Appl. Comput. Harmon. Anal. 67 (2023) 101572

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha

Letter to the Editor

A constructive approach for computing the proximity operator of the p-th power of the ℓ_1 norm

Ashley Prater-Bennette[a], Lixin Shen[b,*], Erin E. Tripp[a]

[a] Air Force Research Laboratory, Rome, NY 13441, United States of America
[b] Department of Mathematics, Syracuse University, Syracuse, NY 13244, United States of America

ARTICLE INFO

Article history:
Received 1 October 2022
Received in revised form 17 June 2023
Accepted 22 June 2023
Available online 29 June 2023
Communicated by Radu Balan

Keywords:
Sparsity
Relative sparsity
Proximity operator

ABSTRACT

This note is to study the proximity operator of $h_p = \|\cdot\|_1^p$, the power function of the ℓ_1 norm. For general p, computing the proximity operator requires solving a system of potentially highly nonlinear inclusions. For $p = 1$, the proximity operator of h_1 is the well known soft-thresholding operator. For $p = 2$, the function h_2 serves as a penalty function that promotes structured solutions to optimization problems of interest; the computation of the proximity operator of h_2 has been discussed in recent literature. By examining the properties of the proximity operator of the power function of the ℓ_1 norm, we will develop a simple and well-justified approach to compute the proximity operator of h_p with $p > 1$. In particular, for the squared ℓ_1 norm function, our approach provides an alternative, yet explicit way to finding its proximity operator. We also discuss how the structure of h_p represents a class of relative sparsity promoting functions.

1. Introduction

The power-p ($p \geq 1$) function of the ℓ_1 norm on $\mathbb{R}^n$ is

$$h_p : \mathbb{R}^n \to \mathbb{R} : x \mapsto \|x\|_1^p. \tag{1}$$

Its *proximity operator* (or *proximal mapping*) at $x \in \mathbb{R}^n$ with index $\beta > 0$ is defined as

$$\operatorname{prox}_{\beta h_p}(x) = \operatorname{argmin}\left\{\frac{1}{2}\|u - x\|_2^2 + \beta h_p(u) : u \in \mathbb{R}^n\right\}. \tag{2}$$

* Corresponding author.
E-mail addresses: ashley.prater-bennette@us.af.mil (A. Prater-Bennette), lshen03@syr.edu (L. Shen), erin.tripp.4@us.af.mil (E.E. Tripp).

https://doi.org/10.1016/j.acha.2023.06.007
1063-5203/

For $p = 1$, the function h_1 is simply the ℓ_1 norm whose proximity operator is the well known soft-thresholding operator, which has been extensively used in sparse optimization in the last two decades [3]. For $p = 2$, the computation of $\mathrm{prox}_{\beta h_2}$ is the main computational component in regularized robust portfolio estimation [4]. Unlike the proximity operator of h_1, the proximity operator of h_2 given in [4] does not have a closed form since it requires locating the positive root of a nonlinear function through the bisection method. The key step in the derivation of $\mathrm{prox}_{\beta h_2}$ in [4] is to view h_2 as the optimal value of an optimization problem over the unit simplex set on $\mathbb{R}^n$. Using the technique of Lagrange multipliers, a procedure was presented for the computation of $\mathrm{prox}_{\beta h_2}$ in [6]. It is not obvious how this idea can be adapted for dealing with h_p for $p \neq 2$.

The focus of this paper is the computation and understanding of the proximity operator of h_p for $p > 1$. We propose a method for computing $\mathrm{prox}_{\beta h_p}$ that essentially relies on finding the unique positive root of the triterm function

$$f(r) = r^{p-1} + ar - b,$$

where both a and b are positive. Correspondingly, $\mathrm{prox}_{\beta h_p}(x)$ for any $x \in \mathbb{R}^n$ has an explicit expression as long as the root of the above function has an explicit expression. Fortunately, the positive root of this triterm function has such a form at least for $p \in \{2, 3, 4\}$. In particular, for $p = 2$ our result clearly improves the one in [4] and is the same as the one in [6] for this case, although arrived at by a different process. If an explicit expression of the positive root of this triterm function is not available, we can efficiently find it by either bisection or Newton's method. However, we emphasize that our proposed method for computing $\mathrm{prox}_{\beta h_p}$ works for all $p > 1$.

The ℓ_1 norm, that is h_1, is a typical sparsity promoting function (SPF). Loosely speaking, a function is an SPF if its subdifferential at the origin contains at least one nonzero element; that is, an SPF has a corner or cusp at the origin [8,9]. Viewed in another way, the subdifferential of the function at the origin is a set that defines a threshold for small elements which are considered noise or insignificant. The proximity operator of the SPF will send all elements under this threshold to the origin. According to the above description, the function h_p with $p > 1$ is not an SPF since it is differentiable at the origin. As a result, the proximity operator of h_p maps a nonzero vector to another nonzero vector. However, we see that h_p promotes a type of *relative* sparsity. In particular, rather than a uniform threshold sending elements to zero, the threshold depends on the relationship between the components.

The remaining part of this note is organized as follows. In Section 2, we give a brief review on the proximity operators of h_1 and h_2 to motivate the computation of the proximity operator of h_p with $p > 1$. In Section 3, we present the properties of the proximity operator of h_p with $p > 1$. In particular, based on these properties, we propose an iterative scheme to compute the proximity operator of h_p for vectors with ordered nonnegative elements. A general scheme for computing proximity operator of h_p is given in Section 4. We visually present the relative sparsity promoted by the function h_p for $p > 1$ in Section 5.

2. Brief review on the proximity operators of h_1 and h_2

The proximity operator of h_1 is the well known soft-thresholding operator given as follows:

$$\mathrm{prox}_{\beta h_1}(x) = (\mathrm{sgn}(x_i)(|x_i| - \beta)_+)_{i=1}^n . \tag{3}$$

We denote by $(a)_+$ the hinge function, namely, $(a)_+ = \max\{0, a\}$, and sgn is the signum function which is defined at $a \in \mathbb{R}$ as

$$\mathrm{sgn}(a) = \begin{cases} -1, & \text{if } a < 0; \\ 0, & \text{if } a = 0; \\ 1, & \text{if } a > 1. \end{cases}$$

The formula (3) shows that the operator $\mathrm{prox}_{\beta h_1}$ maps all vectors in the n-dimensional cube with sidelength β to the origin and shrinks every element of a vector outside this cube towards the origin by β.

