ACCELERATING THE HYBRID STEEPEST DESCENT METHOD FOR AFFINELY CONSTRAINED CONVEX COMPOSITE MINIMIZATION TASKS

Konstantinos Slavakis[1] *Isao Yamada*[2] *Shunsuke Ono*[3]

[1]University at Buffalo (SUNY)
Dept. of Electrical Eng.
Buffalo 14260-2500, USA
kslavaki@buffalo.edu

[2]Tokyo Institute of Technology
Dept. of Inform. & Communications Eng.
Tokyo 152-8550, Japan
isao@ict.e.titech.ac.jp

[3]Tokyo Institute of Technology
Lab. Future Interdisciplinary Research of Sc. & Tech.
Yokohama 226-8503, Japan
ono@isl.titech.ac.jp

ABSTRACT

The hybrid steepest descent method (HSDM) [Yamada, '01] was introduced as a low-computational complexity tool for solving convex variational-inequality problems over the fixed-point set of nonexpansive mappings in Hilbert spaces. Motivated by results on decentralized optimization, this study introduces an HSDM variant that extends, for the first time, the applicability of HSDM to affinely constrained composite convex minimization tasks over Euclidean spaces; the same class of problems solved by the popular alternating direction method of multipliers and primal-dual methods. The proposed scheme shows desirable attributes for large-scale optimization tasks that have not been met, partly or all-together, in any other member of the HSDM family of algorithms: tunable computational complexity, a step-size parameter which stays constant over recursions, promoting thus acceleration of convergence, no boundedness constraints on iterates and/or gradients, and the ability to deal with convex losses which comprise a smooth and a non-smooth part, where the smooth part is only required to have a Lipschitz-continuous derivative. Convergence guarantees and rates are established. Numerical tests on synthetic data and on colored-image inpainting underline the rich potential of the proposed scheme for large-scale optimization tasks.

Index Terms— Composite optimization, convexity, nonexpansive mappings, hybrid steepest descent method, variational-inequality problem.

1. INTRODUCTION

Consider the set $\Gamma_0(\mathcal{X})$ of all convex, proper, and lower semicontinuous functions [1], defined on $\mathcal{X} := \mathbb{R}^D$ (D belongs to the set of all positive integers $\mathbb{N}$) with values in $\mathbb{R} \cup \{+\infty\}$, and loss functions $f, g \in \Gamma_0(\mathcal{X})$, where f is differentiable with L-Lipschitz-continuous derivative ∇f: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\|$, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$. This paper introduces the *accelerated hybrid steepest descent method (AHSDM),* a new member of the HSDM family [14, 16, 21, 24–27], to solve the following affinely constrained composite convex minimization task:

$$\min_{\mathbf{x}\in\mathcal{X}} \; f(\mathbf{x}) + g(\mathbf{x}) \text{ subject to (s.to) } \mathbf{H}\mathbf{x} = \mathbf{c}\,, \qquad (1)$$

for some $\mathbf{H} \in \mathbb{R}^{K\times D}$ and $\mathbf{c} \in \mathbb{R}^K$. The celebrated alternating direction method of multipliers (ADMM) [3, 8, 9] solves the same class of problems as in (1):

$$\min_{(\mathbf{z},\mathbf{z}')\in\mathcal{Z}^2} \; F(\mathbf{z}) + G(\mathbf{z}') \text{ s.to } \mathbf{F}\mathbf{z} + \mathbf{G}\mathbf{z}' = \mathbf{c}\,, \qquad (2)$$

for some Euclidean space $\mathcal{Z}$, losses $F, G \in \Gamma_0(\mathcal{Z})$, and matrices $\mathbf{F}, \mathbf{G}$. Indeed, if F satisfies the requirements of (1), (1) and (2) become equivalent, since one can set $\mathcal{X} := \mathcal{Z}^2$, $\mathbf{x} := [\mathbf{z}^\top, \mathbf{z}'^\top]^\top$, $\mathbf{H} := [\mathbf{F}, \mathbf{G}]$, as well as $f(\mathbf{x}) := F(\mathbf{z})$ and $g(\mathbf{x}) := G(\mathbf{z}')$. Even if F is not differentiable, AHSDM can still undertake the minimization task, since f can be set equal to zero, and $g := F + G$ [*cf.* (7)]. In such a way, the ability of AHSDM to solve (2) underlines its rich potential for all the application domains where ADMM has been shown to be successful [3].

For a user-defined parameter $\rho > 0$, ADMM generates the sequence $(\mathbf{z}_n, \mathbf{z}'_n, \mathbf{u}_n)_{n\in\mathbb{N}}$ by running the following steps during its nth iteration ($n \in \mathbb{N}$):

$(n.1)$ $\mathbf{z}_{n+1} := \arg\min_{\mathbf{z}\in\mathcal{H}} F(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{F}\mathbf{z} + \mathbf{G}\mathbf{z}'_n - \mathbf{c} + \mathbf{u}_n\|^2$.

$(n.2)$ $\mathbf{z}'_{n+1} := \arg\min_{\mathbf{z}'\in\mathcal{H}} G(\mathbf{z}') + \frac{\rho}{2}\|\mathbf{F}\mathbf{z}_{n+1} + \mathbf{G}\mathbf{z}' - \mathbf{c} + \mathbf{u}_n\|^2$.

$(n.3)$ $\mathbf{u}_{n+1} := \mathbf{u}_n + \mathbf{F}\mathbf{z}_{n+1} + \mathbf{G}\mathbf{z}'_{n+1} - \mathbf{c}$.

Steps $(n.1)$ and $(n.2)$ are convex-optimization programs themselves. Even in cases where F, for example, is differentiable and takes a simple form, such as the quadratic $F(\mathbf{z}) = \mathbf{z}^\top\mathbf{\Pi}\mathbf{z}$, for some positive semidefinite matrix $\mathbf{\Pi}$, step $(n.1)$ requires the solution of a system of linear equations with a possibly singular coefficient matrix. To surmount such computational obstacles, at the expense of convergence speed, the popular *primal-dual (PD)* methods, *e.g.*, [7, 22], solve (1), or (2), using low-complexity recursions, where solvers for updating variables, as in the ADMM steps $(n.1)$ and $(n.2)$, are not necessary.

