On the Remarkable Efficiency of SMART

Max Kahl[1][0009−0003−1084−5672], Stefania Petra[2][0000−0002−7189−2275], Christoph Schnörr[3][0000−0002−8999−2338], Gabriele Steidl[4][0000−0003−4638−9245], and Matthias Zisler[2][0000−0003−2936−4556]

[1] Astroinformatics Group, HITS, Heidelberg, Germany
[2] Mathematical Imaging Group, Heidelberg University, Germany
[3] Image and Pattern Analysis Group, Heidelberg University, Germany
[4] Institute of Mathematics, TU Berlin, Germany

Abstract. We consider the problem of minimizing the Kullback-Leibler divergence between two unnormalised positive measures, where the first measure lies in a finitely generated convex cone. We identify SMART (simultaneous multiplicative algebraic reconstruction technique) as a Riemannian gradient descent on the parameter manifold of the Poisson distribution. By comparing SMART to recent acceleration techniques from convex optimization that rely on Bregman geometry and first-order information, we demonstrate that it solves this problem very efficiently.

Keywords: KL divergence · accelerated mirror descent · relative smoothness · information geometry · Riemannian gradient descent.

1 Introduction

This paper explores state-of-the-art first-order optimization methods for solving

$$\min_{x\in\mathbb{R}^n_+} f(x), \qquad f(x) = \mathrm{KL}(Ax, b), \qquad A \in \mathbb{R}^{m\times n}_+, \quad b \in \mathbb{R}^m_{++} \tag{1}$$

in order to recover a discretized *nonnegative* function x from linear *nonnegative* measurements $Ax \approx b$ by minimizing the Kullback-Leibler (KL) divergence, instead e.g., the usual least-squares norm, see e.g., [11] and references therein. We exploit the underlying Bregman geometry in a twofold way. First, *convex* optimization methods based on Bregman distances offer the possibility of matching the Bregman distance to the structure of the problem, leading to simple multiplicative gradient-like iterative schemes and enabling a reduced cost of the complexity per iteration. Secondly, by turning the interior of the feasible set into a Riemannian manifold, the geometry of the space allows smooth unconstrained optimization. We examine both aspects in a principled manner for problem (1) and consider discrete tomography as application scenario.

Related work. A prototypical multiplicative iterative algorithm for (1) is *SMART (simultaneous multiplicative algebraic reconstruction technique)* [7]

$$x^{k+1} = x^k e^{-\tau_k \nabla f(x^k)}, \qquad x^0 \in \mathbb{R}^n_{++} \qquad \text{(SMART)} \tag{2a}$$

$$x_j^{k+1} = x_j^k \prod_{i=1}^m \Big(\frac{b_i}{(Ax^k)_i} \Big)^{\tau_k A_{ij}}, \qquad k = 0, 1, \ldots \quad j = 1, \ldots, n. \tag{2b}$$

In [20] SMART was identified as the classical mirror descent algorithm (MDA) [18] for *fixed* steplength τ_k for the particular objective (1), that converges at a (faster) $O(1/k)$ rate (as opposed to $O(1/\sqrt{k})$ for general MDA [5]) due to relative L-smoothness (see below). In addition, a computationally efficient acceleration scheme based on [22] was suggested, however, without a theoretical underpinning of a $O(1/k^2)$ rate. MDA, and thus SMART too, is a special instance of the Bregman proximal gradient (BPG) method [21]. Recent results concerning optimal complexity of Bregman first-order methods [15, 14, 12] including BPG, motivate us to explore the a-posteriori certification of accelerated rates for (1). In [17] the convergence of SMART was analyzed in the context of primal-dual methods. Hence, it is natural to ask how (1) can be solved by state-of-the-art (accelerated) primal dual splitting methods that employ generalized proximal operators defined in terms of a Bregman distance [9, 8]. For a recent overview of Bregman divergences and proximity operators, see [13].

Contribution and organization. Section 2 introduces essential concepts related to Bregman divergences. The acceleration of SMART is discussed from the viewpoint of BPG in section 3. The Riemannian geometry of SMART is introduced in Section 4. In Section 5, we show in large scale experiments that SMART is on par with the state-of-the-art Bregman first order methods, and the $\mathcal{O}(1/k^2)$ rate of its accelerated version cannot be numerically certified.

Basic notation. We denote the set of nonnegative real vectors by $\mathbb{R}^n_+$ and the set of positive ones by $\mathbb{R}^n_{++}$. Let $\langle \cdot, \cdot \rangle$ denote the standard inner product on $\mathbb{R}^n$, ∇h the gradient of a differentiable function $h : \mathbb{R}^n \to \mathbb{R}$ and h^* the Fenchel conjugate $h^*(p) = \sup_{x \in \mathbb{R}^n} \{\langle p, x \rangle - h(x)\}$. Given a sufficiently well-behaved convex function ϕ, we consider the so-called Bregman distance

$$D_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle \tag{3}$$

between x and y. We frequently 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 $e^x, \log x$ 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 $\mathrm{d}_x h : T_x\mathcal{M} \to \mathbb{R}$ the differential of a smooth function $h : \mathcal{M} \to \mathbb{R}$. The Riemannian gradient $\operatorname{grad} h(x) \in T_x\mathcal{M}$ of h is uniquely defined by $\mathrm{d}_x h[\xi] = g_x\left(\operatorname{grad} h(x), \xi\right), \ \forall \xi \in T_x\mathcal{M}$. The (squared) Riemannian norm is denoted by $\|v\|_x^2 = g_x\left(v, v\right), \ \forall v \in T_x\mathcal{M}$.