The computation of $\mathrm{prox}_{\beta h_2}$ was studied for regularized robust portfolio estimation in [4]. Here we present a similar derivation of $\mathrm{prox}_{\beta h_2}$ from the book [2, Lemma 6.70]. The key step in computing $\mathrm{prox}_{\beta h_2}$ is to express $h_2(x) = \|x\|_1^2$ as the optimal value of an optimization problem as follows

$$\|x\|_1^2 = \min_{\lambda \in \Delta_n} \sum_{j=1}^{n} \varphi(x_j, \lambda_j), \tag{4}$$

where

$$\varphi(s,t) = \begin{cases} \frac{s^2}{t}, & t > 0; \\ 0, & s = 0,\ t = 0; \\ \infty, & \text{else} \end{cases} \tag{5}$$

and Δ_n is the unit simplex set, that is, the convex hull of the standard basis of $\mathbb{R}^n$. With (4), for $x \neq 0$, $u = \mathrm{prox}_{\beta h_2}(x)$ is the u-part solution of the optimal solution of

$$\min_{u \in \mathbb{R}^n, \lambda \in \Delta_n} \left\{ \frac{1}{2}\|u - x\|_2^2 + \beta \sum_{i=1}^{n} \varphi(u_i, \lambda_i) \right\}.$$

Fixing λ and minimizing over u for the above optimization problem we obtain that $u_i = \frac{\lambda_i x_i}{\lambda_i + 2\beta}$, and substituting the resulting vector u into the above problem leads to

$$\min_{\lambda \in \Delta_n} \sum_{i=1}^{n} \frac{\beta x_i^2}{\lambda_i + 2\beta}.$$

This constrained optimization problem can be solved through the associated Lagrangian multipliers. Putting the above together, we obtain that

$$\mathrm{prox}_{\beta h_2}(x) = \begin{cases} \left(\frac{\lambda_i x_i}{\lambda_i + 2\beta} \right)_{i=1}^{n}, & x \neq 0; \\ 0, & x = 0, \end{cases} \tag{6}$$

where $\lambda_i = \left(\frac{\sqrt{\beta}|x_i|}{\sqrt{\mu^*}} - 2\beta \right)_+$ with μ^* being any positive root of the nonincreasing function

$$\psi(\mu) = \sum_{i=1}^{n} \left(\frac{\sqrt{\beta}|x_i|}{\sqrt{\mu}} - 2\beta \right)_+ - 1. \tag{7}$$

The root μ^* is found by the bisection search. To explore (6) in more detail, let us write $u = \mathrm{prox}_{\beta h_2}(x)$. If $\lambda_i > 0$, then $\lambda_i = \frac{\sqrt{\beta}|x_i|}{\sqrt{\mu^*}} - 2\beta$ and $u_i = x_i - 2\sqrt{\beta\mu^*}\,\mathrm{sgn}(x_i)$; all nonzero elements u_i thus are shrinkage of x_i towards the origin by $2\sqrt{\beta\mu^*}$. The value of $2\sqrt{\beta\mu^*}$ does not have an explicit expression in terms of β and x since μ^* as a root of ψ in (7) is obtained from the bisection method.

A different way of exploiting Lagrangian multipliers for the computation of $\mathrm{prox}_{\beta h_2}$ was proposed in [6] and will be mentioned in section 4. In the rest of the paper, we will propose a unified procedure to compute $\mathrm{prox}_{\beta h_p}$ for all $p > 1$.

3. The properties of the proximity operator of h_p

All functions in this work are defined on Euclidean space $\mathbb{R}^n$ equipped with the standard inner product $\langle\cdot,\cdot\rangle$ and the induced Euclidean norm $\|\cdot\|_2$. For a function $g:\mathbb{R}^n\to\mathbb{R}\cup\{\infty\}$, the set $\mathrm{dom}(g)=\{x:g(x)<\infty\}$ is the domain of g. We use $\Gamma_0(\mathbb{R}^n)$ to represent the set of proper lower semicontinuous convex functions on $\mathbb{R}^n$. We denote the cone of vectors x in $\mathbb{R}^n$ satisfying $x_1\geq x_2\geq\ldots\geq x_n\geq 0$ by $\mathbb{R}^n_\downarrow$.

For any $g\in\Gamma_0(\mathbb{R}^n)$, the subdifferential of g at $x\in\mathrm{dom}(g)$ is the set

$$\partial g(x)=\{d\in\mathbb{R}^n: g(y)\geq g(x)+\langle d,y-x\rangle,\ \forall y\in\mathbb{R}^n\}.$$

Furthermore, if g is Fréchet differentiable, $\partial g(x)=\{\nabla g(x)\}$. The relationship between the subdifferential and proximity operator of a function $f\in\Gamma_0(\mathbb{R}^n)$ is characterized as follows (see, e.g., [1,7]):

$$\text{for any } \beta>0,\ x\in\beta\partial f(y) \text{ if and only if } y=\mathrm{prox}_{\beta f}(x+y).$$

From this relationship, we get the following characterization on the proximity operator of h_p.

Lemma 1 *(Chain rule and shrinkage). For the power-p function of the ℓ_1 norm h_p given in (1) with $p>1$, the following statements hold.*

(i) The function h_p is convex on $\mathbb{R}^n$ and $\partial h_p(x)=p\|x\|_1^{p-1}\partial\|\cdot\|_1(x)$ for all $x\in\mathbb{R}^n$.
(ii) $\|\mathrm{prox}_{\beta h_p}(x)\|\leq\|x\|$ for all $x\in\mathbb{R}^n$ and $\beta>0$.

Proof. (i) Note that $h_p=(\cdot)_+^p\circ\|\cdot\|_1$, where $(\cdot)_+^p$ is a nondecreasing and convex function, and the ℓ_1 norm $\|\cdot\|_1$ is convex. Hence, h_p is convex. It follows from [5, Theorem 4.3.1] that $\partial h_p(x)=p\|x\|_1^{p-1}\partial\|\cdot\|_1(x)$.