The *hybrid steepest descent method (HSDM)* was introduced in [24] to solve $\min_{\mathbf{x}\in\mathrm{Fix}\,T\subset\mathcal{X}} \; f(\mathbf{x})$, where $\mathcal{X}$ is a potentially infinite-dimensional Hilbert space, f is a differentiable strongly convex function, and $\mathrm{Fix}\,T$ denotes the fixed-point set of a nonexpansive mapping $T : \mathcal{X} \to \mathcal{X}$ (see Definitions 1 and 2). Conjugate-gradient-based variants were introduced in [11–13], offering acceleration of HSDM's convergence. To secure (strong) convergence to an optimal point in Hilbert spaces, step-size parameters are required to be diminishing across recursions in all of [14, 16, 21, 24–27], while boundedness constraints are imposed on iterates and/or gradients [11–13].

Motivated by recent studies on decentralized optimization [18, 19], where a composite convex minimization task is solved by a large number of computer nodes s.to a consensus constraint, and by the similarities those methods share with HSDM, this paper presents AHSDM to tackle (1). AHSDM's step-size parameter stays constant across recursions, promoting thus convergence acceleration, no boundedness constraints are imposed on iterates and/or gradients, and the smooth part f of the loss is only required to have a Lipschitz-continuous derivative, without any strong-convexity requirements.

Along the lines of HSDM, (1) is revisited as a variational-inequality problem over the fixed-point set $\mathrm{Fix}\,T$ of an affine nonexpansive mapping T. Propelled by the numerous ways that

This work was supported in part by the NSF awards 1514056 and 1525194.

the nonexpansive-mappings theory offers to approach points within $\operatorname{Fix} T$ [1, 6], this study introduces AHSDM; a new member of the HSDM family of algorithms which solves (1) with tunable computational complexity. In its simpler form, AHSDM scores a computational complexity similar to that of the PD methods [7,22], while AHSDM can be tuned to reach a complexity similar to that of ADMM for accelerating convergence. In all its forms, AHSDM iterates are guaranteed to converge to a solution of (1). Convergence rate results are also demonstrated. Owing to its structural flexibility, AHSDM is well-suited for large-scale convex optimization tasks. To this end, numerical tests on synthetic data and on colored-image interpolation, *a.k.a.* inpainting [15], are also presented.

2. AFFINE NONEXPANSIVE MAPPINGS AND THE VARIATIONAL-INEQUALITY PROBLEM

Regarding notation, Id stands for the identity map in $\mathcal{X}$, *i.e.*, $\forall \mathbf{x} \in \mathcal{X}$, $\operatorname{Id}\mathbf{x} = \mathbf{x}$, while $\mathbf{I}$ denotes the identity matrix. Given matrices $\mathbf{Q}_1, \mathbf{Q}_2$, $\|\mathbf{Q}_1\|$ and $\|\mathbf{Q}_1\|_{\mathrm{F}}$ stand for the spectral and Frobenius norms of $\mathbf{Q}_1$, respectively, while $\mathbf{Q}_1 \succ (\succeq)\ \mathbf{Q}_2$ iff $\mathbf{Q}_1 - \mathbf{Q}_2$ is positive (semi)definite. Further, $\operatorname{sp}(\mathbf{Q})$ stands for all eigenvalues $\lambda(\mathbf{Q})$ of the symmetric $\mathbf{Q}$. The null space of matrix $\mathbf{Q}$ is defined as $\ker(\mathbf{Q}) := \{\mathbf{x} \in \mathcal{X} \mid \mathbf{Q}\mathbf{x} = \mathbf{0}\}$. Finally, given $g \in \Gamma_0(\mathcal{X})$, the *subdifferential* ∂g is the set-valued mapping defined as $\partial g : \mathbf{x} \mapsto \partial g(\mathbf{x}) := \{\boldsymbol{\xi} \in \mathcal{X} \mid \boldsymbol{\xi}^\top(\mathbf{y} - \mathbf{x}) + g(\mathbf{x}) \leq g(\mathbf{y}), \forall \mathbf{y} \in \mathcal{X}\}$. The proofs of the following results can be found in [20].

Definition 1. The *fixed-point set* of a mapping $T : \mathcal{X} \to \mathcal{X}$ is defined as the set $\operatorname{Fix} T := \{\mathbf{x} \in \mathcal{X} \mid T\mathbf{x} = \mathbf{x}\}$.

Definition 2. Mapping $T : \mathcal{X} \to \mathcal{X}$ is called (i) *nonexpansive (NonExp)* if $\|T\mathbf{x}_1 - T\mathbf{x}_2\| \leq \|\mathbf{x}_1 - \mathbf{x}_2\|$, $\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, and (ii) *$\alpha$-averaged* if there exist an $\alpha \in (0,1)$ and a NonExp mapping $R : \mathcal{X} \to \mathcal{X}$ such that (s.t.) $T = \alpha R + (1-\alpha)\operatorname{Id}$. It can be verified that $\operatorname{Fix} R = \operatorname{Fix} T$. In the case where $\alpha = 1/2$, T is also called *firmly NonExp*.

Example 3. Several examples of α-averaged mappings follow.

(i) [1, Prop. 4.8] Given a non-empty closed convex set $\mathcal{C} \subset \mathcal{X}$, the *(metric) projection mapping onto $\mathcal{C}$,* defined as $P_{\mathcal{C}} : \mathcal{X} \to \mathcal{C} : \mathbf{x} \mapsto \arg\min_{\mathbf{z}\in\mathcal{C}} \|\mathbf{x} - \mathbf{z}\|$, is $(1/2)$-averaged, with $\operatorname{Fix} P_{\mathcal{C}} = \mathcal{C}$.

