# A nonexpansive operator for computationally efficient hierarchical convex optimization

Masao Yamagishi and Isao Yamada
* Dept. of Communications and Computer Engineering, Tokyo Institute of Technology
E-mail: {myamagi, isao}@sp.ce.titech.ac.jp

***Abstract*—*Hybrid steepest descent method* is an algorithmic solution to certain *hierarchical convex optimization* which is a class of two-stage optimization problems: the *first stage problem* is a convex optimization; the *second stage problem* is the minimization of a differentiable convex function over the solution set of the first stage problem. In the application of this method, the solution set of the first stage problem must be expressed as the fixed point set of a certain nonexpansive operator.**

**In this paper, we propose a nonexpansive operator that yields a computationally efficient update in cases where it is plugged into the hybrid steepest descent method. The proposed operator is applicable to characterize the solution set of recent sophisticated convex optimization problems, where multiple proximable convex functions involving linear operators must be minimized. To the best of our knowledge, for such a problem, there was not reported any nonexpansive operator that yields an update free from the inversions of linear operators in cases where it is utilized in the hybrid steepest descent method. Unlike conventional operators, the proposed operator yields an inversion-free update.**

## I. INTRODUCTION

Convex optimization [1], [2], [3] is known as a powerful formalism for diverse application fields. It has significant flexibility to offer desired properties on its solutions by designing the objective function to be minimized (see e.g. [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]). In this context, fixed point characterizations of solutions by an operator provide a unified view for many iterative algorithms for convex optimization [19]. In particular, *nonexpansive operators* are preferred for characterizing all the solutions because many algorithms for finding a solution can be explained as iterative use of a nonexpansive operator (see e.g. [20], [21], [22], [23], [24], [25], [26] for such instances). On the other hand, the aforementioned iterative use of the nonexpansive operator is limited to approximate a single unspecial solution to the corresponding convex optimization problem though it has in general infinitely many solutions that could be considerably different in terms of another criterion.

For pursuing a better solution in some other aspects, *hierarchical convex optimization* [27] has been introduced to specify rigorously the optimums in the infinitely many solutions. The hierarchical convex optimization is a two-stage optimization problem in a Hilbert space: the first stage problem is a convex optimization problem in the Hilbert space; the second stage problem is the minimization of a convex function over the solution set of the first stage problem. In [27], an algorithmic solution to the *hierarchical convex optimization* was presented based on the hybrid steepest descent method (see for related results [19], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40]). For an iterative algorithm for the first stage problem, the hybrid steepest descent method simply concatenates an additional gradient descent step of the second stage criterion. The hybrid steepest descent method has broad applicability because its convergence is guaranteed essentially under only two requirements: the second stage criterion is smooth and the iterative algorithm for the first stage problem can be explained as the iterative use of a nonexpansive operator defined over the Hilbert space.

Meanwhile, the computational burden on each iteration of the hybrid steepest descent method remains controversial if the criterion of the first stage problem involves the composition of a non-smooth convex function and a linear operator. Recently, in [40], it has been investigated that plugging the computationally efficient nonexpansive operator found implicitly in the Condat's primal-dual splitting [26] into the hybrid steepest descent method results in the (possibly computationally intensive) inversion of a linear operator due to employment of a non-standard inner product space. Consequently, in the case where the first stage criterion involves such a composition, there has not been reported any nonexpansive operator that yields an inversion-free update in the hybrid steepest descent method.

In this paper, we present a way to avoid inversions of linear operators in the iterations of the hybrid steepest descent method even if the first stage criterion involves the composition of a nonsmooth convex function and a linear operator. First, we propose an inversion-free nonexpansive operator to characterize the solution set of the first stage problem. To avoid inversions of linear operators, we reformulate the first stage problem into an equivalent problem in a product space. Accordingly, the proposed operator is defined in a product of Hilbert spaces equipped with the standard inner product. Second, we propose an algorithmic solution to certain hierarchical convex optimization problems by plugging the proposed operator into the hybrid steepest descent method. For a given hierarchical convex optimization problem in a Hilbert space, the proposed operator for the first stage problem is defined in a certain product space. To fill this gap for application of the hybrid steepest descent method, we present a translation of the hierarchical convex optimization problem into an equivalent problem in the product space. Preserving some preferable properties on the original second stage criterion, we design carefully a regularization term for the second stage criterion in the product space to satisfy conditions offered by previous convergence results on the hybrid steepest descent method. Consequently, we succeed in avoiding the inversions of linear operators in the update of the hybrid steepest descent method if the first stage criterion involves a single composition.

**Notation:** Let $\mathcal{X}$ be a real Hilbert space equipped with[1] an inner product $\langle\cdot,\cdot\rangle$ and its induced norm $\|\cdot\| = \sqrt{\langle\cdot,\cdot\rangle}$, which is denoted by $(\mathcal{X},\langle\cdot,\cdot\rangle,\|\cdot\|)$. Let $(\mathcal{K},\langle\cdot,\cdot\rangle_{\mathcal{K}},\|\cdot\|_{\mathcal{K}})$ be another real Hilbert space. Let $A\colon \mathcal{X}\to\mathcal{K}$ be a bounded linear operator of which the norm is defined by $\|A\|_{\mathrm{op}} := \sup_{x\in\mathcal{X}\setminus\{0\}}\frac{\|Ax\|_{\mathcal{K}}}{\|x\|}$. Then, the operator $A^*\colon \mathcal{K}\to\mathcal{X}$ denotes its adjoint, i.e.,

$$(\forall(x,u)\in\mathcal{X}\times\mathcal{K})\quad \langle x, A^*u\rangle = \langle Ax, u\rangle_{\mathcal{K}}.$$

Let $\Gamma_0(\mathcal{X})$ be the set of all proper lower-semicontinuous convex functions defined over the real Hilbert space $\mathcal{X}$.

Other mathematical preliminaries are presented in Appendix for readability.

## II. Hybrid Steepest Descent Method

Consider the problem

$$\text{find } x_\star \in \operatorname*{argmin}_{x\in\mathrm{Fix}(T)} \psi(x) =: \Omega \neq \varnothing, \tag{1}$$

where $\psi\colon\mathcal{X}\to\mathbb{R}$ is differentiable and $T\colon\mathcal{X}\to\mathcal{X}$ is a nonexpansive operator (see (29)) with $\mathrm{Fix}(T) := \{x\in\mathcal{X} \mid Tx = x\}\neq\varnothing$. For problem (1), the hybrid steepest descent method

$$x_{k+1} = T(x_k) - \lambda_{k+1}\nabla\psi(T(x_k)) \tag{2}$$

is an algorithmic solution, i.e., the sequence $(x_k)_{k\in\mathbb{N}}$ generated by (2) converges strongly[2] to a solution of problem (1).

