Iterative Shrinkage/Thresholding Algorithm with a Weakly Convex Penalty

İlker Bayram
ibayram@itu.edu.tr

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. Specifically, we show that convergence is possible without any guarantee of strict descent. We also demonstrate experimentally that for the current setup, using large stepsizes in ISTA can accelerate the convergence more than existing schemes proposed for the convex case, like TwIST or FISTA.

1 Introduction

The iterative shrinkage/thresholding algorithm (ISTA) is a popular 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 imaging operator, and $P(x)$ is a penalty term reflecting our prior knowledge about the object to be recovered [21, 16, 13]. For minimizing $C(x)$, ISTA employs iterations of the form

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

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

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

When $P(x)$ is convex, T is also referred to as the proximity operator of P [11]. In the following, we say that T_α is well-defined if the minimization problem in (3) is strictly convex (hence has a unique solution). We also let σ_m and σ_M denote the least and greatest eigenvalue of $H^T\,H$ respectively.

For convex P, ISTA can be derived via different approaches. The majorization-minimization (MM) scheme [15, 18, 20], enforces the constraint $\alpha < 1/\sigma_M$, and implies that the iterates achieve monotone descent in the cost.

Proposition 1. Suppose $P(x)$ is convex. If $0 < \alpha \leq 1/\sigma_M$, then the iterates x^k in (2) monotonely decrease the cost (i.e., $C(x^{k+1}) \leq C(x^k)$) and converge to a global minimizer of $C(x)$. □

The forward-backward splitting algorithm [12] gives exactly the same iterations, but this interpretation allows to double the stepsize α, which in practice accelerates the convergence to a minimizer. However, when $\alpha > 1/\sigma_M$, monotone decrease of the cost is no longer guaranteed. Instead, the distance to the set of minimizers shows a monotone behavior.

Proposition 2. [12, 2] Suppose $P(x)$ is convex. If $0 < \alpha < 2/\sigma_M$, then the iterates x^k in (2) converge to a global minimizer of $C(x)$. $\square$

When $P(x)$ is non-convex, ISTA is still applicable but convergence to a global minimizer is not guaranteed in general. Further, the forward-backward interpretation is not valid. In this paper, instead of an arbitrary non-convex penalty, we consider weakly-convex penalties [27] [1].

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

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

is convex if and only if $s \geq \rho \geq 0$. $\square$

If $P(x)$ is $(-\rho)$-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 [26] for an iterative scheme where each iteration requires solving such a problem, [22] 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 [26, 6, 9, 3].

Through the MM scheme, we obtain the following in this paper (see Sec. 2).

Proposition 3. Suppose P is $(-\rho)$-weakly convex with $0 \leq \rho \leq \sigma_m$. If $\alpha < 1/\sigma_M$, then T_α is well-defined, and the sequence x^k in (2) monotonely decreases the cost $C(x)$, converging to a global minimizer of $C(x)$. $\square$

The step-size obtained by MM not only guarantees convergence but also ensures monotone decrease in the cost. Although this is not an undesirable feature, one is tempted to ask if larger stepsizes could be used to accelerate the algorithm, possibly sacrificing the strict descent property. 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 Fig. 2 in this paper or Fig. 2 in [26]). 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 $(-\rho)$-weakly convex with $\rho \leq \sigma_m$. If $0 < \alpha < 2/(\sigma_M + \rho)$, then the iterates x^k in (2) converge to a global minimizer of $C(x)$. $\square$

As in the convex case, it is also true that the distance to the set of minimizers monotonely decreases with each iteration. 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.

[1] See specifically Defn.1 and Prop. 4.3 in [27]. We note that the definition used here is slightly different than that in [27]. We made a constant factor modification of 1/2 so as to avoid some 1/2 factors that would appear otherwise.

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

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

In this setup, we have the following proposition for when f and P are convex [12, 2].

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

We note that this result 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 $\|\nabla f(x) - \nabla f(y)\| \leq \sigma_M \|x - y\|_2$.

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

Proposition 6. Suppose $P(x)$ is $(-\rho)$-weakly convex, $D(x) = f(x) + P(x)$ 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 $\alpha < 2/(\sigma + \rho)$, then the iterates x^k in (6) converge to a global minimizer of $D(x)$. □

Although Prop. 4 is a corollary of Prop. 6, we start the discussion with Prop. 4 because 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 [25, 10]. 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.

Contribution

The paper generalizes convergence results of ISTA to cases where the penalty is weakly-convex. In that regard, we show that ISTA preserves monotonic descent provided that the stepsize is small enough (Prop. 3). Then we show that the stepsize can be further increased while still ensuring convergence. Although the monotonic descent property can be derived straightforwardly from the MM scheme, the convergence result without monotonic descent in cost (i.e., Props. 4 and 6) requires a study of the mapping that ISTA employs. For weakly convex penalties, such a study has not been performed as far as we are aware. Throughout the paper, we make use of well-known results from convex analysis, but provide the proofs (along with references to sources where alternatives can be found) whenever possible, for the sake of completeness.

