Smoothing technique for nonsmooth composite minimization with linear operator

Quang Van Nguyen[*], Olivier Fercoq[†], and Volkan Cevher[*]

[*]Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[†]Télécom ParisTech, Institut Mines-Télécom Paris, France
quang.nguyen@epfl.ch, olivier.fercoq@telecom-paristech.fr
volkan.cevher@epfl.ch

Abstract

We introduce and analyze an algorithm for the minimization of convex functions that are the sum of differentiable terms and proximable terms composed with linear operators. The method builds upon the recently developed smoothed gap technique. In addition to a precise convergence rate result, valid even in the presence of linear inclusion constraints, this new method allows an explicit treatment of the gradient of differentiable functions and can be enhanced with line-search. We also study the consequences of restarting the acceleration of the algorithm at a given frequency. These new features are not classical for primal-dual methods and allow us to solve difficult large scale convex optimization problems. We numerically illustrate the superior performance of the algorithm on basis pursuit, TV-regularized least squares regression and L1 regression problems against the state-of-the-art.

Key words. composite minimization, forward-backward, multivariate minimization, atomization energies prediction, smoothing technique, total variation regularization.

Mathematics Subject Classifications (2010) 47H05, 49M29, 49M27, 90C25

1 Introduction

Nonlinear and non-smooth convex optimization problems are widely presented in many disciplines, including signal and image processing, operations research, machine learning, game theory, economics, and mechanics. In this paper, we consider the following problem.

Problem 1.1 Let $\mathcal{H}$ and $\mathcal{G}$ be real Hilbert spaces, let $M\colon \mathcal{H} \to \mathcal{G}$ be a bounded linear operator, and let $f\colon \mathcal{H} \to \mathbb{R}$, $g\colon \mathcal{H} \to (-\infty, +\infty]$ and $h\colon \mathcal{G} \to (-\infty, +\infty]$ be proper, closed lower semi-continuous convex functions where f is moreover assumed to have L_f-Lipschitz gradient. Consider the following generic convex minimization problem

$$F^\star = \min_{x\in\mathcal{H}} \big\{f(x) + g(x) + h(Mx)\big\} \tag{1.1}$$

under the assumption that its set of minimizers $\mathcal{P}^\star$ is non-empty.

Following [1, Definition 19.11], if we suppose that $\emptyset \neq M(\operatorname{dom}(f+g)) \cap \operatorname{dom} h = M(\operatorname{dom} g) \cap \operatorname{dom} h$ and set $\mathcal{F}\colon \mathcal{H} \times \mathcal{G}\colon (x, y) \mapsto f(x) + g(x) + h(Mx - y)$ then Problem 1.1 becomes the primal problem associated to $\mathcal{F}$ and its associated dual problem is

$$G^\star = \max_{y\in\mathcal{G}} \big\{G(y) = \min_{x\in\mathcal{H}} \big\{f(x) + g(x) + \big\langle Mx, y\big\rangle - h^*(y)\big\}\big\}, \tag{1.2}$$

where $\operatorname{dom} f = \big\{x \in \mathcal{H} \;\big|\; f(x) < +\infty\big\}$ is the domain of f and $h^*\colon \mathcal{G} \to (-\infty, +\infty] : y \mapsto \max_{\bar{y}\in\mathcal{G}} \big\langle \bar{y}, y\big\rangle - h(\bar{y})$ is the Fenchel-Moreau conjugate function of h. In this case, [1, Corollary 19.19] states that the set of solution $\mathcal{D}^\star$ to (1.2) is non-empty, and furthermore, a point $x^\star \in \mathcal{H}$ is in $\mathcal{P}^\star$ if and only if there exists $y^\star \in \mathcal{D}^\star$ such that $(x^\star, y^\star)$ is a saddle point of the Lagrangian function

$$\mathcal{L}\colon (x, y) \mapsto f(x) + g(x) + \big\langle Mx, y\big\rangle - h^*(y). \tag{1.3}$$

A particular case of Problem 1.1 is when $h = \iota_{\mathbb{K}}(\cdot - c)$ with $c \in \mathcal{G}$, is the indicator function of a non-empty closed convex subset $\mathbb{K} \subset \mathcal{G}$, i.e.,

$$\iota_{\mathbb{K}}\colon \mathcal{G} \to (-\infty, +\infty] : y \mapsto \begin{cases} 0, & \text{if } y \in \mathbb{K}, \\ +\infty, & \text{otherwise.} \end{cases} \tag{1.4}$$

In this case, Problem 1.1 reduces to the following constrained minimization problem

$$\min_{x\in\mathcal{H}} \big\{f(x) + g(x) \;:\; Mx - c \in \mathbb{K}\big\}, \tag{1.5}$$

and furthermore, if $g = \iota_{\mathscr{X}}$ for some $\mathscr{X} \subset \mathcal{H}$,

$$\min_{x\in\mathcal{H}} \big\{f(x) \;:\; x \in \mathscr{X} \text{ such that } Mx - c \in \mathbb{K}\big\}. \tag{1.6}$$

A traditional approach for smooth minimization problems is the gradient descent algorithm together with its accelerated version. This idea has already been adapted for nonsmooth composite minimization problem by linearizing the smooth term before minimizing. For instance, if $h = 0$, then Problem 1.1 can be solved by FISTA (in other words, an accelerated forward-backward algorithm) [3, 5], and this approach can be generalized to the case where h is with Lipschitz gradient. If furthermore, $h = \iota_{\{c\}}$ for some $c \in \mathcal{G}$, then various of alternating direction optimization methods (ADMM) [7] can be used. A linearization technique is recently combined with ADMM in [15] to tackle such cases. However, in the general case, we need a special treatment of $h(Mx)$.

For instance, we may compute approximations to the proximal operator of $(x \mapsto g(x) + h(Mx))$ as in [2]. We obtain an algorithm with a nested loop for this proximal operator computation. Provided we are able to control theaccuracy of the inner loop, we can obtain convergence rates.

Another possibility is to consider primal-dual splitting. By interpreting the optimization problem 1.1 as a saddle point problem, we can derive methods updating primal and dual variables at each iteration, without any nested loop. A primal-dual method able to deal with our composite framework was given in [6, 14].

A powerful smoothing framework was first introduced in [9], which can also be applied to solve Problem 1.1. The main idea isto consider a smoothed approximation to the nonsmooth function h and minimize the resulting problem using an accelerated forward-backward algorithm. This approach has been improved (for the case $f = 0$) in [13] as follows. Instead of considering a fixed smoothed approximation to the nonsmooth function h, the authors set up a homotopy strategy by considering a decreasing smoothing parameter. In doing so, they obtain improved convergence characterizations, and,more importantly, they prove finite-time convergence rates in terms of function value and infeasibility. Indeed, these rates are difficult to obtain in the constrained case when approximately solving the proximity operator or considering classical primal-dual algorithms.

In this paper, we build on this latter, homotopy-based smoothing technique to tackle the more general composite framework, i.e., Problem 1.1. In this scenario, to apply the technique of [9] as in [13], it would require the computation of the proximity operator of $f + g$ which is generally not easy even the case where one knows how to compute the proximity operators of f and g separately. One of our goals is to avoid this computational difficulty by using the smoothness. The second non-smooth part is then smoothed using the idea of [9]. To see this, let us rewrite the objective function of (1.1) as follows

$$F(x) := f(x) + g(x) + h(Mx) = f(x) + g(x) + \max_{y \in \mathcal{G}} \ \langle Mx, y \rangle - h^*(y), \tag{1.7}$$

Instead of minimizing F, we first smooth one of its nonsmooth parts, says h, controlled by a smoothness parameter $\beta \in]0, +\infty[$ and then minimize

$$F_\beta(x) := f(x) + g(x) + \max_{y \in \mathcal{G}} \ \big\{ \langle Mx, y \rangle - h^*(y) - \beta q(y) \big\}, \tag{1.8}$$

with a suitable strongly convex function $q \colon \mathcal{G} \to (-\infty, +\infty]$. We then use the accelerated forward-backward scheme to design algorithms that maintain the decrease of the approximated objective function in the sense that

$$(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star \leqslant (1 - \tau_k)\big(F_{\beta_k}(\bar{x}^k) - F^\star\big) + \psi_k, \tag{1.9}$$

where $(\bar{x}^k)_{k \in \mathbb{N}}$ and the parameters are generated by the algorithm with $(\tau_k)_{k \in \mathbb{N}} \subset [0, 1)^{\mathbb{N}}$ and $(\max(\psi_k, 0))_{k \in \mathbb{N}}$ tends to zero. We will also simultaneously update the β_{k+1} parameter to zero to achieve an $\mathcal{O}(1/k)$ convergence rate. Our approach allows us to consider features that were introduced initially for unconstrained optimization like line-search or the balance of computational power between the steps of the algorithm.

The rest of the paper is organized as follow. In Section 2 we revise some technical facts. The main result is presented in Section 3. Numerical evidence is placed in Section 4.

Notation. The Hilbert spaces $\mathcal{H}$ and $\mathcal{G}$ are equipped with their respective norms and inner products that we will both denote by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ respectively. A positive definite linear operator S on $\mathcal{G}$, i.e., $\exists\sigma\in\,]0,+\infty[$ such that $(\forall y\in\mathcal{G})\ \langle y,Sy\rangle\geqslant\sigma\|y\|^2$, induces a norm $(\forall y\in\mathcal{G}\ \|y\|_S=\sqrt{\langle y,Sy\rangle}$. Given a proper, closed, lower semi-continuous convex function $f\colon\mathcal{H}\to(-\infty,+\infty]$, we denote by $\operatorname{int}\operatorname{dom}f$ the interior of $\operatorname{dom}f$ and by

$$\partial f\colon x\mapsto\big\{v\in\mathcal{H}\;\big|\;(\forall y\in\mathcal{H})\ f(x)+\langle y-x,v\rangle\leqslant f(y)\big\} \tag{1.10}$$

the subdifferential of f. If f is differentiable, then we use ∇f for its gradient and in this case we say that f has L_f-Lipschitz gradient with respect to norm $\|\cdot\|_S$ if

$$(\forall x\in\mathcal{H})(\forall y\in\mathcal{H})\quad f(x)\leqslant f(y)+\langle x-y,\nabla f(y)\rangle+\frac{L_f}{2}\|x-y\|_S^2. \tag{1.11}$$

Finally, we say that f is μ-strongly convex on $\mathcal{H}$ with respect to $\|\cdot\|_S$ if

$$(\forall x\in\mathcal{H})(\forall y\in\mathcal{H})(\forall v\in\partial f(y))\quad f(x)\geqslant f(y)+\langle x-y,v\rangle+\frac{\mu}{2}\|x-y\|_S^2. \tag{1.12}$$

Without indicating the norm, we are assuming that Lipschitz continuity or strong convexity is with a Hilbertian norm.

2 Preliminaries

In this section we revise some basic facts about the proximity operators and functions, and furthermore, the smoothing technique for non-smooth functions using the Fenchel-Moreau conjugate. In the optimization, the following notion of the proximity operators is widely used.

Definition 2.1 [1, Definition 12.23] Let $g\colon\mathcal{H}\to(-\infty,+\infty]$ be a proper closed lower-semicontinuous convex function. The proximity operator of g is

$$\operatorname{prox}_g\colon\mathcal{H}\to\mathcal{H}\colon x\mapsto\operatorname*{argmin}_{z\in\mathcal{H}}\ g(z)+\frac{1}{2}\|z-x\|^2. \tag{2.1}$$

Lemma 2.2 [1, Proposition 12.26] *Let $g\colon\mathcal{H}\to(-\infty,+\infty]$ be a proper closed lower semi-continuous convex function, let $\gamma\in\,]0,+\infty[$, let $x\in\mathcal{H}$ and let $p=\operatorname{prox}_{\gamma g}(x)$. Then, it holds that*

$$(\forall z\in\mathcal{H})\quad\gamma^{-1}\langle z-p,x-p\rangle+g(p)\leqslant g(z). \tag{2.2}$$

The following smoothing technique using the Fenchel-Moreau conjugate and the proximity functions is from [9].

