On the Convergence of the Iterative Shrinkage/Thresholding Algorithm with a Weakly Convex Penalty

İlker Bayram

Abstract

We consider the iterated shrinkage/thresholding algorithm (ISTA) applied to a cost function composed of a data fidelity term and a penalty term. The penalty is non-convex but the concavity of the penalty is accounted for by the data fidelity term so that the overall cost function is convex. We provide a generalization of the convergence result for ISTA viewed as a forward-backward splitting algorithm. We also demonstrate experimentally that for the current setup, using large stepsizes in ISTA can accelerate convergence more than existing schemes proposed for the convex case, like TwIST or FISTA.

I. INTRODUCTION

The iterative shrinkage/thresholding algorithm (ISTA) is a widely-used tool for minimization problems of the form

$$\min_x \left\{ C(x) = \frac{1}{2} \|y - H\,x\|_2^2 + P(x) \right\} \tag{1}$$

where y is an observed signal, H is a linear operator, and $P(x)$ is a penalty term reflecting our prior knowledge about the object to be recovered [17, 20, 28]. For minimizing $C(x)$, ISTA employs iterations of the form

$$x^{k+1} = T_\alpha \left(x^k + \alpha\,H^T\,(y - H\,x) \right), \tag{2}$$

where T_α is the shrinkage/thresholding operator associated with P, defined as,

$$T_\alpha(z) = \arg\min_x \frac{1}{2\alpha}\|x - z\|_2^2 + P(x). \tag{3}$$

When $P(x)$ is convex, T is also referred to as the proximity operator of P [15]. We remark that if the minimization problem in (3) is strictly convex, then it has a unique solution. In that case, T_α will be well-defined. In the following, we denote the least and greatest eigenvalue of $H^T\,H$ with σ_m and σ_M respectively. We also assume throughout the paper that $P(x)$ is lower semi-continuous and $\inf_x C(x) > -\infty$.

For convex P, ISTA can be derived through different approaches. The majorization-minimization (MM) framework [19, 23, 26], enforces the constraint $\alpha < 1/\sigma_M$, and implies that the iterates achieve monotone descent in the cost. The descent property achieved by MM does not directly imply that the algorithm is convergent. However, the following proposition be shown by making use of this descent property [10] and interpreting it as an instance of a sequential unconstrained minimization algorithm (SUMMA) [11].

Proposition 1. Suppose $P(x)$ is convex. If $0 < \alpha \le 1/\sigma_M$, then the sequence x^k in (2) monotonically reduces the cost (i.e., $C(x^{k+1}) \le C(x^k)$) and converges to a global minimizer of $C(x)$. $\qquad\square$

The forward-backward splitting algorithm [16] gives exactly the same iterations as (2), but allows $\alpha < 2/\sigma_M$. In practice, higher step-sizes accelerate the convergence to a minimizer. Although the MM interpretation is not valid when $\alpha > 1/\sigma_M$, it can be shown also for $1/\sigma_M < \alpha < 2/\sigma_M$ that the cost decreases monotonically with each iteration [35]. However, the convergence proofs in [2, 16] do not make use of this monotone descent property.

İ. Bayram is with the Dept. of Electronics and Communication Engineering, Istanbul Technical University, Istanbul, Turkey (e-mail : ibayram@itu.edu.tr).

Rather, the main object of interest in [2, 16] is the distance to the set of minimizers, which shows a monotone behavior.

Proposition 2. [2, 16] Suppose $P(x)$ is convex. If $0 < \alpha < 2/\sigma_M$, then the iterates x^k in (2) converge to a global minimizer of $C(x)$. □

When $P(x)$ is non-convex, ISTA is still applicable but convergence to a global minimizer is not guaranteed in general. Further, the convergence proof for the forward-backward scheme as in [2, 16] is not directly valid. In this paper, instead of an arbitrary non-convex penalty, we consider weakly-convex penalties [36] [1].

Definition 1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be ρ-weakly convex if

$$h(x) = \frac{s}{2} \|x\|_2^2 + f(x) \tag{4}$$

is convex when $s \geq \rho \geq 0$. □

If $P(x)$ is ρ-weakly convex and $0 < \rho \leq \sigma_m$, then $C(x)$ can be shown to be convex, even though $P(x)$ is not. The minimization problem that arises in this case is not merely of academic interest – see for instance [34] for an iterative scheme where each iteration requires solving such a problem, [29] for a binary denoising formulation which employs weakly penalties. More generally, weakly convex penalties are of interest in sparse recovery, because they allow to reduce the bias in the estimates, which arise when convex penalties such as the ℓ_1 norm are utilized [4, 8, 13, 34]. Also, weakly convex penalties are general enough to produce any monotone threshold function [3, 13].

Through the MM scheme, for $\alpha < 1/\sigma_M$, we obtain in Sec. III the iterations in (2) for a weakly convex penalty and show that the algorithm converges to a minimizer by adapting the proof in [10]. Specifically, we show the following.

Proposition 3. Suppose P is ρ-weakly convex with $0 \leq \rho \leq \sigma_m$. If $0 < \alpha < 1/\sigma_M$, then the sequence x^k in (2) monotonically reduces the cost $C(x)$ and converges to a global minimizer of $C(x)$. □

When the penalty is convex, we noted that the step-size allowed by MM can be doubled. Therefore, one is tempted to ask if larger stepsizes could be used to accelerate the algorithm for when P is weakly convex. The answer is not trivial because, when $\rho > 0$, the operator T_α loses some of its properties. In particular, it is not firmly non-expansive [2], a property which is instrumental in proving Prop. 2. Loss of non-expansivity can be visibly seen by noting that the derivatives of the threshold functions exceed unity when $\rho > 0$ (see Figs. 2, 6 in this paper or Fig. 2 in [34]). On the other hand, weak convexity is a mild departure from convexity and one expects some generalization of Prop. 2 to hold. We have the following result in this direction.

Proposition 4. Suppose P is ρ-weakly convex with $0 \leq \rho \leq \sigma_m$. If $0 < \alpha < 2/(\sigma_M + \rho)$, then the sequence x^k in (2) converges to a global minimizer of $C(x)$. □

If we regard ρ as a measure of the deviation of P from being convex, then we see that this deviation from convexity shows itself in the maximum step-size allowed. Note that since $\rho \leq \sigma_m \leq \sigma_M$, the maximum step-size α allowed by Prop. 4 is in general greater than that allowed by Prop 3. We also note that as in the convex case, it can be shown that the cost monotonically decreases with each iteration but this property is not used in the proof of the proposition.

Generalization to an Arbitrary Data Term

Application of ISTA is not restricted to cost functions that employ quadratic data terms. More generally, consider a generic cost function of the form

$$D(x) = f(x) + P(x), \tag{5}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. To minimize D, ISTA constructs a sequence as

$$x^{k+1} = T_\alpha\Big(x^k - \alpha \nabla f(x)\Big). \tag{6}$$

---

[1]See specifically Defn.1 and Prop. 4.3 in [36]. We use a slightly different definition than [36], in order to simplify notation. Our ρ-weakly convex functions correspond to $(-\rho/2)$-convex functions of [36].

In this setup, the following proposition applies for when f and P are convex [2, 16].

Proposition 5. [2, 16] Suppose f and P are convex and for $\sigma > 0$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \sigma \|x - y\|_2 \text{ for all } x, y. \tag{7}$$

If $0 < \alpha < 2/\sigma$, then the iterates x^k in (6) converge to a global minimizer of $D(x)$ in (5). $\qquad\square$

We note that this is a generalization of Prop. 2 since for $f(x) = \frac{1}{2}\|y - Hx\|_2^2$, we have $\nabla f(x) = H^T(Hx - y)$, and thus $\|\nabla f(x) - \nabla f(y)\| \leq \sigma_M \|x - y\|_2$. Also, when $1/\sigma < \alpha < 2/\sigma$, even though the MM interpretation is not valid, it is possible to show monotonic descent of the cost [35].

For weakly convex P, we show in this paper that Prop. 5 generalizes as follows.

Proposition 6. Suppose $P(x)$ is ρ-weakly convex, $f(x) - \frac{\rho}{2}\|x\|_2^2$ is convex and for $\sigma > 0$,

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq \sigma \|x - y\|_2 \text{ for all } x, y. \tag{8}$$

If $0 < \alpha < 2/(\sigma + \rho)$, then the sequence x^k in (6) converges to a global minimizer of $D(x)$. $\qquad\square$

As in the case with a quadratic data fidelity term, it can also be shown [25, 35] that the cost decreases monotonically with each iteration. Although the proof of Prop. 6 does not depend on this descent property, we will provide a proof of the following proposition for the sake of completeness. We note, however, that the proof we present follows [25, 35].

Proposition 7. Suppose the hypotheses of Prop. 6 hold. Then, $D(x^{k+1}) \leq D(x^k)$. $\qquad\square$

Although Prop. 4 is a corollary of Prop. 6, we separately prove Prop. 4 because showing convergence in this special case allows a more elementary proof. Especially, if $\rho < \sigma_m$, rather than just $\rho \leq \sigma_m$, yet a simpler proof is valid. We will present this simple proof before considering the more general case $\rho \leq \sigma$. We also note that the analysis in this paper can be put in a more compact form by resorting to results from non-smooth (non-necessarily convex) analysis [14, 32]. However, the additional technical requirements in non-smooth non-convex analysis can be avoided because we are not interested in an arbitrary non-convex problem. Rather, we work under a weak-convexity assumption and this in turn allows us to derive the results using convex analysis methods, simplifying the discussions.

Related Work and Contribution

Convergence of ISTA with shrinkage/threshold functions other than the soft-threshold has been studied previously in [25, 35, 37]. The results of [25, 35] imply that descent in the cost is achieved for the sequence x^k, in the sense $C(x^{k+1}) \leq C(x^k)$, when the step-size satisfies $\alpha < 2/(\sigma + \rho)$ as in Prop. 6. This property, along with the global convergence theorem (see Thm. 7.2.3 in [5]) implies that accumulation points of x^k minimize the cost function. ISTA with firm-thresholding instead of soft-thresholding is studied in [37] and it is shown by an MM argument that the cost is reduced at each iteration for $\alpha < 1/\sigma$. The authors then conclude that the algorithm is convergent by an argument based on the properties of the penalty function.