Table 1: Frequently Used Terms/Symbols

P :	penalty function
T_α :	the threshold function defined in (3) (said to be well-defined if the cost in (3) is strictly convex)
α :	step-size for ISTA
σ_m, σ_M :	least and greatest eigenvalues of $H^T H$
ρ :	weak-convexity parameter of P (see Defn.1)
σ :	Lipschitz constant for ∇f

Outline

We start our discussion with a majorization-minimization derivation of ISTA and show part of Prop. 3 in Section 2. This discussion is based solely on the descent properties of the algorithm. To provide a rigorous proof of convergence, we study in Section 3 the operator mapping 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 3.1, study the threshold operator for a weakly convex penalty in Section 3.2, provide a short and easier proof for the case where the cost is strictly convex in Section 3.3 and finally provide a proof of Prop. 4 in Section 3.4. The proof of Prop. 6 is provided in Section 4. In order to demonstrate that larger stepsizes may be more favorable, we also present an experiment in Section 5. We conclude with a brief outlook in Section 6.

We note that some of the frequently used terms and symbols are listed in Table 1.

2 ISTA as a Descent Algorithm via Majorization Minimization

In this section, we briefly recall the MM scheme to show that ISTA achieves monotone descent for a weakly-convex penalty provided that the step size is small enough. We refer to [15, 18] for a general discussion on MM.

Suppose that at a certain iteration, we have the estimate x^k. Expanding $\frac{1}{2}\|y - H\,x\|_2^2$, we have,

$$C(x) = \frac{1}{2}\,x^T\,H^T\,H\,x - x^T\,H^T\,y + P(x) + \text{const.} \tag{9}$$

For $\beta > \sigma_M$, we have that $\beta I - H^T\,H$ is positive definite. Let us define,

$$M(x, x^k) = C(x) + \frac{1}{2}(x - x^k)(\beta I - H^T\,H)\,(x - x^k). \tag{10}$$

It can be checked that

(i) $M(x^k, x^k) = C(x^k)$,

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

Thus, we can achieve descent in $C(x)$ by minimizing $M(x, x^k)$. But observe that

$$M(x, x^k) = \frac{\beta}{2}\, x^T x - x^T \left(\beta x^k + H^T (y - H\, x^k)\right) + P(x) + \text{const.} \tag{11}$$

$$= \beta \left[\frac{1}{2} \left\| x - \left(x^k + \frac{1}{\beta} H^T (y - H\, x^k) \right) \right\|_2^2 + \frac{1}{\beta} P(x) \right] + \text{const.} \tag{12}$$

Now since

$$\beta > \sigma_M \geq \sigma_m \geq \rho, \tag{13}$$

the function in (12) is strictly convex and has a unique minimizer given by

$$T_{\beta^{-1}} \left(x^k + \frac{1}{\beta} H^T \left(y - H\, x^k\right) \right). \tag{14}$$

From the foregoing discussion, we obtain,

Proposition 7. Consider a sequence defined by

$$x^{k+1} = T_\alpha \left(x^k + \alpha\, H^T \left(y - H\, x^k\right)\right). \tag{15}$$

If $\alpha < 1/\sigma_M$ then T_α is well-defined and $C(x^k)$ is a monotonely decreasing sequence. □

We remark that Prop. 7 covers only part of Prop. 3. Specifically, Prop. 7 does not claim that the iterations converge. Under certain assumptions on the set of minimizers, this can indeed be shown within the framework of majorization-minimization [20, 19]. However, the discussion in Section 3 in fact shows that convergence occurs for stepsizes that even exceed those permitted by Prop. 7.

3 ISTA as Iterations of an Operator

When $P(x)$ is convex, the forward-backward splitting algorithm [12, 2] leads to iterations that are of the same form as (15). However, the step size required to achieve monotone decrease in cost can be doubled. This accelerates convergence at the expense of producing a sequence that no longer monotonely decreases the cost. Instead, the distance to the minimizers is guaranteed to decrease with each iteration. We now investigate this issue for a non-convex penalty $P(x)$.

In order to simplify our analysis, we decompose the operator in (15). We define

$$U_\alpha(x) = \alpha\, H^T\, y + \left(I - \alpha\, H^T\, H\right) x. \tag{16}$$

Then, we can rewrite the iterations in (15) as,

$$x^{k+1} = T_\alpha \Big(U_\alpha \big(x^k\big) \Big). \tag{17}$$

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 C. 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. 3.4).

3.1 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 proposition.

Proposition 8. Suppose $P(x)$ is $(-\rho)$-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).

For the proof, let us first recall some definitions and results from convex analysis. We refer to [24, 17, 27] 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

$$f(x) + \langle y - x, z\rangle \leq f(y), \text{ for all } y. \tag{18}$$

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

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

Proposition 9. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is convex. x minimizes f if and only if $0 \in \partial f(x)$. □

In order to counter the concavity introduced by a weakly convex penalty function, we will need the data fidelity term to be strongly convex [27].