Definition 2.3 Let $h\colon\mathcal{G}\to(-\infty,+\infty]$ be a convex function, let $\beta\in\,]0,+\infty[$, let S be a positive definite linear operator on $\mathcal{G}$ and let $\dot{y}\in\mathcal{G}$. The β-smooth approximation of h is

$$h_\beta(\cdot;\dot{y})\colon\mathcal{G}\to(-\infty,+\infty]:y\mapsto\max_{\bar{y}\in\mathcal{G}}\big\{\langle y,\bar{y}\rangle-h^*(\bar{y})-\frac{\beta}{2}\|\bar{y}-\dot{y}\|_S^2\big\}. \tag{2.3}$$

Set

$$(\forall y \in \mathcal{G}) \quad y_\beta^*(y;\dot{y}) = \operatorname*{argmax}_{\bar{y}\in\mathcal{G}} \ \langle y, \bar{y}\rangle - h^*(\bar{y}) - \frac{\beta}{2}\|\bar{y}-\dot{y}\|_S^2. \tag{2.4}$$

For instance, if $S = I$ then

$$(\forall y \in \mathcal{G}) \quad y_\beta^*(y;\dot{y}) = \operatorname{prox}_{\beta^{-1}h^*}\big(\beta^{-1}y + \dot{y}\big). \tag{2.5}$$

We summarize important properties of the smooth approximation in the following lemma which will be a crucial key in the analysis of our algorithm.

Lemma 2.4 *Let $h\colon \mathcal{G} \to (-\infty, +\infty]$ be convex, let S be a positive definite linear operator on $\mathcal{G}$ and let $\dot{y} \in \mathcal{G}$. Consider the smooth approximations of h*

$$(\forall \beta \in \left]0, +\infty\right[) \quad h_\beta(\cdot;\dot{y})\colon \mathcal{G} \to (-\infty, +\infty] : y \mapsto \max_{\bar{y}\in\mathcal{G}} \big\{\langle y, \bar{y}\rangle - h^*(\bar{y}) - \frac{\beta}{2}\|\bar{y}-\dot{y}\|_S^2\big\}. \tag{2.6}$$

Then the following hold:

(i) *Denote $\tilde{h}\colon (\beta, y) \mapsto h_\beta(y;\dot{y})$. Then we have the following:*

(a) *$\tilde{h}$ is differentiable with respect to both variables and*

$$(\forall y \in \mathcal{G})(\forall \beta \in \left]0, +\infty\right[) \quad \frac{\partial \tilde{h}}{\partial \beta}(\beta, y) = -\frac{1}{2}\|y_\beta^*(y;\dot{y}) - \dot{y}\|_S^2 = -\frac{1}{2}\|\nabla h_\beta(y;\dot{y}) - \dot{y}\|_S^2. \tag{2.7}$$

(b) *$\tilde{h}$ is convex with respect to first variable and*

$$\begin{aligned}(\forall y \in \mathcal{G})(\forall \bar{\beta} \leqslant \tilde{\beta}) \quad \tilde{h}(\bar{\beta}, y) &\leqslant \tilde{h}(\tilde{\beta}, y) - (\tilde{\beta} - \bar{\beta})\frac{\partial \tilde{h}}{\partial \beta}(\bar{\beta}, y) \\ &= \tilde{h}(\tilde{\beta}, y) + \frac{\tilde{\beta} - \bar{\beta}}{2}\|\nabla h_{\bar{\beta}}(y;\dot{y}) - \dot{y}\|_S^2.\end{aligned} \tag{2.8}$$

(ii) *Let $\beta \in \left]0, +\infty\right[$. Then the function $y \mapsto h_\beta(y;\dot{y})$ is well-defined on $\mathcal{G}$. It is convex with $\frac{1}{\beta}$-Lipschitz gradient in the norm $\|\cdot\|_{S^{-1}}$ and furthermore,*

$$(\forall (\bar{y}, \hat{y}) \in \mathcal{G}^2) \quad h_\beta(\hat{y};\dot{y}) + \big\langle \bar{y} - \hat{y}, \nabla h_\beta(\hat{y};\dot{y})\big\rangle \leqslant h_\beta(\bar{y};\dot{y}) - \frac{\beta}{2}\|\nabla h_\beta(\hat{y};\dot{y}) - \nabla h_\beta(\bar{y};\dot{y})\|_S^2. \tag{2.9}$$

(iii) *The following inequality holds:*

$$(\forall (y, \hat{y}) \in \mathcal{G}^2)(\forall \beta \in \left]0, +\infty\right[) \quad h_\beta(\hat{y};\dot{y}) + \big\langle y - \hat{y}, \nabla h_\beta(\hat{y};\dot{y})\big\rangle \leqslant h(y) - \frac{\beta}{2}\|\nabla h_\beta(\hat{y};\dot{y}) - \dot{y}\|_S^2. \tag{2.10}$$

(iv) *For all* $(\beta,\tau)\in\left]0,+\infty\right[^2$ *and for all* $(\bar{y},\hat{y})\in\mathcal{G}^2$*, one has*

$$\begin{aligned}0&\leqslant\|(1-\tau)(\nabla h_\beta(\hat{y};\dot{y})-\nabla h_\beta(\bar{y};\dot{y})+\tau(\nabla h_\beta(\hat{y};\dot{y})-\dot{y})\|_S^2\\&=(1-\tau)\|\nabla h_\beta(\hat{y};\dot{y})-\nabla h_\beta(\bar{y};\dot{y})\|_S^2+\tau\|\nabla h_\beta(\hat{y};\dot{y})-\dot{y}\|_S^2-\tau(1-\tau)\|\nabla h_\beta(\bar{y};\dot{y})-\dot{y}\|_S^2.\end{aligned}\tag{2.11}$$

Proof. (i)(a) As there is a unique minimizer to the problem defining $h_\beta(\cdot;\dot{y})$, the function is differentiable with respect to β and y.

(i)(b) The function $\beta\mapsto h_\beta(y;\dot{y})$ is convex as it is a maximum of functions, which are linear in β indexed by y and $\dot{y}$. The rest follows by convexity and the first point.

(ii) By the same arguments as in [9, Theorem 1] we deduce that the function $y\mapsto h_\beta(y;\dot{y})$ is convex and finite on $\mathcal{G}$. It also has $\frac{1}{\beta}$-Lipschitz gradient in the norm $\|\cdot\|_{S^{-1}}$. (2.9) is the cocoercivity inequality for convex functions with Lipschitz gradient. We provide the proof for completeness. Define $\phi(z)=h_\beta(z;\dot{y})-\langle\nabla h_\beta(\hat{y};\dot{y}),z\rangle$. The function ϕ is convex, its minimum is attained at $\hat{y}$ and it has a $\frac{1}{\beta}$-Lipschitz gradient in the norm $\|\cdot\|_{S^{-1}}$. Hence

$$\phi(\hat{y})\leqslant\phi(\bar{y}-\beta^{-1}S\nabla\phi(\bar{y}))\leqslant\phi(\bar{y})-\beta^{-1}\langle\nabla\phi(\bar{y}),S\nabla\phi(\bar{y})\rangle+\frac{\beta}{2}\|\beta^{-1}S\nabla\phi(\bar{y})\|_{S^{-1}}^2.$$

We get the result because $\nabla\phi(\bar{y})=\nabla h_\beta(\bar{y};\dot{y})-\nabla h_\beta(\hat{y};\dot{y})$.

(iii) Let $(\bar{y},\hat{y})\in\mathcal{G}^2$ and $\beta\in\left]0,+\infty\right[$ and set $y_\beta^*=y_\beta^*(\hat{y};\dot{y})=\nabla h_\beta(\hat{y};\dot{y})$. We have

$$\begin{aligned}h_\beta(\hat{y};\dot{y})+\left\langle y-\hat{y},\nabla h_\beta(\hat{y};\dot{y})\right\rangle&=\left\langle\hat{y},y_\beta^*\right\rangle-h^*(y_\beta^*)-\frac{\beta}{2}\|y_\beta^*-\dot{y}\|_S^2+\left\langle y-\hat{y},y_\beta^*\right\rangle\\&=\left\langle y,y_\beta^*\right\rangle-h^*(y_\beta^*)-\frac{\beta}{2}\|y_\beta^*-\dot{y}\|_S^2\\&\leqslant\max_{\bar{y}\in\mathcal{G}}\left\{\left\langle y,\bar{y}\right\rangle-h^*(\bar{y})\right\}-\frac{\beta}{2}\|y_\beta^*-\dot{y}\|_S^2\\&\leqslant h(y)-\frac{\beta}{2}\|\nabla h_\beta(\hat{y};\dot{y})-\dot{y}\|_S^2.\end{aligned}\tag{2.12}$$

(iv) This follows from the classical equality

$$\|(1-\tau)a+\tau b\|_S^2=(1-\tau)\|a\|_S^2+\tau\|b\|_S^2-\tau(1-\tau)\|a-b\|_S^2.\tag{2.13}$$

□

3 Main results

3.1 Presentation of the algorithms

In this section, we design new algorithms to solve Problem 1.1 based on the smoothing technique introduced in the previous section. Consider the setting of Problem 1.1. We fix a positive definite

linear operator S on $\mathcal{G}$ and a point $\dot{y} \in \mathcal{G}$. We will need the operator norm of M defined as

$$\|M\|_{S^{-1}} = \sup_{x \neq 0} \frac{\|Mx\|_{S^{-1}}}{\|x\|}$$

Our first algorithm is given as Algorithm 1.

Algorithm 1 Linearized ASGARD

1: Inputs: $\tau_0 = 1$, $\beta_0 > 0$, $\bar{x}^0 \in \mathcal{H}$, $\tilde{x}^0 \in \mathcal{H}$.
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad \hat{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k \tilde{x}^k$
4: $\quad \beta_{k+1} = \frac{\beta_k}{1+\tau_k}$ and $B_{k+1} = L_f + \frac{\|M\|^2_{S^{-1}}}{\beta_{k+1}}$
5: $\quad v^k = M^* y^*_{\beta_{k+1}}(M\hat{x}^k; \dot{y})$
6: $\quad \tilde{x}^{k+1} = \text{prox}_{\frac{1}{\tau_k B_{k+1}} g}\big(\tilde{x}^k - \frac{1}{\tau_k B_{k+1}}(\nabla f(\hat{x}^k) + v^k)\big)$
7: $\quad \bar{x}^{k+1} = \hat{x}^k + \tau_k(\tilde{x}^{k+1} - \tilde{x}^k) = (1 - \tau_k)\bar{x}^k + \tau_k \tilde{x}^{k+1}$
8: $\quad$ Find the unique positive τ_{k+1} such that

$$\frac{B_{k+1} - L_f}{B_{k+1}} \tau^3_{k+1} + \tau^2_{k+1} + \tau^2_k \tau_{k+1} - \tau^2_k = 0$$

9: **end for**
10: **return** $\bar{x}^{k+1}$

Remark 3.1 Let us consider problem (1.5) with $\mathbb{K} = \{0\}$. Recent paper [15] proposed a combination of linearizing technique and alternating direction of multipliers methods to solve (1.5) in which they obtained the $\mathcal{O}(1/k)$-rate (ergodic) convergence for fixed parameters and $\mathcal{O}(1/k^2)$ with adaptive parameters. Recall that in this case $h = \iota_{\{c\}}$ and hence $y^*_{\beta_{k+1}}(M\hat{x}^k)$ in the step 5 of Algorithm 1 (with $\dot{y} = 0$ and $S = \text{I}$) becomes

$$y^*_{\beta_{k+1}}(M\hat{x}^k) = \text{prox}_{\beta^{-1}_{k+1} h^*}(\beta^{-1}_{k+1} M\hat{x}^k) = \beta^{-1}_{k+1}\big(M\hat{x}^k - \text{prox}_{\beta_{k+1} h}(M\hat{x}^k)\big) = \beta^{-1}_{k+1}\big(M\hat{x}^k - c\big), \quad (3.1)$$

where we used Moreau's decomposition [1, Theorem 14.3]. Hence, for problem (1.5) with $\mathbb{K} = \{0\}$, our Algorithm 1 is non-augmented Lagrangian version of [15, Algorithm 1] where in Step 6, instead of using linearized augmented Lagrangian as [15], we only used linearization of $f + g$ and hence this step only requires the computation of proximity of g. We also note that the update rule for parameters of [15] is different from ours.