**Fact 1.** *[28, special case of Theorems 3.2 and 3.3] Let $T\colon\mathcal{X}\to\mathcal{X}$ be a nonexpansive operator. Suppose that $\psi\colon\mathcal{X}\to\mathbb{R}$ be a differentiable convex function and $\nabla\psi$ is $\kappa$-Lipschitzian and $\eta$-strongly monotone over $T(\mathcal{X}) := \{T(x)\in\mathcal{X} \mid x\in\mathcal{X}\}$. Then, with any $x_0$, the sequence $(x_k)_{k\geq 0}$ generated by (2) converges strongly to the uniquely existing solution of problem (1), if the sequence $(\lambda_k)_{k\geq1}\subset[0,1]$ satisfying one of the following two condition tuples:*

$$(L1)\ \lim_{k\to\infty}\lambda_k = 0,$$
$$(L2)\ \sum_{k\geq1}\lambda_k = \infty,$$
$$(L3)\ \lim_{k\to\infty}(\lambda_k-\lambda_{k+1})\lambda_{k+1}^{-2} = 0;$$

*or*

$$(W1)\ \lim_{k\to\infty}\lambda_k = 0,$$
$$(W2)\ \sum_{k\geq1}\lambda_k = \infty,$$
$$(W3)\ \lim_{k\to\infty}|\lambda_k-\lambda_{k+1}| = 0.$$

Fact 1 can be applied to a certain class of *hierarchical convex optimization*. Suppose that a nonexpansive operator $T\colon\mathcal{X}\to\mathcal{X}$ characterizes, as its fixed point set, the solution set of the minimization of a function $\varphi\in\Gamma_0(\mathcal{X})$, i.e.,

$$\mathrm{Fix}(T) = \operatorname*{argmin}_{x\in\mathcal{X}}\varphi(x). \tag{3}$$

In this case, Fact 1 leads to an algorithmic solution to the hierarchical convex optimization problem

$$\operatorname*{minimize}_{x_\star\in\mathcal{X}}\ \psi(x_\star)\quad \text{s.t.}\quad x_\star\in\operatorname*{argmin}_{x\in\mathcal{X}}\varphi(x)\neq\varnothing.$$

In other words, constructing a nonexpansive operator which characterizes the solution set of the first stage problem, i.e., $\operatorname*{argmin}_{x\in\mathcal{X}}\varphi(x)$, is a key to solve hierarchical convex optimization problems. Obviously, a computationally efficient operator is desired because its computation dominates the whole computational cost of the iteration (2). For example, the proximity operator of $\varphi$ satisfies (3) (see (33)). Its computation is efficient if the function $\varphi$ is proximable[3]. Meanwhile, since $\varphi$ is not necessarily proximable, designs of computationally efficient nonexpansive operators are required in general.

## III. Proposed Method

We present a nonexpansive operator to characterize the solution set of the convex optimization problem

$$\text{find}\ \ x_\star\in\operatorname*{argmin}_{x\in\mathcal{X}}\ (f(x)+g(Ax)) =: \mathcal{S}_1(\neq\varnothing) \tag{4}$$

of which the criterion involves a single composition, where $f\in\Gamma_0(\mathcal{X})$ and $g\in\Gamma_0(\mathcal{K})$, and $A\colon\mathcal{X}\to\mathcal{K}$ is a bounded linear operator. Now let us introduce a real Hilbert space $\mathcal{X}\times\mathcal{K}$ equipped with the inner product $\langle\cdot,\cdot\rangle_{\mathcal{X}\times\mathcal{K}} := \langle\cdot,\cdot\rangle_{\mathcal{X}} + \langle\cdot,\cdot\rangle_{\mathcal{K}}$. Then by defining $F\in\Gamma_0(\mathcal{X}\times\mathcal{K})$ as $F\colon\mathcal{X}\times\mathcal{K}\to(-\infty,\infty] : (x,y)\mapsto f(x)+g(y)$, the problem in (4) can be translated into

$$\begin{aligned}\text{find}\ \ (x_\star,y_\star)\in &\operatorname*{argmin}_{\boldsymbol{z}=(x,y)\in\mathcal{X}\times\mathcal{K}}(F(\boldsymbol{z})+\iota_{\{0\}}(\mathcal{A}\boldsymbol{z})) && (5)\\ &= \{(x_\star,Ax_\star)\in\mathcal{X}\times\mathcal{K}\mid x_\star\in\mathcal{S}_1\}\\ &=: \mathcal{Z}_1(\neq\varnothing), && (6)\end{aligned}$$

where a bounded linear operator $\mathcal{A}\colon\mathcal{X}\times\mathcal{K}\to\mathcal{K}$ is assumed to satisfy

$$\mathcal{A}\boldsymbol{z} = 0 \Leftrightarrow Ax = y. \tag{7}$$

[1] Often $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ denotes $\langle\cdot,\cdot\rangle$ to explicitly describe its domain.

[2] (Strong and weak convergences) A sequence $(x_k)_{k\in\mathbb{N}}\subset\mathcal{X}$ is said to converge strongly to a point $x\in\mathcal{X}$ if the real number sequence $(\|x_k - x\|)_{k\in\mathbb{N}}$ converges to 0, and to converge weakly to $x\in\mathcal{X}$ if for every $y\in\mathcal{X}$ the real number sequence $(\langle x_k - x, y\rangle)_{k\in\mathbb{N}}$ converges to 0. If $(x_k)_{k\in\mathbb{N}}$ converges strongly to $x$, then $(x_k)_{k\in\mathbb{N}}$ converges weakly to $x$. The converse is true if $\mathcal{X}$ is finite dimensional, hence in finite dimensional case we do not need to distinguish these convergences.

[3] In this paper, $f\in\Gamma_0(\mathcal{X})$ is said to be *proximable* if the proximity operator $\mathrm{prox}_f$ (see (32)) is available as a computable operator. Note that the sum of two proximable convex functions is not necessarily proximable. Moreover, for a bounded linear operator $A\colon\mathcal{X}\to\mathcal{K}$, the composition $g\circ A\in\Gamma_0(\mathcal{H})$ of a proximable function $g\in\Gamma_0(\mathcal{K})$ is also not necessarily proximable.

For the problem in (5), we introduce an operator $\mathcal{T}\colon \mathcal{H} \to \mathcal{H} : (\boldsymbol{z}, \nu) \mapsto (\boldsymbol{z}_{\mathcal{T}}, \nu_{\mathcal{T}})$ with

$$\begin{cases} \boldsymbol{z}_{\mathcal{T}} = \mathrm{prox}_F(\boldsymbol{z} - \mathcal{A}^*\mathcal{A}\boldsymbol{z} + \mathcal{A}^*\nu) \\ \nu_{\mathcal{T}} = \nu - \mathcal{A}\boldsymbol{z}_{\mathcal{T}} \end{cases} \tag{8}$$

defined over a real Hilbert space $\mathcal{H} := \mathcal{X} \times \mathcal{K} \times \mathcal{K}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}} := \langle \cdot, \cdot \rangle_{\mathcal{X}} + \langle \cdot, \cdot \rangle_{\mathcal{K}} + \langle \cdot, \cdot \rangle_{\mathcal{K}}$. Then its fixed point set characterizes $\mathcal{Z}_1$ over $\mathcal{H}$, and its nonexpansivity is guaranteed as follows.