2 Preliminaries

Throughout this paper, we assume $A \in \mathbb{R}_+^{m\times n}$ and $b \in \mathbb{R}_{++}^m$. We consider the KL divergence as the Bregman divergence

$$\mathrm{KL}(x,y) := D_\varphi(x,y) = \langle x, \log x - \log y\rangle - \langle \mathbb{1}, x-y\rangle, \tag{4}$$

defined on $\mathbb{R}_+^n \times \mathbb{R}_{++}^n$, which plays a distinguished role among all divergence functions [2, Section 3.4] and is induced by the Bregman kernel

$$\varphi(x) = \langle x, \log x\rangle - \langle \mathbb{1}, x\rangle, \qquad x \in \mathbb{R}_+^n. \tag{5}$$

The specific function φ in (5) is of Legendre type [4, Def. 2.8] which implies that both gradients $\nabla\varphi$ and $\nabla\varphi^*$ are one-to-one and inverses of each other. In particular, φ induces a dual structure induced by the Legendre transform

$$u := \nabla\varphi(x) = \log x, \qquad x = \nabla\varphi^*(u) = e^u, \qquad \varphi^*(u) = \langle \mathbb{1}, e^u\rangle. \tag{6}$$

and a corresponding dual divergence function due to (3) reads

$$D_{\varphi^*}(v,u) = \varphi^*(v) - \varphi^*(u) - \langle \nabla\varphi^*(u), v-u\rangle = D_\varphi(x,y)\big|_{x=e^u,\, y=e^v}. \tag{7}$$

We now briefly discuss the attainment of minima in (1), that are related to the unique Bregman projection onto the cone $K_A := \{Ax \colon x \geq 0\}$, generated by the columns of A, that is a closed and convex set.

Theorem 1 ([4, Thm. 3.12]). *Suppose ϕ is closed proper convex and differentiable on* $\operatorname{int}(\operatorname{dom}\phi)$*, C is closed convex with $C \cap \operatorname{int}(\operatorname{dom}\phi) \neq \emptyset$, and $b \in \operatorname{int}(\operatorname{dom}\phi)$. If ϕ is Legendre, then the Bregman projection $\overline{y}$ of b is unique and contained in* $\operatorname{int}(\operatorname{dom}\phi)$*,*

$$\operatorname*{argmin}_{y\in C\cap \operatorname{dom}\phi} D_\phi(y,b) = \{\overline{y}\}, \qquad \overline{y} \in \operatorname{int}(\operatorname{dom}\phi). \tag{8}$$

As K_A is nonempty, closed and convex, φ in (5) is Legendre with $\operatorname{dom}\varphi = \mathbb{R}_+^m$ and $K_A \cap \mathbb{R}_{++}^m \neq \emptyset$, in view of $A \in \mathbb{R}_+^{m\times n}$. Hence, the assumptions of the theorem above are satisfied. Hence $\overline{y}$ exists and is unique and all $\overline{x} \in \mathbb{R}_+^n$ with $\overline{y} = A\overline{x}$ are minimizers of (1). One can prove a similar result as in (8) for

$$\operatorname*{argmin}_{x\in C\cap \operatorname{dom}\phi} \big\{D_\phi(x,x^0) + \langle c, x\rangle\big\} = \{z\}, \qquad z \in \operatorname{int}(\operatorname{dom}\phi), \tag{9}$$

with $c \in \mathbb{R}^n$ arbitrary and $\|c\| \leq \infty$.

Lemma 1. *Let f and φ be given by* (1) *and* (5)*, respectively. Then $D_f(x,y) = D_\varphi(Ax, Ay)$.*

Proof. From (5) and (3) it follows

$$\begin{aligned} D_f(x,y) &= f(x) - f(y) - \langle \nabla f(y), x-y\rangle \\ &= D_\varphi(Ax,b) - D_\varphi(Ay,b) - \big\langle A^\top\big(\log(Ay) - \log b\big), x-y\big\rangle \\ &= \varphi(Ax) - \varphi(b) - \langle \nabla\varphi(b), Ax-b\rangle - \big(\varphi(Ay) - \varphi(b) - \langle \nabla\varphi(b), Ay-b\rangle\big) \\ &\qquad - \langle \log(Ay) - \log b, A(x-y)\rangle \\ &= \varphi(Ax) - \varphi(Ay) - \langle \nabla\varphi(b), A(x-y)\rangle - \langle \log(Ay) - \log b, A(x-y)\rangle \\ &\overset{\nabla\varphi(b)=\log b}{=} \varphi(Ax) - \varphi(Ay) - \langle \log(Ay), A(x-y)\rangle \\ &= D_\varphi(Ax,Ay). \qquad \square \end{aligned} \tag{10}$$