For some problems, for instance when f encodes some data-fitting term, computing ∇f requires much more computational power than prox_{h^*} or prox_g. To circumvent this issue and concentrate on the non-smoothness of the objective, we propose Algorithm 2, a variant of the standard ASGARD in which we use old gradients.

Algorithm 2 Linearized ASGARD with old gradients

Inputs: $\tau_0 = 1$, $\beta_0 > 0$, $\bar{x}^0 \in \mathcal{H}$, $\tilde{x}^0 \in \mathcal{H}$, $\hat{\tilde{x}}^0 \in \mathcal{H}$, $\delta > 0$ and $\sigma > 0$.
for $k = 0, 1, \ldots$ **do**
  $\hat{x}^k = (1 - \tau_k)\bar{x}^k + \tau_k \tilde{x}^k$
  $\beta_{k+1} = \frac{\beta_k}{1+\tau_k}$ and $B_{k+1} = L_f + \frac{\|M\|^2_{S^{-1}}}{\beta_{k+1}}$
  $v^k = M^* y^*_{\beta_{k+1}}(M\hat{x}^k; \dot{y})$
  $\tilde{x}^{k+1} = \mathrm{prox}_{\frac{1}{\tau_k B_{k+1}} g}\big(\tilde{x}^k - \frac{1}{\tau_k B_{k+1}}(\nabla f(\hat{\tilde{x}}^k) + v^k)\big)$
  $\bar{x}^{k+1} = \hat{x}^k + \tau_k(\tilde{x}^{k+1} - \tilde{x}^k) = (1 - \tau_k)\bar{x}^k + \tau_k \tilde{x}^{k+1}$
  if $\frac{1}{2}\|\bar{x}^{k+1} - \hat{\tilde{x}}^k\|^2 - \frac{1}{2}\|\bar{x}^{k+1} - \hat{x}^k\|^2 \leqslant \sigma\big(\frac{\tau_k^2 B_{k+1}}{L_f}\big)^{2+\delta}$ **then**
    $\hat{\tilde{x}}^k = \hat{\tilde{x}}^{k-1}$
  else
    $\hat{\tilde{x}}^k = \hat{x}^k$
    Goto 6 and recompute $(\bar{x}^{k+1}, \tilde{x}^{k+1})$ with the true gradient.
  end if
  Find the unique positive τ_{k+1} such that

$$\frac{B_{k+1} - L_f}{B_{k+1}}\tau_{k+1}^3 + \tau_{k+1}^2 + \tau_k^2 \tau_{k+1} - \tau_k^2 = 0$$

end for
return $\bar{x}^{k+1}$

Remark 3.2 Let us consider the case when $\hat{\hat{x}}^{k+1} = \hat{x}^k$. We have

$$\begin{aligned}\frac{1}{2}\|\bar{x}^{k+2} - \hat{\hat{x}}^{k+1}\|^2 - \frac{1}{2}\|\bar{x}^{k+2} - \hat{x}^{k+1}\|^2 &= \langle \bar{x}^{k+2} - \frac{\hat{\hat{x}}^{k+1} + \hat{x}^{k+1}}{2}, \hat{x}^{k+1} - \hat{\hat{x}}^{k+1}\rangle \\ &= \langle \tau_{k+1}(\tilde{x}^{k+2} - \tilde{x}^{k+1}) - \frac{\hat{\hat{x}}^{k+1} - \hat{x}^{k+1}}{2}, \hat{x}^{k+1} - \hat{\hat{x}}^{k+1}\rangle \\ \hat{x}^{k+1} - \hat{\hat{x}}^{k+1} = \hat{x}^{k+1} - \hat{x}^k &= \tau_{k+1}(\tilde{x}^{k+1} - \hat{x}^k) + \tau_k(1-\tau_{k+1})(\tilde{x}^{k+1} - \tilde{x}^k)\end{aligned}$$

and so assuming boundedness of the iterates, $\frac{1}{2}\|\bar{x}^{k+2} - \hat{\hat{x}}^{k+1}\|^2 - \frac{1}{2}\|\bar{x}^{k+2} - \hat{x}^{k+1}\|^2$ is at most of the order of $\tau_k^2 \in \mathcal{O}(1/k^2)$. It is thus likely that it may sometimes be smaller than $\sigma\big(\frac{\tau_k^2 B_{k+1}}{L_f}\big)^{2+\delta}$ if δ is small enough.

The last variant of our method is Algorithm 3. It is equipped with a line search inspired by the line search for the accelerated universal gradient method [10]. It is particularly useful when the Lipschitz constant of ∇f is difficult to estimate. Moreover, it automatically adapts to the case when h is smooth by preventing the current estimate of the Lipschitz constant B_{k+1} to increase to infinity. Note that because of the line search test on Step 12, the use of old gradients is not compatible with this line search. Unlike the line search of Malitsky et al [8], the goal here is not only to adaptively estimate $\|M\|^2$ but also the whole $L_f + \|M\|^2/\beta$. Our line search is more computationally demanding but may take profit of some local smoothness of the nonsmooth function h.

Algorithm 3 Linearized ASGARD with line search

1: Inputs: $\tau_0 = 1$, $\beta_0 > 0$, $\bar{x}^0 \in \mathcal{H}$, $\tilde{x}^0 \in \mathcal{H}$, $B_0 \leqslant a(L_f + \frac{\|M\|^2_{S^{-1}}}{\beta_0})$, $a > 1$.
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad B_{k+1} = a^{-1}B_k$
4: $\quad$ **repeat**
5: $\quad\quad B_{k+1} = aB_{k+1}$
6: $\quad\quad$ Find the unique positive τ_k such that $\frac{1-\tau_k}{\tau_k^2 B_{k+1}} = \frac{1}{\tau_{k-1}^2 B_k}$ $\quad$ (except $\tau_0 = 1$)
7: $\quad\quad \hat{x}^k = (1-\tau_k)\bar{x}^k + \tau_k\tilde{x}^k$
8: $\quad\quad \beta_{k+1} = \frac{\beta_k}{1+\tau_k}$
9: $\quad\quad v^k = M^* y^*_{\beta_{k+1}}(M\hat{x}^k; \dot{y})$
10: $\quad\quad \tilde{x}^{k+1} = \mathrm{prox}_{\frac{1}{\tau_k B_{k+1}} g}\big(\tilde{x}^k - \frac{1}{\tau_k B_{k+1}}(\nabla f(\hat{x}^k) + v^k)\big)$
11: $\quad\quad \bar{x}^{k+1} = \hat{x}^k + \tau_k(\tilde{x}^+ - \tilde{x}^k) = (1-\tau_k)\bar{x}^k + \tau_k\tilde{x}^{k+1}$
12: $\quad$ **until** $f(\bar{x}^{k+1}) + h_{\beta_{k+1}}(M\bar{x}^{k+1}; \dot{y}) \leqslant f(\hat{x}^k) + h_{\beta_{k+1}}(M\hat{x}^k, \dot{y}) + \langle \nabla f(\hat{x}^k) + v^k, \bar{x}^{k+1} - \hat{x}^k\rangle + \frac{B_{k+1}}{2}\|\bar{x}^{k+1} - \hat{x}^k\|^2$
13: **end for**
14: **return** $\bar{x}^{k+1}$

3.2 Convergence of the parameters to 0

To analyze the convergence of Algorithm 1, we need the following result.

Lemma 3.3 *Let $(\tau_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$ and $(B_{k+1})_{k\in\mathbb{N}}$ be the positive sequences determined by Algorithm 1. Then the following hold:*

$$(\forall k \in \mathbb{N}\backslash\{0\}) \quad \frac{1-\tau_k}{\tau_k^2 B_{k+1}} = \frac{1}{\tau_{k-1}^2 B_k}. \tag{3.2}$$

Furthermore,

$$(\forall k \in \mathbb{N}) \quad \frac{1}{k+1} \leqslant \tau_k \leqslant \frac{2}{k+2}, \quad \beta_k \leqslant \frac{\beta_0}{k+1} \quad \text{and} \quad \tau_k^2 B_{k+1} \leqslant \frac{\tau_0^2 B_1}{k+1}. \tag{3.3}$$

Proof. First it follows from Step 4 of Algorithms 1 and 2 that

$$(\forall k \in \mathbb{N}\backslash\{0\}) \quad \frac{B_k + (B_k - L_f)\tau_k}{B_{k+1}} = \frac{L_f + \frac{\|M\|_{S^{-1}}^2}{\beta_k} + \frac{\|M\|_{S^{-1}}^2}{\beta_k}\frac{(\beta_k-\beta_{k+1})}{\beta_{k+1}}}{L_f + \frac{\|M\|_{S^{-1}}^2}{\beta_{k+1}}} = \frac{L_f + \frac{\|M\|_{S^{-1}}^2}{\beta_{k+1}}}{L_f + \frac{\|M\|_{S^{-1}}^2}{\beta_{k+1}}} = 1, \tag{3.4}$$

and Step 8 yields

$$(\forall k \in \mathbb{N}\backslash\{0\}) \quad \frac{1-\tau_k}{\tau_k^2 B_{k+1}} = \frac{1}{\tau_{k-1}^2 B_k}\frac{B_k + (B_k - L_f)\tau_k}{B_{k+1}} = \frac{1}{\tau_{k-1}^2 B_k}. \tag{3.5}$$

Next the definitions of $(\tau_k)_{k\in\mathbb{N}}$ and $(\beta_k)_{k\in\mathbb{N}}$ imply that

$$(\forall k \in \mathbb{N}) \quad \alpha_{k+1}\tau_{k+1}^3 + \tau_{k+1}^2 + \tau_k^2\tau_{k+1} - \tau_k^2 = 0 \quad \text{and} \quad (1+\tau_k)\beta_{k+1} = \beta_k, \tag{3.6}$$

where we define

$$(\forall k \in \mathbb{N}) \quad \alpha_{k+1} = \frac{B_{k+1} - L_f}{B_{k+1}} \in [0,1]. \tag{3.7}$$

For $\alpha \in [0,1]$ and $\tau > 0$, let us consider the following cubic function

$$(\forall t \in \mathbb{R}_+) \quad P(t) = \alpha t^3 + t^2 + \tau^2 t - \tau^2. \tag{3.8}$$

On the one hand, since $(\forall t > 0)\ P'(t) > 0$ and $P(0) = -\tau^2 < 0$ and $P(1) = \alpha + 1 > 0$, we deduce that P has a unique root $t_+ \in]0,1[$. On the other hand, let us define

$$(\forall t \in \mathbb{R}_+) \quad Q(t) = t^2 + \tau^2 t - \tau^2. \tag{3.9}$$

Then $(\forall t \in \mathbb{R}_+)\ Q(t) \leqslant P(t)$, and in particular, $P(t^\star) \geqslant Q(t^\star) = 0 = P(t_+)$, where $t^\star$ is the unique positive root of Q, that is $t^\star = (-\tau^2 + \sqrt{\tau^4 + 4\tau^2})/2$. As P is nondecreasing on $]0, +\infty[$, we get $t_+ \leqslant t^\star$. Consequently, since $\tau_0 > 0$, we deduce that $(\tau_k)_{k\in\mathbb{N}}$ is well-defined and furthermore,

$$(\forall k \in \mathbb{N}) \quad \tau_{k+1} \leqslant \frac{-\tau_k^2 + \sqrt{\tau_k^4 + 4\tau_k^2}}{2}. \tag{3.10}$$

This inequality and induction on $k \in \mathbb{N}$ easily yields

$$(\forall k \in \mathbb{N}) \quad \tau_k \leqslant \frac{2}{k+2}. \tag{3.11}$$

Now the two equalities in (3.6) imply that

$$(\forall k \in \mathbb{N}) \quad (\beta_k - \beta_{k+1}) = \beta_{k+1}\tau_k \quad \text{and} \quad \tau_k^2 = \tau_{k+1}^2 \frac{1 + \alpha_{k+1}\tau_{k+1}}{1 - \tau_{k+1}} < \tau_{k+1}^2 \frac{1 + \tau_{k+1}}{1 - \tau_{k+1}}. \tag{3.12}$$