**Theorem 1.** *(a) A qualification condition*[4]

$$0 \in \mathrm{sri}(\mathcal{A}\,\mathrm{dom}(F)) \tag{9}$$

*for the problem in* (5) *implies*

$$\mathrm{Fix}(\mathcal{T}) = \mathcal{Z}_1 \times \mathcal{Z}_1^*,$$

*where*

$$\mathcal{Z}_1^* := \operatorname*{argmin}_{\nu\in\mathcal{K}} \left(F^*(\mathcal{A}^*\nu) + \iota^*_{\{0\}}(-\nu)\right) = \operatorname*{argmin}_{\nu\in\mathcal{K}} F^*(\mathcal{A}^*\nu)$$

*is the solution set of the dual problem of* (5).
*(b) The operator $\mathcal{T}$ is nonexpansive if*[5] $\|\mathcal{A}\|_{\mathrm{op}} \leq 1$, *i.e.,*

$$\|\mathcal{A}\boldsymbol{z}\|_{\mathcal{K}} \leq \|\boldsymbol{z}\|_{\mathcal{X}\times\mathcal{K}} \quad (\forall \boldsymbol{z} \in \mathcal{X} \times \mathcal{K}). \tag{10}$$

**Remark 1:** *The operator $\mathcal{T}$ in Theorem 1 can be computed efficiently if $f \in \Gamma_0(\mathcal{X})$ and $g \in \Gamma_0(\mathcal{K})$ are proximable implying thus $F$ is also proximable due to*

$$(\forall (x, y) \in \mathcal{X} \times \mathcal{K}) \quad \mathrm{prox}_F(x, y) = (\mathrm{prox}_f(x), \mathrm{prox}_g(y)).$$

*Note that inversions involving $\mathcal{A}$ are not necessary for computation of $T$.*
**Remark 2:** *A simply designed linear operator*

$$\mathcal{A} : \mathcal{X} \times \mathcal{K} \to \mathcal{K} : (x, y) \mapsto Ax - y \tag{11}$$

*satisfies the condition* (7). *In this case, the qualification condition* (9) *for the problem in* (5) *is equivalent to that of the problem in* (4), *i.e.,*

$$(9) \;\Leftrightarrow\; 0 \in \mathrm{sri}(\mathrm{dom}(g) - A\,\mathrm{dom}(f)).$$

**Remark 3:** *The proposed operator relates to the* linearized augmented Lagrangian method *(LALM) [15], [18]. In fact, an iterative use of $\mathcal{T}$*

$$(\boldsymbol{z}_{k+1}, \nu_{k+1}) = \mathcal{T}(\boldsymbol{z}_k, \nu_k)$$

*reproduces the iteration of the LALM for the problem in* (5) *(see [18] for its convergence in a finite dimensional space). Meanwhile, since Theorem 1 guarantees the nonexpansivity of $\mathcal{T}$, the* Krasnosel'skiĭ-Mann (KM) iteration *in Fact 7 specialized for $\mathcal{T}$ results in algorithmic solutions to finding a point in* $\mathrm{Fix}(\mathcal{T})$.

[4]This condition holds true in most of practical situations.

[5]Without loss of generality, we can assume that the condition (10) holds true because the conditions (7) and (9) are invariant for scaling factor, i.e., (7) and (9) imply that

$$\alpha\mathcal{A}\boldsymbol{z} = 0 \;\Leftrightarrow\; Ax = y,$$
$$0 \in \mathrm{sri}(\alpha\mathcal{A}\,\mathrm{dom}(F))$$

hold true for any $\alpha \neq 0$.

**Remark 4:** *Applicability of the proposed operator can be significantly extended by incorporating Pierra's idea [41], [42]. Roughly speaking, by introducing a product space, the sum of compositions of proximable convex functions and linear operators can be reduced to a single composition. Hence convex optimization problems whose criteria involve the sum of the compositions can be reduced to special cases of problem* (4). *Therefore, the proposed operator can be utilized to characterize the solution set of convex optimization problems even if their criteria involve the sum of the compositions.*

We shall plug $\mathcal{T}$ into Fact 1 to propose an algorithmic solution to certain hierarchical convex optimization problems. Consider

$$\text{find} \;\; x_{\star\star} \in \operatorname*{argmin}_{x\in\mathcal{S}_1} \psi(x) \;\; =: \mathcal{S}_2, \tag{12}$$

where we suppose that $\psi\colon \mathcal{X} \to \mathbb{R}$ is a differentiable convex function and $\nabla\psi$ is $\kappa$-Lipschitzian and $\eta$-strongly monotone. Note that there has not yet been reported any algorithmic solution which does not require inversion of linear operators (see Sec. IV for detail). Theorem 1 leads to an equivalent problem of (12) in the product space $\mathcal{H}$

$$\text{find} \;\; (\boldsymbol{z}_{\star\star}, \nu_{\star\star}) \in \operatorname*{argmin}_{(\boldsymbol{z},\nu)\in\mathrm{Fix}(\mathcal{T})} \Psi(\boldsymbol{z}, \nu), \tag{13}$$

where

$$\Psi\colon \mathcal{H} \to \mathbb{R} : (x, y, \nu) \mapsto \psi(x). \tag{14}$$

Hence solving (13) results in a solution of the hierarchical problem (12). Fortunately, problem (13) is a special case of (1), so that Fact 1 presents its algorithmic solution if the criterion of the second stage problem has the strongly monotone gradient. Although $\Psi$ in (14) violates this condition[6], a carefully designed regularizer defines the new function $\Upsilon$ whose gradient $\nabla\Upsilon$ is strongly monotone and translates $\mathcal{S}_2$ into the solution set of problem (15) below.