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

$$h(x) = f(x) - \frac{s}{2}\,\|x\|_2^2 \tag{19}$$

is convex if and only if $s \leq \rho$. □

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

Lemma 1. If h is convex and for some $\beta > 0$,

$$h(x) \leq \beta\|z - x\|_2^2 + h(z), \qquad \text{for all } z, \tag{20}$$

then h achieves its minimum at x.

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

$$0 \in \Big(2\beta(z - x) + \partial h(z)\Big)\Big|_{z=x} \tag{21}$$

or, $0 \in \partial h(x)$. □

We now prove a more general version of Prop. 8 using convex analysis methods. This general form will also be referred to in the proof of Prop. 6 in Sec. 4.

Proposition 10. Suppose $P : \mathbb{R}^n \to \mathbb{R}$ is a $(-\rho)$-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,

$$x = T_\alpha\Big(x - \alpha\,\nabla f(x)\Big), \tag{22}$$

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

Proof. ($\Rightarrow$) Suppose (22) 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

$$\frac{1}{2}\,\|x - u\|_2^2 + \alpha\,P(x) \leq \frac{1}{2}\,\|z - u\|_2^2 + \alpha\,P(z), \text{ for all } z. \tag{23}$$

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

$$\frac{1}{2}\,\|\alpha\,\nabla f(x)\|_2^2 + \alpha\,P(x) \leq \frac{1}{2}\,\|z - x\|_2^2 + \frac{1}{2}\,\|\alpha\,\nabla f(x)\|_2^2 + \langle z - x, \alpha\,\nabla f(x)\rangle + \alpha\,P(z), \text{ for all } z. \tag{24}$$

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

$$\alpha\,P(x) \leq \frac{1}{2}\,\|z - x\|_2^2 + \alpha\,f(z) - \alpha\,f(x) + \alpha\,P(z), \text{ for all } z. \tag{25}$$

Rearranging,

$$f(x) + P(x) \leq \frac{1}{2\,\alpha}\,\|z - x\|_2^2 + f(z) + P(z), \text{ for all } z. \tag{26}$$

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

($\Leftarrow$) Now the converse. 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,

$$0 \in \underbrace{\{\nabla f(x) - \rho\,x\}}_{\partial f_\rho(x)} + \partial P_\rho(x) \tag{27}$$

or,

$$\rho\,x - \nabla f(x) \in \partial P_\rho(x) \tag{28}$$

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

$$s\,x - \nabla f(x) \in \partial P_s(x) \tag{29}$$

By the convexity of P_s we then obtain,

$$P_s(x) + \langle z - x, s\,x - \nabla f(x)\rangle \leq P_s(z), \text{ for all } z. \tag{30}$$

Rearranging, we have

$$\frac{s}{2}\|x\|_2^2 - \langle x, s\,x - \nabla f(x)\rangle + P(x) \leq \frac{s}{2}\|z\|_2^2 - \langle z, s\,x - \nabla f(x)\rangle + P(x), \text{ for all } z. \tag{31}$$

Equivalently, for $s \geq \rho$,

$$\frac{1}{2}\left\|x - \left(x - s^{-1}\nabla f(x)\right)\right\|_2^2 + \frac{1}{s}\,P(x) \leq \frac{1}{2}\left\|z - \left(x - s^{-1}\nabla f(x)\right)\right\|_2^2 + \frac{1}{s}\,P(z), \text{ for all } z. \tag{32}$$

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

$$\frac{1}{2}\left\|x - \left(x - \alpha\,\nabla f(x)\right)\right\|_2^2 + \alpha\,P(x) \leq \frac{1}{2}\left\|z - \left(x - \alpha\,\nabla f(x)\right)\right\|_2^2 + \alpha\,P(z), \text{ for all } z, \tag{33}$$

for $\alpha\,\rho \leq 1$. Note that equality in (33) may be achieved by setting $z = x$. But we know that for $\alpha\,\rho < 1$, the right hand side is uniquely minimized by $z = T_\alpha\big(x - \alpha\,\nabla f(x)\big)$. Thus, $x = T_\alpha\big(x - \alpha\,\nabla f(x)\big)$ for $\alpha\,\rho < 1$. □

Prop. 8 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 $\rho \leq \sigma_m$.

3.2 Threshold Operators Associated with Weakly Convex Penalties

We now study the operator T_α. We only assume that T_α is associated with a $(-\rho)$-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,

$$\|Sx - Sz\|_2 \leq \|x - z\|_2. \tag{34}$$

We will make use of the following result (see [11, 2] for instance).

Lemma 2. 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,

$$\|S(x) - S(z)\|_2 \leq \|x - z\|_2. \tag{35}$$

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

$$0 \in u - x + \partial f(u), \tag{36a}$$

$$0 \in z - v + \partial f(v). \tag{36b}$$

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

$$x - z = u - v + (f_u - f_v). \tag{37}$$

Thus,

$$\|x - z\|_2^2 = \|u - v\|_2^2 + \|f_u - f_v\|_2^2 + \underbrace{\langle u - v, f_u - f_v\rangle}_{c}. \tag{38}$$

It follows by the monotonicity of the subdifferential that $c \geq 0$. Therefore, $\|u - v\|_2^2 \leq \|x - z\|_2^2$. □

We now study T_α.