3 SMART and Convex Acceleration

As mentioned, the SMART iteration (2) was studied in [20] in the context of mirror descent (aka Bregman proximal gradient), as investigated by Beck and Teboulle [5]. Specifically, using the Bregman divergence (3) as distance function, the update scheme with stepsize $\tau_k > 0$ reads

$$x^{k+1} = \operatorname*{argmin}_{x\in\mathbb{R}^n_+} f(x^k) + \langle \nabla f(x^k), x\rangle + \frac{1}{\tau_k} D_\phi(x, x^k), \qquad x^0 \in \mathbb{R}^n_{++}, \tag{11}$$

which is well defined according to (9). For D_φ given by (5), one has

$$\nabla D_\varphi(x, x^k) = \nabla\varphi(x) - \nabla\varphi(x^k) \overset{(6)}{=} \log x - \log x^k, \tag{12}$$

so that evaluating the optimality condition with respect to (11) yields

$$0 = \nabla f(x^k) + \frac{1}{\tau_k}\big(\log x^{k+1} - \log x^k\big) \tag{13a}$$

$$\Leftrightarrow \qquad x^{k+1} = x^k e^{-\tau_k A^\top \log \frac{Ax^k}{b}}, \qquad x^0 \in \mathbb{R}^n_{++}, \tag{13b}$$

which is the SMART update (2). Below we summarize the main convergence results based on [7, Thm. 2] and [20, Thm. 2].

Theorem 2. *Let S be the solution set of (1) and $L = \|A\|_1$. For $(x^k)_{k\in\mathbb{N}}$ generated by (2) with starting point $x^0 \in \mathbb{R}^n_{++}$ and $\tau_k = \tau \le 1/L$ we have*

(a) The sequence $(x^k)_{k\in\mathbb{N}}$ converges to a unique point in S, that is

$$\overline{x} = \arg\min_{x\in S} D_\varphi(x, x^0). \tag{14}$$

(b) For every k

$$f(x^k) - f(\overline{x}) \le \frac{L D_\varphi(\overline{x}, x^0)}{k}\ . \tag{15}$$

The following lemma, adapted from [20, Prop. 2], allows to establish the basic convergence rate $\mathcal{O}(1/k)$ of SMART without assuming that ∇f is L-Lipschitz.

Lemma 2. *Suppose $A \in \mathbb{R}_+^{m\times n}$ and consider φ in (5). Then*

$$\forall x, y \in \mathbb{R}_+^n, \qquad D_\varphi(Ax, Ay) \le \|A\|_1 D_\varphi(x, y). \tag{16}$$

Acceleration in Bregman First-Order Convex Optimization. In [12] the $\mathcal{O}(1/k)$ rate is shown to be optimal for a broad class of Bregman proximal gradient (BPG) algorithms under *general* assumptions on the objective function f and the Bregman kernel ϕ. In particular, it is not required that ∇f is L-Lipschitz. Rather, f has merely to be L-smooth *relative* to ϕ, i.e.

$$D_f(x, y) \le L D_\phi(x, y), \qquad \forall x, y \in \operatorname{dom} f \subset \operatorname{dom} \phi. \tag{17}$$

Accelerated Bregman proximal gradient (ABPG) algorithms can *only* be obtained under additional assumptions. The authors in [15] consider an assumption which yields a $\mathcal{O}(1/k^\gamma)$ rate with $\gamma \in [1, 2]$. In particular, they consider the *triangle-scaling property* with *uniform triangle-scaling exponent (TSE)* γ

$$D_\phi\big((1-\theta)x+\theta z, (1-\theta)x+\theta\widetilde{z}\big) \le \theta^\gamma D_\phi(z, \widetilde{z}), \quad \forall\theta \in [0, 1], \quad \forall x, z, \widetilde{z} \in \operatorname{rint}\operatorname{dom}\phi. \tag{18}$$

The focus is on *jointly* convex Bregman divergences D_ϕ since then (18) holds with $\gamma = 1$. Note that KL is jointly convex. Further, the *intrinsic TSE* of D_ϕ is defined by

$$\gamma_{\text{in}} = \limsup_{\theta\searrow 0} \frac{D_\phi\big((1-\theta)x + \theta z, (1-\theta)x + \theta\widetilde{z}\big)}{\theta^\gamma} < \infty, \qquad \forall x, z, \widetilde{z} \in \operatorname{rint}\operatorname{dom}\phi. \tag{19}$$

A broad class of Bregman divergences has $\gamma_{\text{in}} = 2$ which is the value the largest uniform TSE cannot exceed. The analysis in [15] rests upon the *triangle-scaling gain* $G(x, z, \widetilde{z})$ defined by the relaxed triangle-scaling inequality

$$D_\phi\big((1-\theta)x + \theta z, (1-\theta)x + \theta\widetilde{z}\big) \le G(x, z, \widetilde{z})\theta^\gamma D_\phi(z, \widetilde{z}), \qquad \forall\theta \in [0, 1]. \tag{20}$$

$G(x, z, \widetilde{z})$ is bounded based on the relative scaling of the Hessian of ϕ at different points. In particular, *adaptive* ABPG algorithms are proposed based on (20) for problems of the form

$$\min_{x\in C} F(x), \qquad F(x) = f(x) + \Psi(x), \tag{21}$$

with f being L-smooth relative to ϕ, C closed, and C, Ψ convex and simple, in the sense that the key step of the ABPG method

$$z_{k+1} = \operatorname{argmin}_{x\in C}\Big\{f(y_k) + \langle\nabla f(y_k, x - y_k\rangle + \theta_k^{\gamma-1} L D_\phi(x, z_k) + \Psi(x)\Big\}, \tag{22}$$

can be solved efficiently. The convergence analysis of ABPG uses basic relations derived by [10] and [22] in order to relate two subsequent updates.