**Theorem 2.** *Consider*

$$\mathrm{minimize}_{(\boldsymbol{z},\nu)\in\mathrm{Fix}(\mathcal{T})} \quad \Upsilon(\boldsymbol{z}, \nu), \tag{15}$$

*where* $\Upsilon\colon \mathcal{H} \to \mathbb{R} : (\boldsymbol{z}, \nu) \mapsto \Psi(\boldsymbol{z}, \nu) + \frac{\eta_{\boldsymbol{z}}}{2}\|\mathcal{A}\boldsymbol{z}\|^2_{\mathcal{K}} + \frac{\eta_\nu}{2}\|\nu\|^2_{\mathcal{K}}$ *with user-defined positive constants* $\eta_{\boldsymbol{z}}, \eta_\nu > 0$.
*Then (a) the solution set of problem (15) is*

$$\{(x_{\star\star}, Ax_{\star\star}) \mid x_{\star\star} \in \mathcal{S}_2\} \times \{P_{\mathcal{Z}_1^*}(0)\}$$

*under the qualification condition (9) for the problem in* (5).
*(b) $\nabla\Upsilon$ can be expressed as*

$$\nabla\Upsilon(\boldsymbol{z}, \nu) = \begin{pmatrix} G(\boldsymbol{z}) \\ 0 \end{pmatrix} + \begin{pmatrix} \eta_z\mathcal{A}^*\mathcal{A} & O \\ O & \eta_\nu I \end{pmatrix} \begin{pmatrix} \boldsymbol{z} \\ \nu \end{pmatrix},$$

*where*

$$G\colon \mathcal{X} \times \mathcal{K} \to \mathcal{X} \times \mathcal{K} : \boldsymbol{z} = (x, y) \mapsto (\nabla\psi(x), 0).$$

*(c) $\nabla\Upsilon$ is strongly monotone if $\nabla\psi$ is $\eta$-strongly monotone and there exists constant $\eta_{\mathcal{Z}} > 0$ s.t.*

$$(\forall \boldsymbol{z} = (x, y) \in \mathcal{X} \times \mathcal{K}) \quad \eta\|x\|^2_{\mathcal{X}} + \eta_{\boldsymbol{z}}\|\mathcal{A}\boldsymbol{z}\|^2_{\mathcal{K}} \geq \eta_{\mathcal{Z}}\|\boldsymbol{z}\|^2_{\mathcal{X}\times\mathcal{K}}. \tag{16}$$

[6]Since for any $(x, y_1, \nu_1), (x, y_2, \nu_2) \in \mathcal{H}\colon (y_1, \nu_1) \neq (y_2, \nu_2)$ we have $\Psi(x, y_1, \nu_1) - \nabla\Psi(x, y_2, \nu_2) = 0$, $\nabla\Psi$ is not strongly monotone.

*For example, if $\mathcal{A}$ is defined as in* (11),

$$\eta_{\mathcal{Z}} = \min\left\{\frac{\eta}{2}, \left(\frac{\eta\eta_{\boldsymbol{z}}\|A\|_{\mathrm{op}}^{-2}}{2+\eta\eta_{\boldsymbol{z}}\|A\|_{\mathrm{op}}^{-2}}\right)\eta_{\boldsymbol{z}}\right\} > 0$$

*satisfies* (16).

*(d) Let $\mathcal{T}_\tau := (1-\tau)I + \tau\mathcal{T}$ with $\tau \in (0,1]$. Assume that $\nabla\Upsilon$ is strongly monotone over $\mathcal{T}_\tau(\mathcal{H})$, the qualification condition (9) for the problem in* (5)*, and the condition* (10) *hold true. Then, for any $(\boldsymbol{z}_0, \nu_0) \in \mathcal{H}$, the sequence generated by*

$$(\boldsymbol{z}_{k+1}, \nu_{k+1}) = \mathcal{T}_\tau(\boldsymbol{z}_k, \nu_k) - \lambda_{k+1}\nabla\Upsilon(\mathcal{T}_\tau(\boldsymbol{z}_k, \nu_k)) \quad (17)$$

*converges strongly to the unique solution of problem (15) if the conditions on $(\lambda_k)_{k\geq 1}$ in Fact 1(I) are satisfied. Therefore (17) is an algorithmic solution to problem (12) too and does not require inversions of linear operators.*

## IV. Discussions

We clarify practical advantage of the proposed operator $\mathcal{T}$ over other nonexpansive operators found implicitly in related algorithms. Here, we consider three iterative algorithms applicable to the first stage problem (4): the Alternating Direction Method of Multipliers (ADMM) [22], [23], [25], the Augmented Lagrangian Method (ALM) (essentially date back to [20], [21] and a recent paper [25]), and the Condat's primal-dual splitting algorithm [26]. We shall find that the nonexpansive operators found in the first two algorithms cannot be plugged directly into Fact 1 and the operator in the last algorithm leads possibly to an intensive computation in the exact evaluation of the gradient of the second stage criterion in the use in Fact 1.

### A. Alternating Direction Method of Multipliers (ADMM)

The ADMM

$$\begin{cases} x_{k+1} \in \underset{x\in\mathcal{X}}{\operatorname{argmin}}\left(f(x) + \frac{1}{2}\|Ax - y_k - \nu_k\|_{\mathcal{K}}^2\right) \\ y_{k+1} = \underset{y\in\mathcal{K}}{\operatorname{argmin}}\left(g(y) + \frac{1}{2}\|Ax_{k+1} - y - \nu_k\|_{\mathcal{K}}^2\right) \\ \nu_{k+1} = \nu_k - Ax_{k+1} + y_{k+1} \end{cases} \quad (18)$$

is an algorithmic solution to problem (4). It is well-known that, under a mild condition, the ADMM can be interpreted as an iterative use of a nonexpansive operator consisting of two proximity operators, whose fixed point set characterizes the solution set of the dual problem of (4) (see [23], [24], [25]).

**Fact 2.** *Consider a dual problem of* (4)*, i.e.,*

$$\underset{\nu\in\mathcal{K}}{\text{minimize}}\ f^*(A^*\nu) + g^*(-\nu).$$

*Under a qualification condition*

$$0 \in \operatorname{sri}(\operatorname{dom}(g) - A\operatorname{dom}(f))$$

*for* (4) *and*

$$0 \in \operatorname{sri}(\operatorname{dom}(f^*) - \operatorname{ran}(A^*)),$$

*the iterative process*

$$\upsilon_{k+1} = \left[\frac{I + \operatorname{rprox}_{\theta_1}\circ\operatorname{rprox}_{\theta_2}}{2}\right]\upsilon_k$$
$$\nu_{k+1} = \operatorname{prox}_{\theta_2}(\upsilon_{k+1})$$

*reproduces the update of the ADMM, where $\theta_1 := f^*\circ A^*$ and $\theta_2 := g^*\circ(-I)$, by letting $(x_k)_{k\in\mathbb{N}} \subset \mathcal{X}$ and $(y_k)_{k\in\mathbb{N}} \subset \mathcal{K}$ such that*

$$\begin{cases} x_{k+1} \in \partial f^*(A^*(\upsilon_{k+1} - \upsilon_k + \nu_k)) \\ Ax_{k+1} = \nu_k - \upsilon_{k+1} \\ y_{k+1} = \nu_{k+1} - \upsilon_{k+1}. \end{cases}$$

*Fact 7 guarantees that $(\upsilon_k)_{k\in\mathbb{N}}$ converges weakly to a point in $\operatorname{Fix}(\operatorname{rprox}_{\theta_1}\operatorname{rprox}_{\theta_2})$. In addition,*

$$\underset{\nu\in\mathcal{K}}{\operatorname{argmin}}\, f^*(A^*\nu)+g^*(-\nu) = \operatorname{prox}_{\theta_2}\left(\operatorname{Fix}\left(\operatorname{rprox}_{\theta_1}\operatorname{rprox}_{\theta_2}\right)\right).$$

