Global Convergence of Model Function Based Bregman Proximal Minimization Algorithms

Mahesh Chandra Mukkamala, Jalal M. Fadili, Peter Ochs

▶ **To cite this version:**

Mahesh Chandra Mukkamala, Jalal M. Fadili, Peter Ochs. Global Convergence of Model Function Based Bregman Proximal Minimization Algorithms. Journal of Global Optimization, In press, 83 (4), pp.753-781. 10.1007/s10898-021-01114-y . hal-03452326

HAL Id: hal-03452326
https://hal.science/hal-03452326

Submitted on 26 Nov 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Journal of Global Optimization manuscript No.
(will be inserted by the editor)

Global Convergence of Model Function Based Bregman Proximal Minimization Algorithms

Mahesh Chandra Mukkamala · Jalal Fadili · Peter Ochs

Received: date / Accepted: date

Abstract Lipschitz continuity of the gradient mapping of a continuously differentiable function plays a crucial role in designing various optimization algorithms. However, many functions arising in practical applications such as low rank matrix factorization or deep neural network problems do not have a Lipschitz continuous gradient. This led to the development of a generalized notion known as the L-smad property, which is based on generalized proximity measures called Bregman distances. However, the L-smad property cannot handle nonsmooth functions, for example, simple nonsmooth functions like $|x^4 - 1|$ and also many practical composite problems are out of scope. We fix this issue by proposing the MAP property, which generalizes the L-smad property and is also valid for a large class of structured nonconvex nonsmooth composite problems. Based on the proposed MAP property, we propose a globally convergent algorithm called Model BPG, that unifies several existing algorithms. The convergence analysis is based on a new Lyapunov function. We also numerically illustrate the superior performance of Model BPG on standard phase retrieval problems and Poisson linear inverse problems, when compared to a state of the art optimization method that is valid for generic nonconvex nonsmooth optimization problems.

Keywords composite minimization · Bregman proximal minimization algorithms · model function framework · Bregman distance · global convergence · Kurdyka–Łojasiewicz property

Mahesh Chandra Mukkamala and Peter Ochs thank German Research Foundation for providing financial support through DFG Grant OC 150/1-1.

Mahesh Chandra Mukkamala
University of Tübingen, Tübingen, Germany,
E-mail: mamu@math.uni-tuebingen.de

Jalal Fadili
Normandie Univ, ENSICAEN, CNRS, GREYC, Caen Cedex, France,
E-mail: jalal.fadili@greyc.ensicaen.fr

Peter Ochs
University of Tübingen, Tübingen, Germany,
E-mail: ochs@math.uni-tuebingen.de

1 Introduction

We solve possibly nonsmooth and nonconvex optimization problems of the form

$$(\mathcal{P}) \qquad \inf_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})\,. \tag{1}$$

where $f : \mathbb{R}^N \to \overline{\mathbb{R}}$ is a proper lower semicontinuous function that is bounded from below. Special instances of the above mentioned problem include two broad classes of problems, namely, additive composite problems (Section 4.1) and composite problems (Section 4.2). Such problems arise in numerous practical applications such as, quadratic inverse problems Bolte et al. (2018), low-rank matrix factorization problems Mukkamala and Ochs (2019), Poisson linear inverse problems Bauschke et al. (2016), robust denoising problems Mukkamala et al. (2020), deep linear neural networks Mukkamala et al. (2019), and many more.

In this paper, we design an abstract framework for provable globally convergent algorithms based on a quality measure for suitable approximation of the objective. A classical special case is that of a continuously differentiable $f : \mathbb{R}^N \to \mathbb{R}$, whose gradient mapping is Lipschitz continuous over $\mathbb{R}^N$. Such functions enjoy the well-known Descent Lemma (cf. Lemma 1.2.3 of Nesterov (2004))

$$-\frac{\underline{L}}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2 \leq f(\mathbf{x})-f(\bar{\mathbf{x}})-\langle\nabla f(\bar{\mathbf{x}}),\mathbf{x}-\bar{\mathbf{x}}\rangle \leq \frac{\bar{L}}{2}\|\mathbf{x}-\bar{\mathbf{x}}\|^2\,,\ \text{for all } \mathbf{x},\bar{\mathbf{x}}\in\mathbb{R}^N\,, \tag{2}$$

which describes the approximation quality of the objective f by its linearization $f(\bar{\mathbf{x}}) + \langle\nabla f(\bar{\mathbf{x}}),\mathbf{x}-\bar{\mathbf{x}}\rangle$ in terms of a quadratic error estimate with certain $\underline{L}, \bar{L} > 0$. Such inequalities play a crucial role in designing algorithms that are used to minimize f. Gradient Descent is one such algorithm. We illustrate Gradient Descent in terms of sequential minimization of suitable approximations to the objective, based on the first order Taylor expansion – the linearization of f around the current iterate $\mathbf{x}_k \in \mathbb{R}^N$. Consider the following *model function* at the iterate $\mathbf{x}_k \in \mathbb{R}^N$:

$$f(\mathbf{x};\mathbf{x}_k) := f(\mathbf{x}_k) + \langle\nabla f(\mathbf{x}_k),\mathbf{x}-\mathbf{x}_k\rangle\,, \tag{3}$$

where $\langle\cdot,\cdot\rangle$ denotes the standard inner product in the Euclidean vector space $\mathbb{R}^N$ of dimension N and $f(\cdot;\mathbf{x}_k)$ is the linearization of f around $\mathbf{x}_k$. Set $\tau > 0$. Now, the Gradient Descent update can be written equivalently as follows:

$$\mathbf{x}_{k+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^N}\left\{f(\mathbf{x};\mathbf{x}_k) + \frac{1}{2\tau}\|\mathbf{x}-\mathbf{x}_k\|^2\right\} \quad\Leftrightarrow\quad \mathbf{x}_{k+1} = \mathbf{x}_k - \tau\nabla f(\mathbf{x}_k)\,. \tag{4}$$

Its convergence analysis is essentially based on the Descent Lemma (2), which we reinterpret as a bound on the linearization error (model approximation error) of f. However, obviously (2) imposes a quadratic error bound, which cannot be satisfied in general. For example, functions like x^4 or $(x^3+y^3)^2$ or $(1-xy)^2$ do not have a Lipschitz continuous gradient. The same is true in several of the above-mentioned practical applications.

This issue was recently resolved in Bolte et al. (2018), based on the initial work in Bauschke et al. (2016), by introducing a generalization of the Lipschitz continuity assumption for the gradient mapping of a function, which was termed the "L-smad property". In convex optimization, similar notion coined "relative smoothness" was

proposed in Lu et al. (2018). Such a notion was also independently considered in Birnbaum et al. (2011), before Lu et al. (2018). However, all these approaches rely on the model function (3), which is the linearization of the function. In this paper, we generalize to arbitrary model functions (Definition 5) instead of the linearization of the function.

We briefly recall the "L-smad property". The main limitation of the Lipschitz continuous gradient notion is that it can only allow for quadratic approximation model errors. To go far beyond this setting, it then appears natural to invoke more general proximity measures as afforded by Bregman distances Bregman (1967). Several variants of Bregman distances exist in the literature Censor and Lent (1981); Bauschke and Borwein (1997); Bolte et al. (2018); Lu et al. (2018). We focus on those distances that are generated from so-called Legendre functions (Definition 3). Consider a Legendre function h, then the Bregman distance between $\mathbf{x} \in \operatorname{dom} h$ and $\mathbf{y} \in \operatorname{int} \operatorname{dom} h$ is given by

$$D_h(\mathbf{x}, \mathbf{y}) := h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \ . \tag{5}$$

A continuously differentiable function $f : \mathbb{R}^N \to \mathbb{R}$ is L-smad with respect to a Legendre function $h : \mathbb{R}^N \to \mathbb{R}$ over $\mathbb{R}^N$ with $\bar{L}, \underline{L} > 0$, if we have

$$-\underline{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \leq f(\mathbf{x}) - f(\bar{\mathbf{x}}) - \langle \nabla f(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle \leq \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \, , \forall \mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^N \, . \tag{6}$$

Note that with $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ in (6) we recover (2). We interpret the inequalities in (6) as a generalized distance measure for the linearization error of f. Similar to the Gradient Descent setting, minimization of $f(\bar{\mathbf{x}}) + \langle \nabla f(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle + \frac{1}{\tau} D_h(\mathbf{x}, \bar{\mathbf{x}})$ results in the Bregman proximal gradient (BPG) algorithm's update step Bolte et al. (2018) (a.k.a. Mirror Descent Beck and Teboulle (2003)).

However, the L-smad property relies on the continuous differentiability of the function f, thus nonsmooth functions as simple as $|x^4 - 1|$ or $|1 - (xy)^2|$ or $\log(1 + |1 - (xy)^2|)$ cannot be captured under the L-smad property. This lead us to the development of the MAP property (Definition 7), where MAP abbreviates Model Approximation Property. Consider a function $f : \mathbb{R}^N \to \mathbb{R}$ that is proper lower semicontinuous (lsc), and a Legendre function $h : \mathbb{R}^N \to \mathbb{R}$ with $\operatorname{dom} h = \mathbb{R}^N$. For certain $\bar{\mathbf{x}} \in \mathbb{R}^N$, we consider generic model function $f(\mathbf{x}; \bar{\mathbf{x}})$ that is proper lsc and approximates the function around the model center $\bar{\mathbf{x}}$, while preserving the local first order information (Definition 5). The MAP property is satisfied with constants $\bar{L} > 0$ and $\underline{L} \in \mathbb{R}$ if for any $\bar{\mathbf{x}} \in \mathbb{R}^N$ the following holds:

$$-\underline{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \leq f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}}) \leq \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \, , \quad \forall \mathbf{x} \in \mathbb{R}^N \, . \tag{7}$$

Note that we do not require the continuous differentiability of the function f. Our MAP property is inspired from Davis et al. (2018). However, their work considers only the lower bound with a weakly convex model function. Similar to the BPG setting, minimization of $f(\mathbf{x}; \bar{\mathbf{x}}) + \frac{1}{\tau} D_h(\mathbf{x}, \bar{\mathbf{x}})$ essentially results in Model BPG algorithm's update step. We illustrate the MAP property with a simple example. Consider a composite problem $f(x) = g(F(x)) := |x^4 - 1|$, where $F(x) := x^4 - 1$ and $g(x) := |x|$. Note that neither the Lipschitz continuity of the gradient nor the L-smad property is valid for this problem. However, the MAP property is valid with $\bar{L} = \underline{L} = 4$ using $f(x; \bar{x}) := g(F(\bar{x}) + \nabla F(\bar{x})(x - \bar{x}))$, where $\nabla F(\bar{x})$ is the Jacobian of F at $\bar{x}$, and $D_h(x, \bar{x}) = \frac{1}{4}x^4 - \frac{1}{4}\bar{x}^4 - \bar{x}^3(x - \bar{x})$, generated by $h(x) = \frac{1}{4}x^4$. We provide further details in Example 6 and in Example 9.

1.1 Contributions and relations to prior work

Our main contributions are the following.

- We introduce the MAP property, which generalizes the Lipschitz continuity assumption of the gradient mapping and the L-smad property Bolte et al. (2018); Bauschke et al. (2016). Earlier proposed notions were restricted to additive composite problems. The MAP property is essentially an extended Descent Lemma that is valid for generic composite problems (see Section 4) and beyond, based on Bregman distances. MAP like property was considered in Davis et al. (2018), however with focus on stochastic optimization and lower bounds of their MAP like property. The MAP property relies on the notion of *model function*, that serves as a function approximation, and preserves the local first order information of the function. Our work extends the foundations laid by Drusvyatskiy et al. (2019); Davis et al. (2018) based on generic model functions (potentially nonconvex), and Ochs et al. (2019) based on convex model functions. Taking inspiration from the update steps used in Davis et al. (2018) and based on the MAP property, we propose the Model based Bregman Proximal Gradient (Model BPG) algorithm (Algorithm 1). Apart from the work in Davis et al. (2018), another close variant of Model BPG is the line search based Bregman proximal gradient method Ochs et al. (2019), however, both the works do not consider the convergence of the full sequence of iterates.
- The global convergence analysis typically relies on the descent property of the function values. However, using function values can be restrictive, and alternatives are sought Pauwels (2016). We fix this issue by introducing a new Lyapunov function. We show that the (full) sequence generated by Model BPG converges to a critical point of the objective function. Notably, the usage of a Lyapunov function is popular for analysis of inertial algorithms Attouch et al. (2000); Ochs et al. (2014); Mukkamala et al. (2020); Attouch et al. (2020) and through our work we aim to popularize Lyapunov functions also for noninertial algorithms.
- The global convergence analysis of Bregman proximal gradient (BPG) Bolte et al. (2018) relies on the full domain of the Bregman distance, which contradicts their original purpose to represent the geometry of the constraint set. Our convergence theorem relaxes this restriction under certain assumptions that are typically satisfied in practice. In general, this requires the limit points of the sequence to lie in the interior of domain of the employed Legendre function. While this is certainly still a restriction, nevertheless, the considered setting is highly nontrivial and novel in the general context of nonconvex nonsmooth optimization. Moreover, it allows us to avoid the common restriction of requiring (global) strong convexity of the Legendre function, a severe drawback that rules out many interesting applications in related approaches (Section 5.2). In the context of convex optimization, works such as Lu (2019); Gutman and Peña (2018) use the reference functions (notion similar to the Legendre function) that are not strongly convex. In nonconvex nonsmooth optimization, Legendre functions that are not strongly convex are considered in Davis et al. (2018).
- We validate our theory with a numerical section showing the flexibility and the superior performance of Model BPG compared to a state of the art optimization algorithm, namely, Inexact Bregman Proximal Minimization Line Search (IBPM-

LS) Ochs et al. (2013), on standard phase retrieval problems and Poisson linear inverse problems.

