# NEW OPERATORS FOR FIXED-POINT THEORY: THE SPARSITY-AWARE LEARNING CASE

*Konstantinos Slavakis*[1] *Yannis Kopsinis*[2] *Sergios Theodoridis*[2]

[1]University of Minnesota
Digital Technology Center, Minneapolis, USA
Email: slavakis@dtc.umn.edu

[2]University of Athens
Dept. Informatics & Telecomms., Athens, Greece
Emails: kopsinis@ieee.org, stheodor@di.uoa.gr

## ABSTRACT

The present paper offers a link between fixed point theory and thresholding; one of the key enablers in sparsity-promoting algorithms, associated mostly with non-convex penalizing functions. A novel family of operators, the partially quasi-nonexpansive mappings, is introduced to provide the necessary theoretical foundations. Based on such fixed point theoretical ground, and motivated by hard thresholding, the generalized thresholding (GT) mapping is proposed that encompasses hard, soft, as well as recent advances of thresholding rules. GT is incorporated into an online/time-adaptive algorithm of linear complexity that demonstrates competitive performance with respect to computationally thirstier, state-of-the-art, RLS- and proportionate-type sparsity-aware methods.

***Index Terms***— Thresholding, sparsity, fixed point theory, adaptive filtering.

## 1. INTRODUCTION

Thresholding, the operation of nullifying small components of an $L \times 1$ vector $\boldsymbol{a}$ while shrinking or leaving intact the others, is one of the key enablers of sparsity-promoting algorithms [1,2]. It is by now well-established that hard thresholding (HT), a discontinuous operator, tends to introduce large variance on estimates [3–5]. On the other hand, the continuous soft thresholding (ST) operator has the tendency to increase bias [3–5]. To overcome these drawbacks, alternative thresholding rules have been proposed [4–9]. These advances in thresholding operators are strongly connected to optimization tasks; they are obtained by penalizing squared error terms by, usually, non-convex losses.

This paper establishes a link between thresholding and fixed point theory [10] by introducing a novel family of operators; the partially quasi-nonexpansive mappings. Motivated by HT and the established theoretical framework, a generalized thresholding (GT) operator is introduced that encompasses HT, ST, as well as recent advances in [3,5,7,8]. GT is incorporated into the adaptive projection-based generalized thresholding (APGT) algorithm to address system identification tasks in online/time-adaptive settings, i.e., the scenario where training data arrive sequentially, they are only utilized for a limited number of times, and the system to be identified may be time-variant. Extensive numerical examples suggest that APGT offers competitive performance to RLS-type sparsity-promoting algorithms, while it outperforms computationally thirstier, state-of-the-art proportionate-type techniques.

## 2. MODEL DEFINITION

Assume a (separable) Hilbert space $\mathcal{H}_*$, equipped with an inner product $\langle\cdot,\cdot\rangle$, and induced norm $\|\cdot\| := \sqrt{\langle\cdot,\cdot\rangle}$.

Discussion revolves around the following model, linear in an unknown system $f_* \in \mathcal{H}_*$,

$$y_n = \langle f_*, h_n\rangle + \eta_n, \quad n \in \mathbb{N}, \tag{1}$$

where $(y_n, h_n)_{n\in\mathbb{N}} \in \mathbb{R} \times \mathcal{H}_*$ is the set of training data, and $(\eta_n)_{n\in\mathbb{N}}$ stands for the noise process. The objective of this short paper is identification of the generally non-linear object $f_*$, given the side information that $f_*$ admits a sparse representation in some linear subspace $\mathcal{H}$ of $\mathcal{H}_*$. An example is the case where $f_*$ is sparse in the subspace of band-limited functions. If $P_{\mathcal{H}}$ stands for the orthogonal projection onto $\mathcal{H}$, then by $\langle f_*, h_n\rangle = \langle f_*, P_{\mathcal{H}}(h_n)\rangle$ [10, Coroll. 3.22.ii], it suffices to perform the following discussion in $\mathcal{H}$, instead of the larger $\mathcal{H}_*$, and by abusing notation, $h_n$ to denote $P_{\mathcal{H}}(h_n)$.