In this paper, we present two independent proofs of convergence for ISTA with weakly-convex penalties. The first proof (Prop. 3) uses the MM and SUMMA interpretations of the algorithm and relies on the monotonic descent of the cost with each iteration. This proof is obtained by adapting the proof of convergence presented in [10, 11] (where it is assumed that the penalty is convex, unlike the case considered here). Unfortunately, in the context of ISTA, this approach does not extend to the case where the algorithm falls out of the MM framework. Our second proof of convergence (for Prop. 4 and Prop. 6), which covers cases that fall out of the MM framework, does not rely on such a descent property. We instead study the mapping that ISTA employs, following the schema of [16]. For weakly convex penalties, such a study has not appeared as far as we are aware.

Outline

In Section II, we recall some definitions and results from convex analysis. We provide an MM derivation of ISTA and prove Prop. 3 in Section III. To address the cases that fall out of the MM framework, we study in Section IV the operator that maps x^k to x^{k+1} in (2). Specifically, we show the relation between the fixed points

of this operator and the minimizers of the cost in Section IV-A, study the threshold operator for a weakly convex penalty in Section IV-B, provide a short and simple convergence proof for the case where the cost is strictly convex in Section IV-C and finally provide the proof of Prop. 4 in Section IV-D. The proofs of Prop. 6 and Prop. 7 are provided in Section V. In order to demonstrate that larger stepsizes may be more favorable, we also present two experiments in Section VI. We conclude with a brief outlook in Section VII.

Some of the frequently used terms/symbols are listed in Table I. We also note that our discussion is restricted to functions defined on $\mathbb{R}^n$ for simplicity.

II. DEFINITIONS AND RESULTS FROM CONVEX ANALYSIS

For later reference, we briefly recall some definitions and results from convex analysis in this section. We refer to [22, 31, 36] for further discussion.

Definition 2. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex. The subdifferential of f at $x \in \mathbb{R}^n$ is denoted by $\partial f(x)$ and is defined to be the set of $z \in \mathbb{R}^n$ that satisfy

$$f(x) + \langle y - x, z \rangle \leq f(y), \text{ for all } y. \tag{9}$$

Any element of $\partial f(x)$ is said to be a subgradient of f at x. $\qquad \square$

Using the notion of a subdifferential, the minimizer of a convex function can be easily characterized.

Proposition 8. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex. x minimizes f if and only if $0 \in \partial f(x)$. $\qquad \square$

In order to counter the concavity introduced by a weakly convex penalty function, we will need the data fidelity term to be strongly convex [36].

Definition 3. For $\rho \geq 0$, a function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be ρ-strongly convex if

$$h(x) = f(x) - \frac{s}{2} \|x\|_2^2 \tag{10}$$

is convex when $s \leq \rho$. $\qquad \square$

The following lemma, which is of interest in the proximal algorithm, will be used in the sequel.

Lemma 1. If $h : \mathbb{R}^n \to \mathbb{R}$ is convex and there exists some x and some $\beta > 0$ such that

$$h(x) \leq \beta \|z - x\|_2^2 + h(z) \text{ for all } z, \tag{11}$$

then h achieves its minimum at x. $\qquad \square$

Proof. Note that the function $f(z) = \|z - x\|_2^2 + h(z)$ achieves its minimum at $z = x$. This means that

$$0 \in \left(2\beta(z - x) + \partial h(z) \right)\Big|_{z=x} \tag{12}$$

or, $0 \in \partial h(x)$. $\qquad \square$

III. CONVERGENCE OF ISTA VIA MAJORIZATION MINIMIZATION AND SUMMA (PROOF OF PROP. 3)

In this section, we briefly recall the MM scheme [19, 23] to show that ISTA achieves monotone descent for a weakly-convex penalty provided that the step size is small enough. We remark however that achieving descent does not automatically imply that the algorithm converges [24, 27]. In order to show convergence to a minimizer in this context, we follow the approach presented in [10], based on sequential unconstrained minimization algorithms (SUMMA) [11].

A. Descent Property via MM

Suppose that at the kth iteration, we have the estimate x^k. Let us define,

$$g(x, x^k) = \frac{1}{2} \left\langle x - x^k, \left(\frac{1}{\alpha} I - H^T H \right) (x - x^k) \right\rangle, \tag{13}$$

$$M(x, x^k) = C(x) + g(x, x^k). \tag{14}$$

Observe that if $\sigma_M < 1/\alpha$, then the matrix $\frac{1}{\alpha} I - H^T H$ is positive definite. Therefore in this case,

(i) $g(x^k, x^k) = 0 \iff M(x^k, x^k) = C(x^k)$,

(ii) $g(x, x^k) \geq 0 \iff M(x, x^k) \geq C(x)$.

Thus, we can achieve descent in $C(x)$ by minimizing $M(x, x^k)$. After rearranging, $M(x, x^k)$ can be written as,

$$M(x, x^k) = \frac{1}{2\alpha} \left\| x - z^k \right\|_2^2 + P(x) + \text{const.} \tag{15}$$

where $z^k = x^k - \alpha \, H^T(Hx^k - y)$ and 'const.' is independent of x. Now since

$$\rho \leq \sigma_m \leq \sigma_M < 1/\alpha, \tag{16}$$

the function in (15) is strictly convex with respect to x (for fixed x^k) and its unique minimizer is $T_\alpha(z^k)$.

From the foregoing discussion, we conclude that if $\sigma_M < 1/\alpha$ and

$$x^{k+1} = T_\alpha \left(x^k + \alpha \, H^T \left(y - H \, x^k \right) \right), \tag{17}$$

then $C(x^k)$ is a non-increasing sequence. However, this observation alone does not directly imply that x^k converges to a minimizer. We now consider the convergence proof for Prop. 3.

B. Convergence to a Minimizer via SUMMA

We start with a key observation.

Lemma 2. For g, M, x^k defined as in (13), (14), (17), if $\sigma_M < 1/\alpha$, then for any x,

$$M(x, x^k) - M(x^{k+1}, x^k) \geq \left(\frac{1}{2\alpha} - \frac{\rho}{2} \right) \|x - x^{k+1}\|_2^2 \tag{18a}$$

$$\geq g(x, x^{k+1}). \tag{18b}$$

For the proof of this lemma, we use the following auxiliary result, which we will also refer to later.

Lemma 3. For a given z, let $\hat{x} = T_\alpha(z)$, where P is ρ-weakly convex. Then, for any x,

$$P(x) - P(\hat{x}) \geq - \left\{ \frac{\rho}{2} \|x - \hat{x}\|_2^2 + \frac{1}{\alpha} \langle \hat{x} - z, x - \hat{x} \rangle \right\}. \tag{19}$$

Proof. Let $P_\rho(x) = P(x) + \frac{\rho}{2} \|x\|_2^2$. Note that P_ρ is convex. From the definition of T in (3) we have,

$$\hat{x} = \arg\min_x \frac{1}{2\alpha} \|x - z\|_2^2 + P(x) \tag{20a}$$

$$= \arg\min_x \left(\frac{1}{2\alpha} - \frac{\rho}{2} \right) \|x - z\|_2^2 - \rho \langle x, z \rangle + P_\rho(x) \tag{20b}$$

By the optimality conditions, we obtain

$$\left(\frac{1}{2\alpha} - \frac{\rho}{2}\right)(z - \hat{x}) + \rho z \in \partial P_\rho(\hat{x}). \tag{21}$$

But by the definition of subdifferential, this implies that

$$P_\rho(x) \geq P_\rho(\hat{x}) + \left\langle \left(\frac{1}{2\alpha} - \frac{\rho}{2}\right)(z - \hat{x}) + \rho z, x - \hat{x} \right\rangle. \tag{22}$$

Plugging in the definition of P_ρ and rearranging, we obtain (19). □

Using this lemma, we obtain the proof of Lemma 2 as follows.

Proof of Lemma 2. Let $f(x) = \frac{1}{2}\|y - Hx\|_2^2$. Observe that $\nabla f(x) = H^T(Hx - y)$. Noting that $x^{k+1} = T_\alpha(x^k - \alpha \nabla f(x^k))$ and using Lemma 3, we obtain,

$$M(x, x^k) - M(x^{k+1}, x^k)$$

$$= \left\{\frac{1}{2\alpha}\left(\|x - x^k\|_2^2 - \|x^{k+1} - x^k\|_2^2\right)\right\} + \langle x - x^{k+1}, \nabla f(x^k)\rangle + \left\{p(x) - p(x^{k+1})\right\} \tag{23}$$

$$\geq \left\{\frac{1}{2\alpha}\|x - x^{k+1}\|_2^2 + \frac{1}{\alpha}\langle x - x^{k+1}, x^{k+1} - x^k\rangle\right\} + \langle x - x^{k+1}, \nabla f(x^k)\rangle \tag{24}$$

$$\quad - \left\{\frac{\rho}{2}\|x - x^{k+1}\|_2^2 + \frac{1}{\alpha}\langle x^{k+1} - x^k + \alpha \nabla f(x^k), x - x^{k+1}\rangle\right\}$$

$$= \left(\frac{1}{2\alpha} - \frac{\rho}{2}\right)\|x - x^{k+1}\|_2^2. \tag{25}$$

To see (18b), note that $H^T H - \rho I \geq 0$ and observe

$$\frac{1}{\alpha}I - H^T H = \left(\frac{1}{\alpha} - \rho\right)I - \left(H^T H - \rho I\right) \tag{26}$$

$$\leq \left(\frac{1}{\alpha} - \rho\right)I. \tag{27}$$

□

We are now ready to present the proof of convergence. Although there are some variations, the main idea follows the proof of Theorem 4.1 in [10].

Proof of convergence for Prop. 3. By assumption we have $\inf_x C(x) = c > -\infty$. Thus, $C(x^k)$ is a non-increasing sequence which is bounded from below and therefore converges to some $b \geq c$.

Now let $\hat{x}$ be a point such that $C(\hat{x}) \leq b$. Consider $d_k(\hat{x}) = M(\hat{x}, x^k) - M(x^{k+1}, x^k)$. We know by the definition of x^{k+1} that $d_k(\hat{x})$ is non-negative. We now show that it is also non-increasing with k. Using Lemma 2 (with $g(\hat{x}, x^k) \leq d_{k-1}(\hat{x})$), we write

$$d_k(\hat{x}) = M(\hat{x}, x^k) - M(x^{k+1}, x^k) \tag{28}$$

$$= g(\hat{x}, x^k) + C(\hat{x}) - g(x^{k+1}, x^k) - C(x^{k+1}) \tag{29}$$

$$\leq d_{k-1}(\hat{x}) + \left\{C(\hat{x}) - g(x^{k+1}, x^k) - C(x^{k+1})\right\} \tag{30}$$

$$\leq d_{k-1}(\hat{x}). \tag{31}$$

where the last line follows because $C(\hat{x}) \leq C(x^k)$ for all k and $g(x^{k+1}, x^k)$ is non-negative.