We show by induction that

$$(\forall k \in \mathbb{N}) \quad \tau_k \geqslant \frac{1}{k+1}. \tag{3.13}$$

Note that $\tau_0 = 1 \geqslant \frac{1}{0+1}$. Suppose that there exists $k_0 \in \mathbb{N}$ such that $\tau_{k_0} \geqslant \frac{1}{k_0+1}$ and that $\tau_{k_0+1} < \frac{1}{k_0+2}$. Then we deduce from (3.12) that

$$\frac{1}{(k_0+1)^2} \leqslant \tau_{k_0}^2 < \tau_{k_0+1}^2 \frac{1 + \tau_{k_0+1}}{1 - \tau_{k_0+1}} < \frac{1}{(k_0+2)^2} \frac{1 + \frac{1}{k_0+2}}{1 - \frac{1}{k_0+2}} = \frac{1}{(k_0+2)^2} \frac{k_0+2}{k_0+1} \tag{3.14}$$

which is equivalent to $(k_0+2)^2 < (k_0+1)(k_0+2)$ which never happens. Hence, (3.13) holds true. We then deduce from induction that

$$(\forall k \in \mathbb{N}) \quad \beta_{k+1} = \frac{\beta_k}{1 + \tau_k} \leqslant \beta_k \frac{k+1}{k+2} \leqslant \beta_0 \prod_{l=0}^{k} \frac{l+1}{l+2} = \frac{\beta_0}{k+2}. \tag{3.15}$$

Of course $\beta_0 \leqslant \frac{\beta_0}{0+1}$ and hence,

$$(\forall k \in \mathbb{N}) \quad \beta_k \leqslant \frac{\beta_0}{k+1}. \tag{3.16}$$

It again follows from induction and from (3.13) that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad \tau_k^2 B_{k+1} = (1 - \tau_k)\tau_{k-1} B_k = \ldots = \prod_{l=1}^{k} (1 - \tau_l)\tau_0^2 B_1 \\ \leqslant \prod_{l=1}^{k} \left(1 - \frac{1}{l+1}\right) \tau_0^2 B_1 = \tau_0^2 B_1 \prod_{l=1}^{k} \frac{l}{l+1} = \frac{B_1}{k+1}.\end{aligned} \tag{3.17}$$

□

The convergence of Algorithm 3 is based on the following asymptotic property of parameters.

Lemma 3.4 *Let $(\tau_k)_{k\in\mathbb{N}}$, $(\beta_k)_{k\in\mathbb{N}}$ and $(B_{k+1})_{k\in\mathbb{N}}$ be the positive sequences determined by Algorithm 3. Then $(\forall k \in \mathbb{N})$ $\tau_k \leqslant \frac{2}{k+2}$ and*

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad \frac{\beta_0}{2B_1}\tau_k^2 B_{k+1} \leqslant \beta_{k+1} \leqslant \frac{1}{2}\left(3a\frac{\beta_0}{B_1}\tau_k^2 L_f + \sqrt{(3a\frac{\beta_0}{B_1}\tau_k^2 L_f)^2 + 12a\frac{\beta_0}{B_1}\tau_k^2 \|M\|_{S^{-1}}^2}\right) \\ \sim \sqrt{3a\frac{\beta_0}{B_1}}\|M\|_{S^{-1}}\tau_k = \mathcal{O}\left(\frac{1}{k}\right).\end{aligned} \tag{3.18}$$

Proof. For every $k \in \mathbb{N}$, since Lemma 2.4 (ii) states that the function $y \mapsto h_{\beta_{k+1}}(y; \dot{y})$ has $\frac{1}{\beta_{k+1}}$-Lipschitz gradient, we deduce that the function $x \mapsto f(x) + h_{\beta_{k+1}}(Mx; \dot{y})$ has $L_f + \frac{\|M\|^2_{S^{-1}}}{\beta_{k+1}}$-Lipschitz gradient and thus, in Algorithm 3, when the line search terminates, one necessarily has $B_{k+1} \leqslant a(L_f + \frac{\|M\|^2_{S^{-1}}}{\beta_{k+1}})$. Now we deduce from the definition of $(\tau_k)_{k\in\mathbb{N}}$ that

$$(\forall k \in \mathbb{N}) \quad \tau_k^2 B_{k+1} = (1-\tau_k)\tau_{k-1}^2 B_k = \tau_0^2 B_1 \prod_{l=1}^{k}(1-\tau_l), \tag{3.19}$$

so we need to study the bounds for $(\tau_k)_{k\in\mathbb{N}}$. We now prove by induction that

$$(\forall k \in \mathbb{N}) \quad \tau_k \leqslant \frac{2}{k+2}. \tag{3.20}$$

We clearly have $\tau_0 = 1 \leqslant \frac{2}{0+2}$. Suppose that there exists $k_0 \in \mathbb{N}\backslash\{0\}$ such that $\tau_{k_0-1} \leqslant \frac{2}{k_0+1}$ and that $\tau_{k_0} > \frac{2}{k_0+2}$. Then, as $B_{k_0+1} \geqslant B_{k_0}$, we have $\frac{1-\tau_{k_0}}{\tau_{k_0}^2} = \frac{B_{k_0+1}}{\tau_{k_0-1}^2 B_{k_0}} \geqslant \frac{1}{\tau_{k_0-1}^2}$. This would lead to

$$\frac{k_0^2+2k_0+1}{4} = \frac{(k_0+1)^2}{4} \leqslant \frac{1}{\tau_{k_0-1}^2} \leqslant \frac{1-\tau_{k_0}}{\tau_{k_0}^2} < \frac{(k_0+2)^2}{4} - \frac{k_0+2}{2} = \frac{k_0^2+2k_0}{4}. \tag{3.21}$$

This contradiction proves (3.20). Since we have

$$(\forall k \in \mathbb{N}) \quad \beta_{k+1} = \beta_1 \prod_{l=1}^{k} \frac{1}{1+\tau_l} = \beta_1 \frac{\prod_{l=1}^{k}(1-\tau_l)}{\prod_{l=1}^{k}(1-\tau_l^2)} = \frac{\beta_0}{2}\frac{\prod_{l=1}^{k}(1-\tau_l)}{\prod_{l=1}^{k}(1-\tau_l^2)}, \tag{3.22}$$

and since

$$(\forall k \in \mathbb{N}) \quad 1 \geqslant \prod_{l=1}^{k}(1-\tau_l^2) \geqslant \prod_{l=1}^{k}\left(1 - \frac{4}{(l+2)^2}\right) = \frac{(k+4)(k+3)2.1}{(k+2)(k+1)4.3} \geqslant \frac{1}{6}, \tag{3.23}$$

we get

$$(\forall k \in \mathbb{N}) \quad \frac{\beta_0}{2}\prod_{l=1}^{k}(1-\tau_l) \leqslant \beta_{k+1} \leqslant 3\beta_0 \prod_{l=1}^{k}(1-\tau_l). \tag{3.24}$$

Therefore,

$$(\forall k \in \mathbb{N}) \quad \frac{\beta_0}{2}\frac{\tau_k^2 B_{k+1}}{B_1} \leqslant \beta_{k+1} \leqslant 3\beta_0 \frac{\tau_k^2 B_{k+1}}{B_1} \tag{3.25}$$

Since $(\forall k \in \mathbb{N})\ B_{k+1} \leqslant a(L_f + \frac{\|M\|^2_{S^{-1}}}{\beta_{k+1}})$, it follows that

$$(\forall k \in \mathbb{N}) \quad \beta_{k+1}^2 - 3a\frac{\beta_0}{B_1}\tau_k^2\left[L_f\beta_{k+1} + \|M\|^2_{S^{-1}}\right] \leqslant 0, \tag{3.26}$$

which implies that

$$(\forall k \in \mathbb{N}) \quad \beta_{k+1} \leqslant \frac{1}{2}\left(3a\frac{\beta_0}{B_1}\tau_k^2 L_f + \sqrt{\left[3a\frac{\beta_0}{B_1}\tau_k^2 L_f\right]^2 + 12a\frac{\beta_0}{B_1}\tau_k^2\|M\|_{S^{-1}}^2}\right) \sim \sqrt{3a\frac{\beta_0}{B_1}}\|M\|_{S^{-1}}\tau_k = \mathcal{O}\left(\frac{1}{k}\right) \tag{3.27}$$

and hence,

$$(\forall k \in \mathbb{N}) \quad \tau_k^2 B_{k+1} \leqslant \frac{2B_1}{\beta_0}\beta_{k+1} = \mathcal{O}\left(\frac{1}{k}\right). \tag{3.28}$$

□

3.3 Speed of convergence

The convergence theorem is based on the decrease of the smoothed optimality gap. We prove in Proposition 3.5 that for every iteration $k \in \mathbb{N}$, $F_{\beta_{k+1}} - F^\star$ decreases as $\mathcal{O}(1/k)$. Then, using [13, Lemma 2.1] and the decrease of the smoothness parameter to 0, we get the speed of convergence in function value and infeasibility.

Proposition 3.5 *Consider the setting of Problem 1.1. Let $(\bar{x}^k)_{k\in\mathbb{N}}$ be generated by the ASGARD variants (Algorithms 1, 2, 3) and define*

$$(\forall k \in \mathbb{N}) \quad F_{\beta_k} \colon x \mapsto f(x) + g(x) + h_{\beta_k}(Mx; \dot{y}). \tag{3.29}$$

Then for $x^\star \in \mathcal{P}^\star$, we have:

$$(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star \leqslant \begin{cases} \frac{B_1}{2(k+1)}\|x^\star - \tilde{x}^0\|^2, & \text{for Algorithm 1,} \\ \frac{B_1}{k+1}\left(\frac{1}{2}\|x^\star - \tilde{x}^0\|^2 + \sigma\left(\frac{B_1}{2L_f}\right)^{1+\delta}\zeta(1+\delta)\right), & \text{for Algorithm 2,} \\ \frac{B_1\beta_{k+1}}{\beta_0}\|x^\star - \tilde{x}^0\|^2 = \mathcal{O}\left(\frac{\|x^\star - \tilde{x}_0\|^2}{k}\right), & \text{for Algorithm 3,} \end{cases} \tag{3.30}$$

where $\zeta(s) = \sum_{i=0}^{+\infty}\frac{1}{(i+1)^s}$.