**Example 1.** In the case where $\mathcal{H}$ is a *reproducing kernel Hilbert space (RKHS)* [11], equipped with a kernel $\kappa$, the inner products in (1) can be readily available. Recall that $\mathcal{H}$ is RKHS iff there exists a (unique) kernel function $\kappa(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ ($p \in \mathbb{N}_*$) such that (i) $\kappa(\cdot, \boldsymbol{t}) \in \mathcal{H}$, and (ii) the *reproducing property* holds: $\langle f, \kappa(\cdot, \boldsymbol{t})\rangle = f(\boldsymbol{t})$, $\forall (f, \boldsymbol{t}) \in \mathcal{H} \times \mathbb{R}^p$. Indeed, if $h_n := \kappa(\cdot, \boldsymbol{t}_n)$, then $\langle f_*, h_n\rangle = f_*(\boldsymbol{t}_n)$, which is nothing but the evaluation or *sampling* of $f_*$ at $\boldsymbol{t}_n$, $\forall n$.

**Assumption 1.** $\mathcal{H}$ is assumed to be finite-dimensional. In other words, there exists a finite set of orthonormal $\{\psi_i\}_{i=1}^L \subset \mathcal{H}$ such that $\mathcal{H} = \operatorname{Span}\{\psi_i\}_{i=1}^L$.

Assumption 1 simplifies discussion, since for any $f \in \mathcal{H}$ there exists (unique) $\boldsymbol{a} \in \mathbb{R}^L$ with $f = \sum_{i=1}^L a_i\psi_i =: \Psi\boldsymbol{a}$, where $\Psi : \mathbb{R}^L \to \mathcal{H}$ is the linear operator defined by the previous equation. In such a case, the concept of $f$ being sparse lies on a firm basis; it translates to $\boldsymbol{a}$ being sparse. Due to space limitations, either the case where $\dim(\mathcal{H}) = \infty$ or $\{\psi_i\}_{i=1}^L$ are unknown is not examined here. However, it is worthy to note here that the abundance of training data may be beneficial in such cases; $(h_n)_{n\in\mathbb{N}}$ define a sequence of linear subspaces $(\mathcal{H}_n := \operatorname{Span}\{h_i\}_{i=0}^n)_{n\in\mathbb{N}}$ of non-decreasing dimension. Each $\mathcal{H}_n$ possesses an orthonormal basis, which can be identified by online or sequential variations of the classical Gram-Schmidt process, e.g., [12], by which sparsity on some $f_n \in \mathcal{H}_n$ is translated to sparsity on its finite dimensional rendition $\boldsymbol{a}_n \in \mathbb{R}^{\dim(\mathcal{H}_n)}$.

By the orthonormality of $\{\psi_i\}_{i=1}^L$, $\langle\Psi\boldsymbol{a}_1, \Psi\boldsymbol{a}_2\rangle = \boldsymbol{a}_1^\top\boldsymbol{a}_2$, $\forall\boldsymbol{a}_1, \boldsymbol{a}_2 \in \mathbb{R}^L$. In other words, $\Psi$ is an isometry, so that discussion on (1) can be done in $\mathbb{R}^L$ without any loss of generality. Moreover, symbols $f$ and $\boldsymbol{a}$ will be used interchangeably to denote the same object. Since $h_n = \Psi\boldsymbol{u}_n$, for some $\boldsymbol{u}_n \in \mathbb{R}^L$, then $\langle f_*, h_n\rangle = \boldsymbol{a}_*^\top\boldsymbol{u}_n$.

Following a previous ST-based rationale [13], to model inaccuracies and unknown noise statistics, a *hyperslab* is defined around each datum $(y_n, \boldsymbol{u}_n)$ for some user-defined $\epsilon_n \geq 0$:

$$S_n[\epsilon_n] := \left\{\boldsymbol{a} \in \mathbb{R}^L : \left|\boldsymbol{u}_n^\top\boldsymbol{a} - y_n\right| \leq \epsilon_n\right\},\ \forall n. \tag{2}$$

## 3. FRAGMENTS OF FIXED POINT THEORY

The following discussion holds true also in Hilbert spaces $\mathcal{H}$ with $\dim(\mathcal{H}) = \infty$. A concept of fundamental importance, associated with every mapping $T$, is its *fixed point set* $\operatorname{Fix}(T) := \left\{f \in \mathcal{H} : T(f) = f\right\}$ [10, Chap. 4]. To leave no place for ambiguity, $\operatorname{Fix}(T)$ is assumed nonempty.

**Definition 1** (Partially quasi-nonexpansive mappings)**.** A mapping $T$ is called partially quasi-nonexpansive, if

$$\begin{aligned}\forall f \in \mathcal{H}, \exists Y_f \subset \operatorname{Fix}(T) : \forall g \in Y_f,\\ \|T(f) - g\| \leq \|f - g\|.\end{aligned} \tag{3}$$

$\operatorname{Fix}(T)$ is *not* necessarily convex. An example is the fixed point set of the partially quasi-nonexpansive mapping of Section 5, which is a union of linear subspaces.