(ii) [1, Prop. 12.27] Given $f \in \Gamma_0(\mathcal{X})$ and $\gamma > 0$, the *proximal mapping,* defined as $\operatorname{Prox}_{\gamma f} : \mathcal{X} \to \mathcal{X} : \mathbf{x} \mapsto \arg\min_{\mathbf{z}\in\mathcal{X}} f(\mathbf{z}) + \|\mathbf{x} - \mathbf{z}\|^2/(2\gamma)$, is $(1/2)$-averaged, with $\operatorname{Fix}\operatorname{Prox}_{\gamma f} = \arg\min f$.

(iii) [6, Prop. 2.2] Let Let $\{T_j\}_{j=1}^J$ be a finite family ($J < \infty$) of NonExp mappings from $\mathcal{X}$ to $\mathcal{X}$, and $\{\omega_j\}_{j=1}^J$ be real numbers in $(0,1]$ s.t. $\sum_{j=1}^J \omega_j = 1$. Then, $T := \sum_{j=1}^J \omega_j T_j$ is NonExp. Further, consider real numbers $\{\alpha_j\}_{j=1}^J \subset (0,1)$ s.t. T_j is α_j-averaged, $\forall j$. Define $\alpha := \sum_{j=1}^J \omega_j \alpha_j$. Then, T is α-averaged. In all cases, if $\cap_{j=1}^J \operatorname{Fix} T_j \neq \emptyset$, then $\operatorname{Fix} T = \cap_{j=1}^J \operatorname{Fix} T_j$.

(iv) [6, Prop. 2.5], [14, Thm. 3(b)] Let $\{T_j\}_{j=1}^J$ be a finite family ($J < \infty$) of nonexpansive mappings from $\mathcal{X}$ to $\mathcal{X}$. Then, mapping $T := T_1 T_2 \cdots T_J$ is nonexpansive. Further, consider real numbers $\{\alpha_j\}_{j=1}^J \subset (0,1)$ s.t. T_j is α_j-averaged, $\forall j$. Define $\alpha := [1 + (\sum_{j=1}^J \alpha_j/(1-\alpha_j))^{-1}]^{-1}$. Then, T is α-averaged. In all cases, if $\cap_{j=1}^J \operatorname{Fix} T_j \neq \emptyset$, then $\operatorname{Fix} T = \cap_{j=1}^J \operatorname{Fix} T_j$.

Definition 4. A mapping $T : \mathcal{X} \to \mathcal{X}$ is called *affine* if $T[w\mathbf{x}_1 + (1-w)\mathbf{x}_2] = wT\mathbf{x}_1 + (1-w)T\mathbf{x}_2, \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\forall w \in \mathbb{R}$.

The following assumption is the basic building block of the proposed algorithm.

Assumption 5. Mapping T is defined as $T\mathbf{x} := \mathbf{Q}\mathbf{x} + \boldsymbol{\pi}, \forall \mathbf{x} \in \mathcal{X}$, where $\mathbf{Q}^\top = \mathbf{Q}$, $\mathbf{Q} \succeq \mathbf{0}$, $\|\mathbf{Q}\| \leq 1$, and $\boldsymbol{\pi} \in \mathcal{X}$.

Mapping T of Assumption 5 is clearly affine, and according to [1, Ex. 4.4], it is also NonExp (iff $\|\mathbf{Q}\| \leq 1$). More generally, and as the following proposition shows, convex combinations as well as compositions of NonExp affine mappings still satisfy Assumption 5.

Proposition 6. Consider any finite family of mappings $\{T_j\}_{j=1}^J$ which satisfy Assumption 5. Then, (i) for any set of convex weights $\{\omega_j\}_{j=1}^J$, *i.e.*, $\omega_j \in [0,1]$ with $\sum_{j=1}^J \omega_j = 1$, the convex combination $\sum_j \omega_j T_j$ satisfies Assumption 5, and (ii) given also the affine mapping $T_0\mathbf{x} := \mathbf{Q}_0\mathbf{x} + \boldsymbol{\pi}_0$, $\forall \mathbf{x} \in \mathcal{X}$, with a symmetric $\mathbf{Q}_0 \succeq \mathbf{0}$, $\|\mathbf{Q}_0\| \leq 1$, and $\boldsymbol{\pi}_0 \in \mathcal{X}$, the composition mapping $T_J T_{J-1} \cdots T_1 T_0 T_1 \cdots T_{J-1} T_J$ satisfies Assumption 5.

Proposition 7. For any mapping T that satisfies Assumption 5, its fixed-point set is the following affine set $\operatorname{Fix} T = \ker(\mathbf{I} - \mathbf{Q}) + \mathbf{w}_* = \ker(\mathbf{U}) + \mathbf{w}_*$, where $\mathbf{w}_*$ is any fixed point of T, and $\mathbf{U} \succeq \mathbf{0}$ is defined as the symmetric ($\mathbf{U}^\top = \mathbf{U}$) square root of $\mathbf{I} - \mathbf{Q}$, *i.e.*, $\mathbf{U}^2 = \mathbf{I} - \mathbf{Q}$.

Several examples of mappings which satisfy Assumption 5 are provided here. Let's start with an elementary one: consider a nonzero $\mathbf{a} \in \mathcal{X}$ and a real number b to define the hyperplane $\mathcal{P} := \{\mathbf{x} \in \mathcal{X} \mid \mathbf{a}^\top\mathbf{x} = b\}$. The associated (metric) projection mapping is

$$P_{\mathcal{P}} = \left(\mathbf{I} - \tfrac{1}{\|\mathbf{a}\|^2}\mathbf{a}\mathbf{a}^\top\right)\operatorname{Id} + \tfrac{b}{\|\mathbf{a}\|^2}\mathbf{a}\,, \tag{3}$$

which clearly satisfies Assumption 5.

The prototypical affine set is the one obtained from the solution of a *least-squares (LS)* problem. The following proposition provides several characterizations of such an affine set.