Proof. First we note that the arguments for Algorithms 1 and 3 are similar to those of Algorithm 2 by setting $(\forall k \in \mathbb{N})\ \hat{\hat{x}}^k = \hat{x}^k$. We therefore prove the convergence for Algorithm 2. Now let us fix $x^\star \in \mathcal{P}^\star$. Then, we have

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad f(\bar{x}^{k+1}) + h_{\beta_{k+1}}(M\bar{x}^{k+1}; \dot{y}) &\leqslant f(\hat{\hat{x}}^k) + h_{\beta_{k+1}}(M\hat{x}^k; \dot{y}) + \left\langle \bar{x}^{k+1} - \hat{\hat{x}}^k, \nabla f(\hat{\hat{x}}^k)\right\rangle \\ &+ \left\langle \bar{x}^{k+1} - \hat{x}^k, v^k)\right\rangle + \frac{L_f}{2}\|\bar{x}^{k+1} - \hat{\hat{x}}^k\|^2 + \frac{\|M\|_{S^{-1}}^2}{2\beta_{k+1}}\|\bar{x}^{k+1} - \hat{x}^k\|^2. \end{aligned} \tag{3.31}$$

Because g is convex and because Algorithm 2 yield $(\forall k \in \mathbb{N})\ \bar{x}^{k+1} = (1-\tau_k)\bar{x}^k + \tau_k \tilde{x}^{k+1}$, we get

$$(\forall k \in \mathbb{N}) \quad g(\bar{x}^{k+1}) \leqslant (1-\tau_k) g(\bar{x}^k) + \tau_k g(\tilde{x}^{k+1}). \tag{3.32}$$

It now follows from Lemma 2.2 that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad g(\tilde{x}^{k+1}) &\leqslant g(x^\star) - \tau_k B_{k+1}\big\langle x^\star - \tilde{x}^{k+1}, \tilde{x}^k - \tilde{x}^{k+1} - \tau_k^{-1} B_{k+1}^{-1}\big(\nabla f(\hat{\tilde{x}}^k) + v^k\big)\big\rangle \\ &= g(x^\star) - \tau_k B_{k+1}\big\langle x^\star - \tilde{x}^{k+1}, \tilde{x}^k - \tilde{x}^{k+1}\big\rangle + \big\langle x^\star - \tilde{x}^{k+1}, \nabla f(\hat{\tilde{x}}^k) + v^k\big\rangle \\ &= g(x^\star) + \big\langle x^\star - \tilde{x}^{k+1}, \nabla f(\hat{\tilde{x}}^k) + v^k\big\rangle \\ &\quad + \frac{\tau_k B_{k+1}}{2}\|x^\star - \tilde{x}^k\|^2 - \frac{\tau_k B_{k+1}}{2}\|x^\star - \tilde{x}^{k+1}\|^2 - \frac{\tau_k B_{k+1}}{2}\|\tilde{x}^{k+1} - \tilde{x}^k\|^2.\end{aligned} \tag{3.33}$$

In turn, we obtain from (3.31) and the fact that $(\forall k \in \mathbb{N})\ \bar{x}^{k+1} - \hat{\tilde{x}}^k = \hat{x}^k - \hat{\tilde{x}}^k + \tau_k(\tilde{x}^{k+1} - \tilde{x}^k)$ that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1}) &\leqslant f(\hat{\tilde{x}}^k) + h_{\beta_{k+1}}(M\hat{x}^k; \dot{y}) + (1-\tau_k) g(\bar{x}^k) + \tau_k g(x^\star) + \tau_k\big\langle x^\star - \tilde{x}^k, v^k\big\rangle \\ &\quad + \big\langle \hat{x}^k - \hat{\tilde{x}}^k + \tau_k(x^\star - \tilde{x}^k), \nabla f(\hat{\tilde{x}}^k)\big\rangle + \frac{\tau_k^2 B_{k+1}}{2}\|x^\star - \tilde{x}^k\|^2 \\ &\quad - \frac{\tau_k^2 B_{k+1}}{2}\|x^\star - \tilde{x}^{k+1}\|^2 + \frac{L_f}{2}\|\bar{x}^{k+1} - \hat{\tilde{x}}^k\|^2 - \frac{L_f}{2}\|\bar{x}^{k+1} - \hat{x}^k\|^2.\end{aligned} \tag{3.34}$$

Since Algorithm 2 yield $(\forall k \in \mathbb{N})\ \tau_k(\hat{x}^k - \tilde{x}^k) = (1-\tau_k)(\bar{x}^k - \hat{x}^k)$, we have

$$(\forall k \in \mathbb{N}) \quad \tau_k(x^\star - \tilde{x}^k) = \tau_k(x^\star - \hat{x}^k) + (1-\tau_k)(\bar{x}^k - \hat{x}^k), \tag{3.35}$$

and hence,

$$(\forall k \in \mathbb{N}) \quad \tau_k\big\langle x^\star - \tilde{x}^k, v^k\big\rangle = \tau_k\big\langle x^\star - \hat{x}^k, v^k\big\rangle + (1-\tau_k)\big\langle \bar{x}^k - \hat{x}^k, v^k\big\rangle. \tag{3.36}$$

Moreover, since $(\forall k \in \mathbb{N})\ \hat{x}^k - \hat{\tilde{x}}^k + \tau_k(x^\star - \tilde{x}^k) = \tau_k(x^\star - \hat{\tilde{x}}^k) + (1-\tau_k)(\bar{x}^k - \hat{\tilde{x}}^k)$, we get

$$(\forall k \in \mathbb{N}) \quad \big\langle \hat{x}^k - \hat{\tilde{x}}^k + \tau_k(x^\star - \tilde{x}^k), \nabla f(\hat{\tilde{x}}^k)\big\rangle = \tau_k\big\langle x^\star - \hat{\tilde{x}}^k, \nabla f(\hat{\tilde{x}}^k)\big\rangle + (1-\tau_k)\big\langle \bar{x}^k - \hat{\tilde{x}}^k, \nabla f(\hat{\tilde{x}}^k)\big\rangle. \tag{3.37}$$

Because f is convex, (1.10) yields

$$(\forall k \in \mathbb{N}) \quad f(\hat{\tilde{x}}^k) + \big\langle \bar{x}^k - \hat{\tilde{x}}^k, \nabla f(\hat{\tilde{x}}^k)\big\rangle \leqslant f(\bar{x}^k) \quad \text{and} \quad f(\hat{\tilde{x}}^k) + \big\langle x^\star - \hat{\tilde{x}}^k, \nabla f(\hat{\tilde{x}}^k)\big\rangle \leqslant f(x^\star), \tag{3.38}$$

and hence, we derive from (2.9) that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad &f(\hat{\tilde{x}}^k) + h_{\beta_{k+1}}(M\hat{x}^k; \dot{y}) + \big\langle \bar{x}^k - \hat{\tilde{x}}^k, \nabla f(\hat{\tilde{x}}^k)\big\rangle + \big\langle \bar{x}^k - \hat{x}^k, v^k\big\rangle \\ &\leqslant f(\bar{x}^k) + h_{\beta_{k+1}}(M\bar{x}^k; \dot{y}) - \frac{\beta_{k+1}}{2}\|\nabla h_{\beta_{k+1}}(M\hat{x}^k; \dot{y}) - \nabla h_{\beta_{k+1}}(M\bar{x}^k; \dot{y})\|_S^2.\end{aligned} \tag{3.39}$$

and from (2.10) that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad &f(\hat{\tilde{x}}^k) + h_{\beta_{k+1}}(M\hat{x}^k; \dot{y}) + \big\langle x^\star - \hat{\tilde{x}}^k, \nabla f(\hat{\tilde{x}}^k)\big\rangle + \big\langle x^\star - \hat{x}^k, v^k\big\rangle \\ &\leqslant f(x^\star) + h(Mx^\star) - \frac{\beta_{k+1}}{2}\|\nabla h_{\beta_{k+1}}(M\hat{x}^k; \dot{y}) - \dot{y}\|_S^2.\end{aligned} \tag{3.40}$$

On the other hand, we deduce from (2.8) that

$$(\forall k \in \mathbb{N}) \quad h_{\beta_{k+1}}(M\bar{x}^k;\dot{y}) \leqslant h_{\beta_k}(M\bar{x}^k;\dot{y}) + \frac{\beta_k-\beta_{k+1}}{2}\|\nabla h_{\beta_{k+1}}(M\bar{x}^k;\dot{y})-\dot{y}\|_S^2 \tag{3.41}$$

and from (2.11) that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad -(1-\tau_k)\frac{\beta_{k+1}}{2}\|\nabla h_{\beta_{k+1}}(M\hat{x}^k;\dot{y})-\nabla h_{\beta_{k+1}}(M\bar{x}^k;\dot{y})\|_S^2-\tau_k\frac{\beta_{k+1}}{2}\|\nabla h_{\beta_{k+1}}(M\hat{x}^k;\dot{y})-\dot{y}\|_S^2 \\ \leqslant -\tau_k(1-\tau_k)\frac{\beta_{k+1}}{2}\|\nabla h_{\beta_{k+1}}(M\bar{x}^k;\dot{y})-\dot{y}\|_S^2 \end{aligned} \tag{3.42}$$

Altogether, by combining (3.34) and (3.39)-(3.41), we get

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1}) &\leqslant (1-\tau_k)\big(g(\bar{x}^k)+f(\hat{\hat{x}}^k)+h_{\beta_{k+1}}(M\hat{x}^k;\dot{y})+\big\langle \bar{x}^k-\hat{\hat{x}}^k,\nabla f(\hat{\hat{x}}^k)\big\rangle+\big\langle \bar{x}^k-\hat{x}^k,v^k\big\rangle\big) \\ &\quad+\tau_k\big(g(x^\star)+f(\hat{x}^k)+h_{\beta_{k+1}}(M\hat{x}^k;\dot{y})+\big\langle x^\star-\hat{\hat{x}}^k,\nabla f(\hat{\hat{x}}^k)\big\rangle+\big\langle x^\star-\hat{x}^k,v^k\big\rangle\big) \\ &\quad+\tau_k^2\frac{B_{k+1}}{2}\|x^\star-\tilde{x}^k\|^2-\tau_k^2\frac{B_{k+1}}{2}\|x^\star-\tilde{x}^{k+1}\|^2+\frac{L_f}{2}\|\bar{x}^{k+1}-\hat{\hat{x}}^k\|^2-\frac{L_f}{2}\|\bar{x}^{k+1}-\hat{x}^k\|^2 \\ &\leqslant (1-\tau_k)\big(g(\bar{x}^k)+f(\bar{x}^k)+h_{\beta_k}(M\bar{x}^k;\dot{y})\big)+\tau_k\big(g(x^\star)+f(x^\star)+h(Mx^\star)\big) \\ &\quad+\Big[\frac{(1-\tau_k)(\beta_k-\beta_{k+1})}{2}-\frac{\tau_k(1-\tau_k)\beta_{k+1}}{2}\Big]\|\nabla h_{\beta_{k+1}}(M\bar{x}^k;\dot{y})-\dot{y}\|_S^2 \\ &\quad+\tau_k^2\frac{B_{k+1}}{2}\|x^\star-\tilde{x}^k\|^2-\tau_k^2\frac{B_{k+1}}{2}\|x^\star-\tilde{x}^{k+1}\|^2+\frac{L_f}{2}\|\bar{x}^{k+1}-\hat{\hat{x}}^k\|^2-\frac{L_f}{2}\|\bar{x}^{k+1}-\hat{x}^k\|^2. \end{aligned} \tag{3.43}$$

It follows from the definition of $(\tau_k)_{k\geqslant 0}$ and $(\beta_k)_{k\geqslant 0}$ that

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1})-F^\star+\frac{\tau_k^2B_{k+1}}{2}\|x^\star-\tilde{x}^{k+1}\|^2 \leqslant (1-\tau_k)(F_{\beta_k}(\bar{x}^k)-F^\star)+\frac{\tau_k^2B_{k+1}}{2}\|x^\star-\tilde{x}^k\|^2 \\ +\frac{L_f}{2}\|\bar{x}^{k+1}-\hat{\hat{x}}^k\|^2-\frac{L_f}{2}\|\bar{x}^{k+1}-\hat{x}^k\|^2. \end{aligned} \tag{3.44}$$

On the one hand, (3.44) yields

$$F_{\beta_1}(\bar{x}^1)-F^\star \leqslant \frac{\tau_0^2B_1}{2}\|x^\star-\tilde{x}^0\|^2+\tau_0^2B_1\sum_{i=0}^{0}\frac{L_f}{2\tau_i^2B_{i+1}}\big(\|\bar{x}^{i+1}-\hat{\hat{x}}^i\|^2-\|\bar{x}^{i+1}-\hat{x}^i\|^2\big). \tag{3.45}$$