If we set $Y_f := \operatorname{Fix}(T)$, $\forall f$, in (3), then the mapping $T$ is called *quasi-nonexpansive,* with closed and convex $\operatorname{Fix}(T)$ [10]. Clearly, any quasi-nonexpansive mapping is a partially quasi-nonexpansive one. This is a strict inclusion due to the existence of the mapping in Section 5.

$T$ is called nonexpansive if $\forall f_1, f_2 \in \mathcal{H}$, $\|T(f_1) - T(f_2)\| \leq \|f_1 - f_2\|$. In such a case, $\operatorname{Fix}(T)$ is closed and convex [10]. It is easy to see that any nonexpansive mapping is a quasi-nonexpansive one. Well-known examples of nonexpansive mappings, with principle role in applications, are as follows.

**Example 2** (Metric projection mapping)**.** Given a nonempty closed convex set $C \subset \mathcal{H}$, the metric projection mapping $P_C$ onto $C$ is defined as the operator which assigns to an $f \in \mathcal{H}$ the unique $P_C(f) \in C$ such that $\|f - P_C(f)\| = \min_{g\in C}\|f - g\|$. It is well-known that $P_C$ is a nonexpansive mapping, with $\operatorname{Fix}(P_C) = C$. For example, $S_n[\epsilon_n]$ in (2) is a closed convex set, with $P_{S_n[\epsilon_n]}$ available analytically [13]. It is also known that nonexpansiveness is inherited by compositions and convex combinations of finite number of projection mappings onto intersecting closed convex sets [10, Section 4.4].

**Example 3** (Proximal mapping)**.** Given a function $\varphi : \mathcal{H} \to \mathbb{R}\cup\{\infty\}$, and $\lambda > 0$, define the set-valued proximal mapping $\operatorname{Prox}_{\lambda\varphi} : \mathcal{H} \rightrightarrows \mathcal{H}$ as [14, Def. 1.22]

$$\operatorname{Prox}_{\lambda\varphi}(f) := \arg\inf_{g\in\mathcal{H}} \frac{1}{2\lambda}\|f - g\|^2 + \varphi(g), \quad \forall f \in \mathcal{H}.$$

If $\varphi$ is (lower semicontinuous) convex, then $\operatorname{Prox}_{\lambda\varphi}$ becomes single-valued, with eminent applicability to signal processing tasks [10,15]. Moreover, in such a case, $\operatorname{Prox}_{\lambda\varphi}$ is nonexpansive, with $\operatorname{Fix}(\operatorname{Prox}_{\lambda\varphi}) = \arg\min\varphi(\mathcal{H})$, $\forall\lambda > 0$, provided that $\arg\min\varphi(\mathcal{H}) \neq \emptyset$ [10, Prop. 12.28]. If $\varphi := \iota_C$, where $\iota_C$ attains the value of 0 on the closed convex set $C$, and $+\infty$ elsewhere, then $\operatorname{Prox}_{\lambda\iota_C} = P_C$, $\forall\lambda > 0$.

## 4. PENALIZED LEAST-SQUARES

Going back to (1), choose $N \in \mathbb{N}_*$, and define $\boldsymbol{U}_n := [\boldsymbol{u}_n, \ldots, \boldsymbol{u}_{n-N+1}] \in \mathbb{R}^{L\times N}$, $\boldsymbol{y}_n := [y_n, \ldots, y_{n-N+1}]^\top \in \mathbb{R}^N$, and $\boldsymbol{v}_n := [v_n, \ldots, v_{n-N+1}]^\top \in \mathbb{R}^N$. Then, (1) takes the form of $\boldsymbol{y}_n = \boldsymbol{U}_n^\top\boldsymbol{a}_* + \boldsymbol{v}_n$, $\forall n \in \mathbb{N}$. The mainstream of batch sparsity-promoting algorithms utilize all the gathered $N$ training data to find an exact or approximate solution, in most cases iteratively, to the following *penalized least-squares* minimization task,

$$\min_{\boldsymbol{a}\in\mathbb{R}^L} \frac{1}{2}\|\boldsymbol{y}_n - \boldsymbol{U}_n^\top\boldsymbol{a}\|^2 + \lambda\sum_{i=1}^L p(|a_i|), \tag{4}$$

where $p : \mathbb{R} \to [0, \infty)$ stands for a sparsity-promoting, non-decreasing, and non-convex, in general, penalty function, $\lambda \in (0, \infty)$ is the regularization parameter, and $a_i$ stands for the $i$-th coordinate of the vector $\boldsymbol{a}$.