However, the fixed point characterization of $\mathcal{S}_1$ through $\operatorname{Fix}(\operatorname{rprox}_{\theta_1}\operatorname{rprox}_{\theta_2})$ is hardly applicable to Fact 1 to solve the hierarchical optimization problem (12). Under the same conditions as in Fact 2, Fact 5 and Fact 6 ensure that the solution set $\mathcal{S}_1$ of problem (4) can be expressed as

$$\mathcal{S}_1 = \Pi\left(\operatorname{prox}_{\theta_2}(\operatorname{Fix}\left(\operatorname{rprox}_{\theta_1}\operatorname{rprox}_{\theta_2}\right))\right),$$

where $\Pi\colon \mathcal{K} \to 2^{\mathcal{X}} : \nu \mapsto \partial f^*(A^*\nu) \cap A^{-1}(\partial g^*(-\nu))$. This is certainly a fixed point characterization of $\mathcal{S}_1$ but not the form as in (3).

### B. Augmented Lagrangian method

The augmented Lagrangian method

$$\begin{cases} \boldsymbol{z}_{k+1} \in \underset{\boldsymbol{z}\in\mathcal{X}\times\mathcal{K}}{\operatorname{argmin}}\left(F(\boldsymbol{z}) - \langle\nu_k, \mathcal{A}\boldsymbol{z}\rangle_{\mathcal{K}} + \frac{1}{2}\|\mathcal{A}\boldsymbol{z}\|_{\mathcal{K}}^2\right) \\ \nu_{k+1} = \nu_k - \mathcal{A}\boldsymbol{z}_{k+1} \end{cases} \quad (19)$$

is an algorithmic solution to the problem in (5). As reported in [25], [43], [44], under certain conditions, the augmented Lagrangian method can be interpreted as an iterative use of a proximity operator (or the so-called *proximal point algorithm* in (34)) for the dual problem

$$\underset{\nu\in\mathcal{K}}{\text{minimize}}\ F^*(\mathcal{A}^*\nu) + \iota_{\{0\}}^*(-\nu) = F^*(\mathcal{A}^*\nu)$$

of (5). Indeed, by defining $\theta \in \Gamma_0(\mathcal{K})$ as $\theta\colon \mathcal{K} \to (-\infty,\infty] : \nu \mapsto F^*(\mathcal{A}^*\nu)$, the following fact is obtained under a slightly different condition from [25], [43], [44].

**Fact 3.** *Under the qualification condition* (9) *for the problem in* (5) *and*

$$0 \in \operatorname{sri}(\operatorname{dom}(F^*) - \operatorname{ran}(\mathcal{A}^*)),$$

*the proximal point algorithm (see* (34)*)*

$$\nu_{k+1} = \operatorname{prox}_\theta(\nu_k)$$

*for finding a point in $\operatorname{Fix}(\operatorname{prox}_\theta) = \underset{\nu\in\mathcal{K}}{\operatorname{argmin}}\,\theta(\nu)$ reproduces the augmented Lagrangian method* (19) *by letting $\boldsymbol{z}_{k+1} \in \mathcal{X}\times\mathcal{K}$ such that*

$$\begin{cases} \boldsymbol{z}_{k+1} \in \partial F^*(\mathcal{A}^*\nu_{k+1}) \\ \mathcal{A}\boldsymbol{z}_{k+1} = \nu_k - \nu_{k+1}. \end{cases}$$

However, the fixed point characterization of $\mathcal{S}_1$ through $\operatorname{Fix}(\operatorname{prox}_\theta)$ is hardly applicable to Fact 1 to solve the hierarchical optimization problem (12). Under the same conditions as in Fact 3, Fact 5 and Fact 6 ensure that $\mathcal{S}_1$ can be expressed through $\mathcal{Z}_1$ in (6) as

$$\mathcal{Z}_1 = \partial F^*(\mathcal{A}^*\operatorname{Fix}(\operatorname{prox}_\theta)) \cap \mathcal{A}^{-1}(\{0\}).$$

This is certainly a fixed point characterization of $\mathcal{S}_1$ but not the form as in (3).

### C. Condat's primal-dual splitting

Consider a convex optimization problem

$$\text{find } x_\star \in \underset{x\in\mathcal{X}}{\operatorname{argmin}}(f(x)+g(x)+h(Ax)) =: \mathcal{S}_{\mathrm{PDS}} \neq \varnothing, \quad (20)$$

where $f \in \Gamma_0(\mathcal{X})$ is a smooth convex function with $\beta$-Lipschitzian gradient $\nabla f$, $g \in \Gamma_0(\mathcal{X})$ and $h \in \Gamma_0(\mathcal{K})$ are proximable convex functions, and $A\colon \mathcal{X} \to \mathcal{K}$ is a bounded linear operator. Then, for user-defined parameters $\gamma_1, \gamma_2 > 0$, the Condat's primal-dual splitting

$$\begin{cases} x_{k+1} = \operatorname{prox}_{\gamma_1 g}(x_k - \gamma_1(\nabla f(x_k) - A^*\nu_k)) \\ \nu_{k+1} = \operatorname{prox}_{\gamma_2 h^*}(\nu_k - \gamma_2 A(2x_{k+1} - x_k)) \end{cases} \quad (21)$$

is an algorithmic solution to problem (20). It was reported in [40] that this algorithm can be interpreted as an iterative use of a nonexpansive operator.

**Fact 4.** *Define* $T_{\mathrm{PDS}}\colon \mathcal{X}\times\mathcal{K} \to \mathcal{X}\times\mathcal{K} : (x,\nu) \mapsto (\hat{x},\hat{\nu})$ *with*

$$\hat{x} := \operatorname{prox}_{\gamma_1 g}(x - \gamma_1(\nabla f(x) - A^*\nu))$$
$$\hat{\nu} := \operatorname{prox}_{\gamma_2 h^*}(\nu - \gamma_2 A(2\hat{x} - x)),$$

*where* $\gamma_1, \gamma_2 > 0$ *are chosen to satisfy*

$$\kappa := \frac{1}{\beta}\left(\frac{1}{\gamma_1} - \gamma_2\|A\|_{\mathrm{op}}^2\right) > \frac{1}{2}.$$

*Then (a) the operator* $T_{\mathrm{PDS}}$ *is* $\frac{2\kappa}{4\kappa-1}$*-averaged nonexpansive in a real Hilbert space* $\mathcal{X}\times\mathcal{K}$ *equipped with a non-standard inner product*

$$\langle (x_1,\nu_1),(x_2,\nu_2)\rangle_{\mathcal{P}} := \langle (x_1,\nu_1), \mathcal{P}(x_2,\nu_2)\rangle_{\mathcal{X}\times\mathcal{K}} \quad (22)$$