1.2 Preliminaries and notations.

All the notations are primarily taken from Rockafellar and Wets (1998). We work in a Euclidean vector space $\mathbb{R}^N$ of dimension $N \in \mathbb{N}^*$ equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and induced norm $\| \cdot \|$. For a set $C \subset \mathbb{R}^N$, we define $\|C\|_- := \inf_{\mathbf{s} \in C} \|\mathbf{s}\|$. For any vector $\mathbf{x} \in \mathbb{R}^N$, the i^{th} coordinate is denoted by $\mathbf{x}_i$. We work with extended-valued functions $f \colon \mathbb{R}^N \to \overline{\mathbb{R}}$, $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. The domain of f is $\operatorname{dom} f := \left\{\mathbf{x} \in \mathbb{R}^N \,|\, f(\mathbf{x}) < +\infty\right\}$ and a function f is proper, if $\operatorname{dom} f \neq \emptyset$. It is lower semi-continuous (or closed), if $\liminf_{\mathbf{x} \to \bar{\mathbf{x}}} f(\mathbf{x}) \geq f(\bar{\mathbf{x}})$ for any $\bar{\mathbf{x}} \in \mathbb{R}^N$. Let $\operatorname{int} \Omega$ denote the interior of $\Omega \subset \mathbb{R}^N$. We use the notation of f-attentive convergence $\mathbf{x} \xrightarrow{f} \bar{\mathbf{x}} \Leftrightarrow (\mathbf{x}, f(\mathbf{x})) \to (\bar{\mathbf{x}}, f(\bar{\mathbf{x}}))$ and the notation $k \xrightarrow{K} \infty$ for some $K \subset \mathbb{N}$ to represent $k \to \infty$ where $k \in K$. The indicator function δ_C of a set $C \subset \mathbb{R}^N$ is defined by $\delta_C(\mathbf{x}) = 0$, if $\mathbf{x} \in C$ and $\delta_C(\mathbf{x}) = +\infty$, otherwise. The (orthogonal) projection of $\bar{\mathbf{x}}$ onto C, denoted $\operatorname{proj}_C(\bar{\mathbf{x}})$, is given by a minimizer of $\min_{\mathbf{x} \in C} \|\mathbf{x} - \bar{\mathbf{x}}\|$, which is well defined for a non-empty closed C. A set-valued mapping $T \colon \mathbb{R}^N \rightrightarrows \mathbb{R}^M$ is defined by its graph $\operatorname{Graph} T := \left\{(\mathbf{x}, \mathbf{v}) \in \mathbb{R}^N \times \mathbb{R}^M \,|\, \mathbf{v} \in T(\mathbf{x})\right\}$ with domain given by $\operatorname{dom} T := \left\{\mathbf{x} \in \mathbb{R}^N \,|\, T(\mathbf{x}) \neq \emptyset\right\}$. Following (Rockafellar and Wets, 1998, Def. 6.3), let $\bar{\mathbf{x}} \in C$, a vector $\mathbf{v}$ is regular normal to C, written $\mathbf{v} \in \widehat{N}_C(\bar{\mathbf{x}})$, if $\langle \mathbf{v}, \mathbf{x} - \bar{\mathbf{x}} \rangle \leq o(\|\mathbf{x} - \bar{\mathbf{x}}\|)$ for $\mathbf{x} \in C$. Here, $\mathbf{v}$ would be a normal vector, written $\mathbf{v} \in N_C(\bar{\mathbf{x}})$, if there exist sequences $\mathbf{x}_k \to \bar{\mathbf{x}}$ and $\mathbf{v}_k \to \mathbf{v}$, such that $\mathbf{x}_k \in C$ with $\mathbf{v}_k \in \widehat{N}_C(\mathbf{x}_k)$ for all $k \in \mathbb{N}$. Following (Rockafellar and Wets, 1998, Def. 8.3), we introduce subdifferential notions for nonsmooth functions. The Fréchet subdifferential of f at $\bar{\mathbf{x}} \in \operatorname{dom} f$ is the set $\widehat{\partial} f(\bar{\mathbf{x}})$ of elements $\mathbf{v} \in \mathbb{R}^N$ such that

$$\liminf_{\substack{\mathbf{x} \to \bar{\mathbf{x}} \\ \mathbf{x} \neq \bar{\mathbf{x}}}} \frac{f(\mathbf{x}) - f(\bar{\mathbf{x}}) - \langle \mathbf{v}, \mathbf{x} - \bar{\mathbf{x}} \rangle}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \geq 0 \,.$$

For $\bar{\mathbf{x}} \notin \operatorname{dom} f$, we set $\widehat{\partial} f(\bar{\mathbf{x}}) = \emptyset$. The (limiting) subdifferential of f at $\bar{\mathbf{x}} \in \operatorname{dom} f$ is defined by $\partial f(\bar{\mathbf{x}}) := \left\{\mathbf{v} \in \mathbb{R}^N \,|\, \exists\, \mathbf{y}_k \xrightarrow{f} \bar{\mathbf{x}},\ \mathbf{v}_k \in \widehat{\partial} f(\mathbf{y}_k),\ \mathbf{v}_k \to \mathbf{v}\right\}$, and $\partial f(\bar{\mathbf{x}}) = \emptyset$ for $\bar{\mathbf{x}} \notin \operatorname{dom} f$. As a direct consequence of the definition of the limiting subdifferential, we have the following closedness property at any $\bar{\mathbf{x}} \in \operatorname{dom} f$:

$$\mathbf{y}_k \xrightarrow{f} \bar{\mathbf{x}},\ \mathbf{v}_k \to \bar{\mathbf{v}},\ \text{and for all } k \in \mathbb{N} \colon \mathbf{v}_k \in \partial f(\mathbf{y}_k) \quad \Longrightarrow \quad \bar{\mathbf{v}} \in \partial f(\bar{\mathbf{x}}) \,. \tag{8}$$

A vector $\mathbf{v} \in \mathbb{R}^N$ is a horizon subgradient of f at $\bar{\mathbf{x}}$, if there are sequences $\mathbf{x}_k \xrightarrow{f} \bar{\mathbf{x}}$, $\mathbf{v}_k \in \widehat{\partial} f(\mathbf{x}_k)$, one has $\lambda_k \mathbf{v}_k \to \mathbf{v}$ for some sequence $\lambda_k \searrow 0$. The set of all horizon subgradients $\partial^\infty f(\bar{\mathbf{x}})$ is called horizon subdifferential. A point $\bar{\mathbf{x}} \in \operatorname{dom} f$ satisfying $\mathbf{0} \in \partial f(\bar{\mathbf{x}})$ is a called a critical point, which is a necessary optimality condition (Fermat's rule (Rockafellar and Wets, 1998, Thm. 10.1)) for $\bar{\mathbf{x}}$ being a local minimizer. The set of critical points is denoted by

$$\operatorname{crit} f := \left\{\mathbf{x} \in \mathbb{R}^N :\ \mathbf{0} \in \partial f(\mathbf{x})\right\} .$$

The set of (global) minimizers of a function f is

$$\operatorname*{Argmin}_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) := \left\{ \mathbf{x} \in \mathbb{R}^N \mid f(\mathbf{x}) = \inf_{\bar{\mathbf{x}}\in\mathbb{R}^N} f(\bar{\mathbf{x}}) \right\},$$

and the (unique) minimizer of f by $\operatorname{argmin}_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})$ if $\operatorname{Argmin}_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})$ is a singleton. We also use for short $\operatorname{Argmin} f$ and $\operatorname{argmin} f$.

Our global convergence theory relies on the so-called Kurdyka–Łojasiewicz (KL) property. It is a standard tool that is essentially satisfied by most of the functions that appear in practice. We just state the definition here from Attouch et al. (2013) and refer to Bolte et al. (2006, 2007, 2014); Kurdyka (1998) for more details.

Definition 1 (Kurdyka–Łojasiewicz property). Let $f\colon \mathbb{R}^N \to \overline{\mathbb{R}}$ and let $\bar{\mathbf{x}} \in \operatorname{dom} \partial f$. If there exists $\eta \in (0,\infty]$, a neighborhood U of $\bar{\mathbf{x}}$ and a continuous concave function $\varphi\colon [0,\eta) \to \mathbb{R}_+$ such that

$$\varphi(0) = 0, \quad \varphi \in C^1(0,\eta), \quad \text{and} \quad \varphi'(s) > 0 \text{ for all } s \in (0,\eta),$$

and for all $x \in U \cap [f(\bar{\mathbf{x}}) < f(\mathbf{x}) < f(\bar{\mathbf{x}}) + \eta]$ the Kurdyka–Łojasiewicz inequality

$$\varphi'(f(\mathbf{x}) - f(\bar{\mathbf{x}}))\|\partial f(\mathbf{x})\|_- \geq 1 \tag{9}$$

holds, then the function has the Kurdyka–Łojasiewicz property at $\bar{\mathbf{x}}$. If, additionally, the function is lsc and the property holds at each point in $\operatorname{dom} \partial f$, then f is called a Kurdyka–Łojasiewicz function.

We abbreviate Kurdyka–Łojasiewicz property as KL property. The function φ in the KL property is known as the desingularizing function. It is well known that the class of functions definable in an o-minimal structure van den Dries and Miller (1996) satisfy the KL property (Bolte et al., 2007, Theorem 14). Sets and functions that are semi-algebraic and globally subanalytic (for example, see (Bolte et al., 2007, Section 4), (Ochs, 2015, Section 4.5)) can be defined in an o-minimal structure.

We briefly review the concept of gradient-like descent sequence, that eases the global convergence analysis of Model BPG. We use the following results from Ochs (2019). Let $\mathcal{F}\colon \mathbb{R}^N \times \mathbb{R}^P \to \overline{\mathbb{R}}$ be a proper, lsc function that is bounded from below.

Assumption 1 (Gradient-like Descent Sequence Ochs (2019)). Let $(\mathbf{u}_n)_{n\in\mathbb{N}}$ be a sequence of parameters in $\mathbb{R}^P$ and let $(\varepsilon_n)_{n\in\mathbb{N}}$ be an ℓ_1-summable sequence of non-negative real numbers. Moreover, we assume there are sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$, and $(d_n)_{n\in\mathbb{N}}$ of non-negative real numbers, a non-empty finite index set $I \subset \mathbb{Z}$ and $\theta_i \geq 0$, $i \in I$, with $\sum_{i\in I} \theta_i = 1$ such that the following holds:

(H1) (Sufficient decrease condition) For each $n \in \mathbb{N}$, it holds that

$$\mathcal{F}(\mathbf{x}_{n+1}, \mathbf{u}_{n+1}) + a_n d_n^2 \leq \mathcal{F}(\mathbf{x}_n, \mathbf{u}_n)\,.$$

(H2) (Relative error condition) For each $n \in \mathbb{N}$, it holds that: (set $d_j = 0$ for $j \leq 0$)

$$b_{n+1}\|\partial\mathcal{F}(\mathbf{x}_{n+1}, \mathbf{u}_{n+1})\|_- \leq b \sum_{i\in I} \theta_i d_{n+1-i} + \varepsilon_{n+1}\,.$$

(H3) (Continuity condition) There exists a subsequence $((\mathbf{x}_{n_j}, \mathbf{u}_{n_j}))_{j\in\mathbb{N}}$ and $(\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \in \mathbb{R}^N \times \mathbb{R}^P$ such that $(\mathbf{x}_{n_j}, \mathbf{u}_{n_j}) \xrightarrow{\mathcal{F}} (\tilde{\mathbf{x}}, \tilde{\mathbf{u}}) \quad \text{as} \quad j \to \infty\,.$

(H4) (Distance condition) It holds that $d_n \to 0 \implies \|\mathbf{x}_{n+1} - \mathbf{x}_n\|_2 \to 0$ and $\exists n' \in \mathbb{N}\colon \forall n \geq n'\colon d_n = 0 \implies \exists n'' \in \mathbb{N}\colon \forall n \geq n''\colon \mathbf{x}_{n+1} = \mathbf{x}_n$.

(H5) (Parameter condition) $(b_n)_{n\in\mathbb{N}} \notin \ell_1\,, \sup_{n\in\mathbb{N}} \frac{1}{b_n a_n} < \infty\,, \inf_n a_n =: \underline{a} > 0$.

We now provide the global convergence statement from Ochs (2019), based on Assumption 1. The set of limit points of a bounded sequence $((\mathbf{x}_n, \mathbf{u}_n))_{n\in\mathbb{N}}$ is $\omega(\mathbf{x}_0, \mathbf{u}_0) := \limsup_{n\to\infty} \{(\mathbf{x}_n, \mathbf{u}_n)\}$, and denote the subset of $\mathcal{F}$-attentive limit points by

$$\omega_{\mathcal{F}}(\mathbf{x}_0, \mathbf{u}_0) := \left\{(\bar{\mathbf{x}}, \bar{\mathbf{u}}) \in \omega(\mathbf{x}_0, \mathbf{u}_0) \,|\, (\mathbf{x}_{n_j}, \mathbf{u}_{n_j}) \overset{\mathcal{F}}{\to} (\bar{\mathbf{x}}, \bar{\mathbf{u}}) \text{ for } j \to \infty\right\} .$$

Theorem 2 (Global convergence (Ochs, 2019, Theorem 10)). Suppose $\mathcal{F}$ is a proper lsc KL function that is bounded from below. Let $(\mathbf{x}_n)_{n\in\mathbb{N}}$ be a bounded sequence generated by an abstract algorithm parametrized by a bounded sequence $(\mathbf{u}_n)_{n\in\mathbb{N}}$ that satisfies Assumption 1. Assume that $\mathcal{F}$-attentive convergence holds along converging subsequences of $((\mathbf{x}_n, \mathbf{u}_n))_{n\in\mathbb{N}}$, i.e. $\omega(\mathbf{x}_0, \mathbf{u}_0) = \omega_{\mathcal{F}}(\mathbf{x}_0, \mathbf{u}_0)$. Then, the following holds:

(i) The sequence $(d_n)_{n\in\mathbb{N}}$ satisfies $\sum_{k=0}^{\infty} d_k < +\infty$, i.e., the trajectory of the sequence $(\mathbf{x}_n)_{n\in\mathbb{N}}$ has finite length w.r.t. the abstract distance measures $(d_n)_{n\in\mathbb{N}}$.

(ii) Suppose d_k satisfies $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2 \leq \bar{c} d_{k+k'}$ for some $k' \in \mathbb{Z}$ and $\bar{c} \in \mathbb{R}$, then $(\mathbf{x}_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, and thus $(\mathbf{x}_n)_{n\in\mathbb{N}}$ converges to $\tilde{\mathbf{x}}$ from (H3).

(iii) Moreover, if $(\mathbf{u}_n)_{n\in\mathbb{N}}$ is a converging sequence, then each limit point of the sequence $((\mathbf{x}_n, \mathbf{u}_n))_{n\in\mathbb{N}}$ is a critical point of $\mathcal{F}$, which in the situation of (ii) is the unique point $(\tilde{\mathbf{x}}, \tilde{\mathbf{u}})$ from (H3).

Legendre functions defined below generate Bregman distances, which are generalized proximity measures compared to the Euclidean distance.

Definition 3 (Legendre function (Rockafellar, 1970, Section 26)). Let $h : \mathbb{R}^N \to \overline{\mathbb{R}}$ be a proper lsc convex function. It is called:

(i) essentially smooth, if h is differentiable on $\operatorname{int}\operatorname{dom} h$, with moreover $\|\nabla h(\mathbf{x}_k)\| \to \infty$ for every sequence $(\mathbf{x}_k)_{k\in\mathbb{N}} \in \operatorname{int}\operatorname{dom} h$ converging to a boundary point of $\operatorname{dom} h$ as $k \to \infty$;

(ii) of Legendre type if h is essentially smooth and strictly convex on $\operatorname{int}\operatorname{dom} h$.

Some properties of Legendre function include $\operatorname{dom}\partial h = \operatorname{int}\operatorname{dom} h$, and $\partial h(\mathbf{x}) = \{\nabla h(\mathbf{x})\}$, $\forall \mathbf{x} \in \operatorname{int}\operatorname{dom} h$. Additional properties can be found in (Bauschke and Borwein, 1997, Section 2.3). For the purpose of our analysis, we later require that the Legendre functions are twice continuously differentiable (see Assumption 4). Legendre function is also referred as kernel generating distance Bolte et al. (2018), or a reference function Lu et al. (2018). Generic reference functions used in Lu et al. (2018) are more general compared to Legendre functions, as they do not require essential smoothness. The Bregman distance associated with any Legendre function h is defined by

$$D_h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle\,, \quad \forall\, \mathbf{x} \in \operatorname{dom} h,\, \mathbf{y} \in \operatorname{int}\operatorname{dom} h\,. \tag{10}$$

In contrast to the Euclidean distance, the Bregman distance is lacking symmetry. Prominent examples of Bregman distances can be found in (Bauschke et al., 2016, Example 1, 2) and for additional results, we refer the reader to Bauschke and Borwein (1997); Bauschke et al. (2001, 2003, 2016). We provide some examples below.

- Bregman distance generated from $h(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2$ is the Euclidean distance.
- Let $\mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^N_{++}$, the Legendre function $h(\mathbf{x}) = -\sum_{i=1}^N \log(\mathbf{x}_i)$ (Burg's entropy) is helpful in Poisson linear inverse problems Bauschke et al. (2016).
- Let $\mathbf{x} \in \mathbb{R}^N_{+}$, $\bar{\mathbf{x}} \in \mathbb{R}^N_{++}$, the Legendre function $h(\mathbf{x}) = \sum_{i=1}^N \mathbf{x}_i \log(\mathbf{x}_i)$ (Boltzmann–Shannon entropy), with $0\log(0) := 0$ is helpful to handle simplex constraints Beck and Teboulle (2003).
- Phase retrieval problems Bolte et al. (2018) use the Bregman distance based on the Legendre function $h : \mathbb{R}^N \to \mathbb{R}$ that is given by $h(\mathbf{x}) = 0.25\|\mathbf{x}\|_2^4 + 0.5\|\mathbf{x}\|_2^2$.
- Matrix factorization problems Mukkamala and Ochs (2019); Teboulle and Vaisbourd (2020) use the Bregman distance based on the Legendre function $h : \mathbb{R}^{N_1} \times \mathbb{R}^{N_2} \to \mathbb{R}$ that is given by $h(\mathbf{x}, \mathbf{y}) = c_1(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)^2 + c_2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$ with certain $c_1, c_2 > 0$ and $N_1, N_2 \in \mathbb{N}$.