Choices for $p$ are numerous; if, for example, $p(|a|) := \chi_{\mathbb{R}\setminus\{0\}}(|a|)$, $\forall a \in \mathbb{R}$, where $\chi_{\mathscr{A}}$ stands for the characteristic function with respect to $\mathscr{A} \subset \mathbb{R}$, then the regularization term $\sum_{i=1}^L p(|a_i|)$ becomes the $\ell_0$-norm of $\boldsymbol{a}$. In the case where $p(|a|) := |a|$, $\forall a \in \mathbb{R}$, then the regularization term is the $\ell_1$-norm $\|a\|_1 := \sum_{i=1}^L |a_i|$, and (4) is the LASSO [16]. However, it has been observed that if some of the LASSO's regularity conditions are violated, then LASSO is sub-optimal for model selection [5, 9, 17]. Such a behavior has motivated the search for non-convex penalty functions $p$, which bridge the gap between the $\ell_0$- and $\ell_1$-norm; for example, the $\ell_\gamma$-penalty, for $\gamma \in (0,1)$, [8], the log- [8], the SCAD [8], the MC+ [5], and the transformed $\ell_1$-penalties [8].

Recently, sparsity-promoting coordinate-wise optimization techniques for solving (4) are attracting a lot of interest [5,18]. As a justification, assume, for example, that $N = L$, and that $\boldsymbol{U}_n$ is orthogonal. By $\tilde{\boldsymbol{a}}_n := \boldsymbol{U}_n \boldsymbol{y}_n$, (4) is equivalent to the following separable-in-components task [3,8]; find

$$\mathfrak{M}(\tilde{\boldsymbol{a}}) := \arg\min_{\boldsymbol{a}\in\mathbb{R}^L} \left( \sum_{i=1}^{L} \frac{1}{2\lambda} \left(\tilde{a}_i - a_i\right)^2 + p(|a_i|) \right). \quad (5)$$

The connection of (5) with the proximal mapping of Example 3 is evident; in short, $\mathfrak{M}(\tilde{\boldsymbol{a}}) = \bigtimes_{i=1}^{L} \operatorname{Prox}_{\lambda\varphi}(\tilde{a}_i)$, where $\bigtimes$ stands for the cartesian product, and $\varphi := p \circ |\cdot|$. Under regularity conditions on $p$, $\mathfrak{M}(\tilde{\boldsymbol{a}})$ of (5) becomes a singleton [8]. On the other hand, the generalized thresholding mapping in Section 5 does not impose any assumptions on the regularity of $p$, and thus discussion in Section 5 does not confine $\mathfrak{M}(\tilde{\boldsymbol{a}})$ to a singleton.

Figs. 1(b-d), show the PLSTOs associated with some of the most commonly employed penalty functions. For example, if $p(|a|) := \left[\lambda^2 - (|a| - \lambda)^2 \chi_{[0,\lambda)}(|a|)\right]/\lambda$, $\forall a \in \mathbb{R}$, then the resulting PLSTO is the celebrated HT [8], which is depicted in Fig. 1a together with ST, which results in the case where $p(|a|) := |a|$. The rest of the thresholding rules in Fig. 1b correspond to MC+ [5,9] and SCAD [8], respectively. HT is far from being the only discontinuous PLSTO. An example is shown in Fig. 1c, by bridge thresholding (BT) [6], which relates to the $\ell_\gamma$-penalty, $\gamma < 1$. Continuous thresholding functions, with nonlinear parts, are shown in Fig. 1(d). More specifically, the non-negative garrote [7] and representatives of the $n$-degree garrote thresholding are illustrated.

## 5. GENERALIZED THRESHOLDING MAPPING

**Definition 2** (The mapping $T_{\mathrm{GT}}^{(K)}$)**.** Fix a positive integer $K < L$ and define $T_{\mathrm{GT}}^{(K)} : \mathcal{H} \to \mathcal{H}$ as follows. For any $f = \Psi\boldsymbol{a}$, the output $T_{\mathrm{GT}}^{(K)}(f) = \Psi\boldsymbol{b}$ is obtained according to the following steps:

(1) Identify, first, the $K$ largest in absolute value components of $\boldsymbol{a}$, with $J_f^{(K)}$ being the length $K$ ordered tuple which contains their indices. If there are multiple components of $\boldsymbol{a}$ with the same absolute value, choose the one with the smallest index. Then, $\forall i \in M_{J_f^{(K)}}$, set $b_i := a_i$.