Proposition 8 (Least-squares)**.** Given the $M \times 1$ vector $\mathbf{b}$, and the $M \times D$ matrix $\mathbf{A}$, consider the following LS solution set: $\mathcal{A} := \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = \{\mathbf{x} \in \mathcal{X} \mid \mathbf{A}^\top\mathbf{A}\mathbf{x} = \mathbf{A}^\top\mathbf{b}\}$. For the $D \times 1$ vectors $\{\boldsymbol{a}_m\}_{m=1}^M$ defined by the rows of $\mathbf{A}$, *i.e.*, $[\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M] := \mathbf{A}^\top$, as well as the $D \times 1$ vectors $\{\mathbf{g}_d\}_{d=1}^D$: $[\mathbf{g}_1, \ldots, \mathbf{g}_D] := \mathbf{G}$, where $\mathbf{G} := \mathbf{A}^\top\mathbf{A}$ and $\mathbf{c} := \mathbf{A}^\top\mathbf{b}$, let

$$\mathcal{A}_m := \{\mathbf{x} \in \mathcal{X} \mid \boldsymbol{a}_m^\top\mathbf{x} = b_m\}\,, \qquad (m = 1, \ldots, M)\,,$$
$$\mathcal{G}_d := \{\mathbf{x} \in \mathcal{X} \mid \mathbf{g}_d^\top\mathbf{x} = c_d\}\,, \qquad (d = 1, \ldots, D)\,,$$

with associated metric projection mappings $P_{\mathcal{A}_m}$ and $P_{\mathcal{G}_d}$, respectively [*cf.* (3)]. Then, $\mathcal{A}$ becomes the fixed-point set of the following mappings which satisfy Assumption 5:

$$\begin{aligned}
\mathcal{A} &= \operatorname{Fix}\left[\left(\mathbf{I} - \tfrac{\mu}{\|\mathbf{A}\|_{\mathrm{F}}^2}\mathbf{A}^\top\mathbf{A}\right)\operatorname{Id} + \tfrac{\mu}{\|\mathbf{A}\|_{\mathrm{F}}^2}\mathbf{A}^\top\mathbf{b}\right] \\
&\qquad\qquad \left(0 \leq \mu \leq \tfrac{\|\mathbf{A}\|_{\mathrm{F}}^2}{\|\mathbf{A}^\top\mathbf{A}\|}\right) \\
&= \operatorname{Fix}\left[\left(\mathbf{I} - \mathbf{A}^\top\mathbf{A}^{\dagger\top}\right)\operatorname{Id} + \mathbf{A}^\dagger\mathbf{b}\right] \\
&= \operatorname{Fix}\left[\left(\mathbf{I} - \mathbf{G}\mathbf{G}^\dagger\right)\operatorname{Id} + \mathbf{G}^\dagger\mathbf{A}^\top\mathbf{b}\right] \\
&= \operatorname{Fix}\left[\left(\mathbf{I} + \gamma\mathbf{A}^\top\mathbf{A}\right)^{-1}\operatorname{Id} + \gamma\left(\mathbf{I} + \gamma\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top\mathbf{b}\right] \ (\gamma > 0) \\
&= \operatorname{Fix}\left[(1-\theta)\operatorname{Id} + \theta\sum\nolimits_{d=1}^D \omega_d P_{\mathcal{G}_d}\right] \\
&\qquad\qquad \left(0 < \theta \leq 1,\ 0 < \omega_d < 1,\ \sum\nolimits_{d=1}^D \omega_d = 1\right),
\end{aligned}$$

where † denotes the Moore-Penrose pseudoinverse operation [2].

The following definition and fact help revisit (1).

Definition 9 (Variational-inequality problem)**.** For a NonExp mapping $T : \mathcal{X} \to \mathcal{X}$, point $\mathbf{x}_* \in \operatorname{Fix} T$ is said to solve the variational-inequality problem VIP($\nabla f + \partial g, \operatorname{Fix} T$) if there exists $\boldsymbol{\xi}_* \in \partial g(\mathbf{x}_*)$ s.t. $\forall \mathbf{y} \in \operatorname{Fix} T$, $\langle \mathbf{y} - \mathbf{x}_* \mid \nabla f(\mathbf{x}_*) + \boldsymbol{\xi}_* \rangle \geq 0$.

Fact 10 ([1, Prop. 26.5])**.** Point $\mathbf{x}_*$ solves VIP($\nabla f + \partial g, \operatorname{Fix} T$) iff $\mathbf{x}_* \in \arg\min_{\mathbf{x} \in \operatorname{Fix} T} [f(\mathbf{x}) + g(\mathbf{x})]$.

The previous fact suggests that *any* affine NonExp mapping, with fixed-point set equal to the affine set in (1), can be used to revisit (1) as a variational-inequality problem. Examples of such affine NonExp mappings can be found in Proposition 8. This versatility of NonExp mappings, manifested for example in Proposition 6, equips the proposed algorithm of Sec. 3 with a modularity which is desirable in nowadays large-scale convex minimization tasks. Based on Fact 10, the following characterization of minimizers of (1) is made possible.

Proposition 11. Consider any mapping T which satisfies Assumption 5 (see also Proposition 7). Then, point $\mathbf{x}_*$ solves VIP($\nabla f + \partial g, \operatorname{Fix} T$) iff $\exists \boldsymbol{\xi}_* \in \partial g(\mathbf{x}_*)$ and $\forall \lambda \neq 0$, $\exists \mathbf{v}_* \in \mathcal{X}$ s.t. $(\mathbf{x}_*, \mathbf{v}_*) \in \mathcal{O}_*(\boldsymbol{\xi}_*, \lambda) := \{(\mathbf{x}, \mathbf{v}) \in \operatorname{Fix} T \times \mathcal{X} \mid \mathbf{0} = \mathbf{U}\mathbf{v} + \lambda(\nabla f(\mathbf{x}) + \boldsymbol{\xi}_*)\}$.

An additional assumption is needed on the non-smooth part g of (1) to establish the convergence guarantees of AHSDM.