2 Problem setting and Model BPG algorithm

We consider the optimization problem (1) where f satisfies the following assumption, which we impose henceforth.

Assumption 2. $f \colon \mathbb{R}^N \to \overline{\mathbb{R}}$ is proper, lsc (possibly nonconvex nonsmooth) and coercive, i.e., as $\|\mathbf{x}\| \to \infty$ we have $f(\mathbf{x}) \to \infty$.

Due to (Rockafellar and Wets, 1998, Theorem 1.9), the function f satisfying Assumption 2 is bounded from below, and $\operatorname{Argmin}_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})$ is nonempty and compact. Denote $v(\mathcal{P}) := \min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) > -\infty$. We require the following definitions.

Definition 4 (Growth function Drusvyatskiy et al. (2019); Ochs et al. (2019)). A differentiable univariate function $\varsigma \colon \mathbb{R}_+ \to \mathbb{R}_+$ is called *growth function* if it satisfies $\varsigma(0) = \varsigma'_+(0) = 0$, where ς'_+ denotes the one sided (right) derivative of ς. If, in addition, $\varsigma'_+(t) > 0$ for $t > 0$ and equalities $\lim_{t\searrow 0} \varsigma'_+(t) = \lim_{t\searrow 0} \varsigma(t)/\varsigma'_+(t) = 0$ hold, we say that ς is a *proper growth function*.

Example of a proper growth function is $\varsigma(t) = \frac{\eta}{r}t^r$ for $\eta, r > 0$. Lipschitz continuity and Hölder continuity can be interpreted with growth functions or, more generally, with uniform continuity Ochs et al. (2019). We use the notion of a growth function to quantify the difference between a model function (defined below) and the objective.

Definition 5 (Model Function). Let f be a proper lower semi-continuous (lsc) function. A function $f(\cdot, \bar{\mathbf{x}}) \colon \mathbb{R}^N \to \overline{\mathbb{R}}$ with $\operatorname{dom} f(\cdot, \bar{\mathbf{x}}) = \operatorname{dom} f$ is called *model function* for f around the *model center* $\bar{\mathbf{x}} \in \operatorname{dom} f$, if there exists a growth function $\varsigma_{\bar{\mathbf{x}}}$ such that the following is satisfied:

$$|f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}})| \leq \varsigma_{\bar{\mathbf{x}}}(\|\mathbf{x} - \bar{\mathbf{x}}\|), \quad \forall\, \mathbf{x} \in \operatorname{dom} f. \tag{11}$$

A model function is essentially a first-order approximation to a function f, which explains the naming as "Taylor-like model" by Drusvyatskiy et al. (2019). The qualitative approximation property is represented by the growth function. Informally, the model function approximates the function well near the model center. Convex model functions are explored in Ochs et al. (2019); Ochs and Malitsky (2019). However, in our setting, the model functions can be nonconvex. Nonconvex model functions were considered in Drusvyatskiy et al. (2019), however only subsequential convergence was shown.

We refer to (11) as a bound on the model error, and the symbol $\varsigma_{\bar{\mathbf{x}}}$ denotes the dependency of the growth function on the model center $\bar{\mathbf{x}}$. Typically the growth function depends on the model center, as we illustrate below.

Example 6 (Running Example). Let $f(\mathbf{x}) = |g(\mathbf{x})|$ with $g(\mathbf{x}) = \|\mathbf{x}\|^4 - 1$. With $\bar{\mathbf{x}} \in \mathbb{R}^N$ as the model center, and the model function

$$f(\mathbf{x}; \bar{\mathbf{x}}) := |g(\bar{\mathbf{x}}) + \langle \nabla g(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle| \,.$$

With the growth function is $\varsigma_{\bar{\mathbf{x}}}(t) = 24\|\bar{\mathbf{x}}\|^2 t^2 + 8t^4$, the model error obtained is

$$|f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}})| \leq 24\|\bar{\mathbf{x}}\|^2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2 + 8\|\mathbf{x} - \bar{\mathbf{x}}\|^4 \,.$$

It is often of interest to obtain a uniform approximation for the model error $|f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}})|$, where the growth function is not dependent on the model center. In general, obtaining such a uniform approximation is not trivial, and may even be impossible. Moreover, typically finding an appropriate growth function is not trivial. For this purpose, it is preferable to have a global bound on the model error that can be easily verified, the dependency on the model center is more structured, and the constants arising do not have any dependency on the model center. In the context of additive composite problems, previous works such as Bauschke et al. (2016); Lu et al. (2018); Bolte et al. (2018) relied on Bregman distances to upper bound the model error. Based on this idea, we propose the following MAP property, which is valid for a huge class of structured nonconvex problems and also generalizes the previous works.

Definition 7 (MAP: Model Approximation Property). Let h be a Legendre function that is continuously differentiable over $\operatorname{int}\operatorname{dom} h$. A proper lsc function f with $\operatorname{dom} f \subset \operatorname{cl}\operatorname{dom} h$ and $\operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h \neq \emptyset$, and model function $f(\cdot, \bar{\mathbf{x}})$ for f around $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$ satisfy the *Model Approximation Property (MAP) at* $\bar{\mathbf{x}}$, with the constants $\bar{L} > 0, \underline{L} \in \mathbb{R}$, if for any $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$ the following holds:

$$-\underline{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \leq f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}}) \leq \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \,, \quad \forall \mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h \,. \tag{12}$$

Remark 8 (Discussion on Definition 7). (i) The design of a model function is independent of an algorithm. However, algorithms can be governed by the model function (for example, see Model BPG below). The property of a model function is rather an analogue to differentiability or a (uniform) first-order approximation. Note that for $\bar{\mathbf{x}} \in \operatorname{int}\operatorname{dom} h$, the Bregman distance $D_h(\mathbf{x}, \bar{\mathbf{x}})$ is bounded by $o(\|\mathbf{x} - \bar{\mathbf{x}}\|)$, which is a growth function. Therefore, the MAP property requires additional algorithm specific properties of the model function. In particular, we require the constants $\bar{L}$ and $\underline{L}$ to be independent of $\bar{\mathbf{x}}$, which provides a global consistency between the model function approximations.

(ii) The condition $\operatorname{dom} f \subset \operatorname{cl}\operatorname{dom} h$ is a minor regularity condition. For example, if $\operatorname{dom} f = [0, \infty)$ and $\operatorname{dom} h = (0, \infty)$ (e.g., for h in Burg's entropy), such a function h can be used in MAP property. However, the L-smad property Bolte et al. (2018) would require $\mathbf{x}, \bar{\mathbf{x}}$ in (12) to lie in $\operatorname{int}\operatorname{dom} h$ (see also Section 4.1).

(iii) Note that the choice of $\underline{L}$ is unrestricted in MAP property. For nonconvex f, $\underline{L}$ is typically a positive real number. For convex f, typically the condition $\underline{L} \geq 0$ holds true. However, note that the values of $\underline{L}, \bar{L}$ are governed by the model function. For convex additive composite problems, $\underline{L} < 0$ can hold true for relatively strongly convex functions Lu et al. (2018).

Example 9 (Running Example – Contd). We continue Example 6 to illustrate the MAP property. Let $h(\mathbf{x}) = \frac{1}{4}\|\mathbf{x}\|^4$, we clearly have

$$g(\mathbf{x}) - g(\bar{\mathbf{x}}) - \langle \nabla g(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle \leq 4D_h(\mathbf{x}, \bar{\mathbf{x}}), \quad \forall\, \mathbf{x} \in \mathbb{R}^N,$$

which in turn results in the following upper bound for the model error

$$|f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}})| \leq |g(\mathbf{x}) - g(\bar{\mathbf{x}}) - \langle \nabla g(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle| \leq 4D_h(\mathbf{x}, \bar{\mathbf{x}}).$$

The upper bound is obtained in terms of a Bregman distance. Clearly, the constants arising do not have any dependency on the model center.

We now propose the Model BPG algorithm, where the update step relies on the upper bound of the MAP property.

Algorithm 1 (Model BPG: Model based Bregman Proximal Gradient).
- **Initialization:** Select $\mathbf{x}_0 = \mathbf{x}_1 \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$. Choose $\underline{\tau}, \bar{\tau}$ such that $0 < \underline{\tau} < \bar{\tau} < (1/\bar{L})$.
- **For each** $k \geq 1$**:** Choose $\tau_k \in [\underline{\tau}, \bar{\tau}]$ and compute

$$\mathbf{x}_{k+1} \in \operatorname*{Argmin}_{\mathbf{x} \in \mathbb{R}^N} \left\{ f(\mathbf{x}; \mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{x}, \mathbf{x}_k) \right\}. \tag{13}$$

Remark 10. (i) A closely related work in Davis et al. (2018) considers only the lower bound of the MAP property and their algorithm terminates by choosing an iterate based on certain probability distribution. In stark contrast, Model BPG relies on the upper bound of the MAP property and there is no need to invoke any probabilistic argument to choose the final iterate. Also, Davis et al. (2018) considers weakly convex model functions whereas we do not have such a restriction.

(ii) For the global convergence analysis of Model BPG sequences, in addition to the condition $\tau_k \in [\underline{\tau}, \bar{\tau}]$ on step-size, the condition that $\tau_k \to \tau$, as $k \to \infty$ for certain $\tau > 0$ is required (see Theorem 17 , 18).

(iii) We note that Model BPG is applicable to a broad class of structured nonconvex and nonsmooth problems. In particular, Model BPG can be efficiently applied to those nonconvex and nonsmooth problems, for which the update step (13) involving the Bregman distance can be easily computed.

We now collect all the assumptions required for the global convergence analysis of a sequence generated by the Model BPG algorithm.

Assumption 3. Let h be a Legendre function that is $\mathcal{C}^2$ over $\operatorname{int} \operatorname{dom} h$. Moreover, the conditions $\operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h \neq \emptyset$, $\operatorname{crit} f \cap \operatorname{int} \operatorname{dom} h \neq \emptyset$ and $\operatorname{dom} f \subset \operatorname{cl} \operatorname{dom} h$ hold true.

(i) The exist $\bar{L} > 0, \underline{L} \in \mathbb{R}$ such that for any $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$, f and the model function $f(\cdot, \bar{\mathbf{x}})$ satisfy the MAP property at $\bar{\mathbf{x}}$ with constants $\bar{L}, \underline{L}$.

(ii) For any $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$, the following qualification condition holds true:

$$\partial_{\mathbf{x}}^{\infty} f(\mathbf{x}; \bar{\mathbf{x}}) \cap (-N_{\operatorname{dom} h}(\mathbf{x})) = \{\mathbf{0}\}, \quad \forall\, \mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h. \tag{14}$$

(iii) For all $\mathbf{x}, \mathbf{y} \in \operatorname{dom} f$, the conditions $(\mathbf{0}, \mathbf{v}) \in \partial^{\infty} f(\mathbf{x}; \mathbf{y})$ implies $\mathbf{v} = \mathbf{0}$, and $(\mathbf{v}, \mathbf{0}) \in \partial^{\infty} f(\mathbf{x}; \mathbf{y})$ implies $\mathbf{v} = \mathbf{0}$ hold true. Also, $f(\mathbf{x}; \mathbf{y})$ is regular (Rockafellar and Wets, 1998, Definition 7.25) at any $(\mathbf{x}, \mathbf{y}) \in \operatorname{dom} f \times \operatorname{dom} f$.

(iv) The function $f(\mathbf{x};\bar{\mathbf{x}})$ is a proper, lsc function and is continuous over $(\mathbf{x},\bar{\mathbf{x}}) \in \operatorname{dom} f \times \operatorname{dom} f$.

By $\partial_{\mathbf{x}} f(\mathbf{x};\bar{\mathbf{x}})$ we mean the limiting subdifferential of the model function $\mathbf{x} \mapsto f(\mathbf{x};\bar{\mathbf{x}})$ with $\bar{\mathbf{x}}$ fixed and $\partial f(\mathbf{x};\mathbf{y})$ denotes the limiting subdifferential w.r.t $(\mathbf{x},\mathbf{y})$; dito for the horizon subdifferential.

Remark 11 (Discussion on Assumption 3). The qualification condition in (14) is required for the applicability of the subdifferential summation rule (see (Rockafellar and Wets, 1998, Corollary 10.9)). Assumption 3(iii) and (Rockafellar and Wets, 1998, Corollary 10.11) ensure that for all $\mathbf{x},\mathbf{y} \in \operatorname{dom} f$, the following holds true:

$$\partial f(\mathbf{x};\mathbf{y}) = \partial_{\mathbf{x}} f(\mathbf{x};\mathbf{y}) \times \partial_{\mathbf{y}} f(\mathbf{x};\mathbf{y})\,,\ \partial^{\infty} f(\mathbf{x};\mathbf{y}) = \partial^{\infty}_{\mathbf{x}} f(\mathbf{x};\mathbf{y}) \times \partial^{\infty}_{\mathbf{y}} f(\mathbf{x};\mathbf{y})\,. \tag{Assumption 3(iii)'}$$

Our analysis relies on (Assumption 3(iii)'). However, note that Assumption 3(iii) is a sufficient condition for (Assumption 3(iii)') to hold. Certain classes of functions mentioned in Section 4 satisfy (Assumption 3(iii)') directly, instead of Assumption 3(iii). Assumption 3(iv) is typically satisfied in practice and plays a key role in Lemma 30. Based on Assumption 3(iii), for any fixed $\bar{\mathbf{x}} \in \operatorname{dom} f$, the model function $f(\mathbf{x};\bar{\mathbf{x}})$ is regular at any $\mathbf{x} \in \operatorname{dom} f$. Using this fact, we deduce that the model function preserves the first order information of the function, in the sense that for $\mathbf{x} \in \operatorname{dom} f$ the condition $\partial_{\mathbf{y}} f(\mathbf{y};\mathbf{x})|_{\mathbf{y}=\mathbf{x}} = \widehat{\partial} f(\mathbf{x})$ holds true (based on (Ochs and Malitsky, 2019, Lemma 14)).

Many popular algorithms such as Gradient Descent, Proximal Gradient Method, Bregman Proximal Gradient Method, Prox-Linear method are special cases of Model BPG depending on the choice of the model function and the choice of Bregman distance, thus making it a unified algorithm (also c.f. Ochs et al. (2019)). Examples of model functions are provided in Section 4. Let $\tau > 0$, $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$, the update mapping (as in (13)) is defined by

$$T_\tau(\bar{\mathbf{x}}) := \operatorname*{Argmin}_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x};\bar{\mathbf{x}}) + \frac{1}{\tau} D_h(\mathbf{x},\bar{\mathbf{x}})\,. \tag{15}$$

Denote $\varepsilon_k := \left(\frac{1}{\tau_k} - \bar{L}\right) > 0$ and clearly $\underline{\varepsilon} \le \varepsilon_k \le \bar{\varepsilon}$, where $\bar{\varepsilon} := \frac{1}{\underline{\tau}} - \bar{L}$ and $\underline{\varepsilon} := \frac{1}{\bar{\tau}} - \bar{L}$. Well-posedness of the update step (13) is given by the following result.

Lemma 12. Let Assumption 2, 3 hold true and let $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$. Then, for all $0 < \tau < \frac{1}{\bar{L}}$ the set $T_\tau(\bar{\mathbf{x}})$ is a nonempty compact subset of $\operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$.

Proof. As a consequence of MAP property due to Assumption 3(i) and nonnegativity of Bregman distances, the following property is satisfied

$$f(\mathbf{x}) \le f(\mathbf{x};\bar{\mathbf{x}}) + \frac{1}{\tau} D_h(\mathbf{x},\bar{\mathbf{x}})\,, \forall\, \mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h\,.$$

Coercivity of f transfers to that of the objective in (15), and we get the conclusion from standard arguments; see (Rockafellar and Wets, 1998, Theorem 1.9). □

The conclusion of the lemma remains true under other sufficient conditions. For instance, if the model has an affine minorant and h is supercoercive (for example, see (Bolte et al., 2018, Section 3.1)). We now show that Model BPG results in monotonically nonincreasing function values.