(2) For the rest of the components $i \notin J_f^{(K)}$, set $b_i := \operatorname{Shr}(a_i)$, where the function $\operatorname{Shr} : \mathbb{R} \to \mathbb{R}$ satisfies the following conditions: (i) $\tau \operatorname{Shr}(\tau) \geq 0$, (ii) $|\operatorname{Shr}(\tau)| \leq |\tau|$, and (iii) given any sufficiently small $\epsilon > 0$, there exists a $\delta > 0$, and an interval $\mathcal{D} \subset \mathbb{R}$ such that $\forall \tau \in \mathcal{D}\setminus(-\epsilon, \epsilon)$, $|\operatorname{Shr}(\tau)| \leq |\tau| - \delta$. In other words, $(\delta, \mathcal{D})$ could be user-defined parameters to guarantee that Shr acts as a *strict* shrinkage operator on $\mathcal{D} \setminus (-\epsilon, \epsilon)$. The $\epsilon$ parameter is introduced to exclude 0 from the picture, since, usually, $\operatorname{Shr}(0) = 0$ (see Fig. 1).

**Theorem 1.** (1) $T_{\mathrm{GT}}^{(K)}$ is partially quasi-nonexpansive; more specifically, $\forall f \in \mathcal{H}, \forall g \in M_{J_f^{(K)}}$, $\|f - T_{\mathrm{GT}}^{(K)}(f)\|^2 \leq \|f - g\|^2 - \|T_{\mathrm{GT}}^{(K)}(f) - g\|^2$, where $M_{J_f^{(K)}} := \{\Psi\boldsymbol{a} : a_i = 0, \forall i \notin J_f^{(K)}\}$.

(2) $\operatorname{Fix}(T_{\mathrm{GT}}^{(K)}) = \bigcup_{J\in\mathscr{T}(K,L)} M_J$, where $\mathscr{T}(K,L)$ stands for all the ordered tuples of length $K$ out of $\{1, 2, \ldots, L\}$, and $M_J := \{\Psi\boldsymbol{a} : a_i = 0, \forall i \notin J\}$. Notice that $\operatorname{Fix}(T_{\mathrm{GT}}^{(K)})$, as a union of linear subspaces of $\mathcal{H}$, is a non-convex set.

(3) Let a sequence $(f_n)_{n\in\mathbb{N}} \subset \mathcal{H}$ and an $f_* \in \mathcal{H}$. If $\lim_{n\to\infty} f_n = f_*$, and $\lim_{n\to\infty}\big(I - T_{\mathrm{GT}}^{(K)}\big)(f_n) = 0$, then $f_* \in \operatorname{Fix}(T_{\mathrm{GT}}^{(K)})$. This property can be rephrased as $I - T_{\mathrm{GT}}^{(K)}$ being *demiclosed at* 0.

The proof of the previous theorem is omitted due to space limitations. An illustration of GT with a generic Shr can be found in Fig. 1a. Examples where Shr is chosen from the existing rich library of thresholding rules can be found in Fig. 1e.

## 6. ALGORITHM

**Algorithm 1** (The adaptive projection-based generalized thresholding (APGT) algorithm)**.** For an arbitrary initial point, $\boldsymbol{a}_0 \in \mathbb{R}^L$, execute the following for every $n \in \mathbb{N}$:

(1) Define the sliding window $\mathcal{J}_n := \overline{\max\{0, n-q+1\}, n}$ on the time axis, of size at most $q$, where $\overline{j_1, j_2}$ for two integers $j_1 \leq j_2$ stands for $\{j_1, j_1+1, \ldots, j_2\}$. The set $\mathcal{J}_n$ defines all the indices corresponding to the hyperslabs, which are to be processed at the time instant $n$. Among these, identify the "active" hyperslabs $\mathcal{I}_n := \big\{i \in \mathcal{J}_n : P_{S_i[\epsilon_i]}(\boldsymbol{a}_n) \neq \boldsymbol{a}_n\big\}$. Moreover, for every $i \in \mathcal{I}_n$, define the weight $\omega_i^{(n)} > 0$, with $\sum_{i\in\mathcal{I}_n} \omega_i^{(n)} = 1$, to weigh the importance of the information carried by each hyperslab $S_i[\epsilon_i]$.