Again by Lemma 2, we can now conclude that

$$\|\hat{x} - x^k\|_2^2 \leq \left(\frac{1}{2\alpha} - \frac{\rho}{2}\right)^{-1} \left(M(\hat{x}, x^0) - M(x^1, x^0)\right), \tag{32}$$

for all k. Thus x^k is a bounded sequence and it has a convergent subsequence by the Bolzano-Weierstrass theorem [33]. Let x^* be the limit of a convergent subsequence x^{k_n}. Observe that since $C(x^k)$ is non-increasing and $C(\cdot)$ is

lower semi-continuous, we have that $C(x^*) \leq b$. Also, for a given $\epsilon > 0$, we can find some N such that if $n \geq N$, then

$$\left(\frac{1}{2\alpha} - \frac{\rho}{2}\right) \|x^* - x^{k_n}\|_2^2 \leq \epsilon. \tag{33}$$

It thus follows by (18b), along with $C(x^*) \leq C(x^{k_n+1})$ and the non-negativity of g that

$$M(x^*, x^{k_n}) - M(x^{k_n+1}, x^{k_n}) = C(x^*) + g_{k_n}(x^*, x^{k_n}) - C(x^{k_n+1}) + g_{k_n}(x^{k_n+1}, x^{k_n}) \tag{34}$$
$$\leq \epsilon. \tag{35}$$

But since $M(x^*, x^k) - M(x^{k+1}, x^k)$ is a non-increasing, non-negative sequence, this implies that the whole sequence converges to zero. In view of (18a), we thus conclude that

$$\lim_{k \to \infty} \|x^* - x^{k+1}\|_2^2 = 0. \tag{36}$$

Therefore, the whole sequence converges to x^*.

What remains is to show that x^* is actually a global minimizer of C. For that, we show $b = c$. Let z be a minimizer of C, i.e., $C(z) = c$. From Lemma 2, we obtain

$$C(z) + g(z, x^k) - C(x^{k+1}) + g(x^{k+1}, x^k) = M(z, x^k) - M(x^{k+1}, x^k) \tag{37}$$
$$\geq g(z, x^{k+1}). \tag{38}$$

Rearranging, we have,

$$g(z, x^k) - g(z, x^{k+1}) \geq C(x^{k+1}) - C(z) - g(x^{k+1}, x^k) \tag{39}$$
$$\geq b - c. \tag{40}$$

Now if $b > c$, then $g(z, x^k)$ decreases indefinitely with k, which cannot happen since $g(z, x^k) \geq 0$. Since also $c \leq b$, it follows that $b = c$. $\qquad\square$

IV. ISTA AS FIXED-POINT ITERATIONS OF AN OPERATOR

When $P(x)$ is convex, the forward-backward splitting algorithm [2, 16] leads to iterations that are of the same form as (17). However, the results of [16] imply that the maximum step size allowed by MM can in fact be doubled while still ensuring convergence. This in turn accelerates convergence significantly. We now investigate this issue for a weakly-convex penalty $P(x)$.

In order to simplify our analysis, we decompose the operator in (17). We define

$$U_\alpha(x) = \alpha H^T y + \left(I - \alpha H^T H\right) x, \tag{41}$$

and rewrite the iterations in (17) as,

$$x^{k+1} = T_\alpha\Big(U_\alpha(x^k)\Big). \tag{42}$$

In the following, we will first study the fixed points of the composite operator $T_\alpha U_\alpha$ and show an equivalence with the minimizers of the cost $C(x)$. Then, we study the properties of the two operators T_α and U_α. Under a strict convexity assumption, we will see that the composition is actually a contraction mapping. If we lift the strictness restriction from the convexity assumption, the composite operator can be shown to be averaged (see Sec. IV-D).

A. Fixed Points of the Algorithm

We now establish a relation between the fixed points of $T_\alpha U_\alpha$ and the minima of $C(x)$. Specifically, our goal in this subsection is to show the following result.

Proposition 9. Suppose $P(x)$ is ρ-weakly convex, T_α is as defined in (3) and $\alpha\rho < 1$. Then, $x = T_\alpha\Big(U_\alpha(x)\Big)$ if and only if x minimizes $C(x)$ in (1).

Instead of proving Prop. 9 directly, we will prove a more general form using convex analysis methods. This general form will also be referred to in the proof of Prop. 6 in Sec. V.

Proposition 10. Suppose $P : \mathbb{R}^n \to \mathbb{R}$ is a ρ-weakly convex function, and $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable, ρ-strongly convex function. Suppose also that $\alpha\rho < 1$. Under these conditions,

$$x = T_\alpha\Big(x - \alpha\,\nabla f(x)\Big), \tag{43}$$

if and only if x minimizes $f + P$.

Proof. ($\Rightarrow$) Suppose (43) holds. We will show that x minimizes $f + P$. Let $u = x - \alpha\,\nabla f(x)$. By the definition of T_α, $x = T_\alpha(u)$ means that

$$\frac{1}{2}\|x - u\|_2^2 + \alpha\,P(x) \le \frac{1}{2}\|z - u\|_2^2 + \alpha\,P(z), \text{ for all } z. \tag{44}$$

Noting that $x - u = \alpha\,\nabla f(x)$ and $z - u = (z - x) + \alpha\,\nabla f(x)$, we can rewrite this as

$$\frac{1}{2}\|\alpha\,\nabla f(x)\|_2^2 + \alpha\,P(x) \le \frac{1}{2}\|z - x\|_2^2 + \frac{1}{2}\|\alpha\,\nabla f(x)\|_2^2 + \langle z - x, \alpha\,\nabla f(x)\rangle + \alpha\,P(z), \text{ for all } z. \tag{45}$$

Cancelling $\frac{1}{2}\|\alpha\,\nabla f(x)\|_2^2$ from both sides and noting that $\langle z - x, \alpha\,\nabla f(x)\rangle \le \alpha\,f(z) - \alpha\,f(x)$ (because f is convex), we obtain

$$\alpha\,P(x) \le \frac{1}{2}\|z - x\|_2^2 + \alpha\,f(z) - \alpha\,f(x) + \alpha\,P(z), \text{ for all } z. \tag{46}$$

Rearranging,

$$f(x) + P(x) \le \frac{1}{2\,\alpha}\|z - x\|_2^2 + f(z) + P(z), \text{ for all } z. \tag{47}$$

Since $f + P$ is convex, by Lemma 1, we conclude that x minimizes $f + P$.

($\Leftarrow$) Assume that x minimizes $h = f + P$. Let $P_\rho(t) = P(t) + (\rho/2)\|t\|_2^2$ and $f_\rho(t) = f(t) - (\rho/2)\|t\|_2^2$. Note that by assumption f_ρ and P_ρ are both convex and $h = f_\rho + P_\rho$. Since x minimizes h, we have,

$$0 \in \underbrace{\{\nabla f(x) - \rho\,x\}}_{\partial f_\rho(x)} + \partial P_\rho(x) \tag{48}$$

or,

$$\rho\,x - \nabla f(x) \in \partial P_\rho(x) \tag{49}$$

Adding any multiple of x to both sides, we find that for $s \ge \rho$, and $P_s(t) = P(t) + (s/2)\|t\|_2^2$, we have

$$s\,x - \nabla f(x) \in \partial P_s(x) \tag{50}$$

By the convexity of P_s we then obtain,

$$P_s(x) + \langle z - x, s\,x - \nabla f(x)\rangle \le P_s(z), \text{ for all } z. \tag{51}$$

Rearranging, we have

$$\frac{s}{2}\|x\|_2^2 - \langle x, s\,x - \nabla f(x)\rangle + P(x) \le \frac{s}{2}\|z\|_2^2 - \langle z, s\,x - \nabla f(x)\rangle + P(x), \text{ for all } z. \tag{52}$$

Equivalently, for $s \ge \rho$,

$$\frac{1}{2}\Big\|x - \big(x - s^{-1}\nabla f(x)\big)\Big\|_2^2 + \frac{1}{s}P(x) \quad\le \frac{1}{2}\Big\|z - \big(x - s^{-1}\nabla f(x)\big)\Big\|_2^2 + \frac{1}{s}P(z), \text{ for all } z. \tag{53}$$

Now let $\alpha = 1/s$. The inequality above may be written as,

$$\frac{1}{2}\left\|x - (x - \alpha\,\nabla f(x))\right\|_2^2 + \alpha\,P(x) \le \frac{1}{2}\left\|z - (x - \alpha\,\nabla f(x))\right\|_2^2 + \alpha\,P(z), \text{ for all } z, \tag{54}$$

for $\alpha\,\rho \le 1$. Note that equality in (54) may be achieved by setting $z = x$. But we know that for $\alpha\,\rho < 1$, the right hand side is uniquely minimized by $z = T_\alpha(x - \alpha\,\nabla f(x))$. Thus, $x = T_\alpha(x - \alpha\,\nabla f(x))$ for $\alpha\,\rho < 1$. $\qquad\square$

Prop. 9 is a corollary of this proposition. This can be seen by taking $f(x) = \frac{1}{2}\|y - Hx\|_2^2$ and noting that it is ρ-strongly convex since $\rho \le \sigma_m$.

B. Threshold Operators Associated with Weakly Convex Penalties

We now study the operator T_α. We only assume that T_α is associated with a ρ-weakly convex penalty P via (3).

Definition 4. An operator $S : \mathbb{R}^n \to \mathbb{R}^n$ is said to be non-expansive if,

$$\|S(x) - S(z)\|_2 \le \|x - z\|_2. \tag{55}$$

We will make use of the following result (see [2, 15] for instance).