Lemma 13 (Sufficient Descent Property in Function values). Let Assumptions 2, 3 hold. Also, let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a sequence generated by Model BPG, then for $k \geq 1$, the following holds

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \varepsilon_k D_h(\mathbf{x}_{k+1}, \mathbf{x}_k)\,.$$

Proof. Due to (13), we have $f(\mathbf{x}_{k+1};\mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \leq f(\mathbf{x}_k;\mathbf{x}_k) = f(\mathbf{x}_k)$. From MAP property, we have $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_{k+1};\mathbf{x}_k) + \bar{L} D_h(\mathbf{x}_{k+1},\mathbf{x}_k)$. The result follows by combining the previous arguments. □

Remark 14. Under Assumptions 2, 3, the coercivity of f, Lemma 13 implies that the iterates of Model BPG lie in the compact set $\{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$.

Assumption 4.

(i) For any bounded set $B \subset \operatorname{dom} f$. There exists $c > 0$ such that for any $\mathbf{x}, \mathbf{y} \in B$ we have

$$\|\partial_{\mathbf{y}} f(\mathbf{x};\mathbf{y})\|_- \leq c\|\mathbf{x}-\mathbf{y}\|\,.$$

(ii) The function h has bounded second derivative on any compact subset $B \subset \operatorname{int}\operatorname{dom} h$.

(iii) For bounded $(\mathbf{u}_k)_{k\in\mathbb{N}}$, $(\mathbf{v}_k)_{k\in\mathbb{N}}$ in $\operatorname{int}\operatorname{dom} h$, the following holds as $k \to \infty$:

$$D_h(\mathbf{u}_k, \mathbf{v}_k) \to 0 \iff \|\mathbf{u}_k - \mathbf{v}_k\| \to 0\,.$$

We now illustrate Assumption 4(i), which governs the variation of the model function w.r.t. model center.

Example 15. We continue Example 6 to illustrate Assumption 4(i). Note that that $\nabla^2 g(\mathbf{x})$ is bounded over bounded sets. Consider any bounded set $B \subset \mathbb{R}^N$. Define $c := \sup_{\bar{\mathbf{x}}\in B} \|\nabla^2 g(\bar{\mathbf{x}})\|$ and choose any $\bar{\mathbf{x}} \in B$, then consider the model function $f(\mathbf{x};\bar{\mathbf{x}}) := |g(\bar{\mathbf{x}}) + \langle \nabla g(\bar{\mathbf{x}}), \mathbf{x}-\bar{\mathbf{x}}\rangle|$. The subdifferential of the model function is given by $\partial_{\bar{\mathbf{x}}} f(\mathbf{x};\bar{\mathbf{x}}) = \mathbf{u}\nabla^2 g(\bar{\mathbf{x}})(\mathbf{x}-\bar{\mathbf{x}})$, where $\mathbf{u} \in \partial_{g(\bar{\mathbf{x}})+\langle \nabla g(\bar{\mathbf{x}}), \mathbf{x}-\bar{\mathbf{x}}\rangle} |g(\bar{\mathbf{x}}) + \langle \nabla g(\bar{\mathbf{x}}), \mathbf{x}-\bar{\mathbf{x}}\rangle|$. Considering the fact that $\|\mathbf{u}\| \leq 1$ and by the definition of c we have $\|\partial_{\bar{\mathbf{x}}} f(\mathbf{x};\bar{\mathbf{x}})\|_- \leq c\|\mathbf{x}-\bar{\mathbf{x}}\|$, which verifies Assumption 4(i).

In order to exploit the power of KL property in the global convergence analysis of Model BPG, we make the following assumption.

Assumption 5. Let $\mathcal{O}$ be an o-minimal structure. The functions $\tilde{f} : \mathbb{R}^N \times \mathbb{R}^N \to \overline{\mathbb{R}}$, $(\mathbf{x},\bar{\mathbf{x}}) \mapsto f(\mathbf{x};\bar{\mathbf{x}})$ with $\operatorname{dom}\tilde{f} := \operatorname{dom} f \times \operatorname{dom} f$, and $\tilde{h} : \mathbb{R}^N \times \mathbb{R}^N \to \overline{\mathbb{R}}$, $(\mathbf{x},\bar{\mathbf{x}}) \mapsto h(\bar{\mathbf{x}}) + \langle \nabla h(\bar{\mathbf{x}}), \mathbf{x}-\bar{\mathbf{x}}\rangle$ with $\operatorname{dom}\tilde{h} := \operatorname{dom} h \times \operatorname{int}\operatorname{dom} h$ are definable $\mathcal{O}$.

An important feature of our analysis is that the Legendre function h satisfying Assumption 3 is not required to be strongly convex. Instead, we impose a significantly weaker condition in Assumption 6 provided below.

Assumption 6. For any compact convex set $B \subset \operatorname{int}\operatorname{dom} h$, there exists $\sigma_B > 0$ such that h is σ_B-strongly convex over B, i.e., for any $\mathbf{x}, \mathbf{y} \in B$ the condition $D_h(\mathbf{x},\mathbf{y}) \geq \frac{\sigma_B}{2}\|\mathbf{x}-\mathbf{y}\|^2$ holds.

Remark 16 (Discussion on Assumption 4 - 6). Assumption 4(i) is illustrated in Example 15. Assumption 4(ii) is typically used in the analysis of Bregman proximal methods Bolte et al. (2018); Ochs et al. (2019); Mukkamala et al. (2020). Assumption 4(iii) (also see (Ochs et al., 2019, Remark 18)) essentially states that

the asymptotic behavior of vanishing Bregman distance is equivalent to that of vanishing Euclidean distance. Note that Assumption 4(iii) already uses bounded sequences in $\operatorname{int}\operatorname{dom} h$, and thus it is satisfied for many Bregman distances, such as distances based on Boltzmann–Shannon entropy (Ochs et al., 2019, Example 40) and Burg's entropy (Ochs et al., 2019, Example 41). However, such distances may not satisfy Assumption 4(iii) if the sequences are bounded only in $\operatorname{dom} h$ or in $\operatorname{cl}\operatorname{dom} h$ (for example, see Section 5.2). Assumption 5 is used in Lemma 28 to deduce that $F_{\bar{L}}^h$ satisfies KL property. Assumption 6 plays a key role in proving the global convergence of the sequence generated by Model BPG.

3 Global convergence analysis of Model BPG algorithm

3.1 Main results

Our goal is to show that the sequence generated by Model BPG is a gradient-like descent sequence such that Theorem 2 is applicable. The convergence analysis of some popular algorithms (for example, PGM, BPG, PALM Bolte et al. (2014) etc) in nonconvex optimization is based on a descent property. Usually, the objective value is shown to decrease (for example, see (Bolte et al., 2018, Lemma 4.1)). However, techniques used for additive composite setting relying on function values do not work anymore for general composite problems, hence alternatives like Pauwels (2016) are sought after. We analyse Model BPG using a Lyapunov function as our measure of progress. Our Lyapunov function $F_{\bar{L}}^h$ is given by

$$F_{\bar{L}}^h \colon \mathbb{R}^N \times \mathbb{R}^N \to \overline{\mathbb{R}}\,, \quad (\mathbf{x}, \bar{\mathbf{x}}) \mapsto f(\mathbf{x}; \bar{\mathbf{x}}) + \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}})\,, \tag{16}$$

and $\operatorname{dom} F_{\bar{L}}^h = (\operatorname{dom} f)^2 \times (\operatorname{dom} h \times \operatorname{int}\operatorname{dom} h)$. The set of critical points of $F_{\bar{L}}^h$ is given by

$$\operatorname{crit} F_{\bar{L}}^h := \left\{ (\mathbf{x}, \bar{\mathbf{x}}) \in \mathbb{R}^N \times \mathbb{R}^N : \ (\mathbf{0}, \mathbf{0}) \in \partial F_{\bar{L}}^h(\mathbf{x}, \bar{\mathbf{x}}) \right\}. \tag{17}$$

The set of limit points of some sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is denoted as follows $\omega(\mathbf{x}_0) := \left\{ \mathbf{x} \in \mathbb{R}^N \,|\, \exists K \subset \mathbb{N} \colon \mathbf{x}_k \overset{K}{\to} \mathbf{x} \right\}$, and its subset of f-attentive limit points

$$\omega_f(\mathbf{x}_0) := \left\{ \mathbf{x} \in \mathbb{R}^N \,|\, \exists K \subset \mathbb{N} \colon (\mathbf{x}_k, f(\mathbf{x}_k)) \overset{K}{\to} (\mathbf{x}, f(\mathbf{x})) \right\}.$$

To this regard, denote the following

$$\omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) := \omega(\mathbf{x}_0) \cap \operatorname{int}\operatorname{dom} h \quad \text{and} \quad \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) := \omega_f(\mathbf{x}_0) \cap \operatorname{int}\operatorname{dom} h\,.$$

Before we start with the convergence analysis, we present our main results. We defer their proofs to Section 3.2. Informally, the following results state that the sequence generated by Model BPG converges to a point $\mathbf{x}$ such that $(\mathbf{x}, \mathbf{x})$ is the critical point of $F_{\bar{L}}^h$ and $\mathbf{x}$ is a critical point of f.

Theorem 17 (Global convergence to a critical point of the Lyapunov function). Let Assumptions 2, 3, 4, 5, 6 hold. Let the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be generated by Model BPG (Algorithm 1) with $\tau_k \to \tau$ for certain $\tau > 0$ and the condition $\omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) = \omega(\mathbf{x}_0)$ holds true. Then, convergent subsequences are $F_{\bar{L}}^h$-attentive convergent, and $\sum_{k=0}^{\infty} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| < +\infty$ (finite length property). The sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges to $\mathbf{x}$ such that $(\mathbf{x}, \mathbf{x})$ is a critical point of $F_{\bar{L}}^h$.

Theorem 18 (Global convergence to a critical point of the objective function). Under the conditions of Theorem 17, the sequence generated by Model BPG converges to a critical point of f.

It is possible to deduce convergence rates for a certain class of desingularizing functions. Based on Attouch and Bolte (2009); Bolte et al. (2014); Frankel et al. (2014), we provide the following convergence rates for Model BPG sequences.

Theorem 19 (Convergence rates). Under the conditions of Theorem 17, let the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by Model BPG converge to $\mathbf{x} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$, and let F_L^h satisfy KL property with the desingularizing function: $\varphi(s) = cs^{1-\theta}$, for certain $c > 0$ and $\theta \in [0, 1)$. Then, we have the following:

- If $\theta = 0$, then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ converges in finite number of steps.
- If $\theta \in (0, \frac{1}{2}]$, then $\exists\, \rho \in [0, 1)$, $G > 0$ such that $\forall\, k \geq 0$ we have $\|\mathbf{x}_k - \mathbf{x}\| \leq G\rho^k$.
- If $\theta \in (\frac{1}{2}, 1)$, then $\exists\, G > 0$ such that $\forall\, k \geq 0$ we have $\|\mathbf{x}_k - \mathbf{x}\| \leq Gk^{-\frac{1-\theta}{2\theta-1}}$.

The proof is only a slight modification to the proof of (Attouch and Bolte, 2009, Theorem 5), hence we skip it for brevity. In the above theorem θ is the so-called KL exponent (also called Łojasiewicz exponent in classical algebraic geometry) of the Lyapunov function F_L^h and not that of the function f. Thus the KL exponent of F_L^h is nontrivial to deduce even if the KL exponent of f is known, as it has dependency on the model function and the Bregman distance. In this regard, we refer the reader to Li and Pong (2017); Li et al. (2015).

3.2 Additional results and proofs

We now look at some properties of F_L^h.

Proposition 20. The Lyapunov function defined in (16) satisfies the following:

(i) For all $\mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h$ and $\mathbf{y} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$, we have $f(\mathbf{x}) \leq F_L^h(\mathbf{x}, \mathbf{y})$.
(ii) For all $\mathbf{x} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$, we have $F_L^h(\mathbf{x}, \mathbf{x}) = f(\mathbf{x})$.
(iii) Moreover, we have $\inf_{(\mathbf{x},\mathbf{y})\,\in\,\mathbb{R}^N\times\mathbb{R}^N} F_L^h(\mathbf{x}, \mathbf{y}) \geq v(\mathcal{P}) > -\infty$.

Proof. (i) This follows from MAP property and the definition of F_L^h.
(ii) Substituting $\mathbf{y} = \mathbf{x}$ in (16) gives the result.
(iii) By MAP property, we have $v(\mathcal{P}) \leq f(\mathbf{x}) \leq f(\mathbf{x}; \mathbf{y}) + \bar{L}D_h(\mathbf{x}, \mathbf{y})$, for all $(\mathbf{x}, \mathbf{y}) \in \operatorname{dom} F_L^h$. Furthermore, we obtain the following:

$$\inf_{\mathbf{x}\in\operatorname{dom} f\,\cap\,\operatorname{dom} h} f(\mathbf{x}) \leq \inf_{(\mathbf{x},\mathbf{y})\in\operatorname{dom} F_L^h} \left(f(\mathbf{x}; \mathbf{y}) + \bar{L}D_h(\mathbf{x}, \mathbf{y})\right).$$

The statement follows using $\inf_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) = v(\mathcal{P}) > -\infty$ due to Assumption 2. □

We proved the sufficient descent property in terms of function values in Lemma 13. We now prove the sufficient descent property of the Lyapunov function.

Proposition 21 (Sufficient descent property). Let Assumptions 2, 3 hold and let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a sequence generated by Model BPG, then for $k \geq 1$ we have

$$F_L^h(\mathbf{x}_{k+1}, \mathbf{x}_k) \leq F_L^h(\mathbf{x}_k, \mathbf{x}_{k-1}) - \varepsilon_k D_h(\mathbf{x}_{k+1}, \mathbf{x}_k). \tag{18}$$

Proof. From (13), we have $f(\mathbf{x}_{k+1};\mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \leq f(\mathbf{x}_k;\mathbf{x}_k) = f(\mathbf{x}_k)$. From MAP property, we have $f(\mathbf{x}_k) \leq f(\mathbf{x}_k;\mathbf{x}_{k-1}) + \bar{L} D_h(\mathbf{x}_k,\mathbf{x}_{k-1})$. Thus, the result follows from the definition of $F_{\bar{L}}^h$ in (16). □

Proposition 22. Let Assumptions 2, 3 hold and let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a sequence generated by Model BPG. The following assertions hold:

(i) $\left\{F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)\right\}_{k\in\mathbb{N}}$ is nonincreasing and converges to a finite value.
(ii) $\sum_{k=1}^{\infty} D_h(\mathbf{x}_{k+1},\mathbf{x}_k) < \infty$ and $\{D_h(\mathbf{x}_{k+1},\mathbf{x}_k)\}_{k\in\mathbb{N}}$ converges to zero.
(iii) For any $n \in \mathbb{N}$, we have $\min_{1\leq k\leq n} D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \leq \frac{F_{\bar{L}}^h(\mathbf{x}_1,\mathbf{x}_0) - v(\mathcal{P})}{\underline{\varepsilon} n}$.