*for any* $(x_1,\nu_1),(x_2,\nu_2) \in \mathcal{X}\times\mathcal{K}$ *with*

$$\mathcal{P}\colon \mathcal{X}\times\mathcal{K} \to \mathcal{X}\times\mathcal{K} : \begin{pmatrix} x \\ \nu \end{pmatrix} \mapsto \begin{pmatrix} \frac{1}{\gamma_1}x - A^*\nu \\ -Ax + \frac{1}{\gamma_2}\nu \end{pmatrix}.$$

*(b) (Condat's primal-dual splitting method* (21)*) Suppose that the qualification condition*

$$0 \in \operatorname{sri}(\operatorname{dom}(h) - A\operatorname{dom}(f+g)) = \operatorname{sri}(\operatorname{dom}(h) - A\operatorname{dom}(g))$$

*for problem* (20) *holds true, which guarantees that the solution set* $\mathcal{S}^*_{\mathrm{PDS}} := \underset{\nu\in\mathcal{K}}{\operatorname{argmin}}\ ((f+g)^*(A^*\nu) + h^*(-\nu))$ *of the dual problem of* (20) *is non-empty, and* $\operatorname{Fix}(T_{\mathrm{PDS}}) = \mathcal{S}_{\mathrm{PDS}} \times \mathcal{S}^*_{\mathrm{PDS}} (\neq \varnothing)$*. Then, Fact 7 leads to the iterative process*

$$(x_{k+1},\nu_{k+1}) = T_{\mathrm{PDS}}(x_k,\nu_k)$$

*converging weakly to a point in* $\mathcal{S}_{\mathrm{PDS}} \times \mathcal{S}^*_{\mathrm{PDS}}$.

An efficacy of the Condat's primal-dual splitting algorithm (21) is avoiding the computation of $\operatorname{prox}_f$ and inversions involving $A$ in its update (21), which results in an efficient computational cost.

However, unfortunately, plugging $T_{\mathrm{PDS}}$ into Fact 1 results in a requirement of inversion $\mathcal{P}^{-1}$ because the gradient depends on the definition of the inner product (see (25) and also [40]). To see this, consider the hierarchical convex optimization problem

$$\underset{x\in\mathcal{X}}{\operatorname{minimize}}\ \psi(x) \quad \text{s.t.} \quad x \in \mathcal{S}_{\mathrm{PDS}},$$

which can be translated into a problem

$$\underset{(x,\nu)\in\mathcal{X}\times\mathcal{K}}{\operatorname{minimize}}\ \Upsilon_{\mathrm{PDS}}(x,\nu) \quad (23)$$
$$\text{s.t. } (x,\nu) \in \mathcal{S}_{\mathrm{PDS}} \times \mathcal{S}^*_{\mathrm{PDS}} = \operatorname{Fix}(T_{\mathrm{PDS}})$$

over a real Hilbert space $\mathcal{X}\times\mathcal{K}$ equipped with the non-standard inner product (22), where $\Upsilon_{\mathrm{PDS}} \in \Gamma_0(\mathcal{X}\times\mathcal{K})$ defined by $\Upsilon_{\mathrm{PDS}}(x,\nu) = \psi(x) + \frac{1}{2}\|\nu\|^2_{\mathcal{K}}$. Since problem (23) has the same form as (1), plugging $T_{\mathrm{PDS}}$ into Fact 1 results in an algorithmic solution to problem (23)

$$(x_{k+1},\nu_{k+1}) = T_{\mathrm{PDS}}(x_k,\nu_k) - \lambda_{k+1}\mathcal{P}^{-1}\nabla\Upsilon_{\mathrm{PDS}}(T_{\mathrm{PDS}}(x_k,\nu_k)), \quad (24)$$

where $\nabla\Upsilon_{\mathrm{PDS}}(x,\nu) = (\nabla\psi(x),\nu)$ $(\forall(x,\nu) \in \mathcal{X}\times\mathcal{K})$. Therefore, even though the operator $T_{PDS}$ can be efficiently computed, the update (24) involves the inversion[7] $\mathcal{P}^{-1}$.

## Appendix

### A. Selected Elements in Convex Optimization

**(Proper lower semicontinuous convex function)** A function $f\colon \mathcal{X} \to (-\infty,\infty]$ is said to be proper if its effective domain $\operatorname{dom}(f) := \{x \in \mathcal{X} \mid f(x) < \infty\}$ is nonempty. A function $f\colon \mathcal{X} \to (-\infty,\infty]$ is called lower semicontinuous if its lower level set $\operatorname{lev}_{\leq\alpha} f := \{x \in \mathcal{X} \mid f(x) \leq \alpha\}(\subset \mathcal{X})$ is closed for every $\alpha \in \mathbb{R}$. A function $f\colon \mathcal{X} \to (-\infty,\infty]$ is called convex if $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$ for all $x,y \in \mathcal{X}$ and $\lambda \in (0,1)$.

**(Gâteaux derivative)** Let $U$ be an open subset of $\mathcal{X}$. Then a function $f\colon U \to \mathbb{R}$ is called Gâteaux differentiable (or differentiable) at $x \in U$ if there exists $a(x) \in \mathcal{X}$ such that

$$\lim_{\delta\to 0}\frac{f(x+\delta h) - f(x)}{\delta} = \langle a(x), h\rangle \quad (\forall h \in \mathcal{X}). \quad (25)$$

In this case, $\nabla f(x) := a(x)$ is called Gâteaux derivative (or gradient) of $f$ at $x$. If a function $f \in \Gamma_0(\mathcal{X})$ is Gâteaux differentiable, $x_\star \in \mathcal{X}$ is a minimizer of $f$ if and only if $\nabla f(x_\star) = 0$.

**(Subdifferential)** For a function $f \in \Gamma_0(\mathcal{X})$, the subdifferential of $f$ is defined as the set valued operator

$$\partial f\colon\ \mathcal{X} \to 2^{\mathcal{X}}$$
$$x \mapsto \{u \in \mathcal{X} \mid \langle y - x, u\rangle + f(x) \leq f(y), \forall y \in \mathcal{X}\}.$$

Every element $u \in \partial f(x)$ is called a subgradient of $f$ at $x$. For a function $f \in \Gamma_0(\mathcal{X})$, $x_\star \in \mathcal{X}$ is a minimizer of $f$ if and only if $0 \in \partial f(x_\star)$. Note that if $f \in \Gamma_0(\mathcal{X})$ is differentiable at $x \in \mathcal{X}$, then its subdifferential at $x$ becomes singleton $\{\nabla f(x)\}$.