The ABPG-e method with *exponent adaption* starts with a large value $\gamma_k > 2$ and reduces it at each step by some fixed δ, until an inequality (the *local* triangle-scaling property, see below (24) for its specialization to our scenario) as stopping criterion is satisfied. The last value γ_k determines the convergence rate and serves as *empirical certificate*.

The ABPG-g method with *gain adaption* adapts the gain $G_k = G(x_k, z_k, \widetilde{z}_k)$ in an inner loop until a local triangle-scaling inequality is satisfied. In Section 4.1, the authors discuss obstacles for proving a $\mathcal{O}(k^{-2})$ rate, which requires to bound the geometric mean $\overline{G} := (G_0^\gamma G_1 \cdots G_k)^{\frac{1}{k+\gamma}}$ of gains at each step, without additional assumptions. They argue that, in practical situations, one always works with a particular reference function ϕ that may have structural properties yielding fast convergence. Exploiting such a structure for φ (5) is subject to further research.

SMART and Acceleration. Combining Lem. 1 and Lem. 2, we conclude that f in (1) is L-smooth relative to φ (5) with $L = \|A\|_1$. The ABPG iteration (22) leads for $\gamma = 1$ and $\Psi \equiv 0$ to the F(ast)-SMART iteration [20],

$$y^k = (1 - \theta_k)x^k + \theta_k z^k \tag{23a}$$

$$z^{k+1} = z^k \exp\left(-A^\top \log\left(\frac{Ay^k}{b}\right)/L\right) \tag{23b}$$

$$x^{k+1} = (1 - \theta_k)x^k + \theta_k z^{k+1}, \tag{23c}$$

where $x^0 = z^0 \in \operatorname{int}(\operatorname{dom}\varphi)$ and $\theta_k \in (0, 1]$ satisfies $\frac{1-\theta_{k+1}}{\theta_{k+1}^2} \leq \frac{1}{\theta_k^2}$. As the uniform TSE γ equals 1 for our choice φ in (5) only a $\mathcal{O}(1/k)$ rate can be guaranteed according to [15, Thm. 1]. Convergence of the sequence $(x^k)_{k\in\mathbb{N}}$ generated by FSMART, as it is guaranteed for SMART, remains an open issue. As discussed above ABPG-e uses a local triangle-scaling property that, in view of Lem. 1, takes the form

$$D_\varphi(Ax^{k+1}, Ay^{k+1}) < \theta_k^{\gamma_k} L D_\varphi(z^{k+1}, z^k), \tag{24}$$

when specialized to our scenario. Similarly, ABPG-g includes G_k in the r.h.s. above. Hence, satisfying such a condition brings extra cost for each iteration, similar to a line search.

4 SMART: A Geometric Perspective

Riemannian Geometry of the Positive Orthant. In this section, we represent the positive orthant $\mathbb{R}^n_{++}$ as a Riemannian manifold. To this end, we turn the open interval

$$\mathcal{P} := (0, +\infty), \qquad \mathcal{P}_n := \mathcal{P} \times \cdots \times \mathcal{P} = \mathbb{R}^n_{++} \tag{25}$$

into a manifold $(\mathcal{P}, g)$ with metric g and define $(\mathcal{P}_n, g)$ as the corresponding product manifold. In order to specify $(\mathcal{P}, g)$, we apply basic information geometry

[3]. Let points $\lambda \in \mathcal{P}$ parametrize the Poisson distribution $p(z;\lambda) = \frac{\lambda^z e^{-\lambda}}{z!}$ with rate parameter $\lambda > 0$ of a random variable $Z \in \mathbf{N}_0$. Then the metric tensor of the Fisher-Rao metric is a scalar function of λ given by

$$G(\lambda) = 4 \sum_{z \in \mathbf{N}_0} \Big(\frac{d}{d\lambda} \sqrt{p(z;\lambda)} \Big)^2 = \frac{e^{-\lambda}}{\lambda^2} \sum_{z \in \mathbf{N}_0} \frac{\lambda^z (z-\lambda)^2}{z!} = \frac{e^{-\lambda}}{\lambda^2} e^{\lambda} \lambda = \frac{1}{\lambda}. \quad (26)$$

In view of (25), this extends to the metric g on $T\mathcal{P}_n$ in terms of the diagonal matrix

$$G_n(x) = \operatorname{Diag}\Big(\frac{1}{x_1}, \ldots, \frac{1}{x_n} \Big) = \operatorname{Diag}\Big(\frac{\mathbb{1}}{x} \Big) \quad (27)$$

as metric tensor. We naturally identify $T_x\mathcal{P}_n \cong \mathbb{R}^n,\ \forall x \in \mathcal{P}_n$ and denote this metric interchangeably by

$$g_x(v, v') = \langle v, v' \rangle_x = \langle v, G_n(x) v' \rangle, \qquad \forall v, v' \in T_x\mathcal{P}_n. \quad (28)$$

We point out that this geometry differs from the standard geometry of the positive orthant which underlies interior point methods [19].