(ii) From item (i), we have $\partial h_p(0)=\{0\}$. Hence $\mathrm{prox}_{\beta h_p}(0)=0$. It leads to $\|\mathrm{prox}_{\beta h_p}(x)\|=\|\mathrm{prox}_{\beta h_p}(x)-\mathrm{prox}_{\beta h_p}(0)\|\leq\|x\|$ due to the nonexpansiveness of $\mathrm{prox}_{\beta h_p}$. □

Item (i) of Lemma 1 implies that h_p with $p>1$ is differentiable at the origin while it is not differentiable at nonzero vectors having at least one zero element. Item (ii) of Lemma 1 says that $\mathrm{prox}_{\beta h_p}$ is a shrinkage operator.

Now let $P_{(-)}(n)$ denote the set of all $n\times n$ signed permutation matrices: those matrices that have only one nonzero element in every row or column, which is ±1. Since $h_p(x)=h_p(P_{(-)}x)$ for all $P_{(-)}$ in $P_{(-)}(n)$ and $x\in\mathbb{R}^n$, we immediately have that

$$\mathrm{prox}_{\beta h_p}(x)=P_{(-)}^{-1}\,\mathrm{prox}_{\beta h_p}(P_{(-)}x). \tag{8}$$

For every $x\in\mathbb{R}^n$, there exists a signed permutation matrix $P_{(-)}$ such that $P_{(-)}x\in\mathbb{R}^n_\downarrow$. Thus, with the identity (8), we should focus on the computation of the operator $\mathrm{prox}_{\beta h_p}$ over the set $\mathbb{R}^n_\downarrow$.

Lemma 2 *(Order preservation and nonzero elements). Let x be a nonzero vector in $\mathbb{R}^n_\downarrow$.*

(i) For any $p\geq1$ and $\beta>0$, $\mathrm{prox}_{\beta h_p}(x)$ is also in $\mathbb{R}^n_\downarrow$.
(ii) $\mathrm{prox}_{\beta h_p}(x)$ is a nonzero vector in $\mathbb{R}^n_\downarrow$ for any $p>1$ and $\beta>0$.

Proof. (i) Recall that $\mathrm{prox}_{\beta h_p}(x)=\mathrm{argmin}\left\{\frac{1}{2}\|u-x\|_2^2+\beta\|u\|_1^p: u\in\mathbb{R}^n\right\}$. For any given $u\in\mathbb{R}^n$, we have $\|u\|_1^p=\|P_{(-)}u\|_1^p$ for any signed permutation matrix $P_{(-)}$ in $P_{(-)}(n)$. Furthermore, there exists a signed permutation matrix $Q\in P_{(-)}(n)$ such that $Qu\in\mathbb{R}^n_\downarrow$. Since $x\in\mathbb{R}^n_\downarrow$, we know that $\|Qu-x\|^2\leq\|P_{(-)}u-x\|^2$, hence, $\|Qu-x\|^2+\beta\|Qu\|_1^p\leq\|P_{(-)}u-x\|^2+\beta\|P_{(-)}u\|_1^p$ for all $P_{(-)}\in P_{(-)}(n)$. We conclude from the above discussion that $\mathrm{prox}_{\beta h_p}(x)\in\mathbb{R}^n_\downarrow$.

(ii) Let us write $u^\star = \mathrm{prox}_{\beta h_p}(x)$ for the nonzero vector x in $\mathbb{R}^n_\downarrow$. Then, by item (i) of Lemma 2, $u^\star \in \mathbb{R}^n_\downarrow$. To show $u^\star$ is a nonzero vector, it is sufficient to show $u_1^\star > 0$. We prove it by contradiction.

Suppose that $u_1^\star = 0$, that is, $u^\star$ is a zero vector. Define $g(u) = \frac{1}{2}\|u-x\|_2^2 + \beta\|u\|_1^p$. By the definition of proximity operator, we have that

$$\frac{1}{2}\|x\|_2^2 = g(u^\star) \le g(u)$$

holds for all $u \in \mathbb{R}^n_\downarrow$. In particular, plugging $u = (a, 0, \ldots, 0)^\intercal \in \mathbb{R}^n_\downarrow$ with $a \ge 0$ into the above inequality yields $\frac{1}{2}x_1^2 \le \frac{1}{2}(a - x_1)^2 + \beta a^p$, or equivalently,

$$\frac{1}{2}a^2 - x_1 a + \beta a^p \ge 0 \tag{9}$$

for all $a \ge 0$. Write $f(a) = \frac{1}{2}a^2 - x_1 a + \beta a^p$. Since $f(0) = 0$ and $f'(0) = -x_1 < 0$, we conclude that equation (9) does not hold for all $a \ge 0$, therefore, contradicting the assumption. Hence, $u^\star$ must be a nonzero vector. □

Recall that a function $f \in \Gamma_0(\mathbb{R}^n)$ is said to be an SPF provided that (i) $f(0) = 0$ and f achieves its global minimum at the origin; and (ii) the set $\partial f(0)$ contains at least one nonzero element. More discussions on this concept can be found in [8]. Moreover, SPFs are characterized by the thresholding behavior of their proximity operators, which item (ii) of Lemma 2 seems to contradict. Clearly, h_1 is an SPF while h_p with $p > 1$ is not due to $\partial h_p(0) = \{0\}$ by Lemma 1.