On the other hand, since

$$(\forall k \in \mathbb{N}\backslash\{0\}) \quad \frac{1-\tau_k}{\tau_k^2B_{k+1}}=\frac{1}{\tau_{k-1}^2B_k}, \tag{3.46}$$

it follows from (3.44) and (3.45) that

$$
\begin{aligned}
(\forall k \in \mathbb{N}\backslash\{0\}) \quad & \frac{1}{\tau_k^2 B_{k+1}}(F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star) + \frac{1}{2}\|x^\star - \tilde{x}^{k+1}\|^2 \\
& \leqslant \frac{1-\tau_k}{\tau_k^2 B_{k+1}}(F_{\beta_k}(\bar{x}^k) - F^\star) + \frac{1}{2}\|x^\star - \tilde{x}^k\|^2 + \frac{L_f}{2\tau_k^2 B_{k+1}}\big(\|\bar{x}^{k+1} - \hat{\bar{x}}^k\|^2 - \|\bar{x}^{k+1} - \hat{x}^k\|^2\big) \\
& = \frac{1}{\tau_{k-1}^2 B_k}(F_{\beta_k}(\bar{x}^k) - F^\star) + \frac{1}{2}\|x^\star - \tilde{x}^k\|^2 + \frac{L_f}{2\tau_k^2 B_{k+1}}\big(\|\bar{x}^{k+1} - \hat{\bar{x}}^k\|^2 - \|\bar{x}^{k+1} - \hat{x}^k\|^2\big) \\
& \leqslant \frac{1}{\tau_0^2 B_1}(F_{\beta_1}(\bar{x}^1) - F^\star) + \frac{1}{2}\|x^\star - \tilde{x}^1\|^2 + \sum_{i=1}^{k} \frac{L_f}{2\tau_i^2 B_{i+1}}\big(\|\bar{x}^{i+1} - \hat{\bar{x}}^i\|^2 - \|\bar{x}^{i+1} - \hat{x}^i\|^2\big) \\
& \leqslant \frac{1}{2}\|x^\star - \tilde{x}^0\|^2 + \sum_{i=0}^{k} \frac{L_f}{2\tau_i^2 B_{i+1}}\big(\|\bar{x}^{i+1} - \hat{\bar{x}}^i\|^2 - \|\bar{x}^{i+1} - \hat{x}^i\|^2\big).
\end{aligned} \tag{3.47}
$$

Altogether, (3.45) and (3.47) yield

$$
(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star \leqslant \frac{\tau_k^2 B_{k+1}}{2}\|x^\star - \tilde{x}^0\|^2 + \tau_k^2 B_{k+1} \sum_{i=0}^{k} \frac{L_f}{2\tau_i^2 B_{i+1}}\big(\|\bar{x}^{i+1} - \hat{\bar{x}}^i\|^2 - \|\bar{x}^{i+1} - \hat{x}^i\|^2\big). \tag{3.48}
$$

We note that in the above inequalities when $(\forall k \in \mathbb{N})\ \hat{\bar{x}}^k = \hat{x}^k$ then the second term in the right hand side of the last line vanishes. Otherwise, the test

$$
\frac{1}{2}\|\bar{x}^{i+1} - \hat{\bar{x}}^i\|^2 - \frac{1}{2}\|\bar{x}^{i+1} - \hat{x}^i\|^2 \leqslant \sigma\big(\frac{\tau_i^2 B_{i+1}}{L_f}\big)^{2+\delta} \tag{3.49}
$$

ensures that the additional sum is uniformly bounded since $(\forall k \in \mathbb{N})\ \tau_k^2 B_{k+1} \in \mathcal{O}(1/k)$. To conclude, we combine this estimate with Lemma 3.3 and Lemma 3.4. □

We are now ready to state the convergence result for Algorithm 1. The convergence characterizations for the other variants of ASGARD follow mutatis mutandis for the other variants of ASGARD using the same arguments. As in [13], we consider two important particular cases: the case of equality constraints ($h = \iota_{\{c\}}$) and the case where h is Lipschitz continuous.

Theorem 3.6 *Let $(\bar{x}^k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 1.*

(i) *Suppose that $h = \iota_{\{c\}}$ for some $c \in \mathcal{G}$ and that $\operatorname{ri}(\operatorname{dom} g) \cap \{x \in \mathcal{H} \mid Mx = c\} \neq \emptyset$. Denote $F\colon x \mapsto f(x) + g(x)$. Then the following bounds hold for all $k \in \mathbb{N}$ and $(x^\star, y^\star) \in \mathcal{P}^\star \times \mathcal{D}^\star$:*

$$
\begin{aligned}
& F(\bar{x}^{k+1}) - F(x^\star) \geqslant -\|y^\star\|_S \|M\bar{x}^{k+1} - c\|_{S^{-1}} \\
& F(\bar{x}^{k+1}) - F(x^\star) \leqslant \frac{1}{k+1}\left(\frac{L_f}{2} + \frac{\|M\|_{S^{-1}}^2}{2\beta_0}\right)\|\tilde{x}^0 - x^\star\|^2 + \|y^\star\|_S \|M\bar{x}^{k+1} - c\|_{S^{-1}} + \frac{\beta_0}{2(k+1)}\|y^\star - \dot{y}\|_S^2 \\
& \|M\bar{x}^{k+1} - c\|_{S^{-1}} \leqslant \frac{\beta_0}{k+1}\left[\|y^\star - \dot{y}\|_S + \left(\|y^\star - \dot{y}\|_S^2 + \frac{2}{\beta_0}\left(\frac{L_f}{2} + \frac{\|M\|_{S^{-1}}^2}{2\beta_0}\right)\|\tilde{x}^0 - x^\star\|^2\right)^{1/2}\right].
\end{aligned}
$$

(3.50)

(ii) *Suppose that h is $D_{\mathcal{Y}}$-Lipschitz continuous in the S-norm, and denote $F(x) = f(x) + g(x) + h(Mx)$. Then the following bound holds*

$$(\forall k \in \mathbb{N}) \quad F(\bar{x}^{k+1}) - F^\star \leqslant \frac{1}{k+1}\left(\frac{L_f}{2} + \frac{\|M\|^2_{S^{-1}}}{2\beta_0}\right)\|\tilde{x}^0 - x^\star\|^2 + \frac{\beta_0}{k+1}\left[D^2_{\mathcal{Y}} + \|\dot{y}\|^2\right].$$

Proof. (i) Fix $k \in \mathbb{N}$. Using [13, Lemma 2.1], we get

$$\begin{aligned}
F(\bar{x}^{k+1}) - F^\star &\geqslant -\|y^\star\|_S\|M\bar{x}^{k+1} - c\|_{S^{-1}} \\
F(\bar{x}^{k+1}) - F^\star &\leqslant F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star + \|y^\star\|_S\|M\bar{x}^{k+1} - c\|_{S^{-1}} + \frac{\beta_{k+1}}{2}\|y^\star - \dot{y}\|^2_S \\
\|M\bar{x}^{k+1} - c\|_{S^{-1}} &\leqslant \beta_{k+1}\left[\|y^\star - \dot{y}\|_S + \left(\|y^\star - \dot{y}\|^2_S + 2\beta^{-1}_{k+1}(F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star)\right)^{1/2}\right].
\end{aligned}$$

The first inequality is now proved. By Proposition 3.5, $F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star \leqslant \frac{B_1}{2(k+1)}\|x^\star - \tilde{x}^0\|^2$ and by Lemma 3.3, $\beta_{k+1} \leqslant \frac{\beta_0}{k+2} \leqslant \frac{\beta_0}{k+1}$. Hence we get the second inequality. For the last inequality,

$$\begin{aligned}
&\beta_{k+1}\left[\|y^\star - \dot{y}\|_S + \left(\|y^\star - \dot{y}\|^2_S + 2\beta^{-1}_{k+1}(F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star)^{1/2}\right]\right. \\
&\qquad = \beta_{k+1}\|y^\star - \dot{y}\|_S + \left(\beta^2_{k+1}\|y^\star - \dot{y}\|^2_S + 2\beta_{k+1}(F_{\beta_k}(\bar{x}^{k+1}) - F^\star)\right)^{1/2} \qquad (3.51)
\end{aligned}$$

so we just need to use again the inequalities $F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star \leqslant \frac{B_1}{2(k+1)}\|x^\star - \tilde{x}^0\|^2$ and $\beta_{k+1} \leqslant \frac{\beta_0}{k+1}$ to conclude.

(ii) For the second case, since h is $D_{\mathcal{Y}}$-Lipschitz continuous, it follows from [1, Corollary 17.19] that $\operatorname{dom} h^* \subset B[0, D_{\mathcal{Y}}]$, the ball centered at 0 and with radius $D_{\mathcal{Y}}$. Therefore,

$$\begin{aligned}
(\forall k \in \mathbb{N}) \quad h(M\bar{x}^{k+1}) &= \sup_{y \in \operatorname{dom} h^*}\left\{\langle M\bar{x}^{k+1}, y\rangle - h^*(y)\right\} \\
&\leqslant \max_{y \in B[0,D_{\mathcal{Y}}]}\left\{\langle M\bar{x}^{k+1}, y\rangle - h^*(y)\right\} \\
&\leqslant \max_{y \in B[0,D_{\mathcal{Y}}]}\left\{\langle M\bar{x}^{k+1}, y\rangle - h^*(y) - \frac{\beta_{k+1}}{2}\|y - \dot{y}\|^2_S\right\} + \frac{\beta_{k+1}}{2}\max_{y \in B[0,D_{\mathcal{Y}}]}\|y - \dot{y}\|^2_S \\
&\leqslant h_{\beta_{k+1}}(M\bar{x}^{k+1}; \dot{y}) + \frac{\beta_{k+1}}{2}\max_{y \in B[0,D_{\mathcal{Y}}]}\|y - \dot{y}\|^2_S \\
&\leqslant h_{\beta_{k+1}}(M\bar{x}^{k+1}; \dot{y}) + \beta_{k+1}\left[\max_{y \in B[0,D_{\mathcal{Y}}]}\|y\|^2_S + \|\dot{y}\|^2_S\right] \\
&\leqslant h_{\beta_{k+1}}(M\bar{x}^{k+1}; \dot{y}) + \beta_{k+1}\left[D^2_{\mathcal{Y}} + \|\dot{y}\|^2_S\right].
\end{aligned} \qquad (3.52)$$

The conclusion now follows from Proposition 3.5. □

Finally, we extend the above approach for the following multivariate minimization problem.

Problem 3.7 Let m be a strictly positive integer, let $\mathcal{H}$ and $(\mathcal{G})_{1\leqslant i\leqslant m}$ be real Hilbert spaces, let $(\forall i \in \{1,\dots,m\})$ $M_i\colon \mathcal{H}\to\mathcal{G}_i$ be bounded linear operators, and let $f\colon \mathcal{H}\to\mathbb{R}$, $g\colon \mathcal{H}\to(-\infty,+\infty]$ and $h_i\colon \mathcal{G}_i\to(-\infty,+\infty]$ be proper, closed lower semi-continuous convex functions where f is moreover assumed to have L_f-Lipschitz gradient. Consider the following problem

$$F^\star = \inf_{x\in\mathcal{H}} \; f(x)+g(x)+\sum_{i=1}^m h_i(M_ix) \tag{3.53}$$

and suppose that its set of minimizers is non-empty.

For each $i\in\{1,\dots,m\}$, let us choose S_i to be a positive definite linear operator and $\dot{y}_i\in\mathcal{G}_i$. Define

$$(\forall \beta\in\left]0,+\infty\right[) \quad h_{\beta;i}(\cdot;\dot{y}_i)\colon y_i\mapsto \max_{\bar{y}_i\in\mathcal{G}_i}\{\langle y_i,\bar{y}_i\rangle - h_i^*(\bar{y}_i)-\frac{\beta}{2}\|\bar{y}_i-\dot{y}_i\|_{S_i}^2\}. \tag{3.54}$$

The following algorithm is an extension of Algorithm 1 to solve Problem 3.7. The other variants are similar.