Retraction. Retractions [1, Def. 4.1.1] are basic ingredients of first-order optimization algorithms on Riemannian manifolds. The main motivation is to replace the exponential map with respect to the metric (Levi Civita) connection by an approximation that can be efficiently evaluated or even computed in closed form. Below, we compute the exponential map with respect to the e-connection of information geometry [3] and show subsequently that it is a retraction.

Proposition 1. *The exponential maps on $\mathcal{P}$ resp. $\mathcal{P}_n$ with respect to the e-connection are given by*

$$\exp\colon \mathcal{P} \times T\mathcal{P} \to \mathcal{P}, \qquad \exp_\lambda(tv) = \lambda e^{t\frac{v}{\lambda}}, \quad t > 0, \quad (29a)$$

$$\exp\colon \mathcal{P}_n \times T\mathcal{P}_n \to \mathcal{P}_n, \qquad \exp_x(tv) = \big(\exp_{x_j}(tv_j) \big)_{j \in [n]}. \quad (29b)$$

Proof. By definition of the product manifold, it suffices to show (29a). A key concept of information geometry is to replace the metric connection by a pair of connections that are dual to each other with respect to the Riemannian metric g [3, Section 3.1]. In particular, under suitable assumptions, the parameter space of a probability distribution becomes a Riemannian manifold that is dually flat, i.e. two distinguished coordinate systems (the so-called m- and e-coordinates) exist with affine geodesics. We consider the case $(\mathcal{P}, g)$.

First, we rewrite the density of the Poisson distribution as distribution of the exponential family [6]

$$p(z;\theta) = h(z) \exp\big(z\theta - \psi(\theta) \big), \qquad \theta = \theta(\lambda) = \log \lambda \quad (30)$$

with exponential parameter θ, base measure $h(z) = \frac{1}{z!}$ and log-partition function $\psi(\theta) = e^\theta$ that is convex and of Legendre type. The aforementioned two coordinates are λ and θ with affine geodesics

$$t \mapsto \lambda_v(t) = \lambda + tv \in \mathcal{P}, \qquad t \mapsto \theta_u(t) = \theta + tu \in \mathbb{R}. \quad (31)$$

Note that unlike λ, v, the coordinate θ and the tangent u are unconstrained. Using (30), the e-geodesic reads

$$\lambda\big(\theta_u(t)\big) = e^{\theta_u(t)} = e^{\theta} e^{tu} \in \mathcal{P}. \tag{32}$$

We wish to express this curve in terms of λ and v using $\theta = \theta(\lambda)$ by (30) and the relation between the tangents u and v given by the differential

$$u = u(\lambda, v) = \frac{d}{dt}\theta\big(\lambda_v(t)\big)\big|_{t=0} = \frac{d}{dt}\theta(\lambda + tv)\big|_{t=0} = \frac{d}{dt}\log(\lambda + tv)\Big|_{t=0} = \frac{v}{\lambda}. \tag{33}$$

Substituting into (32) yields (29a)

$$\exp_\lambda(tv) := e^{\theta(\lambda)} e^{tu(\lambda, v)} = \lambda e^{t\frac{v}{\lambda}}. \qquad \square \tag{34}$$

Remark 1 (g is a Hessian metric). The dual nature of the exponential parametrization (30) is also highlighted by recovering the metric tensor (26) as Hessian metric from the potential $\varphi(\lambda)$ that is conjugate to the log-partition function $\psi(\theta) = e^{\theta}$, $\varphi(\lambda) = \psi^*(\lambda) = \lambda \log \lambda - \lambda$, to obtain $\varphi''(\lambda) = G(\lambda) = \frac{1}{\lambda}$. The dual coordinate chart and potential yield the inverse metric tensor $\psi''(\theta) = e^{\theta} = G(\lambda)^{-1}|_{\lambda=\lambda(\theta)}$.

Retractions provide a proper class of surrogate mappings for replacing the canonical exponential map corresponding to the metric connection.

Proposition 2 (exp **is a retraction).** *The mapping* $\exp\colon T\mathcal{P} \to \mathcal{P}$ *is a retraction in the sense of [1, Def. 4.1.1.].*

Proof. We check the two criteria that characterize retractions. First, by (29a) we have $\exp_\lambda(0) = \lambda$ for all $\lambda \in \mathcal{P}$. Second, the so-called local rigidity condition $d\exp_\lambda(0) = 1 = \mathrm{id}_{T_\lambda \mathcal{P}}$, $\forall \lambda \in \mathcal{P}$, holds as well, in view of the relation $d\exp_\lambda(u)v = e^{\frac{u}{\lambda}} v$ obtained from (29a). $\square$

SMART as Riemannian Gradient Descent. The *Riemannian* gradient with respect to the metric g from (28), generally defined by [16, p. 89] here specifically reads

$$\operatorname{grad} f(x) := {G_n}^{-1}(x) \nabla f(x) \overset{(27)}{=} x \nabla f(x), \qquad x \in \mathcal{P}_n. \tag{35}$$

The retraction in Prop. 1 allows us to compute updates on the manifold based on numerical operations in the tangent space. Due to the simple structure of the constraints, this can be done in parallel for each coordinate. Furthermore, as a consequence of the choice (28) for g, the corresponding Riemannian gradient (35) exactly matches the exponent in the expression for (29b). Thus, applying $\exp_x$ to the Riemannian gradient simplifies to

$$\exp_x\big(-\tau \operatorname{grad} f(x)\big) = x e^{-\tau \frac{\operatorname{grad} f(x)}{x}} = x e^{-\tau \nabla f(x)}. \tag{36}$$

This results in the following representation of the SMART iteration.