To illustrate the behavior of h_p, consider $x \in \mathbb{R}^2$. As noted before, h_p is differentiable at the origin with $\nabla h_p(0) = 0$. However, when one component is zero, say $x_2 = 0$, the subdifferential is as follows

$$\partial h_p(x_1, 0) = \{p|x_1|^{p-1} \cdot (\mathrm{sgn}(x_1), \eta)^\intercal : \eta \in [-1, 1]\}.$$

With this, we can consider the differential inclusion defined by the proximity operator; i.e. if $(u, v) \in \mathrm{prox}_{\beta h_p}(x_1, x_2)$, then

$$0 \in (u - x_1, v - x_2)^\intercal + \beta\partial h_p(u, v). \tag{10}$$

By Lemma 1, we know that $\mathrm{prox}_{\beta h_p}(x) \ne 0$, but it may still be more sparse than x. For example, prox will send the second component to zero if

$$\begin{aligned} x_1 &= u + \beta p\,\mathrm{sgn}(u)|u|^{p-1} \\ -\beta p|u|^{p-1} &\le x_2 \le \beta p|u|^{p-1} \end{aligned} \tag{11}$$

This highlights the dependence of the thresholding on the relationship between x_1 and x_2. We refer to this as relative sparsity, and defer further discussion to Section 5.

The next result will give a characterization on the number of nonzero elements produced by the proximity operator of h_p. To this end, for a given nonzero vector $x \in \mathbb{R}^n_\downarrow$, scalars $\beta > 0$ and $p > 1$, and an integer $1 \le m \le n$, we define a triterm function $g_{(m,\beta,p)} : [0, \infty) \to \mathbb{R}$ as follows:

$$g_{(m,\beta,p)}(r) := m\beta p r^{p-1} + r - \sum_{i=1}^m x_i. \tag{12}$$

Since $g_{(m,\beta,p)}(0) = -\sum_{i=1}^m x_i < 0$, $g_{(m,\beta,p)}(\|x\|_1) > 0$, and $g'_{(m,\beta,p)}(r) = m\beta p(p-1)r^{p-2} + 1 > 0$ for all $r \ge 0$, $g_{(m,\beta,p)}$ has one and only one positive root located in the interval $(0, \|x\|_1)$. We denote by $g^{-1}_{(m,\beta,p)}(0)$ the positive root of $g_{(m,\beta,p)}$.

Lemma 3 *(Counting nonzero elements). Let x be a nonzero vector in $\mathbb{R}^n_\downarrow$ and let $\beta > 0$. The following statements are equivalent.*

(i) The proximity operator of h_p with index β at x has m nonzero elements.
(ii) $x_m > \beta p r^{p-1}$ and $x_{m+1} \le \beta p r^{p-1}$ if $1 \le m \le n-1$ or $x_m > \beta p r^{p-1}$ if $m = n$, where r is the only positive root of $g_{(m,\beta,p)}$ defined in (12).

Proof. (i)⇒(ii) Write $u = \operatorname{prox}_{\beta h_p}(x)$. From x being a nonzero vector and item (ii) of Lemma 2, we have $m \ge 1$, and the first m elements of u are positive and the remaining $u_i = 0$ for $i = m+1, \ldots, n$. Since $u = \operatorname{prox}_{\beta h_p}(x)$, we know $x - u \in \partial(\beta h_p)(u)$, that is, $x - u \in \beta p \|u\|_1^{p-1} \partial \|\cdot\|_1(u)$ by Lemma 1. Therefore, we obtain the following

$$x_i - u_i = \beta p \|u\|_1^{p-1} \tag{13}$$

for $i = 1, 2, \ldots, m$. Summing (13) over $i = 1, 2, \ldots, m$ yields

$$\sum_{i=1}^{m} x_i - \|u\|_1 = m\beta p \|u\|_1^{p-1},$$

which implies $g_{(m,\beta,p)}(\|u\|_1) = 0$. Hence $r = \|u\|_1$ is the only positive root of $g_{(m,\beta,p)}$. From (13), we have that $u_i = x_i - \beta p r^{p-1} > 0$ for $i = 1, 2, \ldots, m$.

From the inclusion $x - u \in \beta p \|u\|_1^{p-1} \partial \|\cdot\|_1(u)$, due to $u_i = 0$, we have $x_i \le \beta p r^{p-1}$ for $i = m+1, \ldots, n$. Hence, item (ii) holds.

(ii)⇒(i) We construct a vector $u \in \mathbb{R}^n$ as follows: $u_i = x_i - \beta p r^{p-1}$ for $i = 1, 2, \ldots, m$ and $u_i = 0$ for $i = m+1, \ldots, n$ if $1 \le m \le n-1$, and $u_i = x_i - \beta p r^{p-1}$ for $i = 1, 2, \ldots, n$ if $m = n$. By the assumption, $u \in \mathbb{R}^n_\downarrow$ with m nonzero elements. Summing the first m elements of u yields $\|u\|_1 = \sum_{i=1}^{m} x_i - m\beta p r^{p-1}$. We conclude $r = \|u\|_1$. Hence $u_i = x_i - \beta p \|u\|_1^{p-1}$ for $i = 1, 2, \ldots, m$ and $x_i \le \beta p \|u\|_1^{p-1}$ for $i = m+1, \ldots, n$ if $1 \le m \le n-1$, and $u_i = x_i - \beta p \|u\|_1^{p-1}$ for $i = 1, 2, \ldots, n$ if $m = n$. This implies $x - u \in \beta p \|u\|_1^{p-1} \partial \|\cdot\|_1(u)$. By Fermat's rule, $u = \operatorname{prox}_{\beta h_p}(x)$, which has exactly m nonzero elements. Hence, item (i) holds. □

A direct consequence of Lemma 3 is that for any nonzero vector $x \in \mathbb{R}^n_\downarrow$ there is a unique integer m such that

$$u_i = \begin{cases} x_i - \beta p r^{p-1}, & \text{if } i = 1, 2, \ldots, m; \\ 0, & \text{if } i = m+1, \ldots, n, \end{cases} \tag{14}$$

where r is the only positive root of $g_{(m,\beta,p)}$ defined in (12).

It is clear from (14) that one needs to determine both the number of nonzero elements in $\operatorname{prox}_{\beta h_p}(x)$ and the positive root of the function $g_{(m,\beta,p)}$. A complete scheme for computing the proximity of h_p at arbitrary point in $\mathbb{R}^n$ and the discussion on computing the positive root of the function $g_{(m,\beta,p)}$ will be postponed in section 4. We provide an iterative scheme that counts one nonzero element in $\operatorname{prox}_{\beta h_p}(x)$ at a time in the rest of this section.