Lemma 4. Suppose $q(x)$ is convex and the operator $S(x)$ is defined as

$$S(x) = \arg\min_t \; \frac{1}{2}\,\|x - t\|_2^2 + q(t).$$

Then,

$$\|S(x) - S(z)\|_2 \le \|x - z\|_2. \tag{56}$$

Proof. Let $u = S(x)$, $v = S(z)$. We have, by the optimality conditions,

$$0 \in u - x + \partial f(u), \tag{57a}$$
$$0 \in z - v + \partial f(v). \tag{57b}$$

Summing and rearranging, we obtain that for some $f_u \in \partial f(u)$ and $f_v \in \partial f(v)$,

$$x - z = u - v + (f_u - f_v). \tag{58}$$

Thus,

$$\|x - z\|_2^2 = \|u - v\|_2^2 + \|f_u - f_v\|_2^2 + \underbrace{\langle u - v, f_u - f_v\rangle}_{c}. \tag{59}$$

It follows by the monotonicity of the subdifferential that $c \ge 0$. Therefore, $\|u - v\|_2^2 \le \|x - z\|_2^2$. $\qquad\square$

Proposition 11. Suppose $P(x)$ is ρ-weakly convex and $\alpha\,\rho < 1$. Then,

$$\|T_\alpha(x) - T_\alpha(z)\|_2 \le \frac{1}{1 - \alpha\,\rho}\|x - z\|_2. \tag{60}$$

Proof. Note that the function

$$\frac{\rho}{2}\|x\|_2^2 + P(x) \tag{61}$$

is convex. Therefore,

$$S(x) = \arg\min_t \frac{1}{2}\|x - t\|_2^2 + c\left(\frac{\rho}{2}\|t\|_2^2 + P(t)\right) \tag{62}$$

is non-expansive by Lemma 4 for any $c > 0$. But we have

$$S(x) = \arg\min_t \frac{1}{2}\|x - t\|_2^2 + c\left(\frac{\rho}{2}\|t\|_2^2 + P(t)\right) \tag{63}$$

$$= \arg\min_t \frac{1 + c\rho}{2}\|t\|_2^2 - \langle x, t \rangle + c\,P(t) \tag{64}$$

$$= \arg\min_t \frac{1}{2}\left\|t - \frac{1}{1 + c\rho}x\right\|_2^2 + \frac{c}{1 + c\rho}P(t) \tag{65}$$

$$= T_{c\,(1+c\rho)^{-1}}\left(\frac{1}{1 + c\rho}x\right) \tag{66}$$

Therefore we deduce that

$$\left\|T_{c\,(1+c\rho)^{-1}}\left(\frac{1}{1 + c\rho}x\right) - T_{c\,(1+c\rho)^{-1}}\left(\frac{1}{1 + c\rho}z\right)\right\|_2 = \|S(x) - S(z)\|_2 \le \|x - z\|_2. \tag{67}$$

Now set $\alpha = c\,(1 + c\rho)^{-1}$. Observe that $\alpha\rho < 1$ for any $c > 0$. Solving for c, we have $c = \alpha\,(1 - \alpha\rho)^{-1}$. Plugging this in (67), we obtain

$$\left\|T_\alpha\left((1 - \alpha\rho)x\right) - T_\alpha\left((1 - \alpha\rho)x\right)\right\|_2 \le \|x - z\|_2. \tag{68}$$

Making a change of variables, we finally obtain

$$\|T_\alpha(x) - T_\alpha(z)\|_2 \le \frac{1}{1 - \alpha\,\rho}\|x - z\|_2. \tag{69}$$

$\square$

C. ISTA as Iterations of a Contraction Mapping

In this section, we derive a convergence result that is relatively simple. However, simplicity comes at the expense of the requirement $\rho < \sigma_m$. We desire not to exclude the case $\rho = \sigma_m$, because for $\rho = 0$ (a convex penalty function), we would like to allow $\sigma_m = 0$, which corresponds to an operator H with a non-trivial null-space. In Sec. IV-D, we will also allow the case $\rho = \sigma_m$, leading to a generalization of Prop. 2.

Proposition 12. Suppose that the eigenvalues of $H^T H$ are contained in the interval $[\sigma_m, \sigma_M]$, $P(x)$ is ρ-weakly convex and $T_\alpha(\cdot)$ is as given in (3). If

$$\rho < \sigma_m, \tag{70a}$$

$$\alpha < \frac{2}{\sigma_M + \rho}, \tag{70b}$$

then, the iterations in (17) converge to the unique minimizer of $C(x)$.

Proof. For U_α in (41), we have that

$$\|U_\alpha(x) - U_\alpha(z)\|_2 \le \max\left(|1 - \alpha\,\sigma_M|, |1 - \alpha\,\sigma_m|\right)\|x - z\|_2. \tag{71}$$

Therefore,

$$\left\|T_\alpha\left(U_\alpha(x)\right) - T_\alpha\left(U_\alpha(z)\right)\right\|_2 \le \frac{\max\left(|1 - \alpha\,\sigma_M|, |1 - \alpha\,\sigma_m|\right)}{1 - \alpha\,\rho}\|x - z\|_2, \tag{72}$$

by Prop. 11. When $\rho < \sigma_m$, $C(x)$ is strictly convex and hence has a unique minimizer. By Prop. 9, this unique minimizer is in fact the unique fixed point of $T_\alpha U_\alpha$. Let us denote this point as z. Now if $\alpha\rho < 1$,

$$\|T_\alpha\left(U_\alpha\left(x\right)\right) - z\|_2 \le \frac{\max\left(|1 - \alpha\,\sigma_M|, |1 - \alpha\,\sigma_m|\right)}{1 - \alpha\,\rho}\|x - z\|_2. \tag{73}$$

Thus the iterations converge (geometrically) to z if the three conditions below hold

$$\frac{|1 - \alpha\,\sigma_M|}{1 - \alpha\,\rho} < 1, \tag{74a}$$

$$\frac{|1 - \alpha\,\sigma_m|}{1 - \alpha\,\rho} < 1, \tag{74b}$$

$$\alpha\,\rho < 1. \tag{74c}$$

(74a) is equivalent to

$$\rho < \sigma_M, \tag{75a}$$

$$\alpha < \frac{2}{\sigma_M + \rho}. \tag{75b}$$

(74b) is equivalent to

$$\rho < \sigma_m, \tag{76a}$$

$$\alpha < \frac{2}{\sigma_m + \rho}. \tag{76b}$$

Noting that $\sigma_M \geq \sigma_m$, and $2/(\sigma_M + \rho) \leq 2/(\sigma_m + \rho)$ we deduce that (70) implies (75), and (76). Finally, by (70a), we also have $2/(\sigma_M + \rho) < 1/\rho$. Then, from (70b), (74c) follows, completing the proof. $\qquad\square$

Remark 1. Observe that since $\rho < \sigma_m \leq \sigma_M$, we have

$$\frac{2}{\sigma_M + \rho} > \frac{1}{\sigma_M}. \tag{77}$$

Thus the generalized forward backward algorithm converges for stepsizes greater than that allowed by majorization-minimization. We also note that the convergence proof given for Prop. 3 does not straightforwardly extend to the case considered in this subsection.

D. ISTA as Iterations of an Averaged Operator

Definition 5. [2] An operator $S : \mathbb{R}^n \to \mathbb{R}^n$ is said to be β-averaged with $\beta \in (0, 1)$ if S can be written as

$$S = (1 - \beta)\,I + \beta\,U, \tag{78}$$

for a non-expansive U. $\qquad\square$

As a corollary of the definition, we have,

Lemma 5. $S : \mathbb{R}^n \to \mathbb{R}^n$ is β-averaged if and only if

$$\frac{1}{\beta}\Big(S - (1 - \beta)\,I\Big) \tag{79}$$

is non-expansive. $\qquad\square$

Averaged operators are of interest because they behave more desirably concerning convergence. Further, as we will note, averaged-ness is preserved under composition, which is instrumental in proving the convergence of the forward backward splitting algorithm [2, 16]. To demonstrate this discussion let us consider a scenario as follows. Let U be non-expansive with a unique fixed point z, and x an arbitrary point. Then, $\|U(x) - z\|_2 \leq \|x - z\|_2$, but $U(x)$ is not guaranteed to be closer to z than x. In the worst case, both x and $U(x)$ might be equidistant to z. This is illustrated in Fig. 1. Now let $S = (1 - \beta)I + \beta\,U$ for $\beta \in (0, 1)$. z is also the unique fixed point of S. But now, since $S(x)$ lies somewhere on the open segment between x and $U(x)$, we will have $\|S(x) - z\|_2 < \|x - z\|_2$ (see Fig. 1). This discussion is of course not a proof of convergence for iterated applications of S but it provides some intuition. The following proposition, which is also known as the Krasnosels'kii-Mann theorem (see Thm. 5.14 in [2]) provides a convergence result for averaged operators.

Proposition 13. Suppose $S : \mathbb{R}^n \to \mathbb{R}^n$ is β-averaged and its set of fixed points is non-empty. Given x^0, let $x^{n+1} = S(x^n)$. Then the sequence $\{x^n\}_{n \in \mathbb{Z}}$ converges and the limit is a fixed point of S.
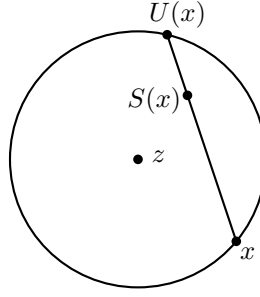
Fig. 1. A non-expansive operator U might not always take a point x closer to its fixed point z. However, an averaged operator $S = (1 - \beta)I + \beta U$ derived from U will have such a property.

Proof. See the appendix. $\qquad\square$

To prove the convergence of ISTA, we will show that $T_\alpha\big(U_\alpha(\cdot)\big)$ is an averaged operator. In order to show that this operator is averaged, we will show that T_α and U_α can be regarded as averaged operators by proper scaling (see S_α and V_α below) and invoke the following result.

Proposition 14. If S_1 and S_2 are averaged, then $S_1\big(S_2(\cdot)\big)$ is also averaged. $\qquad\square$

Proof. Suppose $S_k = (1 - \beta_k)\,I + \beta_k\,U_k$ where U_k is non-expansive and $\beta_k \in (0,1)$, for $k = 1, 2$. Then,

$$S = S_1\,S_2$$
$$= (1 - \beta_1)\,(1 - \beta_2)\,I + \underbrace{(1 - \beta_1)\,\beta_2\,U_2 + (1 - \beta_2)\,\beta_1\,U_1 + \beta_1\,U_1\Big(\beta_2\,U_2\Big)}_{U}$$

Note that

$$\|U(x) - U(z)\|_2 \le (1 - \beta_1)\,\beta_2\,\|x - z\|_2$$
$$+ (1 - \beta_2)\,\beta_1\,\|x - z\|_2 + \beta_1\,\beta_2\,\|x - z\|_2$$
$$= \Big(1 - (1 - \beta_1)(1 - \beta_2)\Big)\,\|x - z\|_2$$

Thus for $\beta = \big(1 - (1 - \beta_1)(1 - \beta_2)\big) \in (0,1)$ and $V = \beta^{-1}\,U$, we can write $S = (1 - \beta)I + \beta V$. Since V is non-expansive, the claim follows by Lemma 5. $\qquad\square$

We will also need the following result (see Cor.23.8 in [2]).