Proposition 3. *Let* $(\mathcal{P}_n, g)$ *be endowed with the Riemannian metric* (28)*. Then the SMART iteration* (2) *equals*

$$x^{k+1} = \exp_{x^k}\big(-\tau_k \operatorname{grad} f(x^k)\big). \tag{37}$$

The choice of τ_k can now be adapted to the current iterate by line search and we still obtain a global convergence result.

Theorem 3. *[1, Thm. 4.3.1] Let $(x^k)_{k\in\mathbb{N}}$ be a sequence generated by the iteration (37) with step-size $\tau_k = \beta^m \alpha$ and scalars $\alpha > 0$, $\beta, \sigma \in (0,1)$, where m is the smallest nonnegative integer such that*

$$f(x^k) - f\big(\exp_{x^k}\big(-\tau_k \operatorname{grad} f(x^k)\big)\big) \geq \sigma\tau_k \|\operatorname{grad} f(x^k)\|_{x^k}^2. \tag{38}$$

Then every cluster point $\hat{x}$ is a critical point of f, i.e. $\operatorname{grad} f(\hat{x}) = 0$.

The above statement assumes existence of a critical point. In view of (35) it is characterized by $\hat{x} > 0$ and $\nabla f(\hat{x}) = 0$. Clearly, such a critical point satisfies the optimality conditions $0 \leq \hat{x} \perp \nabla f(\hat{x}) \geq 0$ of (1). Hence, convergence of a subsequence of the iterates $(x_k)_{k\in\mathbb{N}}$ to a solution on the boundary, i.e., when $\hat{x}_i = 0$, is not covered by Thm. 3. In this paper, we are interested in assessing numerically the convergence speed of the iterates $(x_k)_{k\in\mathbb{N}}$ in the manifold $\mathcal{P}_n$ (25). An analysis of the behavior of these sequences close to the boundary of $\mathcal{P}_n$ will be reported in follow-up work.

5 Experiments

We compare SMART to the state-of-the-art accelerated Bregman proximal gradient methods in [15], which we adapt as described in Section 3, and to its geometric version employing Armijo line search and the retraction in (29). The latter version of SMART is denoted as Riemannian gradient (RG). In addition, we include a state-of-the-art primal dual Bregman method [9] in our comparison. For results and discussions we refer to Fig. 2, 3 and 4.

Problem and data setup. We consider large scale tomographic reconstruction as problem class, where we reconstruct the three phantoms shown in Fig. 1. We generated tomographic projection matrices A using the ASTRA-toolbox[5], with parallel beam geometry and equidistant angles in the range $[0, \pi]$. Each entry in A is nonnegative as it corresponds to the length of the intersection of a ray with a pixel. The undersampling rate was chosen to be 20%. None of the images in Fig. 1 is the unique nonnegative solution to $Ax = b$. Hence, different algorithms might converge to different nonnegative solutions. For the noisy setting we applied Poisson noise to b with a signal-to-noise ratio of SNR = 20 db.

Implementation details. We implemented six different algorithms for solving (1) iteratively. In order to avoid numerical issues, we clip each component x_i to $\max\{x_i, \varepsilon\}$ with $\varepsilon = 10^{-10}$ before applying the logarithm. The maximum number of iterations was set to $n = 1000$, which also serves as a termination criterion. We always used the initialization $x^0 = \mathbb{1}$. For each algorithm the same set of parameters was used across all experiment instances. The algorithms and corresponding parameter choices are listed below:

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

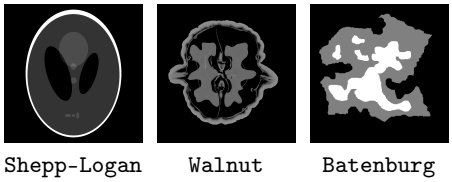

Fig. 1. The phantoms (1024×1024) used for the numerical evaluation.