Given a nonzero vector x in $\mathbb{R}^n_\downarrow$, $\beta > 0$, and $p > 1$, for an integer m between 1 and $n-1$, let r and $\hat{r}$ be the only positive root of $g_{(m,\beta,p)}$ and $g_{(m+1,\beta,p)}$, respectively. From (12), the difference of $g_{(m,\beta,p)}(r) = 0$ and $g_{(m+1,\beta,p)}(\hat{r}) = 0$ is

$$m\beta p(\hat{r}^{p-1} - r^{p-1}) + (\hat{r} - r) + (\beta p \hat{r}^{p-1} - x_{m+1}) = 0. \tag{15}$$

We conclude from (15) that $x_{m+1} \leq \beta p \hat{r}^{p-1}$ if and only if $\hat{r} \leq r$, and $x_{m+1} > \beta p \hat{r}^{p-1}$ if and only if $\hat{r} > r$. With the above preparation, an iterative scheme to determine the number of nonzero elements in $\mathrm{prox}_{\beta h_p}(x)$ for a nonzero vector x in $\mathbb{R}^n_{\downarrow}$ is described as follows. We begin with $m = 1$ since $\mathrm{prox}_{\beta h_p}(x)$ has at least one nonzero element by Lemma 2. If $x_2 \leq \beta p r^{p-1}$, then $x_1 - \beta p r^{p-1}$, the first element of $\mathrm{prox}_{\beta h_p}(x)$, is the only nonzero element of $\mathrm{prox}_{\beta h_p}(x)$ by Lemma 3; If $x_2 > \beta p r^{p-1}$, $x_1 > \beta p r^{p-1}$ holds automatically, hence, the number of nonzero elements in $\mathrm{prox}_{\beta h_p}(x)$ is at least two. Moreover, $\beta p \hat{r}^{p-1}$ is closer to x_i than $\beta p r^{p-1}$ for $i = 1, 2$ due to $\hat{r} > r$. If $x_3 \leq \beta p \hat{r}^{p-1}$, the number of nonzero elements in $\mathrm{prox}_{\beta h_p}(x)$ is exactly two; otherwise, we should update $m = 2$ and repeat the previous procedure until the nonzero elements in $\mathrm{prox}_{\beta h_p}(x)$ are identified.

4. Computing the proximity operator of h_p

In this section, we will present a complete scheme for computing the proximity operator of h_p. First, an iterative scheme given in Algorithm 1 is used to compute the proximity operator of h_p at a nonzero vector $x \in \mathbb{R}^n_{\downarrow}$. The main idea behind Algorithm 2, the general algorithm, is equation (8), which converts the proximity operator of h_p at an arbitrary point as one at a point in $\mathbb{R}^n_{\downarrow}$ so that the latter can be handled by Algorithm 1. In this way, the proximity operator of h_p at an arbitrary point as a highly nonlinear process is viewed as the composition of relatively simpler nonlinear processes. The detailed description of the scheme is outlined in Algorithm 2.

Algorithm 1: Computing $\mathrm{prox}_{\beta h_p}(x)$ for a nonzero vector $x \in \mathbb{R}^n_{\downarrow}$.

Input: $p > 1$, $\beta > 0$, and a nonzero vector $x \in \mathbb{R}^n_{\downarrow}$;

1.1 **Initialization:** $m = 1$ and r is the positive root of $g_{(1,\beta,p)}$;

1.2 **while** $m < n$ **do**

1.3   **if** $x_{m+1} \leq \beta p r^{p-1}$ **then**

1.4
$$u_i = \begin{cases} x_i - \beta p r^{p-1}, & \text{for } i = 1, \ldots, m; \\ 0, & \text{for } i = m+1, \ldots, n. \end{cases}$$
        break ;

1.5   **else**

1.6   Update $m \leftarrow m + 1$ and $r \leftarrow g^{-1}_{(m,\beta,p)}(0)$, the positive root of $g_{(m,\beta,p)}$;

Output: $\mathrm{prox}_{\beta h_p}(x) \leftarrow u$.

Line 1 of Algorithm 1 is ensured by Lemma 2 while the "while-loop" from line 2 to line 6 is due to Lemma 3 and its following discussion. The only remaining issue is finding the positive root of $g_{(m,\beta,p)}$. Fortunately, an explicit expression for the positive root of $g_{(m,\beta,p)}$ exists at least for $p \in \{2, 3, 4\}$. In more details, we have the following.

- For $p = 2$, we have $g_{(m,\beta,2)}(r) = (2m\beta + 1)r - \sum_{i=1}^{m} x_i$, whose positive root is

$$g^{-1}_{(m,\beta,2)}(0) = \frac{\sum_{i=1}^{m} x_i}{2m\beta + 1}. \tag{16}$$

  In this case, Algorithm 1 is the same as the one given in [6]. We point it out that our approach is developed differently from the one in [6].

- For $p = 3$, we have $g_{(m,\beta,3)}(r) = 3m\beta r^2 + r - \sum_{i=1}^m x_i$, whose positive root is

$$g^{-1}_{(m,\beta,3)}(0) = \frac{1}{6m\beta}\left(-1 + \sqrt{1 + 12m\beta \sum_{i=1}^m x_i}\right). \tag{17}$$

- For $p = 4$, we have $g_{(m,\beta,4)}(r) = 4m\beta r^3 + r - \sum_{i=1}^m x_i$, whose positive root is

$$g^{-1}_{(m,\beta,4)}(0) = \sqrt[3]{c + \sqrt{c^2 + b^3}} + \sqrt[3]{c - \sqrt{c^2 + b^3}}, \tag{18}$$

  where $b = \frac{1}{12m\beta}$ and $c = \frac{1}{8m\beta}\sum_{i=1}^m x_i$.
- For $p \in (1, 2)$, since $g_{(m,\beta,p)}(0) < 0$ and $g_{(m,\beta,p)}(\sum_{i=1}^m x_i) > 0$ we can use the bisection method to locate the only root of $g_{(m,\beta,p)}$ in the interval $(0, \sum_{i=1}^m x_i)$.
- $p \in (2, \infty)$, in addition to $g_{(m,\beta,p)}(0) < 0$ and $g_{(m,\beta,p)}(\sum_{i=1}^m x_i) > 0$, since $g_{(m,\beta,p)}$ is at least 2-differentiable on $[0, \infty)$ and convex on $[0, \infty)$, Newton's method, starting with $\sum_{i=1}^m x_i$, can generate a decreasing sequence that converges to the root of $g_{(m,\beta,p)}$ in the interval $(0, \sum_{i=1}^m x_i)$. We comment that the bisection method can be used, but Newton's method is preferred due to its faster convergence rate.