**(Conjugate function)** For a function $f \in \Gamma_0(\mathcal{X})$, the conjugate of $f$ is defined by

$$f^*\colon \mathcal{X} \to [-\infty,\infty] : u \mapsto \sup_{x\in\mathcal{X}}(\langle x,u\rangle - f(x)).$$

Let $f \in \Gamma_0(\mathcal{X})$. Then, we have

$$(\forall(x,u) \in \mathcal{X}\times\mathcal{X}) \quad u \in \partial f(x) \Leftrightarrow x \in \partial f^*(u),$$

[7] For this issue, a practical solution with use of the Neumann series is discussed in [40].

which implies that $(\partial f)^{-1}(u) := \{x \in \mathcal{X} \mid u \in \partial f(x)\} = \partial f^*(u)$ and $(\partial f^*)^{-1}(x) = \{u \in \mathcal{X} \mid x \in \partial f^*(u)\} = \partial f(x)$.

**(Conical hull, span, convex sets)** For a given nonempty set $C \subset \mathcal{X}$, $\mathrm{cone}(C) := \{\lambda x \mid \lambda > 0, x \in C\}$ is called the conical hull of $C$, and $\mathrm{span}(C)$ denotes the smallest linear subspace of $\mathcal{X}$ containing $C$, i.e., the intersection of all the linear subspaces of $\mathcal{X}$ containing $C$. The closure of $\mathrm{span}(C)$ is denoted by $\overline{\mathrm{span}}(C)$. A set $C \subset \mathcal{X}$ is called convex if $\lambda x + (1-\lambda)y \in C$ for all $x, y \in C$ and $\lambda \in (0,1)$. Strong relative interior of a convex set $C \subset \mathcal{X}$ is defined by

$$\mathrm{sri}(C) := \{x \in C \mid \mathrm{cone}(C - x) = \overline{\mathrm{span}}(C - x)\}.$$

**(Indicator function)** For a nonempty closed convex set $C \subset \mathcal{X}$, the indicator function of $C$ is defined by

$$\iota_C \colon \mathcal{X} \to (-\infty, \infty] : x \mapsto \begin{cases} 0, & \text{if } x \in C; \\ +\infty, & \text{if } x \notin C, \end{cases}$$

which belongs to $\Gamma_0(\mathcal{X})$. In particular, the indicator function $\iota_{\{0\}} \in \Gamma_0(\mathcal{X})$ of $\{0\} \subset \mathcal{X}$ satisfies that

$$(\forall u \in \mathcal{X}) \quad \iota_{\{0\}}^*(u) = \sup_{x \in \mathcal{X}} (\langle x, u\rangle - \iota_{\{0\}}(x)) = 0.$$

**(Fenchel-Rockafellar duality)** Let $f \in \Gamma_0(\mathcal{X})$, $g \in \Gamma_0(\mathcal{K})$, and $A \colon \mathcal{X} \to \mathcal{K}$ a bounded linear operator. The primal problem associated with the composite function $f + g \circ A$ is

$$\underset{x \in \mathcal{X}}{\mathrm{minimize}}\ f(x) + g(Ax), \tag{26}$$

its dual problem is

$$\underset{u \in \mathcal{K}}{\mathrm{minimize}}\ f^*(A^*u) + g^*(-u), \tag{27}$$

$\mu := \inf_{x \in \mathcal{X}}(f(x) + g(Ax))$ is called the primal optimal value, and $\mu^* = \inf_{u \in \mathcal{K}}(f^*(A^*u) + g^*(-u))$ the dual optimal value.

**Fact 5** ([3, Theorem 15.23])**.** *The condition*

$$0 \in \mathrm{sri}(\mathrm{dom}(g) - A\,\mathrm{dom}(f)), \tag{28}$$

*the so-called* qualification condition *for problem* (26)*, implies the existence of a minimizer of problem* (27) *and* $\mu = -\min_{u \in \mathcal{K}}(f^*(A^*u) + g^*(-u)) = -\mu^*$.

**Fact 6** ([3, Theorem 19.1])**.** *Suppose that* $\mathrm{dom}(g) \cap A\,\mathrm{dom}(f) \neq \varnothing$ *(note: this is not sufficient for* (28)*). Let* $(x, u) \in \mathcal{X} \times \mathcal{K}$*. Then the following are equivalent:*

(i) $x$ *is a solution of the primal problem,* $u$ *is a solution of the dual problem, and* $\mu = -\mu^*$.
(ii) $A^*u \in \partial f(x)$ *and* $-u \in \partial g(Ax)$.
(iii) $x \in \partial f^*(A^*u) \cap A^{-1}(\partial g^*(-u))$.

### *B. Monotone operators, nonexpansive operators, Krasnosel'skiĭ-Mann iteration*

**(Monotone operator)** A set-valued operator $T \colon \mathcal{X} \to 2^{\mathcal{X}}$ is called *monotone* over $S (\subset \mathcal{X})$ if

$$(\forall x, y \in S)(\forall u \in Tx)(\forall v \in Ty) \quad \langle u - v, x - y\rangle \geq 0.$$

In particular, it is called *$\eta$-strongly monotone* over $S$ if there exists some $\eta > 0$ s.t.

$$(\forall x, y \in S)(\forall u \in Tx)(\forall v \in Ty)\, \langle u - v, x - y\rangle \geq \eta \|x - y\|^2.$$

**(Nonexpansive operator)** An operator $T \colon \mathcal{X} \to \mathcal{X}$ is called nonexpansive if

$$(\forall x, y \in \mathcal{X}) \quad \|Tx - Ty\| \leq \|x - y\|. \tag{29}$$

A nonexpansive operator $T$ is said to be $\alpha$-averaged if there exists $\alpha \in (0,1)$ and a nonexpansive mapping $\widehat{T} \colon \mathcal{X} \to \mathcal{X}$ such that

$$T = (1 - \alpha)I + \alpha \widehat{T}. \tag{30}$$

**Fact 7. (Krasnosel'skiĭ-Mann (KM) Iteration [45], [46])** *For a nonexpansive operator* $T \colon \mathcal{X} \to \mathcal{X}$ *with* $\mathrm{Fix}(T) \neq \varnothing$ *and any initial point* $x_0 \in \mathcal{X}$*, the sequence* $(x_k)_{k \in \mathbb{N}}$ *generated by*

$$x_{k+1} = (1 - \alpha_k)x_k + \alpha_k T x_k$$

*converges weakly to a point in* $\mathrm{Fix}(T)$ *if* $(\alpha_k)_{k \in \mathbb{N}} \subset [0,1]$ *satisfies* $\sum_{k \in \mathbb{N}} \alpha_k(1 - \alpha_k) = \infty$*. Note that the weak limit of* $(x_k)_{k \in \mathbb{N}}$ *depends on the choices of* $x_0$ *and* $(\alpha_k)_{k \in \mathbb{N}}$*. In particular, if* $T$ *is* $\alpha$*-averaged (see* (30)*), a simple iteration*

$$x_{k+1} = T x_k = (1 - \alpha)x_k + \alpha \widehat{T} x_k \tag{31}$$

*converges weakly a point in* $\mathrm{Fix}(T) = \mathrm{Fix}(\widehat{T})$.