Assumption 12. The *graph* $\operatorname{gra} \partial g := \{(\mathbf{x}, \boldsymbol{\xi}) \in \mathcal{X}^2 \mid \boldsymbol{\xi} \in \partial g(\mathbf{x})\}$ of ∂g is closed.

As the following proposition demonstrates, Assumption 12 is loose enough to cover a plethora of well-known non-smooth losses (*cf.* Sec. 4).

Proposition 13. (i) Any $g \in \Gamma_0(\mathcal{X})$ with values in $\mathbb{R}$ satisfies Assumption 12. A celebrated example of such a function is the ℓ_1-norm $g := \|\cdot\|_1$. (ii) For a nonempty closed convex set $\mathcal{C} \subset \mathcal{X}$, the indicator function $\iota_{\mathcal{C}} \in \Gamma_0(\mathcal{X})$, defined as $\iota_{\mathcal{C}}(\mathbf{x}) := 0$, if $\mathbf{x} \in \mathcal{C}$, while $\iota_{\mathcal{C}}(\mathbf{x}) = +\infty$, if $\mathbf{x} \notin \mathcal{C}$, satisfies Assumption 12.

3. ALGORITHM, CONVERGENCE GUARANTEES AND RATES

Consider any mapping T which satisfies Assumption 5. Examples are given in Proposition 8. Many more such mappings T can be generated by combining the "elementary" ones of Proposition 8 in the ways demonstrated by Proposition 6. Given $\alpha \in (0, 1)$, define the α-averaged mapping

$$T_\alpha \mathbf{x} := (\alpha T + (1 - \alpha) \operatorname{Id})\mathbf{x} = \mathbf{Q}_\alpha \mathbf{x} + \alpha \boldsymbol{\pi}\,, \tag{5}$$

where $\mathbf{Q}_\alpha := \alpha \mathbf{Q} + (1 - \alpha)\mathbf{I}$.

Algorithm 1 (AHSDM)**.** Fix $\alpha \in (0, 1)$ and $\lambda > 0$. Then, for an arbitrarily fixed $\mathbf{x}_0 \in \mathcal{X}$, and for all $n \in \mathbb{N}$, AHSDM is stated as follows ($\mathbf{x}_{n+1/2}$ and $\mathbf{x}_{n+3/2}$ are auxiliary variables):

$$\mathbf{x}_{n+\frac{1}{2}} := T_\alpha \mathbf{x}_n - \lambda \nabla f(\mathbf{x}_n)\,, \tag{6a}$$

$$\mathbf{x}_{n+1} := \operatorname{Prox}_{\lambda g}(\mathbf{x}_{n+\frac{1}{2}})\,, \tag{6b}$$

$$\mathbf{x}_{n+\frac{3}{2}} := T\mathbf{x}_{n+1} - \lambda \nabla f(\mathbf{x}_{n+1})\,, \tag{6c}$$

$$\mathbf{x}_{n+2} := \operatorname{Prox}_{\lambda g}(\mathbf{x}_{n+\frac{3}{2}})\,. \tag{6d}$$

In the case where $f = 0$, the previous recursions take the form

$$\mathbf{x}_{n+\frac{1}{2}} := T_\alpha \mathbf{x}_n\,, \qquad \mathbf{x}_{n+1} := \operatorname{Prox}_{\lambda g}(\mathbf{x}_{n+\frac{1}{2}})\,, \tag{7a}$$

$$\mathbf{x}_{n+\frac{3}{2}} := T\mathbf{x}_{n+1}\,, \qquad \mathbf{x}_{n+2} := \operatorname{Prox}_{\lambda g}(\mathbf{x}_{n+\frac{3}{2}})\,. \tag{7b}$$

Moreover, in the case where $g := 0$, (6) takes the special form

$$\mathbf{x}_{n+1} := T_\alpha \mathbf{x}_n - \lambda \nabla f(\mathbf{x}_n)\,, \tag{8a}$$

$$\mathbf{x}_{n+2} := T\mathbf{x}_{n+1} - \lambda \nabla f(\mathbf{x}_{n+1})\,. \tag{8b}$$

The following theorem establishes convergence guarantees for the most general form (6) of AHSDM.

Theorem 14. Consider any mapping T, with $\operatorname{Fix} T \neq \emptyset$, that satisfies Assumption 5. If $\alpha \in [0.5, 1)$ and $\lambda \in (0, 2(1 - \alpha)/L)$, with L being the Lipschitz constant of ∇f, and if the graph $\operatorname{gra} \partial g$ satisfies Assumption 12, then the sequence $(\mathbf{x}_n)_{n \in \mathbb{N}}$ of (6) converges to an $\mathbf{x}_*$ which solves VIP($\nabla f + \partial g, \operatorname{Fix} T$).

The following theorems establish AHSDM's rates of convergence, which appear to be of the same order as that of ADMM [10]. Assumptions for deriving the following results, as well as $\mathbf{x}_*$, are adopted from Theorem 14.

Theorem 15. Considering (6), $\exists (\boldsymbol{\xi}_n, \mathbf{v}_n) \in \partial g(\mathbf{x}_n) \times \mathcal{X}$, $\forall n$, s.t.

$$\tfrac{1}{n+1} \sum_{\nu=0}^{n} (\mathbf{x}_{\nu+1} - \mathbf{x}_*)^\top (\mathbf{I} - \mathbf{Q})(\mathbf{x}_{\nu+1} - \mathbf{x}_*) = O(\tfrac{1}{n+1})\,,$$
$$\tfrac{1}{n+1} \sum_{\nu=0}^{n} \|\mathbf{U}\mathbf{v}_{\nu+1} + \lambda[\nabla f(\mathbf{x}_\nu) + \boldsymbol{\xi}_{\nu+1}]\|^2 = O(\tfrac{1}{n+1})\,,$$
$$\tfrac{1}{n+1} \sum_{\nu=0}^{n} \|(\operatorname{Id} - T)\mathbf{x}_{\nu+1}\|^2 = O(\tfrac{1}{n+1})\,,$$

where $a_n = O(b_n)$, $b_n > 0$, means that $(|a_n|/b_n)_{n \in \mathbb{N}}$ is bounded.