Line 2 of Algorithm 2 is the proximity operator of h_p with $p = 1$, which is the well known soft-thresholding operator. Line 3 is due to item (ii) of Lemma 1. The core part of the algorithm is in lines 4 to 7 which implement equation (8) regarding the computation of $\mathrm{prox}_{\beta h_p}(x)$ for $p > 1$ and any nonzero vector x in $\mathbb{R}^n$. The vector $\tilde{x}$ in line 5 is the sorted absolute values of elements of x and is obtained through a signed permutation matrix $P_{(-)} \in \mathcal{P}_{(-)}(n)$. It is followed by computing $\mathrm{prox}_{\beta h_p}(\tilde{x})$ via Algorithm 1 in line 6. Finally, line 7 follows directly from equation (8).

Algorithm 2: Computing $\mathrm{prox}_{\beta h_p}(x)$ for a vector $x \in \mathbb{R}^n$.

Input: $p \geq 1$, $\beta > 0$, and $x \in \mathbb{R}^n$;
2.1 **case** $p = 1$ **do**
2.2 | $\mathrm{prox}_{\beta h_p}(x) = (\mathrm{sgn}(x_i)(|x_i| - \beta)_+)_{i=1}^n$;
2.3 **case** $p > 1$ *and* $x = 0$ **do** $\mathrm{prox}_{\beta h_p}(x) = 0$;
2.4 **case** $p > 1$ *and* $x \neq 0$ **do**
2.5 | Find a signed permutation matrix $P_{(-)} \in \mathcal{P}_{(-)}(n)$ such that $\tilde{x} = P_{(-)}x \in \mathbb{R}^n_{\downarrow}$;
2.6 | Compute $\mathrm{prox}_{\beta h_p}(\tilde{x})$ via Algorithm 1;
2.7 | Compute $\mathrm{prox}_{\beta h_p}(x) = P_{(-)}^{-1}\,\mathrm{prox}_{\beta h_p}(\tilde{x})$.
Output: $\mathrm{prox}_{\beta h_p}(x)$;

For $p = 2$, Algorithm 2 gives a constructive way to accurately compute $\mathrm{prox}_{\beta h_2}$, thus improving the existing result in [4] in the sense that the bisection method is not required for evaluating $\mathrm{prox}_{\beta h_2}$ at a point.

5. Relative sparsity

In this section, we visually present the relative sparsity promoted by the function h_p for $p > 1$. As mentioned in the introduction, the function h_1, i.e., the ℓ_1 norm, is a sparsity promoting function. Its proximity operator $\mathrm{prox}_{\beta h_1}$ sends the element of a vector to zero if its absolute value is smaller than the threshold β; otherwise, shrinks this element toward zero by β. Similarly, the function h_p with $p > 1$ promotes a type of *relative* sparsity in the sense that its proximity operator maps a nonzero vector to another nonzero vector and sends some elements of the vector to zero according to a threshold depending on the relationship between the elements. Next, we discuss the proximity operator $\mathrm{prox}_{\beta h_p}$ for $p \in \{1, 2, 3\}$ in $\mathbb{R}^2$.

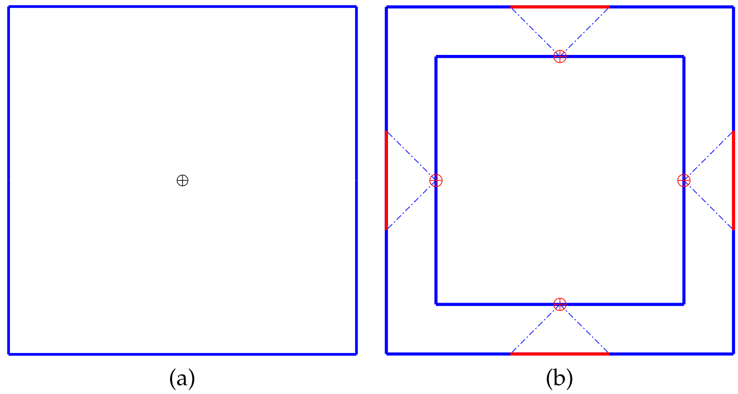

Fig. 5.1. Thresholding and shrinkage of $\text{prox}_{\beta h_1}$. The side length of the square is (a) 2β and (b) 7β (outer square). All of the blue segments in (a) map to the origin, marked by $\oplus$, and the red segments in (b) map to the red $\oplus$ while the blue segments shift to the corresponding blue segments. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Fig. 5.1 shows the properties of $\text{prox}_{\beta h_1}$ in $\mathbb{R}^2$. The side lengths of the squares centered at the origin in Fig. 5.1(a) and (b) are 2β and 7β, respectively. Since the threshold is β, the operator $\text{prox}_{\beta h_1}$ maps all points on the boundary of the square in Fig. 5.1(a) to the origin marked by $\oplus$. Fig. 5.1(b) has two concentric squares with side lengths of 5β and 7β. There are four points marked by the symbol $\oplus$ in the middle of each side of the smaller square while there are four red line segments with length 2β lying in the middle of each side of the bigger square. The operator $\text{prox}_{\beta h_1}$ maps all points in the red line to the closest point marked by $\oplus$. The other points in blue lines of the larger square will be moved toward the origin by the amount β for each coordinate, falling on the smaller square.