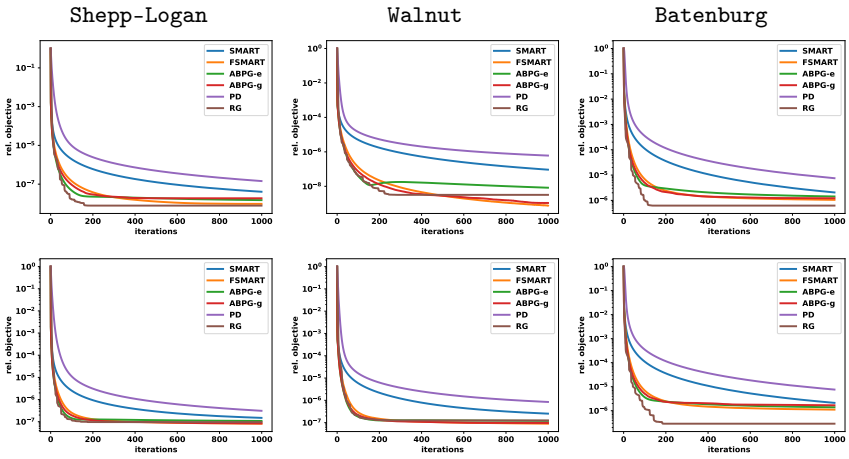

Fig. 2. Comparison of decreasing objective function values per iteration between SMART, FSMART, ABPG-e, ABPG-g, PD (Chambolle-Pock) and RG (Riemannian gradient descent). The i-th column shows the i-th image in Fig. 1, in noiseless (top row) and noisy scenarios (bottom row). Overall, RG (i.e. SMART with line search) aggressively minimizes the objective and in general outperforms the accelerated variants.

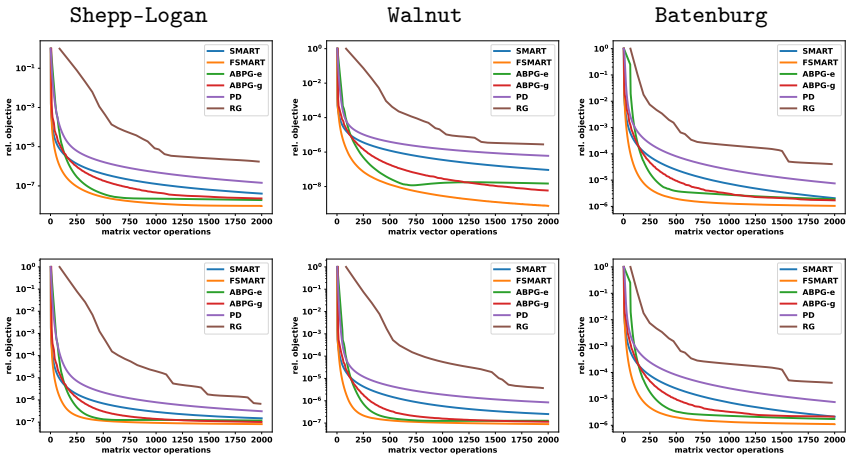

Fig. 3. Comparison of decreasing objective function values as function of costly operations between SMART, FSMART, ABPG-e, ABPG-g, PD (Chambolle-Pock) and RG (Riemannian gradient descent). The i-th column shows the i-th image in Fig. 1, in noiseless (top row) and noisy scenarios (bottom row). Checking the local triangle-scaling property incurs computational overhead for ABPG-e and ABPG-g. RG, the fastest method in terms of iterations, is the most expensive in terms of matrix vector operations due to the line search.

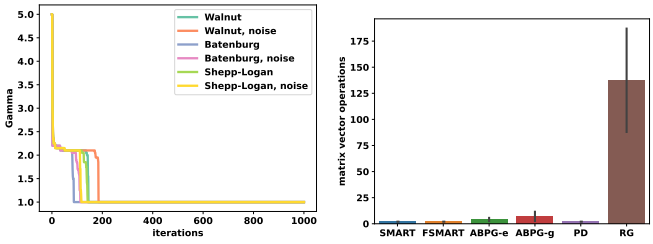

Fig. 4. A-posteriori certificates are obtained from ABPG-e by observing γ_k-values over iterations. These values are shown **left** for all problem instances. We observe that γ_k drops to 1 in all instances. We explored the drop-down-point for each instance and observed that it occurs when ABPG-e approaches the solution. Similar conclusions can be drawn from inspecting G_k in ABPG-g, which we omit here. **Average of matrix vector operations** are shown **right** for each algorithm over all iterates and tomography instances. By definition SMART, FSMART and PD always employ just two matrix vector operations per iteration. ABPG-e and ABPG-g require more such operations as they employ the local triangle scaling property, see (24), to guarantee sufficient decrease of the objective. Similarly, RG employs line search (38).

SMART solely performs the multiplicative update specified in (2) with its step-size fixed to $\tau_k = \frac{1}{L}$, where again f is L-smooth relative to φ.

FSMART is based on the iteration suggested in [20], where initially $\theta_0 = 1$ is chosen, which is then subsequently updated via $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$, as suggested in [22].

ABPG-e as described in [15, Algorithm 2] was applied to (1) with parameters $\gamma_{\min} = 1$, $\gamma_0 = 5$ and $\delta = 0.05$. The choices for γ_0 and δ deviate slightly from the recommendation in [15], but were chosen to facilitate fastest possible convergence on the selected problem instances. To ensure comparability restarting mechanisms and stopping criteria based on the divergence of iterates were foregone. Updates for θ were conducted via Newton's method.

ABPG-g specified in [15, Algorithm 3] to (1) is used with parameters: $\rho = 1.2$, $\gamma = 2$ and $G_{\min} = 10^{-3}$. Restarting, stopping criteria and updating θ was handled analogously to ABPG-e.

RG is a SMART iteration with Armijo line search for choosing the step size τ_k via the retraction in (29) to iterate according to (37). The line search parameters are $\sigma = 0.5$, $\beta = 0.8$, $\alpha = 5.0$.