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

$$\|T_\alpha(x) - T_\alpha(z)\|_2 \leq \frac{1}{1-\alpha\,\rho}\|x - z\|_2. \tag{39}$$

Proof. Note that the function

$$\rho\|x\|_2^2 + P(x) \tag{40}$$

is convex. Therefore,

$$S(x) = \arg\min_t \frac{1}{2}\|x - t\|_2^2 + c\Big(\rho\|t\|_2^2 + P(t)\Big) \tag{41}$$

is non-expansive by Lemma 2 for any $c > 0$. But we have

$$\begin{aligned} S(x) &= \arg\min_t \frac{1}{2}\|x - t\|_2^2 + c\left(\rho\|t\|_2^2 + P(t)\right) &(42)\\ &= \arg\min_t \frac{1+c\rho}{2}\|t\|_2^2 - \langle x, t\rangle + c\,P(t) &(43)\\ &= \arg\min_t \frac{1}{2}\left\|t - \frac{1}{1+c\rho}x\right\|_2^2 + \frac{c}{1+c\rho}\,P(t) &(44)\\ &= T_{c\,(1+c\rho)^{-1}}\left(\frac{1}{1+c\rho}x\right) &(45)\end{aligned}$$

Therefore we deduce that

$$\left\|T_{c\,(1+c\rho)^{-1}}\left(\frac{1}{1+c\rho}x\right) - T_{c\,(1+c\rho)^{-1}}\left(\frac{1}{1+c\rho}z\right)\right\|_2 = \|S(x) - S(z)\|_2 \leq \|x - z\|_2. \tag{46}$$

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 (46), we obtain

$$\left\|T_\alpha\Big((1-\alpha\rho)x\Big) - T_\alpha\Big((1-\alpha\rho)x\Big)\right\|_2 \leq \|x - z\|_2. \tag{47}$$

Making a change of variables, we finally obtain

$$\|T_\alpha(x) - T_\alpha(z)\|_2 \leq \frac{1}{1-\alpha\,\rho}\,\|x - z\|_2. \tag{48}$$

□

3.3 ISTA as Iterations of a Contraction Mapping

In this section, we derive a convergence result that is relatively easier to obtain. However, facility 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. 3.4, 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

$$\rho < \sigma_m, \tag{49a}$$

$$\alpha < \frac{2}{\sigma_M + \rho}. \tag{49b}$$

then, the iterations in (15) converge to the unique minimizer of $C(x)$.

Proof. For U_α in (16), we have that

$$\|U_\alpha(x) - U_\alpha(z)\|_2 \leq \max\big(|1 - \alpha\,\sigma_M|, |1 - \alpha\,\sigma_m|\big)\,\|x - z\|_2. \tag{50}$$

Therefore,

$$\Big\|T_\alpha\left(U_\alpha(x)\right) - T_\alpha\left(U_\alpha(z)\right)\Big\|_2 \leq \frac{\max\big(|1 - \alpha\,\sigma_M|, |1 - \alpha\,\sigma_m|\big)}{1 - \alpha\,\rho}\,\|x - z\|_2, \tag{51}$$

by Prop. 11. When $\rho < \sigma_m$, $C(x)$ is strictly convex and hence has a unique minimizer. By Prop. 8, this unique minimizer is in fact the unique fixed point of $T_\alpha\,U_\alpha$. Let us denote this point as z. Now if $\alpha\,\rho < 1$,

$$\|T_\alpha\,U_\alpha\,x - z\|_2 \leq \frac{\max\big(|1 - \alpha\,\sigma_M|, |1 - \alpha\,\sigma_m|\big)}{1 - \alpha\,\rho}\,\|x - z\|_2. \tag{52}$$

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

$$\frac{|1 - \alpha\,\sigma_M|}{1 - \alpha\,\rho} < 1, \tag{53a}$$

$$\frac{|1 - \alpha\,\sigma_m|}{1 - \alpha\,\rho} < 1, \tag{53b}$$

$$\alpha\,\rho < 1. \tag{53c}$$

(53a) is equivalent to

$$\rho < \sigma_M, \tag{54a}$$

$$\alpha < \frac{2}{\sigma_M + \rho}. \tag{54b}$$

(53b) is equivalent to

$$\rho < \sigma_m, \tag{55a}$$

$$\alpha < \frac{2}{\sigma_m + \rho}. \tag{55b}$$

Noting that $\sigma_M \geq \sigma_m$, and $2/(\sigma_M + \rho) \leq 2/(\sigma_m + \rho)$ we deduce that (49) implies (54), and (55). Finally, by (49a), we also have $2/(\sigma_M + \rho) < 1/\rho$. Then, from (49b), (53c) follows, proving convergence. ☐

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

$$\frac{2}{\sigma_M + \rho} > \frac{1}{\sigma_M}. \tag{56}$$

Thus the generalized forward backward algorithm converges for stepsizes greater than that allowed by majorization-minimization.

