstruction. J. Math. Imaging Vis. **49**(2) (2014) 335–351
8. Dabravolski, A., Batenburg, K., Sijbers, J.: A multiresolution approach to discrete tomography using DART. PLoS One **9**(9) (2014) e106090
9. Plier, J.: Theoretical and numerical approaches to co-/sparse recovery in discrete tomography. PhD thesis, Heidelberg University (2020)
10. Trottenberg, U., Oosterlee, C., Schüller, A.: Multigrid. Academic Press (2001)
11. Gratton, S., Mouffe, M., Toint, P.L., Weber-Mendonca, M.: A recursive formula-trust-region method for bound-constrained nonlinear optimization. IMA Journal of Numerical Analysis **28**(4) (2008) 827–861
12. Ulbrich, M.: Nonmonotone trust-region methods for bound-constrained semismooth equations with applications to nonlinear mixed complementarity problems. SIAM J. Optim **11** (2001) 889–916
13. Iusem, A.N.: On the convergence properties of the projected gradient method for convex optimization. Computational and Applied Mathematics **22** (2003) 37–52
14. Alvarez, F., Bolte, J., Brahic, O.: Hessian Riemannian gradient flows in convex programming. SIAM Journal on Control and Optimization **43**(2) (2004) 477–501
15. Rockafellar, R.T.: Convex Analysis. Princeton University Press (1997)
16. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (2008)
17. Amari, S.I., Nagaoka, H.: Methods of Information Geometry. Amer. Math. Soc. and Oxford Univ. Press (2000)
18. Ay, N., Jost, J., Lê, H.V., Schwachhöfer, L.: Information Geometry. Springer (2017)

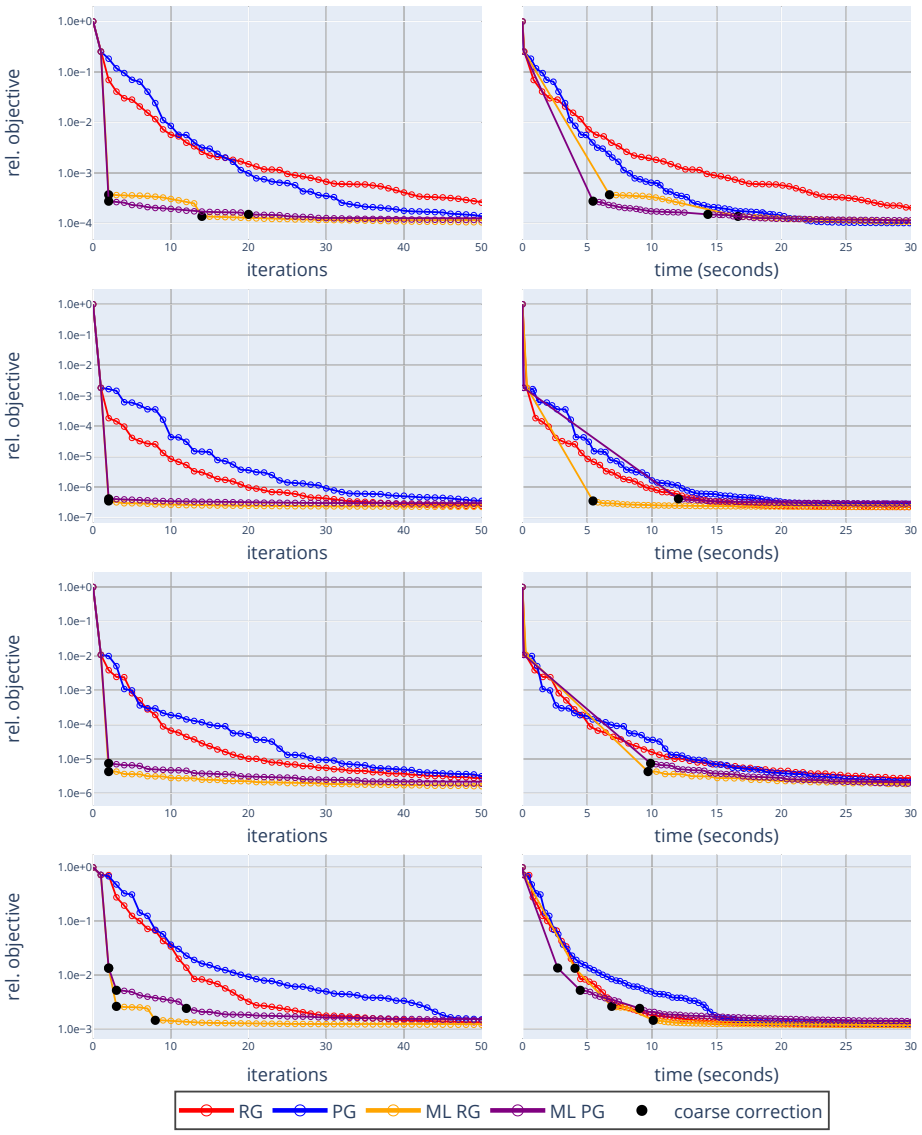

Fig. 4: **Comparison of decreasing objective function values (left column) and runtime (right column)** for single-level resp. multilevel (ML) versions of projected gradient descent (PG, ML PG) and Riemannian gradient descent (RG, ML RG). The i-th row corresponds to the i-th image in Figure 3. Black dots indicate when descent directions were computed on coarser grids. The multilevel schemes aggressively minimize the objective. The computational overhead (checking feasibility of coarse grid descent directions, grid transfer) takes some computation time. Yet, in view of the objective function decrease, multilevel iterations can be terminated earlier than the single-level schemes.