Algorithm 4 Parallel ASGARD

Inputs: $\|M\|_{S^{-1}}^2=\sum_{i=1}^m\|M_i\|_{S_i^{-1}}^2$, $\tau_0=1$, $\beta_0>0$, $\bar{x}^0\in\mathcal{H}$, $\tilde{x}^0\in\mathcal{H}$.
for $k=0,1,\dots$ **do**
  $\hat{x}^k=(1-\tau_k)\bar{x}^k+\tau_k\tilde{x}^k$
  $\beta_{k+1}=\frac{\beta_k}{1+\tau_k}$ and $B_{k+1}=L_f+\frac{\|M\|_{S^{-1}}^2}{\beta_{k+1}}$
  for $i=1,\dots,m$ **do**
    $y_{k;i}^*=\operatorname*{argmax}_{y_i\in\mathcal{G}_i}\; \langle M_i\hat{x}^k,y_i\rangle - h_i^*(y_i)-\frac{\beta_{k+1}}{2}\|y_i-\dot{y}_i\|_{S_i}^2$
  end for
  $v^k=\sum_{i=1}^m M_i^*y_{k;i}^*$
  $\tilde{x}^{k+1}=\operatorname{prox}_{\frac{1}{\tau_kB_{k+1}}g}\big(\tilde{x}^k-\frac{1}{\tau_kB_{k+1}}(\nabla f(\hat{x}^k)+v^k)\big)$
  $\bar{x}^{k+1}=\hat{x}^k+\tau_k(\tilde{x}^{k+1}-\tilde{x}^k)=(1-\tau_k)\bar{x}^k+\tau_k\tilde{x}^{k+1}$
  Find the unique positive τ_{k+1} such that

$$\frac{B_{k+1}-L_f}{B_{k+1}}\tau_{k+1}^3+\tau_{k+1}^2+\tau_k^2\tau_{k+1}-\tau_k^2=0$$

end for
return $\bar{x}^{k+1}$

Proposition 3.8 *Consider the setting of Problem 3.7, let $(\bar{x}^k)_{k\in\mathbb{N}}$ be the sequence generated by Algorithm 4 and define*

$$(\forall k\in\mathbb{N}) \quad F_{\beta_k}\colon x\mapsto f(x)+g(x)+\sum_{i=1}^m h_{\beta_k;i}(M_ix;\dot{y}_i). \tag{3.55}$$

Then for any solution $x^\star$ to Problem 3.7, we have

$$(\forall k \in \mathbb{N}) \quad F_{\beta_{k+1}}(\bar{x}^{k+1}) - F^\star \leqslant \frac{B_1}{2(k+1)}\|x^\star - \tilde{x}^0\|^2. \tag{3.56}$$

Proof. Let $\mathcal{G} = \bigoplus_{i=1}^m \mathcal{G}_i$ be the Hilbert direct sum of $(\mathcal{G}_i)_{1\leqslant i\leqslant m}$. For $y = (y_i)_{1\leqslant i\leqslant m} \in \mathcal{G}$ and $\hat{y} = (\hat{y}_i)_{1\leqslant i\leqslant m} \in \mathcal{G}$, we define $\|y\| = \sqrt{\sum_{i=1}^m \|y_i\|^2}$ and $\langle y, \hat{y}\rangle = \sum_{i=1}^m \langle y_i, \hat{y}_i\rangle$. Set

$$\begin{cases} M\colon \mathcal{H} \to \mathcal{G}\colon x \mapsto (M_i x)_{1\leqslant i\leqslant m}, \\ h\colon \mathcal{G} \to (-\infty, +\infty]\colon (y_i)_{1\leqslant i\leqslant m} \mapsto \sum_{i=1}^m h_i(y_i), \\ S\colon \mathcal{G} \to (-\infty, +\infty]\colon (y_i)_{1\leqslant i\leqslant m} \mapsto \sum_{i=1}^m S_i y_i. \end{cases} \tag{3.57}$$

Then Problem 3.7 reduces to Problem 1.1. It is easy to see that S is a positive definite operator on $\mathcal{G}$ and it induce the norm

$$(\forall y = (y_i)_{1\leqslant i\leqslant m} \in \mathcal{G}) \quad \|y\|_S = \sqrt{\sum_{i=1}^m \|y_i\|_{S_i}^2}. \tag{3.58}$$

Moreover, for any $\beta \in]0, +\infty[$ and any $y \in \mathcal{G}$, we have

$$\begin{aligned} h_\beta(y; \dot{y}) &= \max_{\bar{y}\in\mathcal{G}} \left\{\langle y, \bar{y}\rangle - h^*(\bar{y}) - \frac{\beta}{2}\|\bar{y} - \dot{y}\|_S^2\right\} \\ &= \max_{\bar{y}\in\mathcal{G}} \left\{\sum_{i=1}^m \langle y_i, \bar{y}_i\rangle - \sum_{i=1}^m h_i^*(\bar{y}_i) - \frac{\beta}{2}\sum_{i=1}^m \|\bar{y}_i - \dot{y}_i\|_{S_i}^2\right\} \\ &= \sum_{i=1}^m \max_{\bar{y}_i\in\mathcal{G}_i} \left\{\langle y_i, \bar{y}_i\rangle - h_i^*(\bar{y}_i) - \frac{\beta}{2}\|\bar{y}_i - \dot{y}_i\|_{S_i}^2\right\} \\ &= \sum_{i=1}^m h_{\beta;i}(y_i; \dot{y}_i), \end{aligned} \tag{3.59}$$

where $\dot{y} = (\dot{y}_1, \ldots, \dot{y}_m)$. The assertion then follows from Proposition 3.5. □

3.4 Restarting

It is possible to restart our variants of ASGARD using a fixed iteration restarting strategy, i.e., restart every q iterations, as follows:

$$\begin{cases} \hat{x}^{k+1} \leftarrow \bar{x}^{k+1}, \\ \dot{y} \leftarrow y^*_{\beta_{k+1}}(M\hat{x}^k; \dot{y}), \\ \beta_{k+1} \leftarrow \beta_0, \\ \tau_{k+1} \leftarrow 1. \end{cases} \tag{3.60}$$

The performance of different variants of ASGARD with restarting will be illustrated in the next section.

4 Numerical experiments

4.1 Sparse and TV regularized least squares

We consider the following regularized least squares problem

$$\min_{x\in\mathbb{R}^{100}} \frac{1}{2}\|Ax-b\|^2 + \|x\|_1 + \|Dx\|_1$$

where A is a randomly generated matrix of size 50 $\times$ 100 (Gaussian distribution, covariance $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.95$), b is randomly generated (b_i iid, with uniform distribution on $[1,2]$) and D is the explicit 1D discrete gradient operator. This problem is a special case of (1.1) with

$$f(x) = \frac{1}{2}\|Ax-b\|^2, \quad g(x) = \|x\|_1, \quad h(x) = \|x\|_1 \quad \text{and} \quad M = D. \tag{4.1}$$

In this case,

$$\text{prox}_{\gamma g}(x) = \text{soft}_{[-\gamma,\gamma]}(x) \quad \text{and} \quad \text{prox}_{\gamma h^*}(x) = x - \gamma\,\text{soft}_{[-\gamma^{-1},\gamma^{-1}]}(\gamma^{-1}x), \tag{4.2}$$

where

$$\text{soft}_{[-t,t]}(x) = \text{sign}(x) \otimes \max\{|x| - t, 0\}, \tag{4.3}$$

here $\otimes$ denotes component-wise multiplication.

When the plot has dash-dotted line, we consider the constraint $z = Dx$ and the augmented primal variable (x, z), otherwise, we directly split with $h = \|\cdot\|_1$. For ASGARD with restart, we restart the momentum in the algorithm every 100 iterations; vu-condat is Vu-Condat's algorithm and ladmm is the linearized ADMM method. For each algorithm we plot the difference between the current function value and the best function value encountered in the experiment (Figure 1).

We also considered a medium-scale sparse and TV regularized problem on functional MRI data [12]. For given regularization parameters $\alpha > 0$ and $r \in [0,1]$, we would like to solve the following regression problem with regularization given by the sum of Total Variation (TV) and the ℓ_1 norm:

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax-b\|_2^2 + \alpha\big(r\|x\|_1 + (1-r)\|Mx\|_{2,1}\big).$$

The problem takes place on a 3D image of the brains of size $40\times 48\times 34$. The optimization variable x is a real vector with one entry in each voxel, that is n = 65280. Matrix M is the discretized 3D gradient. This is a sparse matrix of size 195840 $\times$ 65280 with 2 nonzero elements in each row. The matrix $A \in \mathbb{R}^{768\times 65280}$ and the vector $b \in \mathbb{R}^{768}$ correspond to 768 labeled experiments where each line of A gathers brains activity for the corresponding experiment. Parameter r tunes the tradeoff between the two regularization terms. We chose $r = 0.1$ and $\alpha = 0.1$.

In this scenario, we set the objective as $f(x) = \frac{1}{2}\|Ax-b\|_2^2$, $g(x) = \alpha r\|x\|_1$ and $h(y) = \alpha(1-r)\|y\|_{2,1}$. On Figure 2, we compared our algorithms against FISTA [3] with an inexact resolution of

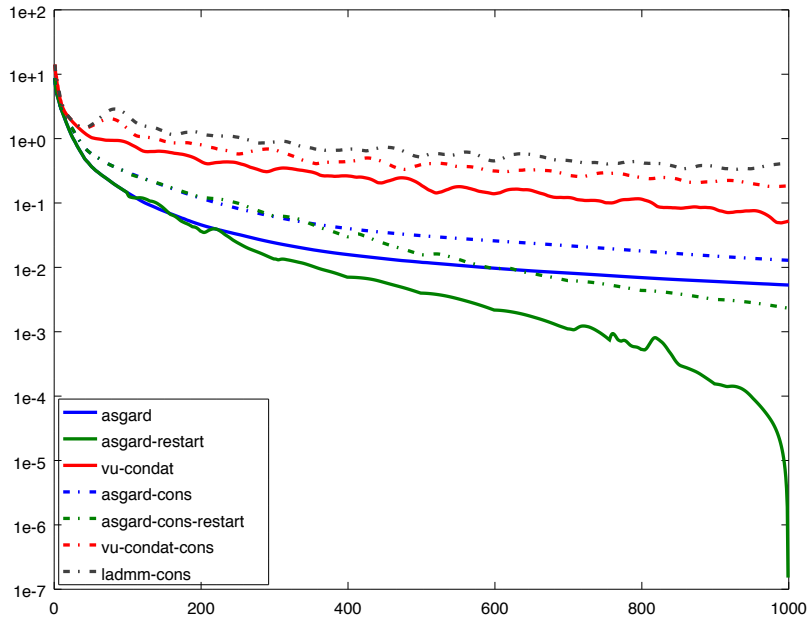

Figure 1: Behavior of various algorithms for the synthetic sparse and TV regularized least squares problem: we plot the difference between the current function value and the best function value encountered vs iterations.

the proximal operator of TV, FISTA restarted every 30 iterations, and Vũ-Condat's algorithm [14, 6]. We can see that on this problem, ASGARD outperforms Vũ-Condat's algorithm but not FISTA. After careful inspection, we realize that ASGARD (and also Vũ-Condat's algorithm) spends too much time computing gradients of f while FISTA spends much of its time to compute the proximity operator of g (Figures 3 and 4). Our framework allows us to consider useful variants in this setting. For instance, the use of old gradients makes ASGARD much faster in this problem in time. We can see on Figure 2 that `ASGARD_Old_Gradients` outperforms FISTA. Moreover, combined with a restart every 400 iterations, we obtain the best performance among the algorithms we test.