3.4 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

$$S = (1-\beta)\, I + \beta\, U, \tag{57}$$

for a non-expansive U. □

As a corollary of the definition, we have,

Lemma 3. $S : \mathbb{R}^n \to \mathbb{R}^n$ is β-averaged if and only if

$$\frac{1}{\beta}\Big(S - (1-\beta)\, I\Big) \tag{58}$$

is non-expansive. □

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, 12]. 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, $\|Ux - z\|_2 \leq \|x - z\|_2$, but Ux is not guaranteed to be closer to z than x. In the worst case, both x and Ux might have the same distance 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 Sx lies somewhere on the open segment between x and Ux, we will have $\|Sx - z\|_2 < \|x - z\|_2$ (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. We provide a proof in the appendix for the sake of completeness.

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 $x^{n+1} = Sx^n$. Then the sequence $\{x^n\}_{n\in\mathbb{Z}}$ converges and the limit is a fixed point of S. □

To prove the convergence of ISTA, we will show that $T_\alpha U_\alpha$ is an averaged operator. In order to show the averaged-ness of this operator, we will show that both T_α and U_α is averaged and invoke the following result (for an alternative statement and proof of this result, see Prop. 4.32 in [2]).

Proposition 14. Composition of two averaged operators is also averaged.

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,

$$\begin{aligned} S &= S_1\, S_2 \qquad (59)\\ &= (1-\beta_1)\,(1-\beta_2)\, I + \underbrace{(1-\beta_1)\,\beta_2\, U_2 + (1-\beta_2)\,\beta_1\, U_1 + \beta_1\, U_1\Big(\beta_2\, U_2\Big)}_{U} \end{aligned}$$

Note that

$$\begin{aligned} \|Ux - Uz\|_2 &\leq (1-\beta_1)\,\beta_2\,\|x-z\|_2 + (1-\beta_2)\,\beta_1\,\|x-z\|_2 + \beta_1\,\beta_2\,\|x-z\|_2 \qquad (60)\\ &= \Big(1 - (1-\beta_1)(1-\beta_2)\Big)\,\|x-z\|_2 \qquad (61) \end{aligned}$$

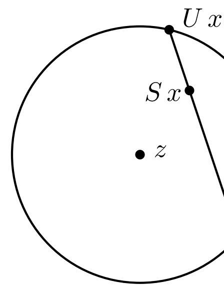

Figure 1: A non-expansive operator U might not always take a point x closer to its fixed point z. However, an averaged operator $S = (1-\beta)I + \beta\,U$ derived from U will have such a property.

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 3. $\square$

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

Lemma 4. Suppose $q : \mathbb{R}^n \to \mathbb{R}$ is convex and $S(x)$ is defined as

$$S(x) = \arg\min_t \frac{1}{2}\,\|x-t\|_2^2 + q(t). \tag{62}$$

Then, S is 1/2-averaged.

Proof. By Lemma 3, we need to show that $U = 2S - I$ is non-expansive. Let $u = S(x)$ and $v = S(z)$. Recall from the proof of Lemma 2 that we can write, by the optimality conditions that, for some $f_u \in \partial f(u)$ and $f_v \in \partial f(v)$,

$$x - z = u - v + (f_u - f_v). \tag{63}$$

Observing this, we have

$$\begin{aligned}
\|Ux - Uz\|_2^2 &= 4\|u-v\|_2^2 + \|x-z\|_2^2 - 4\langle u-v, x-z\rangle & (64)\\
&= 4\|u-v\|_2^2 + \|x-z\|_2^2 - 4\langle u-v, u-v+f_u-f_v\rangle & (65)\\
&= \|x-z\|_2^2 - 4\,\underbrace{\langle u-v, f_u-f_v\rangle}_{c} & (66)\\
&\leq \|x-z\|_2^2 & (67)
\end{aligned}$$

where the last inequality follows because $c \geq 0$, by the monotonicity of the subgradient. But this shows that U is non-expansive. $\square$

Using this lemma, we can show that composition of the threshold operator T_α with a scaling can be shown to be averaged.

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

$$S_\alpha(x) = T_\alpha\Big((1-\alpha\,\rho)x\Big). \tag{68}$$

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 a 1/2-averaged by Lemma 4. $\square$

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

$$S(x) = M\,x + u, \tag{69}$$

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 $[\sigma_0, \sigma_1] \subset (-1, 1]$. Consider

$$V = \frac{1}{\beta}\Big(S - (1 - \beta)\,I\Big). \tag{70}$$

Then,

$$V(x) - V(z) = \underbrace{\frac{1}{\beta}\Big(M - (1 - \beta)I\Big)}_{U}(x - z). \tag{71}$$

The eigenvalues of U lie in the interval,

$$\left[\frac{\sigma_0 - 1 + \beta}{\beta}, \frac{\sigma_1 - 1 + \beta}{\beta}\right] \tag{72}$$

Note that for any $\beta > 0$,

$$\frac{\sigma_1 - 1 + \beta}{\beta} \le 1. \tag{73}$$

Also, if $\beta \le (1 - \sigma_0)/2 < 1$, we have

$$\frac{\sigma_0 - 1 + \beta}{\beta} \ge -1. \tag{74}$$

Thus U is non-expansive and S is averaged. $\square$

In the convergence proof, to counter the scale factor that appears in (68), we will invoke the following.