Proof. (i) Nonincreasing property follows trivially from Proposition 21 and as $\varepsilon_k > 0$. We know from Proposition 20(iii) that the Lyapunov function is lower bounded, which implies convergence of $\left\{F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)\right\}_{k\in\mathbb{N}}$ to a finite value.
(ii) Summing (18) from $k=1$ to n (a positive integer) and using $\underline{\varepsilon} \leq \varepsilon_k$ we get

$$\sum_{k=1}^{n} D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \leq \frac{1}{\underline{\varepsilon}} \left(F_{\bar{L}}^h(\mathbf{x}_1,\mathbf{x}_0) - v(\mathcal{P})\right), \tag{19}$$

since $F_{\bar{L}}^h(\mathbf{x}_{n+1},\mathbf{x}_n) \geq v(\mathcal{P})$. Taking the limit as $n \to \infty$, we obtain the first assertion, from which we deduce that $\{D_h(\mathbf{x}_{k+1},\mathbf{x}_k)\}_{k\in\mathbb{N}}$ converges to zero.
(iii) Follows from (19) and $n \min_{1\leq k\leq n}(D_h(\mathbf{x}_{k+1},\mathbf{x}_k)) \leq \sum_{k=1}^{n}(D_h(\mathbf{x}_{k+1},\mathbf{x}_k))$. □

Lemma 23 (Relative error). Let Assumptions 2, 3, 4 hold. Let the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be generated by Model BPG, then there exists a constant $C > 0$ such that for certain $k \geq 0$, we have

$$\|\partial F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)\|_- \leq C\|\mathbf{x}_{k+1} - \mathbf{x}_k\|. \tag{20}$$

Proof. As per (Rockafellar and Wets, 1998, Exercise 8.8) or (Mordukhovich, 2018, Theorem 2.19), $\partial F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)$ is given by

$$\partial F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) = \partial f(\mathbf{x}_{k+1};\mathbf{x}_k) + \bar{L}\nabla D_h(\mathbf{x}_{k+1},\mathbf{x}_k), \tag{21}$$

because the Bregman distance is continuously differentiable around $\mathbf{x}_k \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$. Using (Rockafellar and Wets, 1998, Corollary 10.11), Assumption 3(iv), and using the fact that h is $\mathcal{C}^2$ over $\operatorname{int}\operatorname{dom} h$ (cf. Assumption 3) we obtain

$$\begin{aligned}\partial F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) = \Big(&\partial_{\mathbf{x}_{k+1}} f(\mathbf{x}_{k+1};\mathbf{x}_k) + \bar{L}\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big),\\ &\partial_{\mathbf{x}_k} f(\mathbf{x}_{k+1};\mathbf{x}_k) - \bar{L}\nabla^2 h(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k)\Big).\end{aligned} \tag{22}$$

Consider the following:

$$\begin{aligned}\|\partial F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)\|_- &= \inf_{\xi\in\partial f(\mathbf{x}_{k+1};\mathbf{x}_k)} \|\xi + \bar{L}\nabla D_h(x_{k+1};x_k)\|,\\ &= \inf_{(\xi_x,\xi_y)\in\partial f(\mathbf{x}_{k+1};\mathbf{x}_k)} \|(\xi_x,\xi_y) + \bar{L}\nabla D_h(x_{k+1},x_k)\|,\\ &\leq \inf_{\xi_x\in\partial_{\mathbf{x}_{k+1}} f(\mathbf{x}_{k+1};\mathbf{x}_k)} \|(\xi_x + \bar{L}\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big))\|\\ &+ \inf_{\xi_y\in\partial_{\mathbf{x}_k} f(\mathbf{x}_{k+1};\mathbf{x}_k)} \|(\xi_y + \bar{L}\nabla^2 h(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k))\|,\end{aligned} \tag{23}$$

where in the first equality we use (21), in the second equality we use the result in (22) with $\xi := (\xi_x, \xi_y)$ such that $\xi_x \in \partial_{\mathbf{x}_{k+1}} f(\mathbf{x}_{k+1}, \mathbf{x}_k)$ and $\xi_y \in \partial_{\mathbf{x}_k} f(\mathbf{x}_{k+1}, \mathbf{x}_k)$, and in the last step we used $\nabla D_h(\mathbf{x}_{k+1}, \mathbf{x}_k) = (\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k), \nabla^2 h(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k))$. The optimality of $\mathbf{x}_{k+1}$ in (13) implies the existence of $\xi^{k+1}_{\mathbf{x}_{k+1}} \in \partial_{\mathbf{x}_{k+1}} f(\mathbf{x}_{k+1}; \mathbf{x}_k)$ such that the following condition holds: $\xi^{k+1}_{\mathbf{x}_{k+1}} + \frac{1}{\tau_k}(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)) = \mathbf{0}$. Therefore, the first block coordinate in (22) satisfies

$$\xi^{k+1}_{\mathbf{x}_{k+1}} + \bar{L}\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big) = \varepsilon_k\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big)\,. \tag{24}$$

Now consider the first term of the right hand side in (23). We have

$$\begin{aligned}
&\inf_{\xi_x \in \partial_{\mathbf{x}_{k+1}} f(\mathbf{x}_{k+1}; \mathbf{x}_k)} \|(\xi_x + \bar{L}\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big))\| \\
&\leq \|\xi^{k+1}_{\mathbf{x}_{k+1}} + \bar{L}\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big)\|\,, \\
&\leq \varepsilon_k \|\big(\nabla h(\mathbf{x}_{k+1}) - \nabla h(\mathbf{x}_k)\big)\| \leq \varepsilon_k \tilde{L}_h \|\mathbf{x}_{k+1} - \mathbf{x}_k\|\,,
\end{aligned}$$

where in the second step we used (24) and in the last step we applied mean value theorem along with the fact that the entity $\|\nabla^2 h(\mathbf{x}_{k+1} + s(\mathbf{x}_{k+1} - \mathbf{x}_k))\|$ is bounded by a constant $\tilde{L}_h > 0$ for certain $s \in [0, 1]$, due to Assumption 4(ii). Considering the second term of the right hand side in (23), we have

$$\begin{aligned}
&\inf_{\xi_y \in \partial_{\mathbf{x}_k} f(\mathbf{x}_{k+1}; \mathbf{x}_k)} \|(\xi_y + \bar{L}\nabla^2 h(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k))\| \\
&\leq \inf_{\xi_y \in \partial_{\mathbf{x}_k} f(\mathbf{x}_{k+1}; \mathbf{x}_k)} \|\xi_y\| + \|\bar{L}\nabla^2 h(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k)\|\,, \\
&\leq c\|\mathbf{x}_{k+1} - \mathbf{x}_k\| + \bar{L}L_h\|(\mathbf{x}_{k+1} - \mathbf{x}_k)\|\,,
\end{aligned}$$

where in the last step we used Assumption 4(i) and the fact that $\|\nabla^2 h(\mathbf{x}_k)\|$ is bounded by L_h. The result follows from combining the results obtained for (24). □

We now consider results on generic limit points and show that stationarity can indeed be attained for iterates produced by Model BPG.

Proposition 24. For a bounded sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ such that $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \to 0$ as $k \to \infty$, the following holds:

(i) $\omega(\mathbf{x}_0)$ is connected and compact,
(ii) $\lim_{k\to\infty} \operatorname{dist}(\mathbf{x}_k, \omega(\mathbf{x}_0)) = 0$.

The proof relies on the same technique as the proof of (Bolte et al., 2014, Lemma 3.5) (also see (Bolte et al., 2014, Remark 3.3)). We now show that the sequence generated by Model BPG $(\mathbf{x}_k)_{k\in\mathbb{N}}$ indeed attains $\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \to 0$ as $k \to \infty$, which in turn enables the application of Proposition 24 to deduce the properties of the sequence generated by Model BPG crucial for the proof of global convergence.

Proposition 25. Let Assumption 2, 3, 4 hold. Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a sequence generated by Model BPG. Then, $\mathbf{x}_{k+1} - \mathbf{x}_k \to 0$ as $k \to \infty$.

Proof. The result follows as a simple consequence of Proposition 22(ii) along with Assumption 4(iii).

□

Analyzing the full set of limit points of the sequence generated by Model BPG is difficult, as illustrated in Ochs et al. (2019). Obtaining the global convergence is still an open problem. Moreover, the work in Ochs et al. (2019) relies on convex model functions. In order to simplify slightly the setting, we restrict the set of limit points to the set $\operatorname{int}\operatorname{dom} h$. Such a choice may appear to be restrictive, however, Model BPG when applied to many practical problems results in sequences that have this property as illustrated in Section 5. The subset of $F_{\bar{L}}^h$-attentive (similar to f-attentive) limit points is

$$\omega_{F_{\bar{L}}^h}(\mathbf{x}_0) := \left\{(\mathbf{y},\mathbf{x}) \in \mathbb{R}^N \times \mathbb{R}^N \,|\, \exists K \subset \mathbb{N}\colon (\mathbf{x}_k, F_{\bar{L}}^h(\mathbf{x}_k,\mathbf{x}_{k-1})) \overset{K}{\to} (\mathbf{x}, F_{\bar{L}}^h(\mathbf{y},\mathbf{x}))\right\}.$$

Also, we define $\omega_{F_{\bar{L}}^h}^{(\operatorname{int}\operatorname{dom} h)^2} := \omega_{F_{\bar{L}}^h} \cap (\operatorname{int}\operatorname{dom} h \times \operatorname{int}\operatorname{dom} h)$.

Proposition 26. Let Assumptions 2, 3, 4 hold. Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a sequence generated by Model BPG. Then, the following holds:

(*i*) $\omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) = \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$,

(*ii*) $\mathbf{x} \in \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$ if and only if $(\mathbf{x},\mathbf{x}) \in \omega_{F_{\bar{L}}^h}^{(\operatorname{int}\operatorname{dom} h)^2}(\mathbf{x}_0)$.

(*iii*) $F_{\bar{L}}^h$ is constant and finite on $\omega_{F_{\bar{L}}^h}^{(\operatorname{int}\operatorname{dom} h)^2}(\mathbf{x}_0)$ and f is constant and finite on $\omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$ with same value.

Proof. (*i*) We show the inclusion $\omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) \subset \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$ and $\omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) \subset \omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$ is clear by definition. Let $\mathbf{x}^\star \in \omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$, then we obtain

$$\begin{aligned} f(\mathbf{x}^\star) + \left(\underline{L} + \frac{1}{\tau_k}\right) D_h(\mathbf{x}^\star,\mathbf{x}_k) &\overset{(12)}{\geq} f(\mathbf{x}^\star;\mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{x}^\star,\mathbf{x}_k) \\ &\overset{(13)}{\geq} f(\mathbf{x}_{k+1};\mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \\ &\overset{(12)}{\geq} f(\mathbf{x}_{k+1}) - \Big(\bar{L} - \frac{1}{\tau_k}\Big) D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{\varepsilon_k>0}{\geq} f(\mathbf{x}_{k+1}). \end{aligned}$$

By Assumption 4(iii) combined with the fact that $\mathbf{x}_k \overset{K}{\to} \mathbf{x}^\star$, we have $D_h(\mathbf{x}^\star,\mathbf{x}_k) \to 0$ as $k \overset{K}{\to} \infty$, which, together with the lower semicontinuity of f, implies the following: $f(\mathbf{x}^\star) \geq \liminf_{k\overset{K}{\to}\infty} f(\mathbf{x}_{k+1}) \geq f(\mathbf{x}^\star)$, thus $\mathbf{x}^\star \in \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$.

(*ii*) If $\mathbf{x} \in \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$, then we have $\mathbf{x}_k \overset{K}{\to} \mathbf{x}$ for $K \subset \mathbb{N}$, and $f(\mathbf{x}_k) \overset{K}{\to} f(\mathbf{x})$. As a consequence of Proposition 22 and Assumption 4(iii), $D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \to 0$ as $k \to \infty$, which implies that $\mathbf{x}_{k+1} \overset{K}{\to} \mathbf{x}$. The first part of the proof implies $f(\mathbf{x}_{k+1}) \overset{K}{\to} f(\mathbf{x})$. We also have $F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{K}{\to} f(\mathbf{x})$ which we prove below, which implies that $(\mathbf{x},\mathbf{x}) \in \omega_{F_{\bar{L}}^h}^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$. Note that by definition of $F_{\bar{L}}^h$ we have

$$\begin{aligned} F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) &= f(\mathbf{x}_{k+1};\mathbf{x}_k) + \bar{L} D_h(\mathbf{x}_{k+1},\mathbf{x}_k), \\ &= f(\mathbf{x}_{k+1}) + (f(\mathbf{x}_{k+1};\mathbf{x}_k) - f(\mathbf{x}_{k+1})) + \bar{L} D_h(\mathbf{x}_{k+1},\mathbf{x}_k). \end{aligned}$$

MAP property gives $f(\mathbf{x}_{k+1}) \leq F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) \leq f(\mathbf{x}_{k+1}) + (\bar{L}+\underline{L}) D_h(\mathbf{x}_{k+1},\mathbf{x}_k)$. Thus, we have that $F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{K}{\to} f(\mathbf{x})$ as $D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{K}{\to} 0$. Conversely,

suppose $(\mathbf{x},\mathbf{x}) \in \omega_{F_{\bar{L}}^h}^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$ and $\mathbf{x}_k \overset{K}{\to} \mathbf{x}$ for $K \subset \mathbb{N}$. This, together with $D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \to 0$ as $k \overset{K}{\to} \infty$, induces $F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{K}{\to} f(\mathbf{x})$, which further implies $f(\mathbf{x}_{k+1}) \overset{K}{\to} f(\mathbf{x})$ due to the following. Note that we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) &= F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) + (f(\mathbf{x}_{k+1}) - f(\mathbf{x}_{k+1};\mathbf{x}_k)) + \bar{L}D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \\ &\geq F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) + (\bar{L}-\underline{L})D_h(\mathbf{x}_{k+1},\mathbf{x}_k)\,. \end{aligned}$$

Finally we have $F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) + (\bar{L}-\underline{L})D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \leq f(\mathbf{x}_{k+1}) \leq F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)$. Thus, with $D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \to 0$ as $k \overset{K}{\to} \infty$ and $F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{K}{\to} f(\mathbf{x})$, we deduce that $f(\mathbf{x}_{k+1}) \overset{K}{\to} f(\mathbf{x})$. And therefore $\mathbf{x} \in \omega_f^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0)$.

(*iii*) By Proposition 21, the sequence $(F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k))_{k\in\mathbb{N}}$ converges to a finite value $\underline{F}$. Note that $D_h(\mathbf{x}_{k+1},\mathbf{x}_k) \to 0$ as $k \overset{K}{\to} \infty$ due to Proposition 22 (ii), when combined with Assumption 4(iii) implies that $\|\mathbf{x}_{k+1}-\mathbf{x}_k\| \to 0$. For $(\mathbf{x}^\star,\mathbf{x}^\star) \in \omega_{F_{\bar{L}}^h}^{(\operatorname{int}\operatorname{dom} h)^2}(\mathbf{x}_0,\mathbf{x}_0)$ there exists $K \subset \mathbb{N}$ such that $\mathbf{x}_k \overset{K}{\to} \mathbf{x}^\star$ and $F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k) \overset{K}{\to} F_{\bar{L}}^h(\mathbf{x}^\star,\mathbf{x}^\star) = f(\mathbf{x}^\star)$, i.e., the value of the limit point is independent of the choice of the subsequence. The result follows directly and by using (*i*). □

The following result states that $F_{\bar{L}}^h$-attentive sequences converge to a critical point.

Theorem 27 (Sub-sequential convergence). Let Assumptions 2, 3, 4 hold. If the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is generated by Model BPG, then

$$\omega_{F_{\bar{L}}^h}^{(\operatorname{int}\operatorname{dom} h)^2}(\mathbf{x}_0) \subset \operatorname{crit}(F_{\bar{L}}^h)\,. \tag{25}$$

Proof. From (20), we have $\|\partial F_{\bar{L}}^h(\mathbf{x}_{k+1},\mathbf{x}_k)\|_- \leq C\|\mathbf{x}_{k+1}-\mathbf{x}_k\|$ for some constant $C > 0$. Using $\|\mathbf{x}_{k+1}-\mathbf{x}_k\| \to 0$, convergence of $(\tau_k)_{k\in\mathbb{N}}$, and Proposition 26(*i*) yields (25), by the closedness property of the limiting subdifferential (8). □

Discussion. Subsequential convergence to a stationary point was already considered in few works. In particular, the work in Drusvyatskiy et al. (2019) already provides such a result, however, it relies on certain abstract assumptions. Even though such assumptions are valid for some practical algorithms, the authors do not consider a concrete algorithm. Moreover, their abstract update step depends on the minimization of the model function, which can require additional regularity conditions on the problem. For example, if the model function is linear, then the domain must be compact to guarantee the existence of a solution. A related line-search variant of Model BPG was considered in Ochs et al. (2019), for which subsequential convergence to a stationarity point was proven. The subsequential convergence results in Ochs et al. (2019) are more general than our work, as they analyse the behavior of limit points in $\operatorname{dom} h$, $\operatorname{cl}\operatorname{dom} h$, $\operatorname{int}\operatorname{dom} h$ (cf. (Ochs et al., 2019, Theorem 22)). Our analysis is restricted to limit points in $\operatorname{int}\operatorname{dom} h$, as typically such an assumption holds in practice (see Section 5). Though subsequential convergence is satisfactory, proving global convergence is nontrivial, in general.