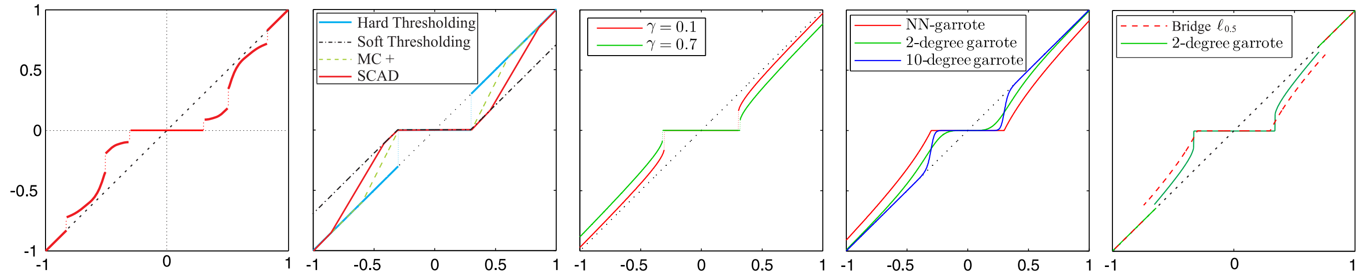
(2) Choose an $\varepsilon' \in (0,1]$, and let the user-defined $\mu_n$ take values within $[\varepsilon'\mathcal{M}_n, (2-\varepsilon')\mathcal{M}_n]$, where

$$\mathcal{M}_n := \begin{cases} \frac{\sum_{i\in\mathcal{I}_n} \omega_i^{(n)} \|P_{S_i[\epsilon_i]}(\boldsymbol{a}_n) - \boldsymbol{a}_n\|^2}{\|\sum_{i\in\mathcal{I}_n} \omega_i^{(n)} P_{S_i[\epsilon_i]}(\boldsymbol{a}_n) - \boldsymbol{a}_n\|^2}, \\ \qquad \text{if } \sum_{i\in\mathcal{I}_n} \omega_i^{(n)} P_{S_i[\epsilon_i]}(\boldsymbol{a}_n) \neq \boldsymbol{a}_n, \\ 1, \qquad \text{otherwise.} \end{cases} \quad (6a)$$

Notice that due to convexity of $\|\cdot\|^2$, $\mathcal{M}_n \geq 1$. In general, the larger the $\mu_n$, the larger the convergence speed of APGT.

(3) Finally, compute the next estimate by

$$\boldsymbol{a}_{n+1} := \begin{cases} T_{\mathrm{GT}}^{(K)}\left(\boldsymbol{a}_n + \mu_n\left(\sum_{i\in\mathcal{I}_n} \omega_i^{(n)} P_{S_i[\epsilon_i]}(\boldsymbol{a}_n) - \boldsymbol{a}_n\right)\right), \\ \qquad\qquad \text{if } \mathcal{I}_n \neq \emptyset, \\ T_{\mathrm{GT}}^{(K)}(\boldsymbol{a}_n), \qquad \text{if } \mathcal{I}_n = \emptyset. \end{cases} \quad (6b)$$



(b) GT with arbitrary shr (b) Piece-wise linear thresholds (c) Bridge, ($\ell_\gamma$, $\gamma < 1$) (d) NN and n-degree garrote (e) GT with PLSTO's as shr

**Fig. 1**. Illustration of PLSTOs for various choices of the regularizing function $p$ in (5), and some examples of GT.

Due to space limitations, the theoretical properties of APGT are deferred to a future work.

## 7. NUMERICAL EXAMPLES

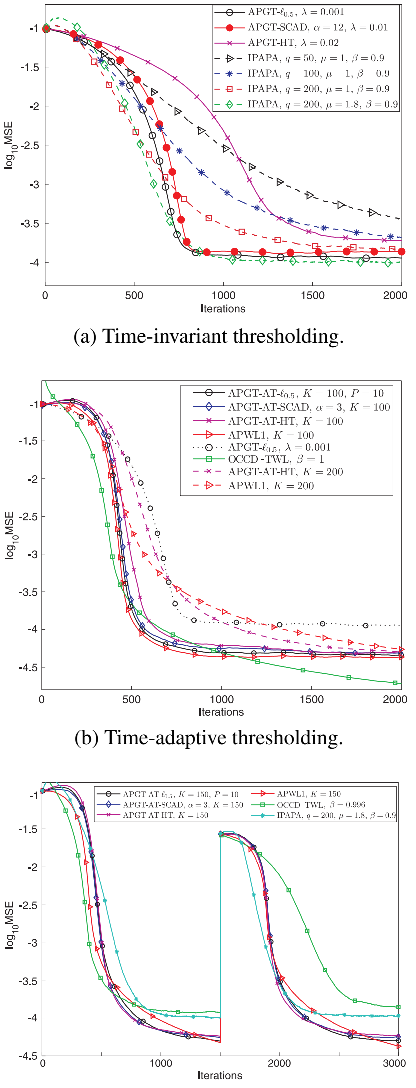
To assess the APGT performance, the HT, SCAD, and the $\ell_\gamma$-penalty ($\gamma < 1$) based thresholding rule, called here BT, are incorporated in GT, that is, HT, SCAD, and BT are used in the place of Shr. In the following experiments, $\mathcal{H} := \mathbb{R}^L$, with $L = 1024$. For Figs. 2a and 2b, $\Psi := \boldsymbol{I}_L$, and $\boldsymbol{a}_*$ is 100-sparse. The sensing vectors $\{\boldsymbol{u}_n\}_{n\in\mathbb{N}_*}$ have independent components drawn from $\mathcal{N}(0, 1)$, and the observations are corrupted by additive white Gaussian noise of variance $\sigma^2 = 0.1$. Regarding APGT, $\mu_n := \mathcal{M}_n$, and $\epsilon_n := 1.3\sigma$, $\forall n$. In this paper, convergence speed is of primal concern. To this end, $q := 390$ since this appeared to be the lowest $q$ value leading to enhanced convergence speed for the specific $L$ and $K$ values. It should be stressed, however, that APGT is not sensitive to $q$. An extensive and complementary experimental study of the APGT performance, for small values of $q$ can be found in [19].