PD is the Chambolle-Pock primal dual algorithm [9, Algorithm 1] for solving convex composite structured optimization problems of the form $f(x) = g(x) + h(Ax)$. For $h(y) := \mathrm{KL}(y, b)$ with $y = Ax$ and $g \equiv 0$ we obtain

$$x^{k+1} = x^k e^{-\tau A^\top y^k} \qquad \text{(primal-step)} \tag{39}$$

$$y^{k+1} = \log\left(\frac{e^{y^k} + \sigma A(2x^{k+1} - x^k)}{\mathbb{1} + \sigma b}\right), \qquad \text{(dual-step)} \tag{40}$$

whereby we compute the primal step using the generalized proximal w.r.t. the KL divergence, defined in (4), and the dual step w.r.t. its dual divergence (7). The selected step size parameters were $\tau = \frac{1}{2L}$ and $\sigma = \frac{2}{L}$.

6 Conclusion

We explored recent acceleration techniques derived in the context of Bregman proximal methods (BPG) for SMART as well as the numerical a-posteriori certification of acceleration for a large scale problem. Even though the $\mathcal{O}(1/k^2)$ rate could not be certified in this way, the heuristically accelerated version FSMART turned out to be remarkably efficient. In addition, we characterized SMART as a Riemannian gradient descent scheme on the parameter manifold induced by the Fisher-Rao geometry which opens up possibilities for connecting the local triangle scaling property — employed by accelerated BPG for certifying convergence rates — with line search methods based on suitable retractions.

Acknowledgement. MK and MZ gratefully acknowledge the generous and invaluable support of the Klaus Tschira Foundation.

References

1. Absil, P.A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press (2008)
2. Amari, S.I., Cichocki, A.: Information Geometry of Divergence Functions. Bull. Polish Acad. Sci **58**(1), 183–195 (2010)
3. Amari, S.I., Nagaoka, H.: Methods of Information Geometry. Amer. Math. Soc. and Oxford Univ. Press (2000)
4. Bauschke, H.G., Borwein, J.M.: Legendre Functions and the Method of Random Bregman Projections. J. Convex Anal. **4**, 27–67 (1997)
5. Beck, A., Teboulle, M.: Mirror Descent and Nonlinear Projected Subgradient Methods for Convex Optimization. Oper. Res. Lett. **31**(3), 167–175 (2003)
6. Brown, L.D.: Fundamentals of Statistical Exponential Families. Institute of Mathematical Statistics, Hayward, CA (1986)
7. Byrne, C.L.: Iterative Image Reconstruction Algorithms based on Cross-Entropy Minimization. IEEE Trans. Image Process. **2**(1), 96–103 (1993)
8. Chambolle, A., Contreras, J.: Accelerated Bregman Primal-Dual Methods Applied to Optimal Transport and Wasserstein Barycenter Problems. SIAM J. Math. Data Sci. **4**(4), 1369–1395 (2022)
9. Chambolle, A., Pock, T.: On the Ergodic Convergence Rates of a First-Order Primal–Dual Algorithm. Math. Program. **159**(1), 253–287 (2016)
10. Chen, G., Teboulle, M.: Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM J. Optim. **3**(3), 538–543 (1993)
11. Csiszár, I.: Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems. Ann. Stat. **19**(4), 2032–2066 (1991)
12. Dragomir, R.A., Taylor, A.B., d'Aspremont, A., Bolte, J.: Optimal Complexity and Certification of Bregman First-Order Methods. Math. Program. **194**, 41–83 (2022)
13. El Gheche, M., Chierchia, G., Pesquet, J.C.: Proximity Operators of Discrete Information Divergences. IEEE Trans. Inf. Theory **64**(2), 1092–1104 (2017)
14. Gutman, D.H., Peña, J.F.: Perturbed Fenchel Duality and First-Order Methods. Math. Program. **198**(1), 443–469 (2023)
15. Hanzely, F., Richtárik, P., Xiao, L.: Accelerated Bregman Proximal Gradient Methods for Relatively Smooth Convex Optimization. Comput. Optim. Appl. **79**, 405–440 (2021)
16. Jost, J.: Riemannian Geometry and Geometric Analysis. Springer, 4th edn. (2005)
17. Lent, A., Censor, Y.: The Primal-Dual Algorithm as a Constraint-Set-Manipulation Device. Math. Program. **50**(1–3), 343–357 (1991)
18. Nemirovski, A., Yudin, D.: Problem Complexity and Method Efficiency in Optimization. Wiley (1983)
19. Nesterov, Y.E., Todd, M.J.: On the Riemannian Geometry Defined by Self-Concordant Barriers and Interior-Point Methods. Found. Comput. Math. **2**(4), 333–361 (2002)
20. Petra, S., Schnörr, C., Becker, F., Lenzen, F.: B-SMART: Bregman-Based First-Order Algorithms for Non-Negative Compressed Sensing Problems. In: Proc. SSVM, LNCS. vol. 7893, pp. 110–124. Springer (2013)
21. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. **170**(1), 67–96 (2018)
22. Tseng, P.: On Accelerated Proximal Gradient Methods for Convex-Concave Optimization (2008), unpublished manuscript