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

$$V_\alpha = \frac{1}{1 - \alpha\,\rho} U_\alpha. \tag{75}$$

Then, V_α is averaged if

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

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

Proof. Observe that V_α is of the form

$$V_\alpha(x) = \underbrace{\frac{1}{1-\alpha\,\rho}\left(I - \alpha\,H^T\,H\right)}_{M}\,x + u, \tag{77}$$

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

$$\left[\frac{1-\alpha\,\sigma_M}{1-\alpha\,\rho}, \frac{1-\alpha\,\sigma_m}{1-\alpha\,\rho}\right]. \tag{78}$$

By (76a), we have

$$\frac{1-\alpha\,\sigma_m}{1-\alpha\,\rho} \leq 1. \tag{79}$$

By (76b), we have

$$\frac{1-\alpha\,\sigma_M}{1-\alpha\,\rho} > -1. \tag{80}$$

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. 8 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. □

4 General Data Fidelity Term

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

$$D(x) = f(x) + P(x), \tag{81}$$

where P is $(-\rho)$-weakly convex and f is ρ-strongly convex which is differentiable and

$$\|\nabla f(x) - \nabla f(y)\| \leq \sigma \|x - y\|, \text{ for all } x, y. \tag{82}$$

Definition 6. We say that f has a σ-Lipschitz continuous gradient, if (82) holds. □

Let us now recall the Baillon-Haddad theorem [2, 1, 8].

Lemma 5. Suppose f is convex, differentiable and

$$\|\nabla f(x) - \nabla f(z)\|_2 \leq \tau \|x - z\|_2. \tag{83}$$

Then,

$$\langle \nabla f(x) - \nabla f(z), x - z\rangle \geq \frac{1}{\tau}\,\|\nabla f(x) - \nabla f(z)\|_2^2. \tag{84}$$

In the setting above, we have,

Lemma 6. 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 $T = \nabla f$ and $U = \nabla g$. Note that $U = T - \rho I$. First observe that, by the Baillon-Haddad theorem applied to T, we have,

$$\langle Tx - Ty, x - y \rangle \geq \frac{1}{\sigma} \|Tx - Ty\|_2^2. \tag{85}$$

Now,

$$\begin{aligned}
\|Ux - Uy\|_2^2 &= \|Tx - Ty\|_2^2 + \rho^2 \|x - y\|_2^2 - 2\rho \langle Tx - Ty, x - y \rangle \\
&\leq \left(1 - 2\frac{\rho}{\sigma}\right) \|Tx - Ty\|_2^2 + \rho^2 \|x - y\|_2^2 \\
&\leq \left(1 - 2\frac{\rho}{\sigma}\right) \sigma^2 \|x - y\|_2^2 + \rho^2 \|x - y\|_2^2 \\
&= (\sigma - \rho)^2 \|x - y\|_2^2.
\end{aligned}$$

Taking square roots, we obtain $\|Ux - Uy\|_2 \leq (\sigma - \rho)\|x - y\|_2$, which is the claimed result. $\square$

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

$$V_\alpha = \frac{1}{1 - \alpha\,\rho} \Big(I - \alpha \nabla f \Big). \tag{87}$$

Then, V_α is averaged if

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

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

$$\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq \frac{1}{\sigma - \rho} \|\nabla g(x) - \nabla g(y)\|_2^2. \tag{89}$$

Also,

$$V_\alpha = \frac{1}{1 - \alpha\,\rho} \Big(I - \alpha \big(\rho I + \nabla g \big) \Big) \tag{90a}$$

$$= I - \frac{\alpha}{1 - \alpha\,\rho} \nabla g. \tag{90b}$$

And now by Prop. 4.33 of [2], it follows that $I - \frac{\alpha}{1-\alpha\,\rho} \nabla g$ is aveararged if

$$\frac{\alpha}{1 - \alpha\,\rho} < \frac{2}{\sigma - \rho}. \tag{91}$$

But this inequality is a rearrangement of (88). $\square$

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

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

$$A\,x = T_\alpha\Big(x - \alpha\,\nabla f(x)\Big), \tag{92}$$

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. □

5 Experiment

To demonstrate the discussion, we performed an experiment. First, we produced a 60×50 matrix H by sampling the columns of a convolution matrix associated with a filter. Using a sparse x as shown in Fig.3a, we took as observations