Lemma 28. Let Assumptions 2, 3, 4, 5 hold. Then, the Lyapunov function $F_{\bar{L}}^h$ is definable in $\mathcal{O}$, and satisfies KL property at any point of $\operatorname{dom}\partial F_{\bar{L}}^h$.

The proof is straightforward application of (Ochs, 2015, Corollary 4.32) and (Bolte et al., 2007, Theorem 14). For additive composite problems, the global convergence analysis of BPG based methods Bolte et al. (2018); Mukkamala et al. (2020) relies on strong convexity of h. However, in our setting we relax such a requirement on h, via Assumption 6. Note that imposing such an assumption is weaker than imposing the strong convexity of h, as we only need the strong convexity property to hold over a compact convex set. Such a property can be satisfied even if h is not strongly convex, for example, Burg's entropy (see Section 5.2). We now present the proof of Theorem 17, result pertaining to the global convergence of the sequence generated by Model BPG.

Proof of Theorem 17. Note that the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by Model BPG is a bounded sequence (see Remark 14). The proof relies on Theorem 2, for which we need to verify the conditions (H1)–(H5). Due to Lemma 28, F_L^h satisfies Kurdyka–Łojasiewicz property at each point of $\operatorname{dom}\partial F_L^h$. Note that as $\omega^{\operatorname{int}\operatorname{dom} h}(\mathbf{x}_0) = \omega(\mathbf{x}_0)$ holds true, there exists a sufficiently small $\varepsilon > 0$ such that $\tilde{B} := \{\mathbf{x} : \operatorname{dist}(\mathbf{x}, \omega(\mathbf{x}_0)) \leq \varepsilon\} \subset \operatorname{int}\operatorname{dom} h$. As $\omega(\mathbf{x}_0)$ is compact due to Proposition 24(i), the set $\tilde{B}$ is also compact. Moreover, the convex hull of the set $\tilde{B}$ denoted by $B := \operatorname{conv}\tilde{B}$ is also compact, as the convex hull of a compact set is also compact in finite dimensional setting. A simple calculation reveals that the set B lies in the set $\operatorname{int}\operatorname{dom} h$. Thus, due to Proposition 25 along with Proposition 24(ii), without loss of generality, we assume that the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by Model BPG lies in the set B. By definition of σ_B as per Assumption 6 we have $D_h(\mathbf{x}_{k+1}, \mathbf{x}_k) \geq \frac{\sigma_B}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$, through which we obtain

$$F_L^h(\mathbf{x}_{k+1}, \mathbf{x}_k) \leq F_L^h(\mathbf{x}_k, \mathbf{x}_{k-1}) - \frac{\varepsilon_k \sigma_B}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2,$$

which is (H1) with $d_k = \frac{\varepsilon_k\sigma_B}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2$ and $a_k = 1$. We also have existence of $\mathbf{w}_{k+1} \in \partial F_L^h(\mathbf{x}_{k+1}, \mathbf{x}_k)$ such that the conclusion of Lemma 23 holds true for some $C > 0$, which is (H2) with $b = C$, since the coefficients for both Euclidean distances are bounded from above. The continuity condition (H3) is deduced from a converging subsequence, whose existence is guaranteed by boundedness of $(\mathbf{x}_k)_{k\in\mathbb{N}}$, and Proposition 26 guarantees that such convergent subsequences are F_L^h-attentive convergent. The distance condition (H4) holds trivially as $\varepsilon_k > 0$ and $\sigma_B > 0$. The parameter condition (H5), holds because $b_n = 1$ in this setting, hence $(b_n)_{n\in\mathbb{N}} \notin \ell_1$ and also we have $\sup_{n\in\mathbb{N}} \frac{1}{b_n a_n} = 1 < \infty\,, \quad \inf_n a_n = 1 > 0\,$. Theorem 2 implies the finite length property from which we deduce that the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by Model BPG converges to a single point, which we denote by $\mathbf{x}$. As $(\mathbf{x}_{k+1})_{k\in\mathbb{N}}$ also converges to $\mathbf{x}$, the sequence $((\mathbf{x}_{k+1}, \mathbf{x}_k))_{k\in\mathbb{N}}$ converges to $(\mathbf{x}, \mathbf{x})$, which is a critical point of F_L^h due to Theorem 27. □

The global convergence result in Theorem 17 shows that Model BPG converges to a point, which in turn can be used to represent a critical point of the Lyapunov function. However, our goal is to find a critical point of the objective function f. Firstly, we need the following result, which establishes the connection between fixed points of the update mapping and critical points of f.

Lemma 29. Let Assumptions 2, 3 hold. For any $0 < \tau < (1/\bar{L})$ and $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h$, the fixed points of the update mapping $T_\tau(\bar{\mathbf{x}})$ are critical points of f.

Proof. Let $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$ be a fixed point of T_τ, in the sense the condition $\bar{\mathbf{x}} \in T_\tau(\bar{\mathbf{x}})$ holds true. By definition of $T_\tau(\bar{\mathbf{x}})$, the following condition holds true: $\mathbf{0} \in \partial f(\mathbf{x}; \bar{\mathbf{x}}) + \frac{1}{\tau}(\nabla h(\mathbf{x}) - \nabla h(\bar{\mathbf{x}}))$ at $\mathbf{x} = \bar{\mathbf{x}}$, which implies that $\mathbf{0} \in \partial f(\bar{\mathbf{x}}; \bar{\mathbf{x}})$. We know that $\partial f(\bar{\mathbf{x}}; \bar{\mathbf{x}}) \subset \partial f(\bar{\mathbf{x}})$, thus $\bar{\mathbf{x}}$ is a critical point of the function f. □

We also require the following technical result. The following lemma proves the sequential closedness property of the update mapping.

Lemma 30 (Continuity property). Let Assumptions 2, 3, 4 hold. Let the sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be bounded such that $\mathbf{x}_k \to \bar{\mathbf{x}}$, where $\mathbf{x}_k \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$ $\forall\, k \in \mathbb{N}$, and $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$. Let $\tau_k \to \tau$, such that $0 < \underline{\tau} \leq \tau_k \leq \bar{\tau} < 1/\bar{L}$. Assume that there exists a bounded set $B \subset \operatorname{int} \operatorname{dom} h$, such that $T_{\tau_k}(\mathbf{x}_k) \subset B$, $\mathbf{x}_k \in B$, $\forall k \in \mathbb{N}$. If $\limsup_{k\to\infty} T_{\tau_k}(\mathbf{x}_k) \subset \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$, then $\limsup_{k\to\infty} T_{\tau_k}(\mathbf{x}_k) \subset T_\tau(\bar{\mathbf{x}})$.

Proof. Consider any sequence $(\mathbf{y}_k)_{k\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$, the condition $\mathbf{y}_k \in T_{\tau_k}(\mathbf{x}_k)$ holds true. Recall that $f(\mathbf{x}; \mathbf{y})$ is continuous on its domain due to Assumption 3(iv). By optimality of $\mathbf{y}_k \in T_{\tau_k}(\mathbf{x}_k)$, for any $\mathbf{z} \in \mathbb{R}^N$ we have

$$f(\mathbf{y}_k; \mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{y}_k, \mathbf{x}_k) \leq f(\mathbf{z}; \mathbf{x}_k) + \frac{1}{\tau_k} D_h(\mathbf{z}, \mathbf{x}_k)\,. \tag{26}$$

As a consequence of boundedness of the sequence $(\mathbf{y}_k)_{k\in\mathbb{N}}$, by Bolzano–Weierstrass Theorem there exists a convergent subsequence. Let $\mathbf{y}_k \overset{K}{\to} \pi$ such that $\pi \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$. Note that $\tau_k \overset{K}{\to} \tau$ for some $K \subset \mathbb{N}$. Applying limit on both sides of (26) using the continuity of the model function and the Bregman distance gives

$$f(\pi; \bar{\mathbf{x}}) + \frac{1}{\tau} D_h(\pi, \bar{\mathbf{x}}) \leq f(\mathbf{z}; \bar{\mathbf{x}}) + \frac{1}{\tau} D_h(\mathbf{z}, \bar{\mathbf{x}})\,, \quad \forall\, \mathbf{z} \in \operatorname{dom} f \cap \operatorname{dom} h\,, \tag{27}$$

which implies that π minimizes the function $f(\cdot; \bar{\mathbf{x}}) + \frac{1}{\tau} D_h(\cdot, \bar{\mathbf{x}})$. This implies that $\pi \in T_\tau(\bar{\mathbf{x}})$ and the result follows. □

We now provide the proof of Theorem 18, that states that the sequence generated by Model BPG indeed converges to a critical point of the objective function.

Proof of Theorem 18. The sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by Model BPG under the assumptions as in Theorem 17 is globally convergent, thus let $\mathbf{x}_k \to \mathbf{x}$ and also $\mathbf{x}_{k+1} \to \mathbf{x}$. As $\mathbf{x}_{k+1} \in T_{\tau_k}(\mathbf{x}_k)$ and τ_k converges to τ, with Lemma 30 we deduce that $\mathbf{x} \in T_\tau(\mathbf{x})$. Additionally, with the result in Lemma 30, we deduce that $\mathbf{x}$ is the fixed point of the mapping $T_\tau(\mathbf{x})$, i.e., $\mathbf{x} \in T_\tau(\mathbf{x})$. Then, using Lemma 29 we conclude that $\mathbf{x}$ is a critical point of the function f. □

4 Examples

In this section we consider special instances of $(\mathcal{P})$, namely, additive composite problems and a broad class of composite problems. The goal is to quantify assumptions for these problems such that the global convergence result (Theorem 18) of Model BPG is applicable. We enforce the following blanket assumptions.

(B1) The function h is a Legendre function that is $\mathcal{C}^2$ over $\operatorname{int}\operatorname{dom} h$. For any compact convex set $B \subset \operatorname{int}\operatorname{dom} h$, there exists $\sigma_B > 0$ such that h is σ_B-strongly convex with bounded second derivative on B. Moreover, for bounded $(\mathbf{u}_k)_{k\in\mathbb{N}}$, $(\mathbf{v}_k)_{k\in\mathbb{N}}$ in $\operatorname{int}\operatorname{dom} h$, the following holds as $k \to \infty$:

$$D_h(\mathbf{u}_k, \mathbf{v}_k) \to 0 \iff \|\mathbf{u}_k - \mathbf{v}_k\| \to 0\,.$$

(B2) The function f is coercive and additionally the conditions $\operatorname{dom} f \cap \operatorname{int}\operatorname{dom} h \neq \emptyset$, $\operatorname{crit} f \cap \operatorname{int}\operatorname{dom} h \neq \emptyset$, $\operatorname{dom} f \subset \operatorname{cl}\operatorname{dom} h$ hold true.

(B3) The functions $\tilde{f} : \mathbb{R}^N \times \mathbb{R}^N \to \overline{\mathbb{R}}$, $(\mathbf{x}, \bar{\mathbf{x}}) \mapsto f(\mathbf{x}; \bar{\mathbf{x}})$ with $\operatorname{dom}\tilde{f} := \operatorname{dom} f \times \operatorname{dom} f$, and $\tilde{h} : \mathbb{R}^N \times \mathbb{R}^N \to \overline{\mathbb{R}}$, $(\mathbf{x}, \bar{\mathbf{x}}) \mapsto h(\bar{\mathbf{x}}) + \langle \nabla h(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}}\rangle$ with $\operatorname{dom}\tilde{h} := \operatorname{dom} h \times \operatorname{int}\operatorname{dom} h$ are definable in an o-minimal structure $\mathcal{O}$.

4.1 Additive composite problems

We consider the following nonconvex additive composite problem:

$$\inf_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x})\,, \quad f(\mathbf{x}) := f_0(\mathbf{x}) + f_1(\mathbf{x})\,, \tag{28}$$

which is a special case of $(\mathcal{P})$. Additive composite problems arise in several applications, such as standard phase retrieval Bolte et al. (2018), low rank matrix factorization Mukkamala and Ochs (2019), deep linear neural networks Mukkamala et al. (2019), and many more. We present below the BPG algorithm, a specialization of Model BPG that is applicable for additive composite problems.

BPG is Model BPG (Algorithm 1) with

$$f(\mathbf{x}; \mathbf{x}_k) := f_0(\mathbf{x}) + f_1(\mathbf{x}_k) + \langle \nabla f_1(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle\ . \tag{29}$$

We impose the following conditions that are common in the analysis of forward–backward algorithms Ochs et al. (2014), which are used to optimize additive composite problems.

(C1) $f_0 : \mathbb{R}^N \to \overline{\mathbb{R}}$ is a proper, lsc function and is regular at any $\mathbf{x} \in \operatorname{dom} f_0$ and

$$\partial^\infty f_0(\mathbf{x}) \cap (-N_{\operatorname{dom} h}(\mathbf{x})) = \{\mathbf{0}\}\,, \quad \forall\, \mathbf{x} \in \operatorname{dom} f_0 \cap \operatorname{dom} h\,. \tag{30}$$

(C2) $f_1 : \mathbb{R}^N \to \overline{\mathbb{R}}$ is a proper, lsc function and is $\mathcal{C}^2$ on an open set that contains $\operatorname{dom} f_0$. Also, there exist $\bar{L}, \underline{L} > 0$ such that for any $\bar{\mathbf{x}} \in \operatorname{dom} f_0 \cap \operatorname{int}\operatorname{dom} h$, the following holds:

$$-\underline{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \leq f_1(\mathbf{x}) - f_1(\bar{\mathbf{x}}) - \langle \nabla f_1(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}}\rangle \leq \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}})\,, \tag{31}$$

for all $\mathbf{x} \in \operatorname{dom} f_0 \cap \operatorname{dom} h$.

Note that with Assumption (C1), (C2) it is easy to deduce that $\operatorname{dom} f_0 = \operatorname{dom} f$. For $\bar{\mathbf{x}} \in \operatorname{dom} f$, the model function $f(\cdot; \bar{\mathbf{x}}) : \mathbb{R}^N \to \overline{\mathbb{R}}$ at $\mathbf{x} \in \operatorname{dom} f$ is given by

$$f(\mathbf{x}; \bar{\mathbf{x}}) := f_0(\mathbf{x}) + f_1(\bar{\mathbf{x}}) + \langle \nabla f_1(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}}\rangle\ . \tag{32}$$

Using the model function in (32) and the condition (31), we deduce that there exist $\underline{L}, \bar{L} > 0$ such that for any $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$, MAP property is satisfied at $\bar{\mathbf{x}}$ with $\underline{L}, \bar{L}$ as the following holds true:

$$-\underline{L} D_h(\mathbf{x}, \bar{\mathbf{x}}) \leq f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}}) \leq \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}}), \quad \forall\, \mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h, \tag{33}$$

as $f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}}) := f_1(\mathbf{x}) - f_1(\bar{\mathbf{x}}) - \langle \nabla f_1(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle$, thus satisfying Assumption 3(i). The condition in (33) is similar to the popular L-smad property in Bolte et al. (2018). The main addition is that $\mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h$ and $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$, whereas the L-smad property requires $\mathbf{x}, \bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$.

Remark. Consider $f_1(x) := \frac{1}{2}x^2$, $f_0(x) := \delta_{[0,\infty)}(x)$ and $h(x) = x \log(x)$ with $\operatorname{dom} h = [0, \infty)$ under $0 \log(0) = 0$. Clearly, $\operatorname{dom} h \subset \operatorname{dom} f_1$ and $\operatorname{dom} f \subset \operatorname{dom} h$ hold true. The function f_1 is differentiable at $x = 0$, and MAP condition in (31) holds true for $x = 0$. This scenario is not considered in the L-smad property.

It is straightforward to verify that Assumptions (C1), (C2), (B1), (B2), (B3) imply Assumptions 2, 3, 4, 5, 6. Thus, due to Theorem 18, the sequence generated by BPG globally converges to a critical point of the function.