The modifier "time-invariant" in Fig. 2a implies that $\lambda$ remains fixed $\forall n$. In all cases, $K := K_* := 100$. The parameter $\lambda$ was optimized leading to the values shown in the figure legend. Moreover, APGT-SCAD, without being considerably sensitive to parameter $\alpha$, appeared to perform best when $\alpha = 12$. For comparison, the improved proportionate adaptive projection algorithm (IPAPA) [20] is employed. The projection order $q$ of IPAPA is the parameter which dictates its performance. The step parameter of the IPAPA is denoted by $\mu$. The best IPAPA performance, i.e., the one depicted with a dashed curve with diamonds, is achieved with $q = 200$ and $\mu = 1.8$. For lower $q$ values, such a large $\mu$ led to unstable performance.

In Fig. 2a, the shape of the thresholding function was determined in advance using fixed values for the parameters $\lambda$, $\gamma$, $\alpha$, etc. This is quite limiting, since APGT has the potential to incorporate time-varying thresholding rules. For this reason, $\lambda$ changes with time $n$ in Fig. 2b. Assuming that an estimate $K$ of the true sparsity level $K_*$ is available at each $n$, $\lambda_n$ is properly tuned to guarantee that after thresholding, a fixed number of components will become zero. Details on how these strategies are determined are deferred to a future work. Moreover, regarding BT, apart from the $K$ larger in magnitude components which remain unaltered, the next, say $P$, smaller in magnitude components could be shrunk according to BT. The performance of APGT, using the time-adaptive thresholding strategies, hereafter abbreviated as APGT-AT, is shown in Fig. 2b. For reference, the dotted curve marked with open circles is the one from Fig. 2a, corresponding to the best APGT performance with fixed $\lambda$. The linear complexity APWL1 of [13] is also employed. For completeness, the online cyclic coordinate descent - time weighted Lasso (OCCD-TWL) [18], an RLS-type algorithm approximating the LASSO solution, is also depicted. It can be seen that APGT ($q = 390$) demonstrates a performance competitive to the $\mathcal{O}(L^2)$-complexity driven OCCD-TWL.

Fig. 2c shows the ability of the tested algorithms to track an abrupt change of the unknown vector $\boldsymbol{a}_*$, which is realized here after 1500 observations. In Fig. 2c, $\Psi$ is a wavelet basis. Prior to time 1500, the signal under consideration is of length $L = 1024$, with $K_* = 100$. However, at the 1500 time instant, ten randomly selected wavelet coefficients change their values from 0 to a randomly selected nonzero one. Since the sparsity level of the signal changes (from 100 to 110, at most) and it is not possible to know $K_*$ exactly in advance, taking into account also that the proposed methods are robust to $K_*$ over-estimations, $K$ is set to 150 throughout the whole experiment. Moreover, $q = 390$. In OCCD-TWL, an RLS-like forgetting factor lower than 1 is adopted to succeed in re-estimating the unknown signal after the abrupt change. More specifically, the value of 0.996 offers a good trade-off between convergence speed and steady-state error floor.