Lemma 6. Suppose $q : \mathbb{R}^n \to \mathbb{R}$ is convex and $S(x)$ is defined as

$$S(x) = \arg\min_t \frac{1}{2}\,\|x - t\|_2^2 + q(t).$$

Then, S is 1/2-averaged. $\qquad\square$

Using this lemma, we can show that the composition of the threshold operator T_α with a proper scaling is averaged.

Proposition 15. For $\alpha\,\rho < 1$, let S_α be defined as

$$S_\alpha(x) = T_\alpha\Big((1 - \alpha\,\rho)x\Big). \tag{80}$$

Then, S_α is 1/2-averaged.

Proof. In the proof of Prop. 11, for $\alpha\,\rho < 1$, we show that S_α can actually be expressed as

$$S_\alpha(x) = \arg\min_t \frac{1}{2}\|x - t\|_2^2 + \frac{\alpha}{1 - \alpha\,\rho}\,\Big(\rho\|t\|_2^2 + P(t)\Big).$$

Since $\rho\|t\|_2^2 + P(t)$ is convex, this means that S_α is 1/2-averaged by Lemma 6. $\qquad\square$

For affine operators that employ a symmetric matrix, such as those used in ISTA with a quadratic data term, averaged-ness is associated with eigenvalues of the matrix.

Proposition 16. Suppose $S : \mathbb{R}^n \to \mathbb{R}^n$ is an affine mapping of the form

$$S(x) = M\,x + u, \tag{81}$$

where M is a symmetric matrix, i.e. $M = M^T$, and u is a constant vector. Then, S is β-averaged for some $\beta \in (0,1)$ if and only if the eigenvalues of M lie in the interval $(-1, 1]$.

Proof. Suppose the eigenvalues of M lie in the interval $[\sigma_0, \sigma_1]$. Consider

$$V = \frac{1}{\beta}\Big(S - (1 - \beta)\,I\Big). \tag{82}$$

Then,

$$V(x) - V(z) = \underbrace{\frac{1}{\beta}\Big(M - (1 - \beta)I\Big)}_{Q}(x - z). \tag{83}$$

The eigenvalues of Q lie in the interval,

$$\left[\frac{\sigma_0 - 1 + \beta}{\beta}, \frac{\sigma_1 - 1 + \beta}{\beta}\right] \tag{84}$$

Therefore, by Lemma 5, S is β-averaged for some $\beta \in (0,1)$ iff Q is non-expansive. But Q is non-expansive for some $\beta \in (0,1)$ iff

$$-1 \le \frac{\sigma_0 - 1 + \beta}{\beta} \le \frac{\sigma_1 - 1 + \beta}{\beta} \le 1,$$

$$\text{for some } \beta \in (0,1). \tag{85}$$

Note that since β is restricted to be positive, (85) can be written equivalently as,

$$1 - 2\beta \le \sigma_0 \le \sigma_1 \le 1, \text{ for some } \beta \in (0,1). \tag{86}$$

But (86) holds iff $-1 < \sigma_0 \le \sigma_1 \le 1$. $\qquad\square$

In the convergence proof, to counter the scale factor that appears in (80), we use the following result.

Proposition 17. For $\alpha\,\rho < 1$, let V_α be defined as

$$V_\alpha = \frac{1}{1 - \alpha\,\rho} U_\alpha. \tag{87}$$

Then, V_α is averaged if

$$\rho \le \sigma_m, \tag{88a}$$

$$\alpha < \frac{2}{\sigma_M + \rho}. \tag{88b}$$

Proof. Observe that V_α is of the form

$$V_\alpha(x) = \underbrace{\frac{1}{1 - \alpha\,\rho}\big(I - \alpha\,H^T H\big)}_{M} x + u, \tag{89}$$

for a constant vector u. The eigenvalues of M are contained in the interval

$$\left[\frac{1 - \alpha\,\sigma_M}{1 - \alpha\,\rho}, \frac{1 - \alpha\,\sigma_m}{1 - \alpha\,\rho}\right]. \tag{90}$$

By (88a), we have

$$\frac{1 - \alpha\,\sigma_m}{1 - \alpha\,\rho} \le 1. \tag{91}$$

By (88b), we have

$$\frac{1 - \alpha \sigma_M}{1 - \alpha \rho} > -1. \tag{92}$$

Thus, it follows by Prop. 16 that V_α is averaged. ☐

We are now ready for the proof of Prop. 4.

Proof of Prop. 4. Note that $T_\alpha U_\alpha = S_\alpha V_\alpha$, where S_α and V_α are as defined in Prop. 15 and Prop. 17. By Prop. 15 and Prop. 17, S_α and V_α are averaged. Then, by Prop. 14 we conclude that $T_\alpha U_\alpha$ is also averaged. We also have by Prop. 9 that the fixed points of $T_\alpha U_\alpha$ coincide with the set of minimizers of $C(x)$, which is non-empty since $C(x)$ is convex. The claim now follows by Prop. 13. ☐

V. GENERAL DATA FIDELITY TERM

In this section, we provide proofs of Prop. 6 and Prop. 7. Let us recall the setup. We consider a cost function of the form

$$D(x) = f(x) + P(x), \tag{93}$$

where P is ρ-weakly convex and f is ρ-strongly convex which is differentiable and

$$\|\nabla f(x) - \nabla f(y)\| \leq \tau \|x - y\|, \text{ for all } x, y. \tag{94}$$

Definition 6. f is said to have a τ-Lipschitz continuous gradient, if (94) holds. ☐

Let us now recall the Baillon-Haddad theorem [1, 2, 12].

Lemma 7. Suppose f is convex, differentiable and it has a τ-Lipschitz continuous gradient. Then,

$$\langle \nabla f(x) - \nabla f(z), x - z \rangle \geq \frac{1}{\tau} \|\nabla f(x) - \nabla f(z)\|_2^2. \tag{95}$$

In the setting above, we have,

Lemma 8. Suppose f is ρ-strongly convex, differentiable and its gradient is σ-Lipschitz continuous with $\sigma > \rho$. Also, let $g(x) = f(x) - \rho \|x\|_2^2/2$. Then, ∇g is $(\sigma - \rho)$-Lipschitz continuous.

Proof. Let $F = \nabla f$ and $G = \nabla g$. Note that $G = F - \rho I$. First observe that, by the Baillon-Haddad theorem applied to F, we have,

$$\langle F(x) - F(y), x - y \rangle \geq \frac{1}{\sigma} \|F(x) - F(y)\|_2^2. \tag{96}$$

Now,

$$\|G(x) - G(y)\|_2^2 = \|F(x) - F(y)\|_2^2 + \rho^2 \|x - y\|_2^2 - 2\rho \langle F(x) - F(y), x - y \rangle \tag{97}$$

$$\leq \left(1 - 2\frac{\rho}{\sigma}\right) \|F(x) - F(y)\|_2^2 + \rho^2 \|x - y\|_2^2 \tag{98}$$

$$\leq \left(1 - 2\frac{\rho}{\sigma}\right) \sigma^2 \|x - y\|_2^2 + \rho^2 \|x - y\|_2^2 \tag{99}$$

$$= (\sigma - \rho)^2 \|x - y\|_2^2. \tag{100}$$

Taking square roots, we obtain

$$\|G(x) - G(y)\|_2 \leq (\sigma - \rho)\|x - y\|_2, \tag{101}$$

which is the claim. ☐

Proposition 18. Suppose f is ρ-strongly convex, differentiable and its gradient is σ-Lipschitz continuous with $\sigma > \rho$. Also, for $\alpha \rho < 1$, let V_α be defined as

$$V_\alpha = \frac{1}{1 - \alpha \rho} \left(I - \alpha \nabla f\right). \tag{102}$$

Then, V_α is averaged if

$$\alpha < \frac{2}{\sigma + \rho}. \tag{103}$$

Proof. Let $g(x) = f(x) - \rho \|x\|_2^2/2$. Then, g is convex, $\nabla g = \nabla f - \rho I$ and ∇g is $(\sigma - \rho)$-Lipschitz. By the Baillon-Haddad theorem, we have

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq \frac{1}{\sigma - \rho} \|\nabla g(x) - \nabla g(y)\|_2^2. \tag{104}$$

Also,

$$V_\alpha = \frac{1}{1 - \alpha \rho} \left(I - \alpha(\rho I + \nabla g) \right) \tag{105a}$$

$$= I - \frac{\alpha}{1 - \alpha \rho} \nabla g. \tag{105b}$$

Assuming (103) holds, set $\beta = \alpha(\sigma + \rho)/2$ and observe that $0 < \beta < 1$. Observe that

$$Q = \frac{1}{\beta} \left(V_\alpha - (1 - \beta)I \right) = I - c \nabla g, \tag{106}$$

where $c = \dfrac{2}{(1 - \alpha \rho)(\sigma + \rho)}$. We now have, by (104),

$$\|Q(x) - Q(y)\|_2^2 \leq \|x - y\|_2^2 + \left(c^2 - \frac{2c}{\sigma - \rho} \right) \|\nabla g(x) - \nabla g(y)\|_2^2. \tag{107}$$

It can be checked that (104) implies

$$\left(c^2 - \frac{2c}{\sigma - \rho} \right) < 0. \tag{108}$$

Therefore Q is non-expansive and by Lemma 5, V_α is β-averaged. $\qquad\square$

We now present the proof of Prop. 6. The argument is similar to that in the proof of Prop. 4.

Proof of Prop. 6. Let us define the operator A through

$$A x = T_\alpha \left(x - \alpha \nabla f(x) \right), \tag{109}$$

Then, for $\alpha \rho < 1$, A can also be written as $A = S_\alpha V_\alpha$ where S_α and V_α are as defined in Prop. 15 and Prop. 18. But if $\alpha < 2/(\sigma + \rho)$, S_α and V_α are averaged by Prop. 15 and Prop. 18. Then, by Prop. 14, we conclude that A is also averaged. We also have by Prop. 10 that the fixed points of A comprise the set of minimizers of the convex function $D(x)$ defined in (5). The claim now follows by Prop. 13. $\qquad\square$

We remark that the convergence proof for ISTA did not use and does not directly imply that the cost decreases monotonically with each iteration (as stated in Prop. 7). However, such a descent property has been shown to hold in [25, 35] previously. We include a proof of this descent property (i.e., Prop. 7) for the sake of completeness. The proof depends on two lemmas.