4.2 Composite problems

We consider the following nonconvex composite problem:

$$\inf_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}), \quad f(\mathbf{x}) := f_0(\mathbf{x}) + g(F(\mathbf{x})), \tag{34}$$

which is a special case of the problem $(\mathcal{P})$. Composite problems arise in robust phase retrieval, robust PCA, censored $\mathbb{Z}_2$ synchronization Drusvyatskiy (2017); Lewis and Wright (2016); Nesterov (2007); Drusvyatskiy and Lewis (2018); Drusvyatskiy and Paquette (2019). We present below Prox-Linear BPG, specialization of Model BPG that is applicable for generic composite problems.

Prox-Linear BPG is Model BPG (Algorithm 1) with

$$f(\mathbf{x}; \mathbf{x}_k) := f_0(\mathbf{x}) + g(F(\mathbf{x}_k) + \nabla F(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)). \tag{35}$$

We require the following conditions.

(D1) $f_0 : \mathbb{R}^N \to \overline{\mathbb{R}}$ is a proper, lsc function and is regular at any $\mathbf{x} \in \operatorname{dom} f_0$ and:

$$\partial^\infty f_0(\mathbf{x}) \cap (-N_{\operatorname{dom} h}(\mathbf{x})) = \{\mathbf{0}\}, \quad \forall\, \mathbf{x} \in \operatorname{dom} f_0 \cap \operatorname{dom} h. \tag{36}$$

(D2) $g : \mathbb{R}^M \to \mathbb{R}$ is a Q-Lipschitz continuous and regular function. There exists $P > 0$ such that at any $\mathbf{x} \in \mathbb{R}^M$, the following holds:

$$\sup_{\mathbf{v} \in \partial g(\mathbf{x})} \|\mathbf{v}\| \leq P. \tag{37}$$

(D3) $F : \mathbb{R}^N \to \mathbb{R}^M$ is $\mathcal{C}^2$ over $\mathbb{R}^N$ and there exist $L > 0$ such that for any $\bar{\mathbf{x}} \in \operatorname{dom} f_0 \cap \operatorname{int} \operatorname{dom} h$, the following condition holds true:

$$\|F(\mathbf{x}) - F(\bar{\mathbf{x}}) - \nabla F(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})\| \leq L D_h(\mathbf{x}, \bar{\mathbf{x}}), \quad \forall\, \mathbf{x} \in \operatorname{dom} f_0 \cap \operatorname{dom} h,$$

where $\nabla F(\bar{\mathbf{x}})$ is the Jacobian of F at $\bar{\mathbf{x}}$.

The properties (D1), (D2), (D3) along with (B2) imply proper, lsc property and lower-boundedness of f, thus satisfying Assumption 2. Note that with Assumption (D1), (D2), (D3) it is easy to deduce that $\operatorname{dom} f_0 = \operatorname{dom} f$. Let $\bar{\mathbf{x}} \in \operatorname{dom} f$ and the model function with $\bar{\mathbf{x}}$ as model center evaluated at $\mathbf{x} \in \operatorname{dom} f$ is given by:

$$f(\mathbf{x}; \bar{\mathbf{x}}) = f_0(\mathbf{x}) + g(F(\bar{\mathbf{x}}) + \nabla F(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}))\,. \tag{38}$$

Using (D2), (D3) we deduce that there exists $\bar{L} := LQ > 0$ such that for any $\bar{\mathbf{x}} \in \operatorname{dom} f \cap \operatorname{int} \operatorname{dom} h$, the following MAP property holds at $\bar{\mathbf{x}}$ with $\bar{L}$:

$$|f(\mathbf{x}) - f(\mathbf{x}; \bar{\mathbf{x}})| = |g(F(\mathbf{x})) - g(F(\bar{\mathbf{x}}) + \nabla F(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}))| \leq \bar{L} D_h(\mathbf{x}, \bar{\mathbf{x}})\,,$$

for all $\mathbf{x} \in \operatorname{dom} f \cap \operatorname{dom} h$, as g is Q-Lipschitz continuous and (D3) holds true. Thus, Assumption 3(i) is satisfied with $\bar{L} = \underline{L} = LQ$.

It is straightforward to verify that Assumptions (D1), (D2), (D3), (B1), (B2), (B3) imply Assumptions 2, 3, 4, 5, 6. Thus, due to Theorem 18, the sequence generated by Prox-Linear BPG globally converges to a critical point of the function.

5 Experiments

For the purpose of empirical evaluation we consider standard phase retrieval problems and Poisson linear inverse problems. We compare our algorithms with Inexact Bregman Proximal Minimization Line Search (IBPM-LS) Ochs et al. (2013), which is a popular algorithm to solve generic nonsmooth nonconvex problems. Before we provide the empirical results, we comment below on a variant of Model BPG based on the backtracking technique, which we used in the experiments.

Model BPG with backtracking. It is possible that the value of $\bar{L}$ in the MAP property is unknown. This issue can be solved by using a backtracking technique, where in each iteration a local constant $\bar{L}_k$ is found such that the following holds:

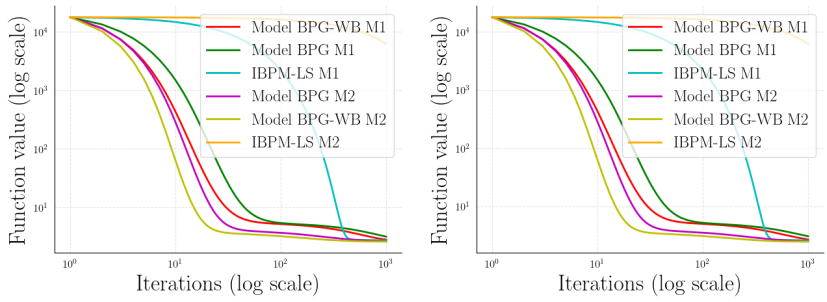
$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_{k+1}; \mathbf{x}_k) + \bar{L}_k D_h(\mathbf{x}_{k+1}, \mathbf{x}_k)\,. \tag{39}$$

The value of $\bar{L}_k$ is found by taking an initial guess $\bar{L}_k^0$. If the condition (39) fails to hold, then with a scaling parameter $\nu > 1$, we set $\bar{L}_k$ to the smallest value in the set $\{\nu \bar{L}_k^0, \nu^2 \bar{L}_k^0, \nu^3 \bar{L}_k^0, \ldots\}$ such that (39) holds true. Enforcing $\bar{L}_k \geq \bar{L}_{k-1}$ for $k \geq 1$ ensures that after finite number of iterations there is no change in the value of $\bar{L}_k$, which takes us to the situation that we analyzed in the paper. The condition $\bar{L}_k \geq \bar{L}_{k-1}$ can be enforced by choosing $\bar{L}_k^0 = \bar{L}_{k-1}$.

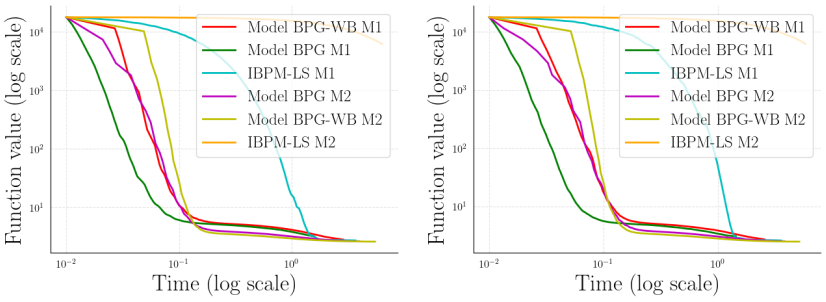
Code. The code is open sourced at the following link: `https://github.com/mmahesh/composite-optimization-code`. It contains the implementation of the algorithms, the random synthetic datasets generation process, the choices for hyper-parameters, the plots generation process and all the other related details.

5.1 Standard phase retrieval

The phase retrieval problem involves approximately solving a system of quadratic equations. Let $b_i \in \mathbb{R}$ and $\mathbf{A}_i \in \mathbb{R}^{N \times N}$ be a symmetric positive semi-definite matrix, for all $i = 1, \ldots, M$. The goal of standard phase retrieval problem is to



(a) ℓ_1 reg **(b)** Squared ℓ_2 reg



(c) ℓ_1 reg **(d)** Squared ℓ_2 reg

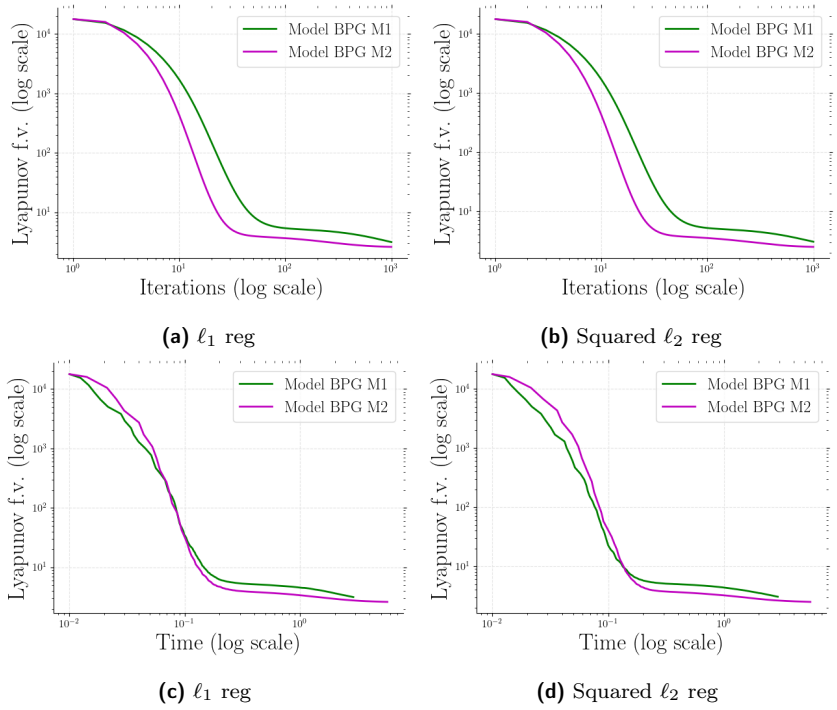
Fig. 1: In this experiment we compare the performance of Model BPG, Model BPG with Backtracking (denoted as Model BPG-WB), and IBPM-LS Ochs et al. (2013) on standard phase retrieval problems, with both ℓ_1 and squared ℓ_2 regularization. For this purpose, we consider M1 model function as in (41) without absolute sign (which is the same setting as Bolte et al. (2018)), and with M2 model function as in (44). Model BPG with M2 (44) is faster in both the settings and Model BPG variants perform significantly better than IBPM-LS. By *reg*, we mean *regularization*.

find $\mathbf{x} \in \mathbb{R}^N$ such that the following system of quadratic equations is satisfied: $\mathbf{x}^T\mathbf{A}_i\mathbf{x} \approx b_i$, for $i = 1, \ldots, M$, where b_i's are measurements and $\mathbf{A}_i$'s are so-called sampling matrices. In the context of Bregman proximal algorithms, regarding the phase retrieval problem, we refer the reader to Bolte et al. (2018); Mukkamala et al. (2020). Further references regarding the phase retrieval problem include Candes et al. (2015); Wang et al. (2018); Luke (2017). The standard technique to solve such system of quadratic equations is to solve the following optimization problem:

$$\min_{\mathbf{x}\in\mathbb{R}^N} \mathcal{P}_0(\mathbf{x}), \quad \mathcal{P}_0(\mathbf{x}) := \frac{1}{M}\sum_{i=1}^{M} (\mathbf{x}^T\mathbf{A}_i\mathbf{x} - b_i)^2 + \mathcal{R}(\mathbf{x}), \tag{40}$$

where $\mathcal{R}(\mathbf{x})$ is the regularization term. We use ℓ_1 regularization with $\mathcal{R}(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ and squared ℓ_2 regularization with $\mathcal{R}(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|^2$, with some $\lambda > 0$. We consider two model functions in order to solve the problem in (40).

Model 1. Here, the analysis falls under the category of additive composite problems (Section 4.1), where we set $f_0(\mathbf{x}) := \mathcal{R}(\mathbf{x})$, and $f_1(\mathbf{x}) := \frac{1}{M}\sum_{i=1}^{M} (\mathbf{x}^T\mathbf{A}_i\mathbf{x} - b_i)^2$.



(a) ℓ_1 reg **(b)** Squared ℓ_2 reg

(c) ℓ_1 reg **(d)** Squared ℓ_2 reg

Fig. 2: We illustrate that when Model BPG applied to standard phase retrieval problem in (40), with model function chosen to be either Model 1 in (41) or Model 2 in (44), result in sequences where the Lyapunov function value evaluations are monotonically nonincreasing. In terms of iterations, Model BPG with Model 2 (Model BPG M2) is better than Model BPG with Model 1 (Model BPG M1). In terms of time, Model BPG M1 and Model BPG M2 perform almost the same, however, towards the end Model BPG M2 is faster in both the cases. By *reg* we mean *regularization*, and by *Lyapunov f.v.* we mean Lyapunov function values.

Consider the standard model for additive composite problems Bolte et al. (2018), where around $\mathbf{y} \in \mathbb{R}^N$, the model function $\mathcal{P}_0(\cdot;\mathbf{y}) : \mathbb{R}^N \to \mathbb{R}$ is given by

$$\mathcal{P}_0(\mathbf{x};\mathbf{y}) := \frac{1}{M}\sum_{i=1}^{M}\Big((\mathbf{y}^T\mathbf{A}_i\mathbf{y} - b_i)^2 + (\mathbf{y}^T\mathbf{A}_i\mathbf{y} - b_i)\left\langle 2\mathbf{A}_i\mathbf{y}, \mathbf{x}-\mathbf{y}\right\rangle\Big) + \mathcal{R}(\mathbf{x})\,. \quad (41)$$

We use the Legendre function: $h(\mathbf{x}) = \frac{1}{4}\|\mathbf{x}\|^4 + \frac{1}{2}\|\mathbf{x}\|^2$. Then, due to (Bolte et al., 2018, Lemma 5.1) the following L-smad/MAP property is satisfied:

$$|\mathcal{P}_0(\mathbf{x}) - \mathcal{P}_0(\mathbf{x};\mathbf{y})| \leq L_0 D_h(\mathbf{x},\mathbf{y})\,, \text{ for all } \mathbf{x},\mathbf{y} \in \mathbb{R}^N, \quad (42)$$

where $L_0 \geq \sum_{i=1}^{M}(3\|\mathbf{A}_i\|_F^2 + \|\mathbf{A}_i\|_F|b_i|)$. In this setting, Model BPG subproblems have closed form solutions (see Bolte et al. (2018); Mukkamala et al. (2020)).

Model 2. The importance of finding better models suited to a particular problem was emphasized in Asi and Duchi (2019). The above provided model function in

(41) is satisfactory, however, we would like take advantage of the structure of the function (40). Taking inspiration from Asi and Duchi (2019), a simple observation that the objective is nonnegative can be exploited to create a new model function. We incorporate such a behavior in our second model function provided below. We use the Prox-Linear setting described in Section 4.2, where for any $\mathbf{x} \in \mathbb{R}^N$ we set

$$f_0(\mathbf{x}) := \mathcal{R}(\mathbf{x}), \quad (F(\mathbf{x}))_i = (\mathbf{x}^T \mathbf{A}_i \mathbf{x} - b_i)^2, \text{ for all } i = 1, \dots, M, \tag{43}$$

and, for any $\tilde{\mathbf{y}} \in \mathbb{R}^M$ we set $g(\tilde{\mathbf{y}}) := \frac{1}{M}\|\tilde{\mathbf{y}}\|_1$, for $\tilde{\mathbf{y}} \in \mathbb{R}^M$. Based on (38), for fixed $\mathbf{y} \in \mathbb{R}^N$, the model function $\mathcal{P}_1(\cdot; \mathbf{y}) : \mathbb{R}^N \to \mathbb{R}$ is given by

$$\mathcal{P}_1(\mathbf{x}; \mathbf{y}) := \frac{1}{M} \sum_{i=1}^{M} |(\mathbf{y}^T \mathbf{A}_i \mathbf{y} - b_i)^2 + (\mathbf{y}^T \mathbf{A}_i \mathbf{y} - b_i) \langle 2\mathbf{A}_i \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle| + \mathcal{R}(\mathbf{x}). \tag{44}$$

Considering the Legendre function $h(\mathbf{x}) = \frac{1}{4}\|\mathbf{x}\|^4 + \frac{1}{2}\|\mathbf{x}\|^2$ and (Bolte et al., 2018, Lemma 5.1) shows that the L-smad (or MAP) property holds true:

$$|\mathcal{P}_0(\mathbf{x}) - \mathcal{P}_1(\mathbf{x}; \mathbf{y})| \leq L_0 D_h(\mathbf{x}, \mathbf{y}), \text{ for all } \mathbf{x}, \mathbf{y} \in \mathbb{R}^N, \tag{45}$$

with $L_0 \geq \sum_{i=1}^{M} (3\|\mathbf{A}_i\|_F^2 + \|\mathbf{A}_i\|_F |b_i|)$. Unlike the Model 1 setting, we do not have closed form solutions for Model BPG subproblems in Model 2 setting. Here, we solve such subproblems using Primal-Dual Hybrid Gradient Algorithm (PDHG) Pock and Chambolle (2011). We use a random synthetic dataset, for which we provide empirical results in Figure 1, where we show superior performance of Model BPG variants compared to IBPM-LS, in particular, with the model function provided in (44). For simplicity, we choose a constant step-size τ in all the iterations, such that $\tau \in (0, 1/L_0)$. We empirically validate Proposition 21 in Figure 2. All the assumptions required to deduce the global convergence of Model BPG are straightforward to verify, and we leave it as an exercise to the reader. Note that here $\operatorname{int} \operatorname{dom} h = \mathbb{R}^N$, thus $\omega^{\operatorname{int} \operatorname{dom} h}(\mathbf{x}_0) = \omega(\mathbf{x}_0)$ holds trivially.