4.2 Quantum properties prediction

In materials science, quantum properties such as energy requires expensive calculations based on the density functional theory (DFT). Machine learning has been recently used to predict such properties for new molecules based on dataset derived by DFT. Let us represent the dataset by $\{(r_i, p_i)\}_{i=1}^N$ where $r_i \in \mathbb{R}^n$ is Coulomb matrix representation [11] of i-th molecule and $p_i \in \mathbb{R}$ is its properties. In this experiment, the Laplacian kernel, i.e., $K(r, r') = \exp(-\|r - r'\|_1/\sigma)$ with $\|\cdot\|_1$ is ℓ_1-norm of $\mathbb{R}^n$, is used to measure the dissimilarity between molecules. A quantum property of a

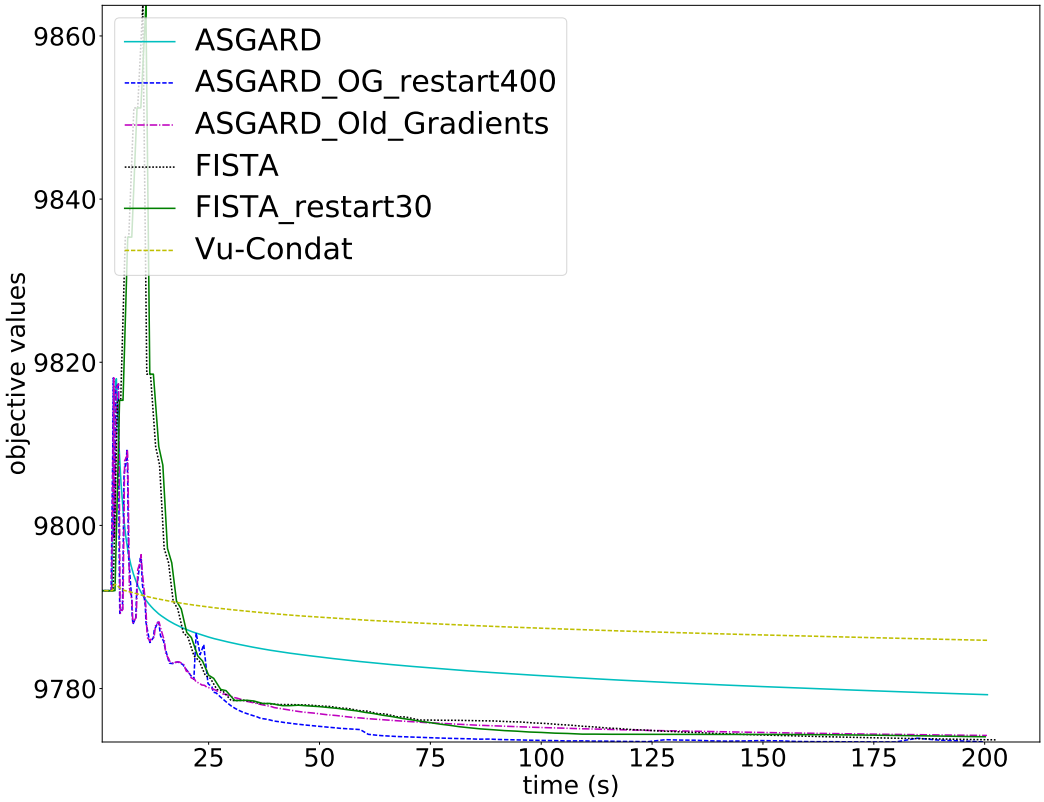

Figure 2: Comparison of various algorithms for the functional MRI problem: function value against computational time.

molecule with representation r is assumed to have the following form

$$e(r) = \sum_{i=1}^{N} x_i K(r, r_i). \tag{4.4}$$

The regression coefficients $x = (x_1, \ldots, x_N)^T$ are obtained by solving the following elastic net regularized minimization problem

$$\underset{x \in \mathbb{R}^N}{\text{minimize}} \;\; \|Kx - p\|_1 + \frac{\lambda}{2} x^T K x + (1-\lambda)\|x\|_1, \tag{4.5}$$

here $p = (p_1, \ldots, p_N)^T$ and $K_{ij} = K(r_i, r_j)$. Note that (4.5) is a particular case of (1.1) with

$$f(x) = \frac{\lambda}{2} x^T K x, \quad g(x) = (1-\lambda)\|x\|_1, \quad h(y) = \|y - p\|_1 \quad \text{and} \quad M \colon x \mapsto Kx. \tag{4.6}$$

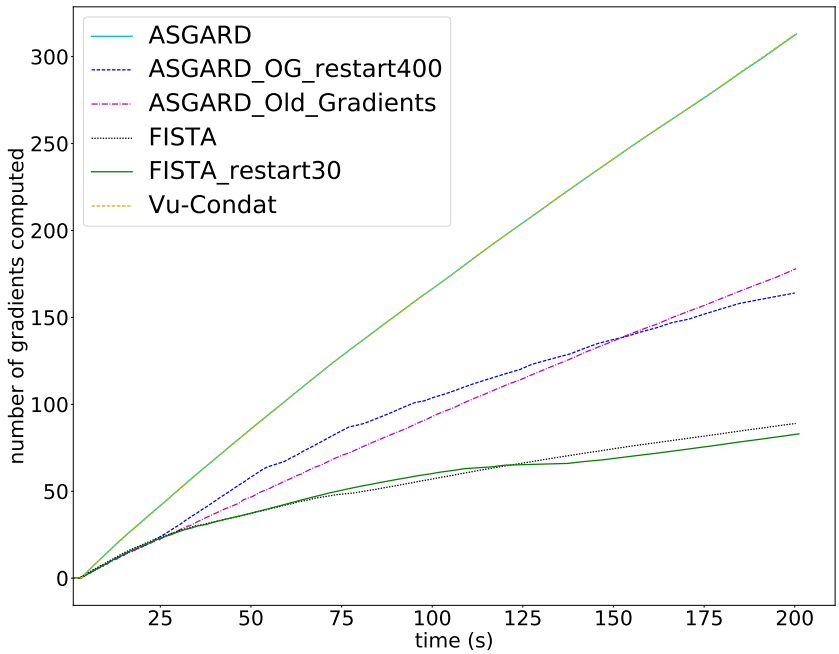

Figure 3: Comparison of various algorithms for the functional MRI problem: number of gradients evaluations against computational time.

In this case,

$$\text{prox}_{\gamma g}(x) = \text{soft}_{[-(1-\lambda)\gamma,(1-\lambda)\gamma]}(x) \quad \text{and} \quad \text{prox}_{\gamma h^*}(x) = x - \gamma\big(p + \text{soft}_{[-\gamma^{-1},\gamma^{-1}]}(\gamma^{-1}x - p)\big). \quad (4.7)$$

In Figure 5, we compare the behavior of different versions of our ASGARD with Vu-Condat's algorithm [6, 14] and Combettes-Pesquet's algorithm [4] on the dataset of 7211 molecules in [11] in which 50% molecules are used to train.[4].

5 Conclusion

In this paper, we build, based on the homotopy-based smoothing and acceleration technique of [ASGARD], a new method to solve a large class of generic convex optimization problems where the objective function is split into a sum of one smooth term and two non-smooth terms, one of which is combined with a linear operator. The variants of our method with line-search and old gradients benefits from the local smoothness of nonsmooth function and can avoid computing the whole gradient of the smooth function. In contrast to the existing methods in the literature, our method

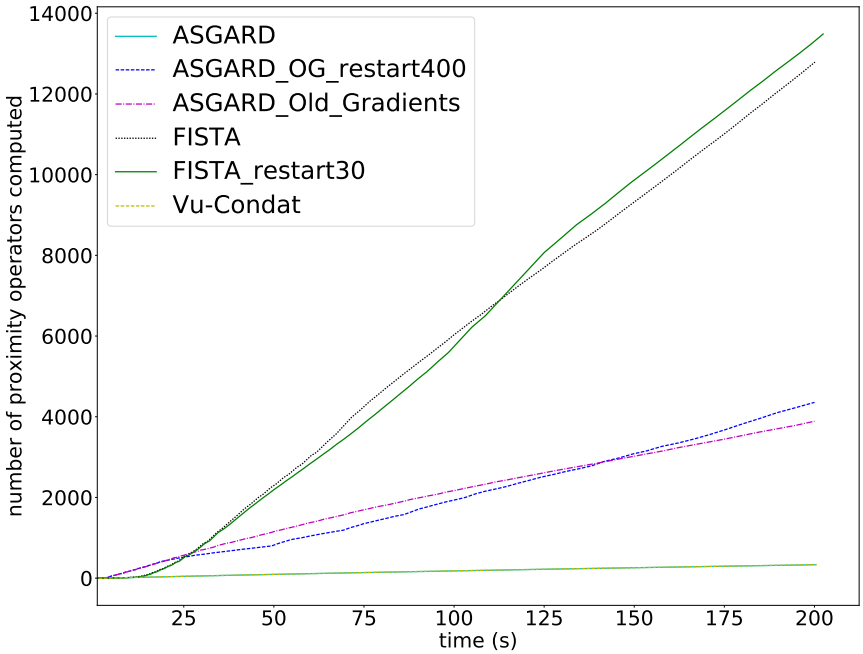

Figure 4: Comparison of various algorithms for the functional MRI problem: number of prox evaluations against computational time.

also features rigorous convergence guarantees. Numerical experiments with real-world problems illustrate the superiority of our method vs. the other state-of-the-art algorithms.

Acknowledgments.

The work of V. Cevher and Q. V. Nguyen is supported by the NCCR MARVEL, funded by the Swiss National Science Foundation. The work of O. Fercoq is supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH and PGMO.

References

[1] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, Cham, second edition, 2017. With a foreword by Hédy Attouch.

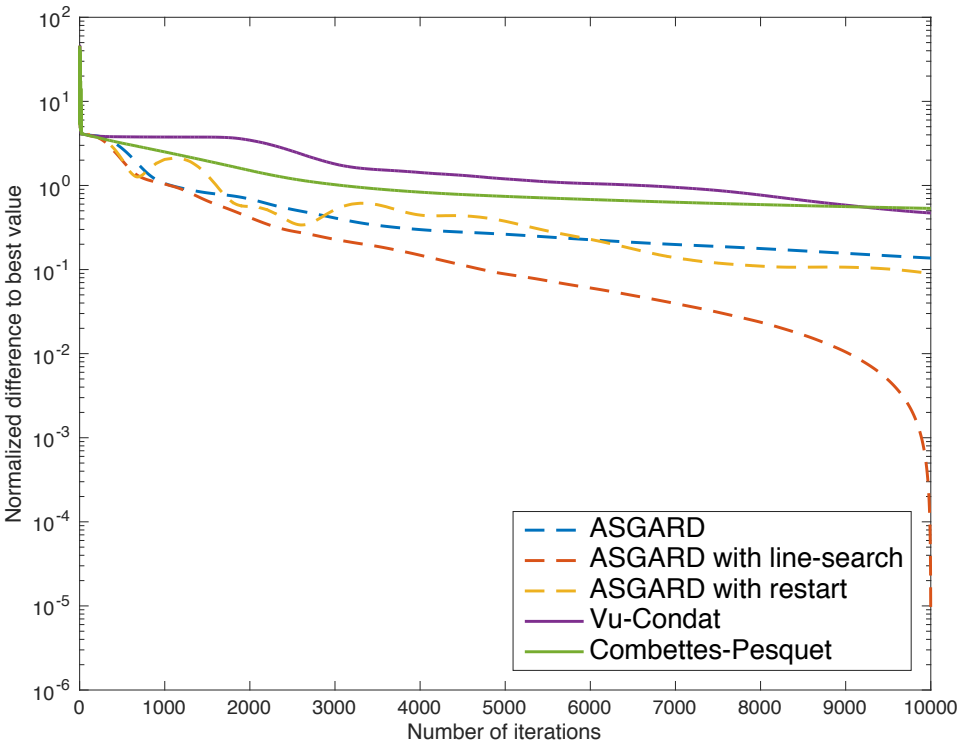

Figure 5: Comparison with existing algorithms ($\sigma = 4000$ and $\lambda = 0.001$).

[2] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[4] Patrick L. Combettes and Jean-Christophe Pesquet. Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. *Set-Valued Var. Anal.*, 20(2):307–330, 2012.

[5] Patrick L. Combettes and Băng C. Vũ. Variable metric forward-backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[6] Laurent Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158(2):460–479, 2013.

[7] Tom Goldstein, Brendan O'Donoghue, Simon Setzer, and Richard Baraniuk. Fast alternating direction optimization methods. *SIAM Journal on Imaging Sciences*, 7(3):1588–1623, 2014.

[8] Yura Malitsky and Thomas Pock. A first-order primal-dual algorithm with linesearch. *arXiv preprint arXiv:1608.08883*, 2016.

[9] Yu. Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1, Ser. A):127–152, 2005.

[10] Yu Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

[11] Matthias Rupp. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry*, 115(16):1058–1073, 2015.

[12] Sabrina M Tom, Craig R Fox, Christopher Trepel, and Russell A Poldrack. The neural basis of loss aversion in decision-making under risk. *Science*, 315(5811):515–518, 2007.

[13] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher. A smooth primal-dual optimization framework for nonsmooth composite convex minimization. *arXiv preprint arXiv:1507.06243*, 2015.

[14] Bằng Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comput. Math.*, 38(3):667–681, 2013.

[15] Yangyang Xu. Accelerated first-order primal-dual proximal methods for linearly constrained composite convex programming. *To appear in SIAM Journal on Optimization*, 2017.