Lemma 9. The sequence constructed in (6) satisfies,

$$P(x^{k+1}) - P(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle \leq \left(\frac{\rho}{2} - \frac{1}{\alpha} \right) \|x^k - x^{k+1}\|_2^2. \tag{110}$$

Proof. This claim follows Lemma 3 by plugging x^k for x, x^{k+1} for $\hat{x}$ and $x^k - \alpha \nabla f(x^k)$ for z (since $x^{k+1} = T_\alpha(x^k - \alpha \nabla f(x^k))$). $\qquad\square$

Lemma 10. Suppose $h : \mathbb{R}^n \to \mathbb{R}$ is a continuous differentiable function whose gradient satisfies a Lipschitz condition as,

$$\|\nabla h(x) - \nabla h(y)\|_2 \leq L \|x - y\|_2. \tag{111}$$

Then,

$$h(x) \le h(y) + \langle \nabla h(y), (x-y) \rangle + \frac{L}{2} \|x - y\|_2^2. \tag{112}$$

Proof. Let $s(t) = h\big(y + t\,(x-y)\big)$, where $t \in \mathbb{R}$. Note that s is a smooth function with $s(0) = h(y)$, $s(1) = h(x)$. Observe also that

$$s'(t) = \Big\langle \nabla h\big(y + t\,(x-y)\big), (x-y) \Big\rangle. \tag{113}$$

Using (111), Cauchy-Schwarz inequality and an expansion of s, we obtain the following chain.

$$h(x) = s(1) \tag{114}$$

$$= s(0) + \int_0^1 s'(u)\,du \tag{115}$$

$$= h(y) + \int_0^1 \Big\langle \nabla h\big(y + u\,(x-y)\big), (x-y) \Big\rangle\,du \tag{116}$$

$$= h(y) + \int_0^1 \langle \nabla h(y), (x-y) \rangle\,du + \int_0^1 \Big\langle \nabla h\big(y + u\,(x-y)\big) - \nabla h(y), (x-y) \Big\rangle\,du \tag{117}$$

$$\le h(y) + \langle \nabla h(y), (x-y) \rangle + \int_0^1 \Big| \Big\langle \nabla h\big(y + u\,(x-y)\big) - \nabla h(y), (x-y) \Big\rangle \Big|\,du \tag{118}$$

$$\le h(y) + \langle \nabla h(y), (x-y) \rangle + \int_0^1 \Big\| \nabla h\big(y + u\,(x-y)\big) - \nabla h(y) \Big\|_2 \|x - y\|_2\,du \tag{119}$$

$$\le h(y) + \langle \nabla h(y), (x-y) \rangle + \int_0^1 L \|u(x-y)\|_2 \|x - y\|_2\,du \tag{120}$$

$$\le h(y) + \langle \nabla h(y), (x-y) \rangle + \frac{L}{2} \|x - y\|_2^2. \tag{121}$$

$\square$

We remark that for the data fidelity term f, the Lipschitz constant is σ.

Proof of Prop. 7. By Lemma 10, we have,

$$f(x^{k+1}) \le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{\sigma}{2} \|x^{k+1} - x^k\|_2^2.$$

Adding $p(x^{k+1})$ to both sides, we can rewrite this as,

$$D(x^{k+1}) \le D(x^k) + \left\{ p(x^{k+1}) - p(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{\sigma}{2} \|x^{k+1} - x^k\|_2^2 \right\} \tag{122}$$

Let us denote the term in curly brackets with c. If we can show that c is non-positive, we are done. By Lemma 9, we have,

$$c \le \left(\frac{\rho}{2} - \frac{1}{\alpha} + \frac{\sigma}{2} \right) \|x^{k+1} - x^k\|_2^2. \tag{123}$$

But since $\alpha < 2/(\sigma + \rho)$, the term in the parentheses is negative and thus $c < 0$. $\square$

VI. Experiments

In this section, we present two different experiments to evaluate the acceleration achieved by increasing the step size. We also compare the speed of convergence with those achieved by acceleration methods like TwIST [7] and FISTA [6]. We note however, that TwIST and FISTA are proposed for when convex penalties are used in the cost function. Therefore, their convergence analyses are not valid in the current setup. Nevertheless, we have observed that they lead to considerable acceleration and included them to give an idea about what can be gained by increasing the step size. We also note that our purpose in these experiments is not to promote weakly convex penalties but to demonstrate the effects of increasing the step size in ISTA. Matlab code for the experiments is available at "http://web.itu.edu.tr/ibayram/NCISTA/".
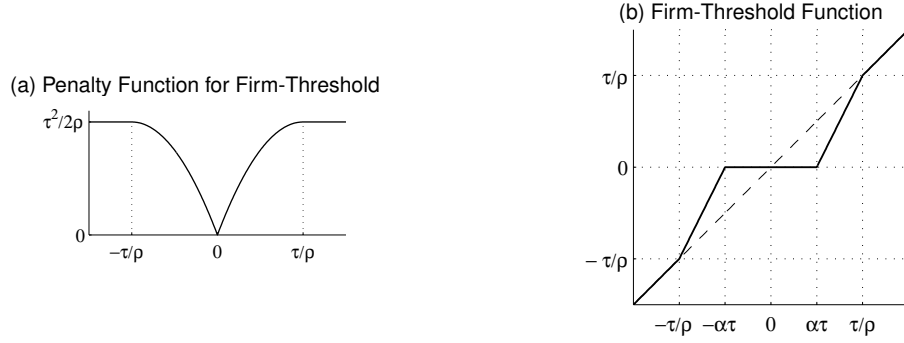
Fig. 2. The penalty and threshold function used in Exp. 1. Observe that this threshold function is not non-expansive. Specifically, the derivative exceeds unity on the intervals $(-\tau/\rho, -\alpha\rho)$ and $(\alpha\rho, \tau/\rho)$
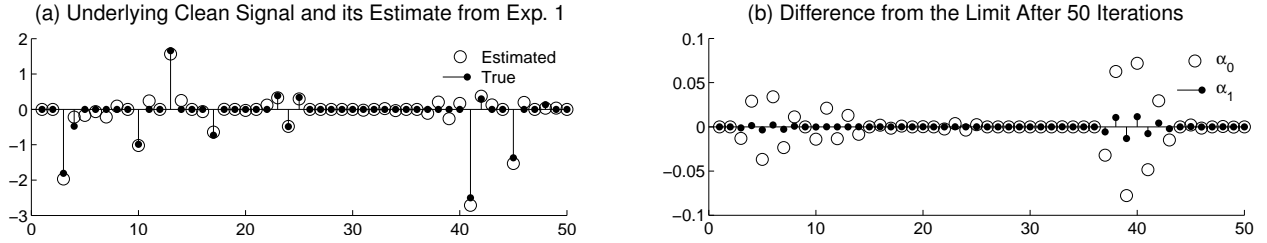

Fig. 3. (a) The true signal and its estimate obtained by solving (127). (b) The difference from the minimizer after 50 iterations of ISTA with stepsizes $\alpha_0 < \alpha_1$.

Experiment 1. Our first experiment involves a sparse signal recovery problem. We use a 60×50 convolution matrix H associated with an invertible filter (so that $H^T H$ is invertible) to construct the observed signal. Using a sparse x as shown in Fig.3a, we produce the observed signal as

$$y = H\,x + u, \tag{124}$$

where u denotes white Gaussian noise. We use a penalty function $P : \mathbb{R} \to \mathbb{R}$ defined as

$$P_{\tau,\rho}(s) = \begin{cases} \tau\,|s| - s^2/(2\rho), & \text{if } |s| < \tau/\rho, \\ \tau^2/(2\rho), & \text{if } |s| \geq \tau/\rho, \end{cases} \tag{125}$$

This function is shown in Fig. 2a. This penalty function is ρ-weakly convex. The threshold function associated with P (see (3)) is known as a firm-threshold [21] and is given by (provided $\alpha\,\rho < 1$),

$$T_\alpha(s) = \begin{cases} 0, & \text{if } |s| < \alpha\,\tau, \\ (1 - \alpha\,\rho)^{-1}\,(s - \alpha\,\tau), & \text{if } \alpha\,\tau \leq |s| < \tau/\rho, \\ s, & \text{if } \tau/\rho \leq |s|. \end{cases} \tag{126}$$

This threshold function is depicted in Fig. 2b. Observe that for $\alpha\,\tau < |s| < \tau/\rho$, the derivative of the firm-threshold function exceeds unity. Therefore the firm-threshold function is not non-expansive.

In the setup described above, we set ρ as the least eigenvalue of $H^T H$, which is the maximum value allowed if a convex cost is desired. We also set $\tau = 3\,\rho\,\mathrm{std}(u)$, where $\mathrm{std}(u)$ denotes the standard deviation of the noise. We obtain the estimate of the sparse signal as,

$$x^* = \arg\min_t \frac{1}{2}\|y - H\,t\|_2^2 + \sum_i P_{\tau,\rho}(t_i). \tag{127}$$

Note that P applies componentwise to an input vector t. The threshold function T_α is also applied componentwise in the realization of ISTA.