$$y = H\,x + n, \tag{93}$$

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

$$P_{\tau,\rho}(s) = \begin{cases} \tau\,|s| - s^2/(2\rho), & \text{if } |s| < \tau/\rho, \\ \tau^2/(2\rho), & \text{if } |s| \geq \tau/\rho, \end{cases} \tag{94}$$

as the penalty function. It can be shown this the penalty function is $(-\rho)$-weakly convex. The threshold function associated with P (see (3)) is given by (provided $\alpha\,\rho < 1$)

$$T_\alpha(x) = \begin{cases} 0, & \text{if } |x| < \alpha\,\tau, \\ (1-\alpha\,\rho)^{-1}\,(x - \alpha\,\tau), & \text{if } \alpha\,\tau \leq |x| < \tau/\rho, \\ x, & \text{if } \tau/\rho \leq |x|. \end{cases} \tag{95}$$

This threshold function is depicted in Fig. 2. Observe that for $\alpha\,\tau < |x| < \tau/\rho$, the derivative of the threshold function exceeds unity. Therefore the 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. For that choice of ρ in the definition of P, we obtained the estimate as

$$x^* = \arg\min_t \frac{1}{2}\|y - H\,t\|_2^2 + \sum_i P(t_i). \tag{96}$$

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 for which monotone descent is guaranteed (see Prop. 3), whereas for α_1, convergence is guaranteed but monotone descent is not (see Prop. 4). For this specific problem, the ratio α_1/α_0 was found to be 1.88.

We ran ISTA with α_0 and α_1, in both cases, starting from zero. In order to better evaluate the convergence properties we also tried FISTA [4] and TwIST [5] which are simple schemes leading to significant acceleration of ISTA. However, both methods require that the components of the cost function be convex. Therefore, the convergence analyses in [4, 5] are not valid in our setting. Nevertheless, both methods are formally applicable. We note that at least in the convex case,

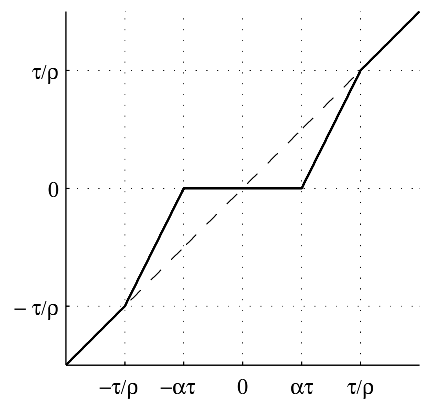

Figure 2: The threshold function in (95). Observe that this threshold function is not non-expansive. Specifically, the derivative exceeds unity on the intervals $(-\tau/\rho, -\alpha\rho)$ and $(\alpha\rho, \tau/\rho)$

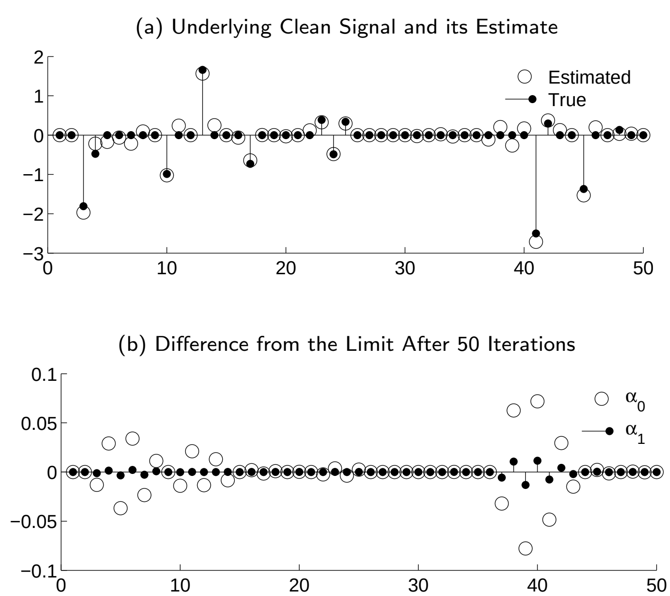

Figure 3: (a) The true signal and its estimate obtained by solving (96). (b) The difference from the minimizer after 50 iterations of ISTA with stepsizes $\alpha_0 < \alpha_1$.