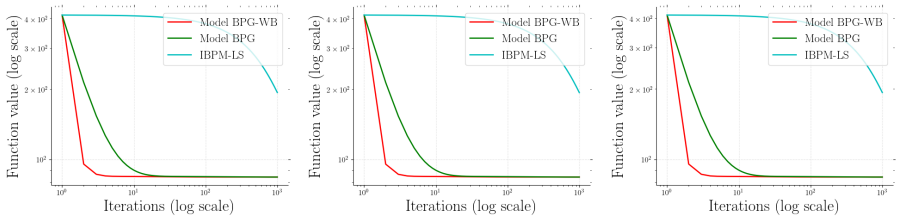
5.2 Poisson linear inverse problems

We now consider a broad class of problems with varied practical applications, known as Poisson inverse problems Bertero et al. (2009); Bauschke et al. (2016); Ochs et al. (2019); Nikolova (2005). For all $i = 1, \dots, M$, let $b_i > 0$, $\mathbf{a}_i \neq 0$ and $\mathbf{a}_i \in \mathbb{R}^N_+$ be known. Moreover, we have for any $\mathbf{x} \in \mathbb{R}^N_+$, $\langle \mathbf{a}_i, \mathbf{x} \rangle > 0$ and $\sum_{i=1}^{M} (\mathbf{a}_i)_j > 0$, for all $j = 1, \dots, N$, $i = 1, \dots, M$. Equipped with these notions, the optimization problem of Poisson linear inverse problems as following:

$$\min_{\mathbf{x} \in \mathbb{R}_+} \left\{ f(\mathbf{x}) := \sum_{i=1}^{M} (\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i \log(\langle \mathbf{a}_i, \mathbf{x} \rangle)) + \phi(\mathbf{x}) \right\}, \tag{46}$$

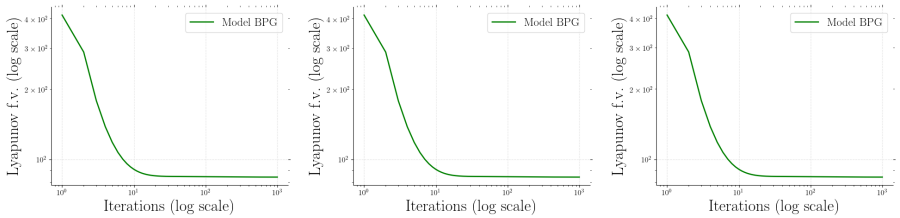
where ϕ is the regularizing function, which is potentially nonconvex. For simplicity, we set $\phi = 0$. The function $f_1 : \mathbb{R}^N \to \overline{\mathbb{R}}$ at any $\mathbf{x} \in \mathbb{R}^N$ is defined as:

$$f_1(\mathbf{x}) := \sum_{i=1}^{M} (\langle \mathbf{a}_i, \mathbf{x} \rangle - b_i \log(\langle \mathbf{a}_i, \mathbf{x} \rangle)).$$



(a) ℓ_1 regularization **(b)** Squared ℓ_2 regularization **(c)** No regularization

Fig. 3: In this experiment we compare the performance of Model BPG, Model BPG with Backtracking (denoted as Model BPG-WB) and IBPM-LS Ochs et al. (2013) on Poisson linear inverse problems with ℓ_1 regularization, squared ℓ_2 regularization and with no regularization. We set the regularization parameter λ to 0.1. The plots illustrate that Model BPG-WB is faster in all the settings, followed by Model BPG.



(a) ℓ_1 regularization **(b)** Squared ℓ_2 regularization **(c)** No regularization

Fig. 4: Under the same setting as in Figure 3, we illustrate here that Model BPG results in sequences that have monotonically nonincreasing Lyapunov function value evaluations. By *Lyapunov f.v.* we mean Lyapunov function values.

Note that the function f_1 is coercive. The Legendre function $h : \mathbb{R}^N_{++} \to \mathbb{R}$ (Burg's entropy) that is given by

$$h(\mathbf{x}) = -\sum_{i=1}^{N} \log(\mathbf{x}_i)\,, \qquad \text{for all } \mathbf{x} \in \mathbb{R}^N_{++}, \tag{47}$$

where $\mathbf{x}_i$ is the i^{th} coordinate of $\mathbf{x}$.

Lemma 31. Let h be defined as in (47). For $L \geq \sum_{i=1}^{M} b_i$, the function $Lh - f_1$ and $Lh + f_1$ is convex on $\mathbb{R}^N_{++}$, or equivalently the following L-smad property or the MAP property holds true:

$$-LD_h(\mathbf{x}, \bar{\mathbf{x}}) \leq f_1(\mathbf{x}) - f_1(\bar{\mathbf{x}}) - \langle \nabla f_1(\bar{\mathbf{x}}), \mathbf{x} - \bar{\mathbf{x}} \rangle \leq LD_h(\mathbf{x}, \bar{\mathbf{x}})\,, \text{ for all } \mathbf{x}, \bar{\mathbf{x}} \in \mathbb{R}^N_{++}\,.$$

Proof. The proof of convexity of $Lh - f_1$ follows from (Bauschke et al., 2016, Lemma 7). The function $Lh + f_1$ is convex as f_1 is convex. □

When Model BPG is applied to solve (46) with h given in (47), if the limit points of the sequence generated by Model BPG lie in $\operatorname{int} \operatorname{dom} h$, our global convergence

result is valid. However, it is difficult to guarantee such a condition. This is because, there can exist subsequences for which certain components of the iterates can tend to zero. In such a scenario, some components of $\nabla^2 h(\mathbf{x}_k)$ will tend to ∞, which will lead to the failure of the relative error condition in Lemma 23. In that case, our analysis cannot guarantee the global convergence of the sequence generated by Model BPG. Thus, in such a scenario it is important to guarantee that the iterates of Model BPG lie in $\mathbb{R}^N_{++}$. To this regard, we could modify the problem (46), by adding certain constraint set, such that all the limit points lie in $\operatorname{int}\operatorname{dom} h$. In particular, with certain $\varepsilon > 0$, we use the constraint set given by $C_\varepsilon = \{\mathbf{x} : \mathbf{x}_i \geq \varepsilon,\, \forall i = 1, \ldots, N\}$,

6 Conclusion

Bregman proximal minimization framework is prominent in solving additive composite problems, in particular, using BPG Bolte et al. (2018) algorithm or its variants Mukkamala et al. (2020). However, extensions to generic composite problems was an open problem. To this regard, based on foundations of Drusvyatskiy et al. (2019); Ochs et al. (2019), we proposed Model BPG algorithm that is applicable to a vast class of structured nonconvex nonsmooth problems, including generic composite problems. Model BPG relies on certain function approximation, known as model function, which preserves first order information about the function. The model error is bounded via certain Bregman distance, which drives the global convergence analysis of the sequence generated by Model BPG. The analysis is nontrivial and requires significant changes compared to the standard analysis of Bolte et al. (2018, 2014); Attouch and Bolte (2009); Attouch et al. (2013). Moreover, we numerically illustrate the superior performance of Model BPG on various real world applications.

References

Asi H, Duchi JC (2019) The importance of better models in stochastic optimization. Proceedings of the National Academy of Sciences 116(46):22924–22930

Attouch H, Bolte J (2009) On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Mathematical Programming 116(1):5–16

Attouch H, Goudou X, Redont P (2000) The heavy ball with friction method, i. the continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Communications in Contemporary Mathematics 2(1):1–34

Attouch H, Bolte J, Svaiter B (2013) Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods. Mathematical Programming 137(1-2):91–129, DOI 10.1007/s10107-011-0484-9

Attouch H, Chbani Z, Fadili J, Riahi H (2020) First-order optimization algorithms via inertial systems with Hessian driven damping. Mathematical Programming pp 1–43

Bauschke H, Borwein J (1997) Legendre functions and the method of random Bregman projections. Journal of Convex Analysis 4(1):27–67

Bauschke H, Borwein J, Combettes P (2001) Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. Communications in Contemporary Mathematics 3(4):615–647

Bauschke H, Borwein J, Combettes P (2003) Bregman monotone optimization algorithms. SIAM Journal on Control and Optimization 42(2):596–636

Bauschke H, Bolte J, Teboulle M (2016) A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Mathematics of Operations Research 42(2):330–348

Beck A, Teboulle M (2003) Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters 31(3):167–175

Bertero M, Boccacci P, Desiderà G, Vicidomini G (2009) Image deblurring with poisson data: from cells to galaxies. Inverse Problems 25(12):123006

Birnbaum B, Devanur NR, Xiao L (2011) Distributed algorithms via gradient descent for Fisher markets. In: Proceedings of the 12th ACM conference on Electronic commerce, ACM, pp 127–136

Bolte J, Daniilidis A, Lewis A (2006) The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization 17(4):1205–1223, DOI 10.1137/050644641, URL `http://dx.doi.org/10.1137/050644641`

Bolte J, Daniilidis A, Lewis A, Shiota M (2007) Clarke subgradients of stratifiable functions. SIAM Journal on Optimization 18(2):556–572, URL `http://dblp.uni-trier.de/db/journals/siamjo/siamjo18.html#BolteDLS07`

Bolte J, Sabach S, Teboulle M (2014) Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming 146(1-2):459–494

Bolte J, Sabach S, Teboulle M, Vaisbourd Y (2018) First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM Journal on Optimization 28(3):2131–2151

Bregman LM (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics 7(3):200–217

Candes EJ, Li X, Soltanolkotabi M (2015) Phase retrieval via Wirtinger flow: Theory and algorithms. IEEE Transactions on Information Theory 61(4):1985–2007

Censor Y, Lent A (1981) An iterative row-action method for interval convex programming. Journal of Optimization Theory and Applications 34(3):321–353

Davis D, Drusvyatskiy D, MacPhee KJ (2018) Stochastic model-based minimization under high-order growth. arXiv preprint arXiv:180700255

van den Dries L, Miller C (1996) Geometric categories and o-minimal structures. Duke Mathematical Journal 84(2):497–540

Drusvyatskiy D (2017) The proximal point method revisited. arXiv preprint arXiv:171206038

Drusvyatskiy D, Lewis AS (2018) Error bounds, quadratic growth, and linear convergence of proximal methods. Mathematics of Operations Research

Drusvyatskiy D, Paquette C (2019) Efficiency of minimizing compositions of convex functions and smooth maps. Mathematical Programming 178(1-2):503–558

Drusvyatskiy D, Ioffe AD, Lewis AS (2019) Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. Mathematical

Programming pp 1–27
Frankel P, Garrigos G, Peypouquet J (2014) Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. Journal of Optimization Theory and Applications 165(3):874–900
Gutman DH, Peña JF (2018) Perturbed Fenchel duality and first-order methods. arXiv preprint arXiv:181210198
Kurdyka K (1998) On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier 48(3):769–783
Lewis AS, Wright SJ (2016) A proximal method for composite minimization. Mathematical Programming 158(1-2):501–546
Li G, Pong T (2017) Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods. Foundations of Computational Mathematics pp 1–34
Li G, Mordukhovich BS, Phạm TS (2015) New fractional error bounds for polynomial systems with applications to hölderian stability in optimization and spectral theory of tensors. Mathematical Programming 153(2):333–362
Lu H (2019) "Relative-Continuity" for non-Lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. INFORMS Journal on Optimization 1(4):288–303
Lu H, Freund RM, Nesterov Y (2018) Relatively smooth convex optimization by first-order methods, and applications. SIAM Journal on Optimization 28(1):333–354
Luke DR (2017) Phase retrieval, What's new? SIAG/OPT Views and News 25(1):1–6
Mordukhovich BS (2018) Variational analysis and applications. Springer
Mukkamala MC, Ochs P (2019) Beyond alternating updates for matrix factorization with inertial Bregman proximal gradient algorithms. In: Advances in Neural Information Processing Systems, pp 4266–4276
Mukkamala MC, Westerkamp F, Laude E, Cremers D, Ochs P (2019) Bregman proximal framework for deep linear neural networks. arXiv preprint arXiv:191003638
Mukkamala MC, Ochs P, Pock T, Sabach S (2020) Convex-Concave backtracking for inertial Bregman proximal gradient algorithms in nonconvex optimization. SIAM Journal on Mathematics of Data Science 2(3):658–682
Nesterov Y (2004) Introductory lectures on convex optimization: a basic course
Nesterov Y (2007) Modified Gauss–Newton scheme with worst case guarantees for global performance. Optimisation methods and software 22(3):469–483
Nikolova M (2005) Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. Multiscale Modeling & Simulation 4(3):960–991
Ochs P (2015) Long term motion analysis for object level grouping and nonsmooth optimization methods. PhD thesis, Albert-Ludwigs-Universität Freiburg, URL `http://lmb.informatik.uni-freiburg.de//Publications/2015/Och15`
Ochs P (2019) Unifying abstract inexact convergence theorems and block coordinate variable metric ipiano. SIAM Journal on Optimization 29(1):541–570, DOI 10.1137/17M1124085
Ochs P, Malitsky Y (2019) Model function based conditional gradient method with Armijo-like line search. In: International Conference on Machine Learning, pp 4891–4900

Ochs P, Dosovitskiy A, Pock T, Brox T (2013) An iterated ℓ_1 algorithm for non-smooth non-convex optimization in computer vision. In: International Conference on Computer Vision and Pattern Recognition (CVPR)

Ochs P, Chen Y, Brox T, Pock T (2014) iPiano: Inertial proximal algorithm for non-convex optimization. SIAM Journal on Imaging Sciences 7(2):1388–1419, URL `http://lmb.informatik.uni-freiburg.de//Publications/2014/OB14`

Ochs P, Fadili J, Brox T (2019) Non-smooth non-convex Bregman minimization: Unification and new algorithms. Journal of Optimization Theory and Applications 181(1):244–278

Pauwels E (2016) The value function approach to convergence analysis in composite optimization. Operations Research Letters 44(6):790–795

Pock T, Chambolle A (2011) Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In: International Conference on Computer Vision, pp 1762–1769

Rockafellar RT (1970) Convex Analysis. Princeton University Press, Princeton

Rockafellar RT, Wets RJB (1998) Variational analysis, vol 317. Springer Berlin Heidelberg, Heidelberg, DOI 10.1007/978-3-642-02431-3

Teboulle M, Vaisbourd Y (2020) Novel proximal gradient methods for nonnegative matrix factorization with sparsity constraints. SIAM Journal on Imaging Sciences 13(1):381–421

Wang G, Giannakis GB, Eldar YC (2018) Solving systems of random quadratic equations via truncated amplitude flow. IEEE Transactions on Information Theory 64(2):773–794