Theorem 16. Considering (7), $\exists (\boldsymbol{\xi}_n, \mathbf{v}_n) \in \partial g(\mathbf{x}_n) \times \mathcal{X}$, $\forall n$, s.t.

$$\langle \mathbf{x}_{n+1} - \mathbf{x}_* \mid (\mathbf{I} - \mathbf{Q})(\mathbf{x}_{n+1} - \mathbf{x}_*) \rangle = O(\tfrac{1}{n+1})\,,$$
$$\|\mathbf{U}\mathbf{v}_{n+1} + \lambda \boldsymbol{\xi}_{n+1}\|^2 = O(\tfrac{1}{n+1})\,,$$
$$\|(\operatorname{Id} - T)\mathbf{x}_{n+1}\|^2 = O(\tfrac{1}{n+1})\,.$$

4. NUMERICAL TESTS

Tests were performed by running MATLAB on a 64-core server, with Intel Xeon CPUs (64bits, 2.30GHz) and 256MB of memory.

4.1. Synthetic data

Given $\mathcal{H} := \mathbb{R}^d$, with $d := 1,000$, define the closed ball $\mathcal{B}[\mathbf{h}_c, r] := \{\mathbf{h} \in \mathcal{H} \mid \|\mathbf{h} - \mathbf{h}_c\| \leq r\}$, for some $\mathbf{h}_c \in \mathcal{H}$ and $r > 0$. Motivated by [12, Prob. 4.1], the following constrained quadratic program:

$$\min_{\mathbf{y} \in \mathcal{B}_1 \cap \mathcal{B}_2} \mathbf{y}^\top \boldsymbol{\Pi} \mathbf{y} = \min_{\mathbf{x} := (\mathbf{y}, \mathbf{z}, \mathbf{w}) \in \mathcal{H}^3} \tfrac{1}{2} \mathbf{y}^\top \boldsymbol{\Pi} \mathbf{y} + \iota_{\mathcal{B}_1}(\mathbf{z}) + \iota_{\mathcal{B}_2}(\mathbf{w})$$
$$\text{s.to } \mathbf{y} = \mathbf{z} = \mathbf{w}\,, \tag{9}$$

where $\mathbf{x} := (\mathbf{y}, \mathbf{z}, \mathbf{w}) := [\mathbf{y}^\top, \mathbf{z}^\top, \mathbf{w}^\top]^\top$, and $\boldsymbol{\Pi}$ is a $d \times d$ diagonal matrix, with (unique) minimum entry $[\boldsymbol{\Pi}]_{11} := 1$, maximum entry equal to 100, while all other diagonal entries are chosen randomly from $(1, 100)$. Moreover, if $\mathbf{e}_1$ denotes the first column of the $d \times d$ identity matrix $\mathbf{I}$, then $\mathcal{B}_1 := \mathcal{B}[2\mathbf{e}_1, 1]$ and $\mathcal{B}_2 := \mathcal{B}[\mathbf{0}, 2]$, while $\iota_{\mathcal{B}_1}$ and $\iota_{\mathcal{B}_2}$ denote the associated indicator functions [*cf.* Proposition 13(ii)]. By construction, $\mathbf{y}^\top \boldsymbol{\Pi} \mathbf{y}$ is strongly convex s.t. $\mathbf{x}_* := \{\mathbf{e}_1\}^3$ is the unique minimizer of (9).

Being a linear subspace of $\mathcal{X} := \mathcal{H}^3$, all points $\mathcal{A}$ which satisfy the constraint in (9) constitute an affine set. A nonexpansive mapping T having $\mathcal{A}$ as its fixed-point set is the metric projection mapping $P_{\mathcal{A}}(\mathbf{y}, \mathbf{z}, \mathbf{w}) = \{(1/3)(\mathbf{y} + \mathbf{z} + \mathbf{w})\}^3$, $\forall \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathcal{H}$. Form (6) of AHSDM was applied to (9), with $f(\mathbf{x}) := (1/2)\mathbf{y}^\top \boldsymbol{\Pi} \mathbf{y}$, $g(\mathbf{x}) := \iota_{\mathcal{B}_1}(\mathbf{z}) + \iota_{\mathcal{B}_2}(\mathbf{w})$, and $\alpha = 0.5$. Notice that for any $\lambda > 0$, $\operatorname{Prox}_{\lambda g} =$

$(\mathrm{Id}, \mathrm{Prox}_{\lambda\iota_{\mathcal{B}_1}}, \mathrm{Prox}_{\lambda\iota_{\mathcal{B}_2}}) = (\mathrm{Id}, P_{\mathcal{B}_1}, P_{\mathcal{B}_2})$, and that the Lipschitz-continuity constant L of ∇f equals the maximum entry of $\mathbf{\Pi}$.

Since the majority of entries of $\mathbf{\Pi}$ were chosen randomly, 100 Monte-Carlo runs were performed, and the uniformly averaged results are demonstrated in Fig. 1. AHSDM is compared with ADMM, PD [7], HSDM [24], as well as [13], [11], and [12], which are denoted by CG-HSDM-I, CG-HSDM-II, and CG-HSDM-III, respectively. All methods were tuned for optimal results. As Fig. 1 shows, ADMM, PD, and AHSDM exhibit similar convergence behavior.

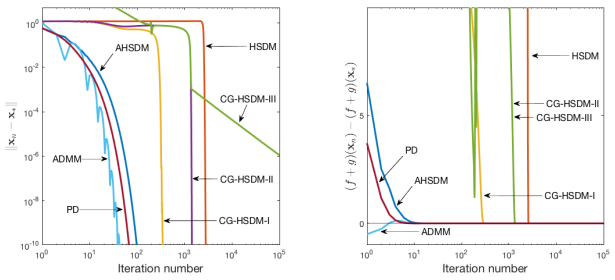

Fig. 1. Deviation of the iterates from the unique minimizer (left), as well as loss function values (right) are demonstrated in the case where the condition number of $\mathbf{\Pi}$ [*cf.* (9)] equals 100.