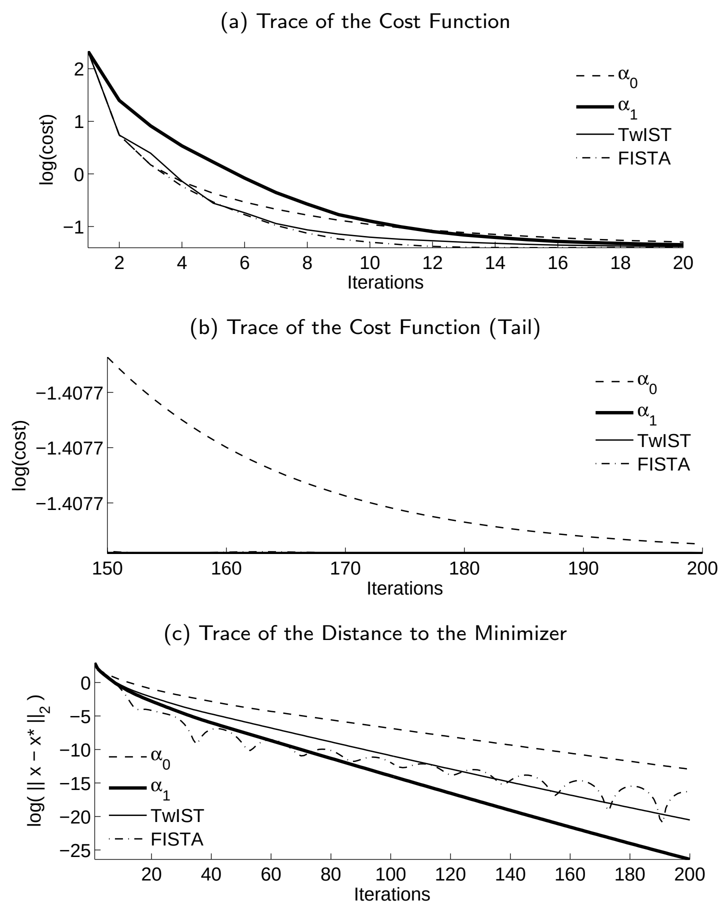

Figure 4: Effects of changing the step size α in (15). (a,b) show how the cost function changes with iterations using the greatest possible α_0 that guarantees strict descent vs. the greatest possible α_1 that merely guarantees convergence. The cost with TwIST and FISTA are also included. (c) shows a comparison of how the distance to the unique minimizer changes with iterations for the same four choices.

FISTA requires that the iterates strictly decrease the cost. Therefore for FISTA, we employed the stepsize α_0, for which monotone descent is guaranteed (we also tried it for α_1, but the sequence diverged). Given the step-size, we applied the algorithm listed as 'FISTA with constant stepsize' in Sec.4 of [4]. For TwIST, there are two parameters to choose, namely 'α' and 'β' – for these we used the suggestions in equations (26) and (27) of [5].

The trace of the cost function with iterations is shown in Fig. 4a,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 (see Fig.4b). One surprising outcome was that the cost decreases monotonely with α_1 even though there is no theoretical guarantee. 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.

6 Conclusion

In this paper, we generalized the ISTA convergence condition for the stepsize to the case where the penalty is weakly-convex. We have 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 c onvex function. Specifically, we have seen that the proximity operator for a weakly convex function is no longer (firmly) non-expansive. However, under proper scaling, the proximity operators become firmly non-expansive. 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 [14], which can in turn be used to derive other algorithms such as the alternative direction method of multipliers (ADMM) [14, 7], or the parallel proximal algorithm (PPXA) [23]. 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

$$\|Sx - Sy\| = \|(1-\beta)(x-y) + \beta(Ux - Uy)\| \tag{97a}$$

$$\leq (1-\beta)\|x-y\| + \beta\|Ux - Uy\| \tag{97b}$$

$$\leq \|x-y\|, \tag{97c}$$

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,

$$\|Sx - Sy\|^2 = (1-\beta)^2\|x-y\|^2 + \beta^2\|Ux - Uy\|^2 + 2\beta(1-\beta)\langle x-y, Ux-Uy\rangle. \tag{98}$$

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

$$\langle x-y, Ux-Uy\rangle \leq \|x-y\|\,\|Ux-Uy\|. \tag{99}$$

Equality holds in (99) 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 (99), we will have that

$$\langle x-y, Ux-Uy\rangle < \|x-y\|^2. \tag{100}$$

Combining this observation with the non-expansiveness of U (i.e., $\|Ux-Uy\|^2 \leq \|x-y\|^2$), we obtain from (98)

$$\|Sx - Sy\|^2 < \|x-y\|^2. \tag{101}$$

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

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 $x^* \in \operatorname{Fix} S$,

$$\|x^n - x^*\| \leq \|x^0 - x^*\|. \tag{102}$$

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,

$$\epsilon = \|z - z^*\| - \|Sz - z^*\| > 0. \tag{103}$$

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

$$\|x^{n_k} - z\| \leq \epsilon/4, \tag{104a}$$

$$\|Sx^{n_k} - Sz\| \leq \epsilon/4. \tag{104b}$$

Now if $k \geq K$, then by nonexpansivity and (104b)

$$\|z^* - x^{n_{k+1}}\| \leq \|z^* - Sx^{n_k}\| \tag{105a}$$

$$\leq \|z^* - Sz\| + \epsilon/4. \tag{105b}$$

But by (104a),

$$\|z^* - x^{n_{k+1}}\| \geq \|z^* - z\| - \epsilon/4 \tag{106}$$

Combining (105) and (107), we obtain a contradiction as

$$\|z^* - z\| \leq \|z^* - Sz\| + \epsilon/2. \tag{107}$$

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. $\square$

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, 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.

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

[5] 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.

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

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] 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.

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

[23] 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.

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

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

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

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