Regarding the computational complexities of the APGT-based methods, these are as follows: (1) APGT-AT-HT: (i) Multiplications: $(qe_1 + e_2 + 1)L + (K + e_1 + 1)q$, (ii) Divisions: $e_2 + 1$, (iii) Sortings: $\mathcal{O}(L)$, (2) APGT-AT-$\ell_{0.5}$: (i) Multiplications: $(qe_1+e_2+1)L+(K+P+e_1+1)q+12P+1$, (ii) Divisions: $P + e_2 + 2$, (iii) Sortings: $\mathcal{O}(L)$, (iv) Powers: $3P + 1$, (3) APGT-AT-SCAD: (i) Multiplications $(qe_1 + e_2 + 1)L+(L+e_1+1)q+(L-K)$, (ii) Divisions: $L-K+e_2+1$, and (iv) Sortings: $\mathcal{O}(L)$, where $e_1$ is either 1 or 2, depending



(a) Time-invariant thresholding.

(b) Time-adaptive thresholding.

(c) Robustness against time variations of the unknown system.

**Fig. 2**. (a) APGT with time-invariant thresholding rules. (b) APGT with time-adaptive thresholding rules. (c) The unknown vector has a sparse wavelet representation which changes abruptly after 1500 observations.

on whether all $\omega_i^{(n)}$ of the APGT attain the same value or not (here $e_1 = 1$), and $e_2$ is either 1, if the $\ell_2$ norm of the input vectors $(\boldsymbol{u}_n)_{n\in\mathbb{N}}$ is not fixed, or 0, if it is normalized to unity. Notice that IPAPA is a $\mathcal{O}(q^3) + (q^2 + 3q + 1)L + q$ multiplication-based strategy, while OCCD-TWL is $\mathcal{O}(L^2)$-complexity driven. Moreover, the complexity of APWL1 is (i) $(qe_1 + e_2 + 1)L + (L + e_1 + 1)q + 3L$ multiplications, (ii) $2L + e_2 + 1$ divisions, and (iv) $\mathcal{O}(L)$ sortings.

## 8. REFERENCES

[1] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied Comput. Harmonic Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[2] S. Foucart, "Hard thresholding pursuit: An algorithm for compressive sensing," *SIAM J. Numer. Anal.*, vol. 49, no. 6, pp. 2543–2563, 2011.

[3] A. Antoniadis, "Wavelet methods in statistics: Some recent developments and their applications," *Statist. Surveys*, vol. 1, pp. 16–55, 2007.

[4] Y. She, "Thresholding-based iterative selection procedures for model selection and shrinkage," *Electr. J. Statist.*, vol. 3, pp. 384–415, 2009.

[5] R. Mazumder, J. H. Friedman, and T. Hastie, "SPARSENET: Coordinate descent with nonconvex penalties," *J. Amer. Statist. Assoc.*, vol. 106, no. 495, pp. 1125–1138, Sept. 2011.

[6] I. E. Frank and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[7] H.-Y. Gao, "Wavelet shrinkage denoising using the non-negative garrote," *J. Comput. Graph. Statist.*, vol. 7, no. 4, pp. 469–488, Dec. 1998.

[8] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *J. Amer. Statist. Assoc.*, vol. 96, pp. 939–967, 2001.

[9] C.-H. Zhang, "Nearly unbiased variable selection under minimax concave penalty," *Annals Statist.*, vol. 38, no. 6, pp. 894–942, 2010.

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

[11] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, no. 3, pp. 337–404, 1950.

[12] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2275–2285, Aug. 2004.

[13] Y. Kopsinis, K. Slavakis, and S. Theodoridis, "Online sparse system identification and signal reconstruction using projections onto weighted $\ell_1$-balls," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 936–952, March 2011.

[14] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin, 2004.

[15] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer-Verlag, 2011.

[16] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. Royal. Statist. Soc. B.*, vol. 58, no. 1, pp. 267–288, 1996.

[17] H. Zou, "The adaptive LASSO and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, pp. 1418–1429, Dec. 2006.

[18] D. Angelosante, J. A. Bazerque, and G. B. Giannakis, "Online adaptive estimation of sparse signals: Where RLS meets the $\ell_1$-norm," *IEEE Trans. Signal Process.*, vol. 58, no. 7, pp. 3436–3447, July 2010.

[19] Y. Kopsinis, K. Slavakis, S. Theodoridis, and S. McLaughlin, "Thresholding-based online algorithms of complexity comparable to sparse LMS methods," in *Proc. ISCAS*, May 2013.

[20] C. Paleologu, S. Ciochina, and J. Benesty, "An efficient proportionate affine projection algorithm for echo cancellation," *IEEE Signal Process. Letters*, vol. 17, no. 2, pp. 165–168, 2010.