Denoting the greatest eigenvalue of $H^T H$ by σ, we set $\alpha_0 = 1/\sigma$ and $\alpha_1 = 2/(\sigma + \rho)$. Note that α_0 is the greatest value of the stepsize allowed by MM (see Prop. 3), whereas for α_1, convergence is guaranteed by Prop. 4. For this specific problem, the ratio α_1/α_0 was found to be 1.88.
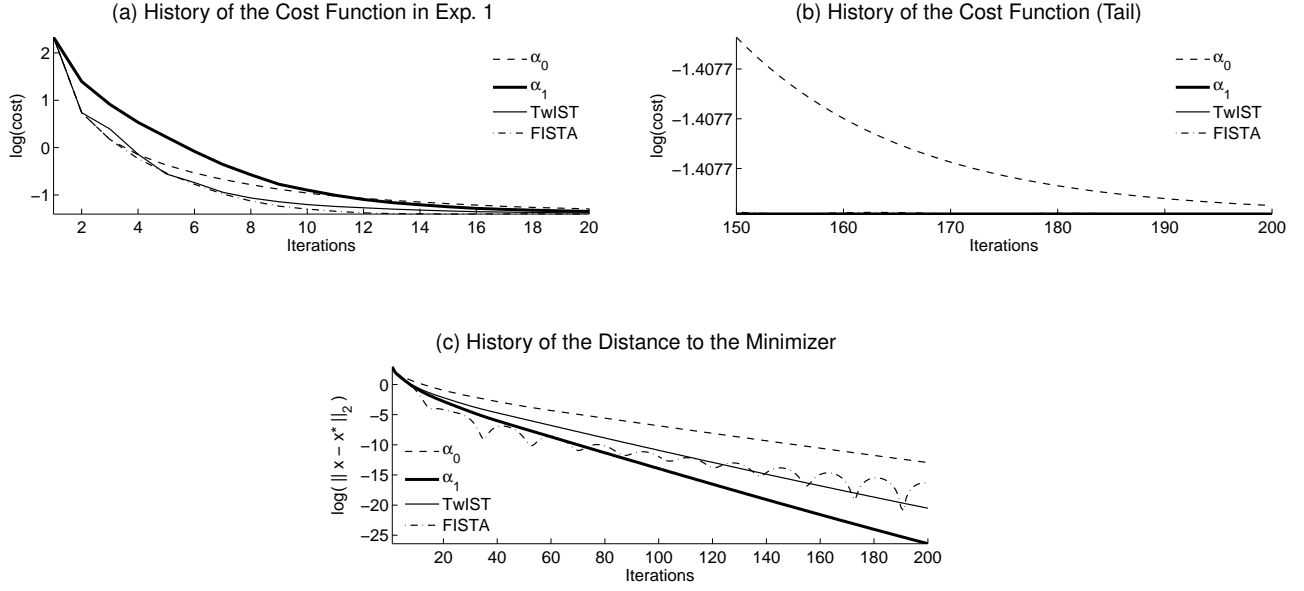
Fig. 4. Effects of changing the step size α in ISTA for Exp. 1. (a,b) show how the cost function evolves with iterations using the greatest possible α_0 that guarantees strict descent and the greatest possible α_1 that merely guarantees convergence. The cost with TwIST and FISTA are also included. (c) shows the distance to the unique minimizer with respect to iterations for the same four choices.

We ran ISTA with α_0 and α_1, in both cases, starting from zero. In order to better evaluate the convergence speed, we also tried FISTA [6] and TwIST [7]. We note that both methods require that the components of the cost function be convex and thus they were not originally proposed for the current setup. Nevertheless, both methods are formally applicable. We note that at least in the convex case, FISTA requires the step-size to be less than or equal to $1/\sigma$. Therefore for FISTA, we employed the stepsize $\alpha_0 = 1/\sigma$ (we also tried it for α_1, but the sequence diverged). Given the step-size, we applied the algorithm referred to as 'FISTA with constant stepsize' in Sec.4 of [6]. For TwIST, there are two parameters to choose, namely 'α' and 'β' – for these, we used the suggestions in equations (26) and (27) of [7].

The history of the cost function with iterations is shown in Fig. 4 a,b. With α_0, initial descent is greater compared to that achieved by α_1 but eventually the cost achieved by α_1 drops below that achieved by α_0. FISTA and TwIST perform quite well in the beginning, dropping the cost faster than ISTA with α_0 or α_1. However, after about 30 iterations, FISTA, TwIST and ISTA (with α_1) achieve almost the same cost values. We found these observations to be quite stable with respect to different noise realizations.

In order to assess the convergence speed, using α_0 we ran ISTA for 10K iterations to obtain an estimate of the minimizer x^*. Then, we reran the algorithms mentioned above. The logarithm of the Euclidean distance to x^* with respect to iterations is shown in Fig. 4c. We see that the the distance to the minimizer decreases faster when larger steps are used in ISTA. TwIST converges with a rate that lies in between those of ISTA with stepsizes α_0 and α_1. The behavior of FISTA is less stable. We see bursts that take the iterate close to the limit followed by slight departures. Especially in the first few iterations, the distance to the minimizer is greatly reduced by FISTA but in the long run, convergence rate is slightly worse than that of TwIST, on average. To conclude, although TwIST and FISTA are very successful in reducing the cost rapidly, this does not necessarily lead to the fastest rate of convergence when we monitor the distance to the minimizer.

Experiment 2. In this experiment, we consider a clean signal $x(n)$ composed of piecewise constant blocks of three samples. That is, the signal satisfies

$$x(3n) = x(3n+1) = x(3n+2), \tag{128}$$

for every integer n. We further know that $x(n)$ takes integer values in the range $[0, 4]$. An example is shown in Fig. 5a (stem plot). We observe a blurred and noisy version of this signal, namely y. The least squares reconstruction of x given the observed y is also shown in Fig. 5a (circles). In order to take (128) into account, we propose to synthesize x as a weighted linear combination of atoms $\{g(n-3\,k)\}_{k \in \mathbb{Z}}$, where $g(n) = \delta(n) + \delta(n-1) + \delta(n-2)$.
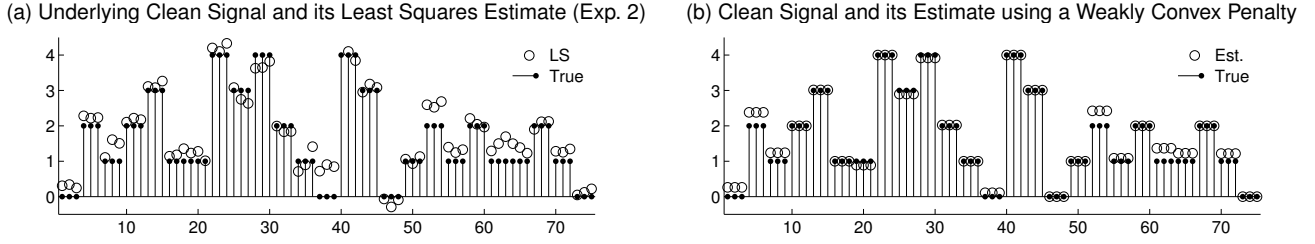
Fig. 5. (a) The true signal and its least squares estimate for Exp. (2). (b) The clean signal and its estimate obtained by solving (130).

That is, we let our estimate $\hat{x}$ be given as,

$$\hat{x} = \sum_k \hat{c}(k)\, g(n - 3k), \tag{129}$$

where $\hat{c}$ denotes the coefficients to be estimated. Let us denote the operator that takes c to $\hat{x}$ as G. Also, let F denote the blurring operator. In this setting, we wish to estimate $\hat{c}$ as

$$\hat{c} = \arg\min_c \frac{1}{2}\|y - H\,c\|_2^2 + \tau \sum_k P\big(c(k)\big), \tag{130}$$

where $H = F\,G$ and $P : \mathbb{R} \to \mathbb{R}$ is a penalty function that applies to the components of c. We expect P to penalize deviations from the integers in the range $[0, 4]$. Notice that this is inherently a non-convex constraint. No useful convex penalty function exists for enforcing such a constraint. However, viable weakly-convex penalties can be found for this task (see also [29] in this context). We propose to use a penalty function given as

$$P(s) = \begin{cases} \infty, & \text{if } s < 0, \\ (s - \lfloor s \rfloor)\,(\lceil s \rceil - s), & \text{if } 0 \le s \le 4, \\ \infty, & \text{if } 4 < s, \end{cases} \tag{131}$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote the floor and ceiling functions respectively. This penalty function is 2-weakly convex and is shown in Fig. 6a. The associated threshold function $T_\alpha(s)$ for $\alpha < 1/2$ is given as

$$\begin{cases} 0, & \text{if } s < 0, \\ \lfloor s \rfloor, & \text{if } 0 \le s \le \lfloor s \rfloor + \alpha \le 4 + \alpha, \\ \lfloor s \rfloor + \frac{s - \lfloor s \rfloor - \alpha}{1 - 2\alpha}, & \text{if } 0 \le \lfloor s \rfloor + \alpha \le s \le \lceil s \rceil - \alpha \le 4, \\ \lceil s \rceil & \text{if } 0 \le \lceil s \rceil - \alpha \le s \le 4, \\ 4, & \text{if } 4 < s. \end{cases}$$

The threshold function is shown in Fig. 6b. Observe that the threshold function allows non-integer values but favors integers (it has integer valued deadzones). Note also that this threshold function applies componentwise.

In this setup, we ran ISTA with the four choices described in Experiment 1, namely ISTA with α_0 (largest step-size allowed by the MM framework), ISTA with α_1 (largest step-size for which convergence is ensured), TwIST and FISTA. We chose the parameters for TwIST and FISTA similarly as in Exp. 1. The history of the cost function and the distance to the limit is shown in Fig. 7a,b respectively. The figures are in agreement with those in Exp. 1. Increased step sizes lead to acceleration that is about the same order (in fact, better in this example) as that achieved by TwIST and FISTA. We again note, however that we do not have a theoretical explanation for this behavior (whereas FISTA is theoretically optimal in terms of reducing the cost for convex penalties).

VII. CONCLUSION

In this paper, we generalized the ISTA convergence condition for the stepsize to the case where the penalty is weakly-convex. We also demonstrated that the larger stepsizes lead to faster convergence, although we do not have a precise theoretical justification at the moment. The generalization in this paper relies on a study of the proximity operator for a weakly convex function. Specifically, we have seen that the proximity operator for a weakly convex function is no longer (firmly) non-expansive and/or averaged. However, under proper scaling, the
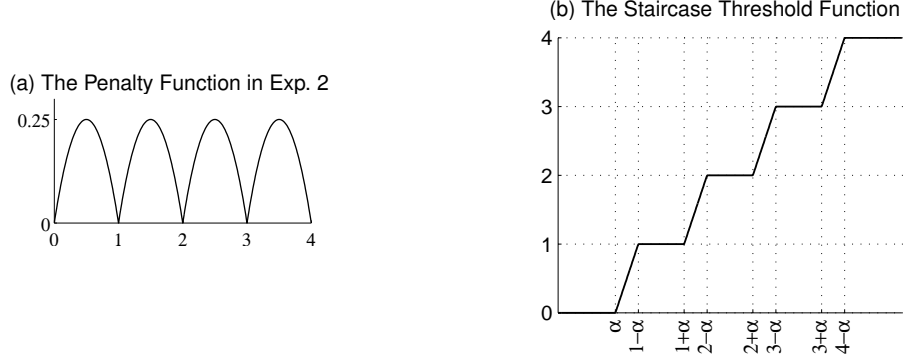
(a) The Penalty Function in Exp. 2

(b) The Staircase Threshold Function

Fig. 6. (a) The penalty function used in Exp. 2 to promote integer values in the range $[0, 4]$. (b) The threshold function T_α per (3) associated with the function in (a).