**(Proximity operator [47], [48])** The proximity operator of $f \in \Gamma_0(\mathcal{X})$ is defined by

$$\mathrm{prox}_f \colon \mathcal{X} \to \mathcal{X} : x \mapsto \underset{y \in \mathcal{X}}{\mathrm{argmin}}\, f(y) + \frac{1}{2}\|y - x\|^2. \tag{32}$$

Note that $\mathrm{prox}_f(x) \in \mathcal{X}$ is well-defined for all $x \in \mathcal{X}$ due to the coercivity and the strict convexity[8] of $f(\cdot) + \frac{1}{2}\|\cdot - x\|^2 \in \Gamma_0(\mathcal{X})$. It is also well-known that $\mathrm{prox}_f$ is nothing but the resolvent of $\partial f$, i.e., $\mathrm{prox}_f = (\partial f + I)^{-1} =: J_{\partial f}$, which implies that

$$x \in \mathrm{Fix}(\mathrm{prox}_f) \Leftrightarrow x \in \underset{y \in \mathcal{X}}{\mathrm{argmin}}\, f(y). \tag{33}$$

Thanks to this fact, the set of all minimizers of $f \in \Gamma_0(\mathcal{X})$ can be characterized in terms of a single-valued map, i.e., $\mathrm{prox}_f$. Moreover, since the proximity operator is $1/2$-averaged nonexpansive, i.e., $\mathrm{rprox}_f := 2\mathrm{prox}_f - I$ is nonexpansive, the iteration

$$x_{k+1} = \mathrm{prox}_f(x_k) \tag{34}$$

converges weakly to a point in $\underset{x \in \mathcal{X}}{\mathrm{argmin}}\, f(x) = \mathrm{Fix}(\mathrm{prox}_f)$ by (31) in Fact 7. The iterative algorithm (34) is known as *proximal point algorithm* [43].

## Acknowledgment

This work was supported in part by JSPS Grants-in-Aid (26730128).

[8](Coercivity and strict convexity) A function $f \in \Gamma_0(\mathcal{X})$ is said to be coercive if

$$\|x\| \to \infty \Rightarrow f(x) \to \infty.$$

Coercivity of $f \in \Gamma_0(\mathcal{X})$ implies $\underset{x \in \mathcal{X}}{\mathrm{argmin}}\, f(x) \neq \varnothing$. A function $f \colon \mathcal{X} \to (-\infty, \infty]$ is called strictly convex if $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$ for all $\lambda \in (0,1)$ and for all $x, y \in \mathcal{X} : x \neq y$. Strict convexity of $f \in \Gamma_0(\mathcal{X})$ implies that the set of minimizers is at most singleton.

## REFERENCES


[1] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[2] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

[3] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert space*. Springer, 2011.

[4] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992.

[5] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.

[6] M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

[7] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical Imaging and Vision*, 20(1-2):89–97, 2004.

[8] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling & Simulation*, 4(2):460–489, 2005.

[9] R. G. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4), 2007.

[10] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[11] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Sciences*, 1(3):248–272, 2008.

[12] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

[13] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[14] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[15] X. Zhang, M. Burger, X. Bresson, and S. Osher. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Sciences*, 3(3):253–276, 2010.

[16] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[17] 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*, 3(1):1–122, 2011.

[18] J. Yang and X. Yuan. Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation*, 82(281):301–329, 2013.

[19] 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*, pages 345–390. Springer, 2011.

[20] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[21] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, chapter 19. Adacemic Press, London, 1969.

[22] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[23] D. Gabay. *Applications of the method of multipliers to variational inequalities*, chapter IX. Augmented Lagrangian methods: Applications to the numerical solution of boundary value problems. North-Holland, Amsterdam, 1983.

[24] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[25] J. Eckstein. Augmented lagrangian and alternating direction methods for convex optimization: A tutorial and some illustrative computational results. Technical Report RRR 32-2012, RUTCOR Research Report, 2012.

[26] L. Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.

[27] I. Yamada, N. Ogura, and N. Shirakawa. A numerically robust hybrid steepest descent method for the convexly constrained generalized inverse problems. *Contemporary Mathematics*, 313:269–305, 2002.

[28] I. Yamada. The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings. *Studies in Computational Mathematics*, 8:473–504, 2001.

[29] N. Ogura and I. Yamada. Non-strictly convex minimization over the fixed point set of an asymptotically shrinking nonexpansive mapping. *Numerical Functional Analysis and Optimization*, 23(1 & 2), 2002.

[30] N. Ogura and I. Yamada. Nonstrictly convex minimization over the bounded fixed point set of a nonexpansive mapping. *Numerical Functional Analysis and Optimization*, 24(1-2):129–135, 2003.

[31] H. K. Xu and T. H. Kim. Convergence of hybrid steepest-descent methods for variational inequalities. *Journal of Optimization Theory and Applications*, 119(1):185–201, 2003.

[32] I. Yamada and N. Ogura. Hybrid steepest descent method for variational inequality problem over the fixed point set of certain quasi-nonexpansive mappings. *Numerical Functional Analysis and Optimization*, 25(7 & 8):619–655, 2004.

[33] F. Murtagh and J. L. Starck. *Astronomical image and data analysis*. Springer, 2006.

[34] S. Iemoto and W. Takahashi. Strong convergence theorems by a hybrid steepest descent method for countable nonexpansive mappings in hilbert spaces. *Sci. Math. Jpn.*, 21:557–570, 2008.

[35] N. Takahashi and I. Yamada. Parallel algorithms for variational inequalities over the cartesian product of the intersections of the fixed point sets of nonexpansive mappings. *Journal of Approximation Theory*, 153(2):139–160, 2008.

[36] P.-E. Maingé. Extension of the hybrid steepest descent method to a class of variational inequalities and fixed point problems with nonself-mappings. *Numerical Functional Analysis and Optimization*, 29(7-8):820–834, 2008.

[37] C. Chidume. *Geometric properties of Banach spaces and nonlinear iterations*, volume 1965. Springer, 2009.

[38] H. Iiduka. Iterative algorithm for solving triple-hierarchical constrained optimization problem. *Journal of Optimization Theory and Applications*, 148(3):580–592, 2011.

[39] H. Iiduka. Fixed point optimization algorithm and its application to power control in CDMA data networks. *Mathematical Programming*, 133(1-2):227–242, 2012.

[40] S. Ono and I. Yamada. Hierarchical convex optimization with primal-dual splitting. *IEEE Trans. Signal Process*, 63(2), 2014.

[41] G. Pierra. *Méthodes de décomposition et croisement d'algorithmes pour des problèmes d'optimisation*. PhD thesis, University of Grenoble, 1976.

[42] G. Pierra. Decomposition through formalization in a product space. *Mathematical Programming*, 28(1):96–115, 1984.

[43] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[44] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.

[45] W. R. Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

[46] M. A. Krasnosel'skiĭ. Two remarks on the method of successive approximations. *Uspekhi Matematicheskikh Nauk*, 10(1), 1955.

[47] J.-J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

[48] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.