For $p > 1$, we call $\mathbb{R}^n_{\downarrow}$ the primitive set for $\text{prox}_{\beta h_p}$ in the sense that

$$\mathbb{R}^n = \cup_{P_{(-)} \in P_{(-)}(n)} P_{(-)} \mathbb{R}^n_{\downarrow}.$$

Therefore, we focus on investigating the operator $\text{prox}_{\beta h_p}$ on the primitive set $\mathbb{R}^n_{\downarrow}$. For any nonzero vector in $\mathbb{R}^n_{\downarrow}$, depending on the relative values of its elements, the number of nonzero elements in the resulting vector from $\text{prox}_{\beta h_p}$ can be $1, 2, \ldots, n-1$, or n by Algorithm 1.

Fig. 5.2 shows the properties of the operator $\text{prox}_{\beta h_2}$ in $\mathbb{R}^2$ through observing where the points on the squares are mapped to. The side lengths of the squares centered at the origin in Fig. 5.2(a) and (b) are 2β and 7β with $\beta = \frac{1}{2}$, respectively. The side lengths of the squares centered at the origin in Fig. 5.2(c) and (d) are 2β and 7β with $\beta = 2$, respectively. We partition vectors in $\mathbb{R}^2_{\downarrow}$ as follows:

$$\{x = (x_1, x_2)^\intercal : x \in \mathbb{R}^2_{\downarrow}\} = I_1 \cup I_2$$

with

$$I_1 = \left\{ x = (x_1, x_2)^\intercal \in \mathbb{R}^2_{\downarrow} : x_2 \leq \frac{2\beta}{2\beta + 1} x_1 \right\}$$

and

$$I_2 = \left\{ x = (x_1, x_2)^\intercal \in \mathbb{R}^2_{\downarrow} : x_2 > \frac{2\beta}{2\beta + 1} x_1 \right\}.$$

For a fixed value of x_1, the red lines on the right and left of each outside square in Fig. 5.2 are formed by I_1 and its reflection across the x_2-axis, and the other red lines are the rotated versions of these red lines. By simply replacing I_1 by I_2, the blue lines are formed. With Lemma 3, we get

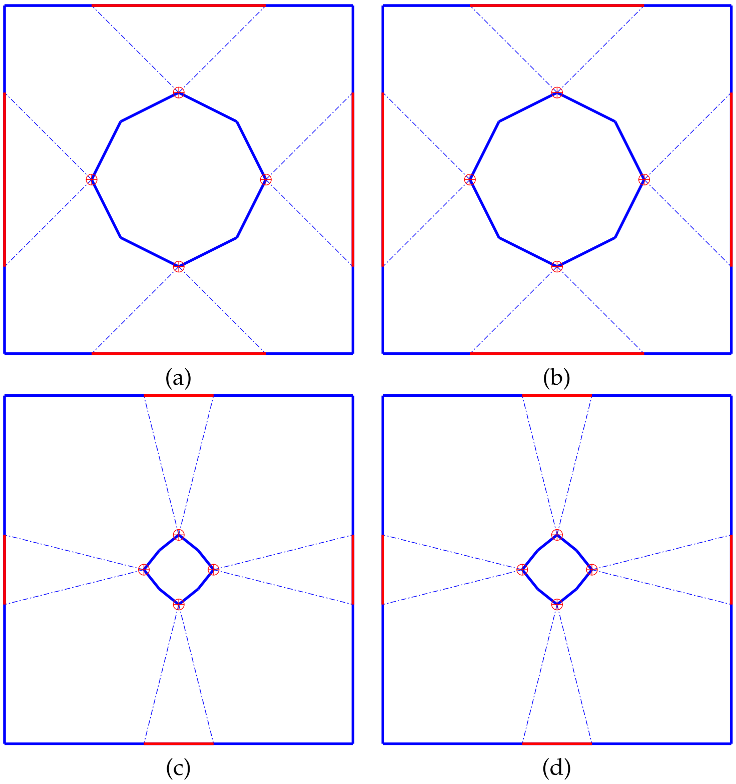

Fig. 5.2. The properties of $\mathrm{prox}_{\beta h_2}$. The side length of the square is (a) 2β and (b) 7β with $\beta = \frac{1}{2}$; (c) 2β and (b) 7β with $\beta = 2$. Red segments send one component to zero, marked by red $\oplus$, while blue segments map to blue segments. The thresholding depends not only on the parameter β but also the relationship between the two components.

$$\mathrm{prox}_{\beta h_2}(x) = \begin{cases} \left(\frac{1}{2\beta+1}x_1, 0\right)^\intercal, & \text{if } x \in I_1; \\ x - \frac{2\beta}{2\beta+1}(x_1 + x_2) \cdot (1,1)^\intercal, & \text{if } x \in I_2. \end{cases}$$

Clearly, $\mathrm{prox}_{\beta h_2}$ might send some components of a vector, for example vectors in I_1, to zero, or shrink the components of a vector by the threshold depending on both β and the components of the vector itself, like the vectors in I_2. Visually, in Fig. 5.2 $\mathrm{prox}_{\beta h_2}$ maps all points in the red lines to a closed point marked by the nearest $\oplus$, and maps the points in the blue lines to the ones on the inner octagon. We further remark that the operator $\mathrm{prox}_{\beta h_2}$ is linear on I_1 and I_2, and the plots of Fig. 5.2(a) and Fig. 5.2(b) are visually similar, while the plots of Fig. 5.2(c) and Fig. 5.2(d) are also visually similar, although each on different scales.

Fig. 5.3 shows the properties of the operator $\mathrm{prox}_{\beta h_3}$ in $\mathbb{R}^2$. The side lengths of the squares centered at the origin in Fig. 5.3(a) and (b) are 2β and 7β with $\beta = \frac{1}{2}$, respectively. The side lengths of the squares centered at the origin in Fig. 5.3(c) and (d) are 2β and 7β with $\beta = 2$, respectively. We partition vectors in $\mathbb{R}^2_\downarrow$ as follows:

$$\{x = (x_1, x_2)^\intercal : x \in \mathbb{R}^2_\downarrow\} = I_3 \cup I_4$$

with

$$I_3 = \left\{ x = (x_1, x_2)^\intercal \in \mathbb{R}^2_\downarrow : x_2 \le x_1 - \frac{1}{6\beta}\sqrt{1 + 12\beta x_1} \right\}$$