The previous setting is repeated, but the minimum entry $\pi_{\min} := [\mathbf{\Pi}]_{11} := 10^{-15}$, while $\pi_{\max} := 10$, resulting into a condition number $\pi_{\max}/\pi_{\min} = 10^{16}$. Form (7) of AHSDM is applied to (9), where $f := 0$ and $g(\mathbf{x}) := (1/2)\mathbf{y}^\top \mathbf{\Pi}\mathbf{y} + \iota_{\mathcal{B}_1}(\mathbf{z}) + \iota_{\mathcal{B}_2}(\mathbf{w})$. Since $f := 0$, any $L > 0$ can be considered here for the Lipschitz-continuity constant of f; tuning yielded $L = 10^{-2}$. All methods were tested for 100 Monte-Carlo runs, and uniformly averaged results are depicted in Fig. 2. As expected, HSDM, and CG-HSDM-I, -II, and -III face problems in converging to the unique minimizer of (9), since their convergence guarantees are provided only for strongly convex losses, while $\mathbf{\Pi}$ is chosen here to be "nearly" singular.

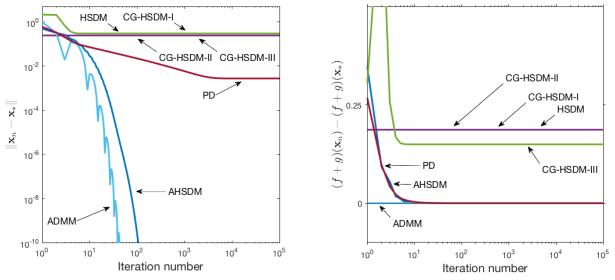

Fig. 2. The setting of this experiment follows that of Fig. 1, but the condition number of the "nearly" singular $\mathbf{\Pi}$ is set equal to 10^{16}. Since the optimal loss value is very small, $0.5 \cdot 10^{-15}$, the ADMM curve appears to lie on the all-zero curve.

4.2. Colored-image inpainting

Given a noisy (vectorized) image $\check{\mathbf{i}} \in \mathbb{R}^{\check{d}}$ with missing entries, observed after a $d \times d$ measurement matrix $\mathbf{\Phi}$ (the noiselet transform [5] was used here), noise $\mathbf{n}$, and a $\check{d} \times d$ sampling matrix $\mathbf{S}$, with $\check{d} < d$, are applied to the original (normalized) image $\mathbf{i} \in [0,1]^d \subset \mathbb{R}^d =: \mathcal{H}$, as in $\check{\mathbf{i}} = \mathbf{S}(\mathbf{\Phi}\mathbf{i} + \mathbf{n})$, the goal is to recover $\mathbf{i}$ by removing noise while estimating the missing entries of $\check{\mathbf{i}}$. Following [15], the previous task is formulated as

$$\min_{\substack{\mathbf{y}\in[0,1]^d \subset \mathcal{H},\\ \|\mathbf{S}\mathbf{\Phi}\mathbf{y}-\check{\mathbf{i}}\|\leq\epsilon}} \|\mathbf{D}\mathbf{y}\|_{1,2} = \min_{(\mathbf{y},\mathbf{z},\mathbf{w})\in\mathcal{H}^3} \|\mathbf{z}\|_{1,2} + \iota_{\mathcal{B}[\check{\mathbf{i}},\epsilon]}(\mathbf{S}\mathbf{w}) + \iota_{[0,1]^d}(\mathbf{y}) \quad \text{s.to } [\mathbf{L}, -\mathbf{I}_{2d}] \begin{bmatrix} \mathbf{y} \\ \mathbf{z} \\ \mathbf{w} \end{bmatrix} = \mathbf{0}\,, \tag{10}$$

where $\mathbf{D} : \mathcal{H} \to \mathcal{H}^2$ stands for the discrete gradient operator, which forms vertical and horizontal differences within an image, $\|\cdot\|_{1,2}$ is the mixed $\ell_{1,2}$-norm [15], s.t. $\|\mathbf{D}\mathbf{w}\|_{1,2}$ denotes the celebrated vectorial total variation (VTV) [4]. For a user-defined $\epsilon > 0$, constraint $\|\mathbf{S}\mathbf{\Phi}\mathbf{y} - \check{\mathbf{i}}\| \leq \epsilon$ accommodates the data-fit requirement. Moreover, $\mathbf{L} := [\mathbf{D}^\top, \mathbf{\Phi}^\top]^\top$ introduces the affine constraint of (10) by splitting variables in the loss function.

Form (7) of AHSDM was applied to (10), where $f = 0$, while, under $\mathbf{x} := (\mathbf{y}, \mathbf{z}, \mathbf{w})$, $g(\mathbf{x}) := \|\mathbf{z}\|_{1,2} + \iota_{\mathcal{B}[\check{\mathbf{i}},\epsilon]}(\mathbf{S}\mathbf{w}) + \iota_{[0,1]^d}(\mathbf{y})$. The affine nonexpansive mapping T used in (7) is $(\mathbf{I} + \gamma[\mathbf{L}, -\mathbf{I}_{2d}]^\top[\mathbf{L}, -\mathbf{I}_{2d}])^{-1}\,\mathrm{Id}$, whose fixed-point set, according to Proposition 8, comprises all points $\mathcal{A}$ which satisfy the linear constraint in (10). Parameter α was set equal to 0.5, $\lambda := 1$, and $\gamma := 0.02$. As in Sec. 4.1, tests on the image of Fig. 3 reveal the rich potential of the advocated AHSDM since it yields similar results to those of ADMM.