(a) History of the Cost Function in Exp. 2
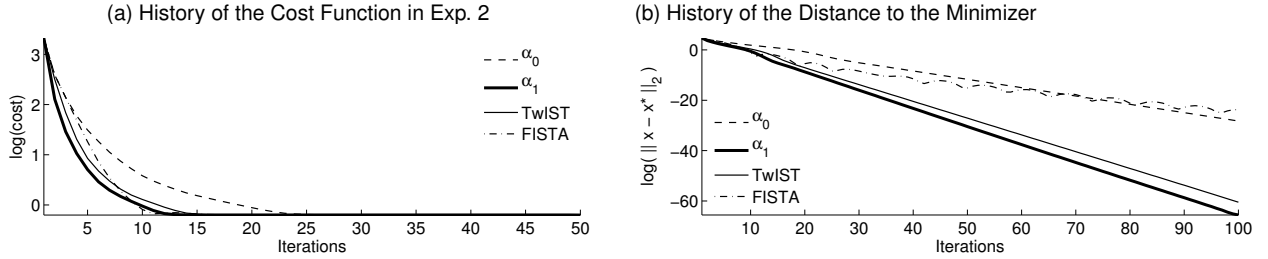
(b) History of the Distance to the Minimizer

Fig. 7. (a) shows the history of the cost function for Exp. 2. The same choices as in Exp. 1 are used for comparison. (b) shows the history of the distance to the unique minimizer.

proximity operators regain these properties. Given this observation, it is also natural to consider extension of this work to other algorithms that employ proximity operators, specifically the Douglas-Rachford algorithm [18], which can in turn be used to derive other algorithms such as the alternative direction method of multipliers (ADMM) [9, 18], or the parallel proximal algorithm (PPXA) [30]. We hope to investigate these extensions in future work.

APPENDIX (PROOF OF PROP. 13)

In this appendix, we provide a proof of Prop. 13. An alternative discussion can be found in [2].

Definition 7. A point x is said to be fixed point of T if $x = Tx$. The set of fixed points of T is denoted as $\operatorname{Fix} T$. □

Definition 8. Suppose T is an operator such that $\operatorname{Fix} T \neq \emptyset$. T is called strictly quasi non-expansive if for any $x \notin \operatorname{Fix} T$ and $y \in \operatorname{Fix} T$, we have $\|Tx - y\| < \|x - y\|$. □

Non-expansiveness and strict quasi non-expansiveness properties do not imply each other. However, it turns out that an averaged operator with a non-empty fixed point set is both non-expansive and strictly quasi non-expansive.

Proposition 19. Suppose S is β-averaged and $\operatorname{Fix} S \neq \emptyset$. Then,

(i) S is non-expansive.
(ii) S is strictly quasi non-expansive.

Proof. Since S is β-averaged, it can be decomposed as $S = (1 - \beta)I + \beta U$ for a non-expansive U. Also note that $\operatorname{Fix} S = \operatorname{Fix} U$.

To see (i), observe that

$$\|Sx - Sy\| = \|(1 - \beta)(x - y) + \beta(Ux - Uy)\| \tag{132a}$$

$$\leq (1 - \beta)\|x - y\| + \beta\|Ux - Uy\| \tag{132b}$$

$$\leq \|x - y\|, \tag{132c}$$

where we have used the nonexpansivity of U in the last line.

For (ii), suppose $x \in \operatorname{Fix} S = \operatorname{Fix} U$ and $y \notin \operatorname{Fix} S = \operatorname{Fix} U$. First, note that,

$$\|Sx - Sy\|^2 = (1 - \beta)^2 \|x - y\|^2 + \beta^2 \|Ux - Uy\|^2 + 2\beta(1 - \beta)\langle x - y, Ux - Uy \rangle. \tag{133}$$

Consider now the last term. By the Cauchy-Schwarz inequality,

$$\langle x - y, Ux - Uy \rangle \leq \|x - y\| \, \|Ux - Uy\|. \tag{134}$$

Equality holds in (134) only if $Ux - Uy = \beta (x - y)$ with $\beta > 0$. Non-expansiveness of U implies that β cannot be greater than or equal to unity. Therefore even if equality holds in (134), we will have that

$$\langle x - y, Ux - Uy \rangle < \|x - y\|^2. \tag{135}$$

Combining this observation with the non-expansiveness of U (i.e., $\|Ux - Uy\|^2 \leq \|x - y\|^2$), we obtain from (133)

$$\|Sx - Sy\|^2 < \|x - y\|^2. \tag{136}$$

Thus S is strictly quasi non-expansive. $\qquad\square$

Fixed point iterations of a non-expansive or a strictly quasi non-expansive operator may not be convergent. However, when an operator possesses both properties (e.g., an β-averaged operator), then we have the following convergence result.

Proposition 20. Suppose S is non-expansive and also strictly quasi non-expansive (implying $\operatorname{Fix} S \neq \emptyset$). Given x^0, let $x^{n+1} = Sx^n$. Then the sequence $\{x^n\}_{n \in \mathbb{Z}}$ converges and the limit is a fixed point of S.

Proof. Since S is non-expansive, we have, for $x^* \in \operatorname{Fix} S$,

$$\|x^n - x^*\| \leq \|x^0 - x^*\|. \tag{137}$$

Thus, the sequence generated by S is bounded. By the Bolzano-Weierstrass theorem, we can therefore extract a convergent subsequence $x^{n_k} \to z$. We will first show that $z \in \operatorname{Fix} S$. Then we will deduce that the whole sequence actually converges to this z.

Suppose now that $z \notin \operatorname{Fix} S$. Since S is non-expansive, it is also continuous. Therefore, $Sx^{n_k} \to Sz$. Now let z^* be an arbitrary point of $\operatorname{Fix} S$. By the quasi strict nonexpansiveness of S, we have,

$$\epsilon = \|z - z^*\| - \|Sz - z^*\| > 0. \tag{138}$$

For K sufficiently large, we will have that if $k \geq K$,

$$\|x^{n_k} - z\| \leq \epsilon/4, \tag{139a}$$
$$\|Sx^{n_k} - Sz\| \leq \epsilon/4. \tag{139b}$$

Now if $k \geq K$, then by nonexpansivity and (139b)

$$\|z^* - x^{n_{k+1}}\| \leq \|z^* - Sx^{n_k}\| \tag{140a}$$
$$\leq \|z^* - Sz\| + \epsilon/4. \tag{140b}$$

But by (139a),

$$\|z^* - x^{n_{k+1}}\| \geq \|z^* - z\| - \epsilon/4 \tag{141}$$

Combining (140) and (142), we obtain a contradiction as

$$\|z^* - z\| \leq \|z^* - Sz\| + \epsilon/2. \tag{142}$$

Therefore, $z = Sz$. But since z is a fixed point, nonexpansivity of S implies that the whole sequence of x^n's must converge to z. $\qquad\square$

Prop. 13 now follows as a corollary of Prop. 19 and Prop. 20.

REFERENCES

[1] H. Bauschke and P. Combettes. The Baillon-Haddad theorem revisited. *Journal of Convex Analysis*, 17:781–787, 2010.

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

[3] İ. Bayram. Penalty functions derived from monotone mappings. *IEEE Signal Processing Letters*, 22(3):265–269, March 2015.

[4] İ. Bayram, P.-Y. Chen, and I. W. Selesnick. Fused lasso with a non-convex sparsity inducing penalty. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2014.

[5] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty. *Nonlinear Programming: Theory and Algorithms*. Wiley, 2nd edition, 1993.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 1:183–202, 2009.

[7] J. M. Bioucas-Dias and M. A. T. Figueiredo. A new TwIST: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Trans. Image Proc.*, 16(12):2992–3003, December 2007.

[8] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, 1987.

[9] 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.

[10] C. L. Byrne. An elementary proof of convergence for the forward-backward splitting algorithm. *Journal of Nonlinear and Convex Analysis*, 15(4):681–691, 2014.

[11] C. L. Byrne. *Iterative Optimization in Inverse Problems*. Taylor and Francis, 2014.

[12] C. L. Byrne. On a generalized Baillon-Haddad theorem for convex functions on Hilbert space, 2015. To appear in Journal of Convex Analysis.

[13] R. Chartrand. Shrinkage mappings and their induced penalty functions. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2014.

[14] F. H. Clarke. *Optimization and Nonsmooth Analysis*. SIAM, 1990.

[15] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, and H. Wolkowicz, editors, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, New York, 2011.

[16] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM J. Multiscale Model. Simul.*, 4(4):1168–1200, November 2005.

[17] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. Pure Appl. Math.*, 57(11):1413–1457, August 2004.

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

[19] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak. Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Trans. Image Proc.*, 16(12):2980–2991, December 2007.

[20] M. A. T. Figueiredo and R. D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, August 2003.

[21] H.-Y. Gao and A. G. Bruce. Waveshrink with firm shrinkage. *Statistica Sinica*, 7(4):855–874, 1997.

[22] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of Convex Analysis*. Springer, 2004.

[23] D. R. Hunter and K. Lange. A tutorial on MM algorithms. *Amer. Statist.*, 58(1):30–37, February 2004.

[24] M. W. Jacobson and J. A. Fessler. An expanded theoretical treatment of iteration-dependent majorize-minimize algorithms. *IEEE Trans. Image Proc.*, 16(10):2411–2422, October 2007.

[25] M. Kowalski. Thresholding rules and iterative shrinkage/thresholding algorithm: A convergence study. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2014.

[26] K. Lange. *Optimization*. Springer, 2004.

[27] R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematicalprogramming algorithms. *J. of Computer and System Sci.*, 12(1):108–121, February 1976.

[28] C. De Mol and M. Defrise. A note on wavelet-based inversion methods. In M. Z. Nashed and O. Scherzer,

editors, *Inverse Problems, Image Analysis and Medical Imaging*. American Mathematical Society, 2002.

[29] M. Nikolova. Estimation of binary images using convex criteria. In *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, 1998.

[30] N. Pustelnik, C. Chaux, and J.-C. Pesquet. Parallel proximal algorithm for image restoration using hybrid regularization. *IEEE Trans. Image Proc.*, 20(9):2450–2462, September 2011.

[31] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1996.

[32] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 1998.

[33] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3rd edition, 1976.

[34] I. W. Selesnick and İ. Bayram. Sparse signal estimation by maximally sparse convex optimization. *IEEE Trans. Signal Processing*, 62(5):1078–1092, March 2014.

[35] Y. She. Thresholding-based iterative selection procedures for model selection and shrinkage. *Electronic Journal of Statistics*, 3:384–415, 2009.

[36] J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 8:231–259, May 1983.

[37] S. Voronin and R. Chartrand. A new generalized thresholding algorithm for inverse problems with sparsity constraints. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP)*, 2013.