and

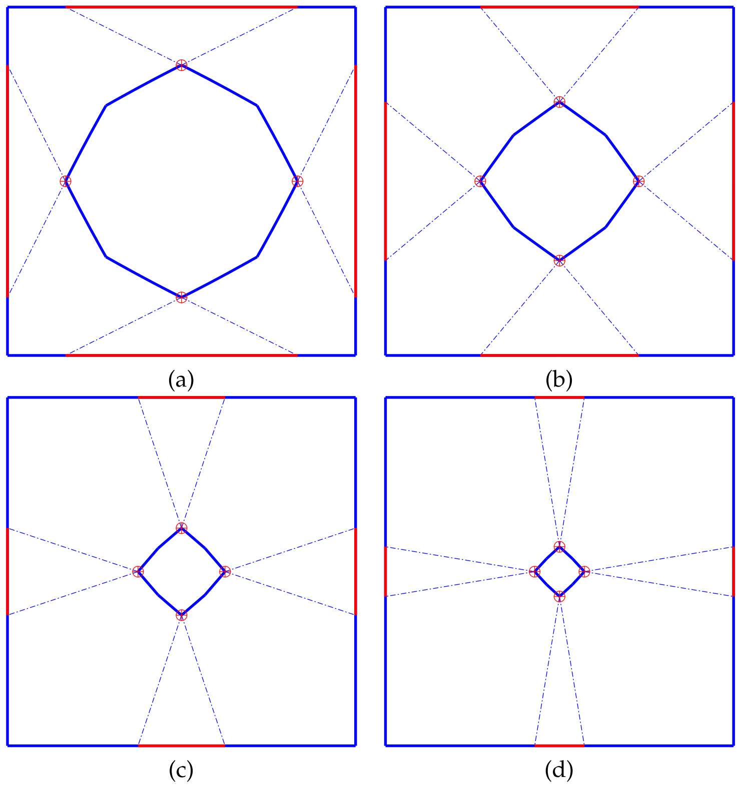

Fig. 5.3. The properties of $\mathrm{prox}_{\beta h_3}$. The side length of the square is (a) 2β and (b) 7β with $\beta = \frac{1}{2}$; (c) 2β and (b) 7β with $\beta = 2$. As before, red segments map to red $\oplus$, while blue segments map to blue segments.

$$I_4 = \left\{ x = (x_1, x_2)^\intercal \in \mathbb{R}^2_\downarrow : x_2 > x_1 - \frac{1}{6\beta}\sqrt{1 + 12\beta x_1} \right\}.$$

Swapping I_1 (I_2) and I_3 (I_4), the red (blue) lines on the squares in Fig. 5.3 are formed as in Fig. 5.2. With Lemma 3, we get

$$\mathrm{prox}_{\beta h_3}(x) = \begin{cases} \left(\frac{1}{6\beta}\sqrt{1+12\beta x_1}, 0\right)^\intercal, & \text{if } x \in I_3; \\ x - \left(\frac{x_1+x_2}{2} - \frac{1}{24\beta}\sqrt{1+24\beta(x_1+x_2)}\right) \cdot (1,1)^\intercal, & \text{if } x \in I_4. \end{cases}$$

Similar to $\mathrm{prox}_{\beta h_2}$, $\mathrm{prox}_{\beta h_3}$ might send some components of a vector, for example vectors in I_3, to zero, or shrink the components of a vector by the threshold depending on both β and the components of the vector itself, like the vectors in I_4. In Fig. 5.3, we see that $\mathrm{prox}_{\beta h_3}$ maps all points in the red lines to a closed point marked by $\oplus$, while it maps the points in the blue lines to the ones on the octagon-like curve.

Similar to the preceding examples for $p \in \{1, 2, 3\}$ in $\mathbb{R}^2$, the function h_p with $p > 1$ promotes *relative* sparsity in that rather than a uniform threshold shrinking elements or mapping them to zero, the threshold depends on the relationship among the components.

6. Conclusions

In this paper, we provide a constructive way to compute the proximity operator of the p-th power of the ℓ_1 norm. For $p \in \{2, 3, 4\}$, explicit expressions of these proximity operators can be derived from the proposed way. Moreover, this construction gives a characterization on the regions where the proximity operator of the p-th power of the ℓ_1 norm can provide sparsity.

Data availability

No data was used for the research described in the article.

Disclaimer and Acknowledgment of Support

This work was funded in part by AFOSR grant 21RICO035. The work of L. Shen was supported in part by the National Science Foundation under grant DMS-1913039 and DMS-2208385. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Air Force Research Laboratory. Cleared for public release 23 August 2022: Case Number AFRL-2022-4045.

References

[1] H.L. Bauschke, P.L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Spaces, AMS Books in Mathematics, Springer, New York, 2011.
[2] A. Beck, First-Order Methods in Optimization, SIAM, 2017.
[3] D. Donoho, I. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika 81 (1994) 425–455.
[4] T. Evgeniou, M. Pontil, D. Spinellis, N. Nassuphis, Regularized robust portfolio estimation, regularization, optimization, kernels, and support vector machines, in: Chapman & Hall/CRC Mach. Learn. Pattern Recogn. Ser., CRC Press, Boca Raton, FL, 2015, pp. 237–256.
[5] J.-B. Hiriart-Urruty, C. Lemarechal, Convex Analysis and Minimization Algorithms II, Grundlehren der Mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences), vol. 306, Springer-Verlag, Berlin, 1993.
[6] A.F.T. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, Online learning of structured predictors with multiple kernels, in: G. Gordon, D. Dunson, M. Dudík (Eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 11–13 April, in: Proceedings of Machine Learning Research, vol. 15, PMLR, 2011, pp. 507–515.
[7] C.A. Micchelli, L. Shen, Y. Xu, Proximity algorithms for image models: denoising, Inverse Probl. 27 (2011) 045009prnse , (30pp).
[8] L. Shen, B.W. Suter, E.E. Tripp, Structured sparsity promoting functions, J. Optim. Theory Appl. 183 (2019) 386–421.
[9] L. Shen, B.W. Suter, E.E. Tripp, Algorithmic versatility of SPF-regularization methods, Anal. Appl. 19 (2021) 43–69.