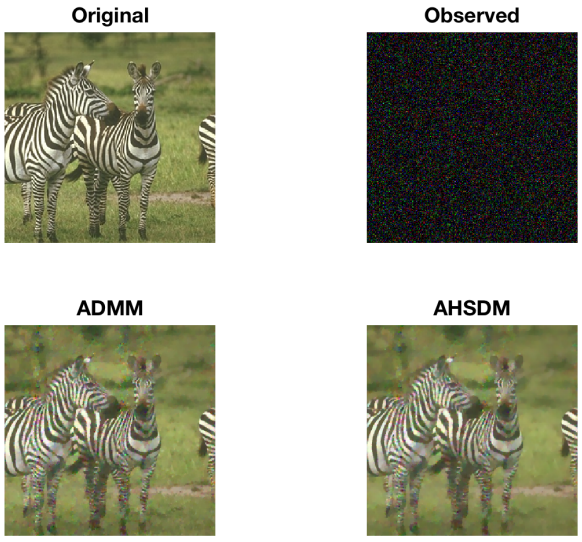

Fig. 3. Original 256×256 colored image, *i.e.*, $\mathbf{i} \in \mathbb{R}^{256\times256\times3}$, and its noisy rendition, observed after Gaussian noise of standard deviation 0.1 was added to $\mathbf{\Phi}\mathbf{i}$, and 80% of the entries of the noisy $\mathbf{\Phi}\mathbf{i}$ were randomly removed. Recovered images by ADMM and AHSDM are also shown.

	PSNR (dB)	CIEDE2000
ADMM	**22.047**	7.737
AHSDM	21.640	**7.372**

Table 1. Uniformly averaged results obtained after 100 Monte Carlo runs on the observed image of Fig. 3. Larger values of *peak signal-to-noise ratio (PSNR)* [23] and smaller values of the color-difference metric *CIEDE2000* [17] indicate "better-quality" reconstructed images.

5. REFERENCES

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. New York: Springer, 2011.

[2] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications*, 2nd ed. New York: Springer-Verlag, 2003.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[4] X. Bresson and T. F. Chan, "Fast dual minimization of the vectorial total variation norm and application to color image processing," *Inverse Problems Imaging*, vol. 2, no. 4, pp. 455–484, 2008.

[5] R. Coifman, F. Geshwind, and Y. Meyer, "Noiselets," *Applied Comput. Harmonic Anal.*, vol. 10, no. 1, pp. 27–44, 2001.

[6] P. L. Combettes and I. Yamada, "Compositions and convex combinations of averaged nonexpansive operators," *J. Math. Anal. Appl.*, vol. 425, pp. 55–70, 2015.

[7] L. Condat, "A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms," *J. Optim. Theory Appl.*, vol. 158, no. 2, pp. 460–479, 2013.

[8] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite-element approximations," *Comp. Math. Appl.*, vol. 2, pp. 17–40, 1976.

[9] R. Glowinski and A. Marrocco, "Sur l'approximation par éléments finis et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires," *Rev. Francaise d'Aut. Inf. Rech. Oper.*, vol. 9, no. 2, pp. 41–76, 1975.

[10] B. He and X. Yuan, "On the $\mathcal{O}(1/n)$ convergence rate of the Douglas-Rachford alternating direction method," *SIAM J. Numerical Analysis*, vol. 50, no. 2, pp. 700–709, 2012.

[11] H. Iiduka, "Three-term conjugate gradient method for the convex optimization problem over the fixed point set of a nonexpansive mapping," *Applied Math. Computation*, vol. 217, pp. 6315–6327, 2011.

[12] ——, "Acceleration method for convex optimization over the fixed point set of a nonexpansive mapping," *Math. Program.*, vol. 149, pp. 131–165, 2015.

[13] H. Iiduka and I. Yamada, "A use of conjugate gradient direction for the convex optimization problem over the fixed point set of a nonexpansive mapping," *SIAM J. Optim.*, vol. 19, no. 4, pp. 1881–1893, 2009.

[14] N. Ogura and I. Yamada, "Non-strictly convex minimization over the fixed point set of the asymptotically shrinking nonexpansive mapping," *Numerical Functional Analysis & Optim.*, vol. 23, pp. 113–137, 2002.

[15] S. Ono and I. Yamada, "Color-line regularization for color artifact removal," *IEEE Trans. Computational Imaging*, vol. 2, no. 3, pp. 204–217, 2016.

[16] ——, "Hierarchical convex optimization with primal-dual splitting," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 373–388, 2015.

[17] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Research & Applications*, vol. 30, no. 1, pp. 21–30, 2005.

[18] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *SIAM J. Optim.*, vol. 25, no. 2, 2014.

[19] ——, "A proximal gradient algorithm for decentralized composite optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 22, pp. 6013–6023, 2015.

[20] K. Slavakis and I. Yamada, "Accelerated hybrid steepest descent method for solving affinely constrained composite convex optimization tasks," *arXiv e-prints*, 2016. [Online]. Available: http://arxiv.org/abs/1608.02500

[21] ——, "Robust wideband beamforming by the hybrid steepest descent method," *IEEE Trans. Signal Process.*, vol. 55, no. 9, pp. 4511–4522, 2007.

[22] B. C. Vũ, "A splitting algorithm for dual monotone inclusions involving cocoercive operators," *Advances in Computational Math.*, vol. 38, no. 3, pp. 667–681, 2013.

[23] Wikipedia, "Peak signal-to-noise ratio," 2016. [Online]. Available: https://en.wikipedia.org/wiki/Peak_signal-to-noise_ratio

[24] I. Yamada, "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings," in *Inherently Parallel Algorithms for Feasibility and Optimization and their Applications*, D. Butnariu, Y. Cencor, and S. Reich, Eds. Elsevier, 2001, pp. 473–504.

[25] I. Yamada, N. Ogura, and N. Shirakawa, "A numerically robust hybrid steepest descent method for the convexly constrained generalized inverse problems," *Contemporary Math.*, vol. 313, pp. 269–305, 2002.

[26] I. Yamada, M. Yukawa, and M. Yamagishi, "Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, 2011, pp. 345–390.

[27] M. Yamagishi and I. Yamada, "Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization," *Inverse Problems*, 2016, to appear.