

# Perspective Functions: Proximal Calculus and Applications in High-Dimensional Statistics\*

Patrick L. Combettes<sup>1</sup> and Christian L. Müller<sup>2</sup>

<sup>1</sup>North Carolina State University  
Department of Mathematics  
Raleigh, NC 27695-8205, USA  
[plc@math.ncsu.edu](mailto:plc@math.ncsu.edu)

<sup>2</sup>Flatiron Institute  
Simons Foundation  
New York, NY 10010, USA  
[cmueller@simonsfoundation.org](mailto:cmueller@simonsfoundation.org)

## Abstract

Perspective functions arise explicitly or implicitly in various forms in applied mathematics and in statistical data analysis. To date, no systematic strategy is available to solve the associated, typically nonsmooth, optimization problems. In this paper, we fill this gap by showing that proximal methods provide an efficient framework to model and solve problems involving perspective functions. We study the construction of the proximity operator of a perspective function under general assumptions and present important instances in which the proximity operator can be computed explicitly or via straightforward numerical operations. These results constitute central building blocks in the design of proximal optimization algorithms. We showcase the versatility of the framework by designing novel proximal algorithms for state-of-the-art regression and variable selection schemes in high-dimensional statistics.

---

\*Contact author: P. L. Combettes, [plc@math.ncsu.edu](mailto:plc@math.ncsu.edu), phone: +1 (919) 515 2671.

# 1 Introduction

Perspective functions appear, often implicitly, in various problems in areas as diverse as statistics, control, computer vision, mechanics, game theory, information theory, signal recovery, transportation theory, machine learning, disjunctive optimization, and physics (see the companion paper [7] for a detailed account). In the setting of a real Hilbert space  $\mathcal{G}$ , the most useful form of a perspective function, first investigated in Euclidean spaces in [24], is the following.

**Definition 1.1** Let  $\varphi: \mathcal{G} \rightarrow ]-\infty, +\infty]$  be a proper lower semicontinuous convex function and let  $\text{rec } \varphi$  be its recession function. The perspective of  $\varphi$  is

$$\tilde{\varphi}: \mathbb{R} \times \mathcal{G} \rightarrow ]-\infty, +\infty] : (\eta, y) \mapsto \begin{cases} \eta\varphi(y/\eta), & \text{if } \eta > 0; \\ (\text{rec } \varphi)(y), & \text{if } \eta = 0; \\ +\infty, & \text{if } \eta < 0. \end{cases} \quad (1.1)$$

Many scientific problems result in minimization problems that involve perspective functions. In statistics, a prominent instance is the modeling of data via “maximum likelihood-type” estimation (or M-estimation) with a so-called concomitant parameter [17]. In this context,  $\varphi$  is a likelihood function,  $\eta$  takes the role of the concomitant parameter, e.g., an unknown scale or location of the assumed parametric distribution, and  $y$  comprises unknown regression coefficients. The statistical problem is then to simultaneously estimate the concomitant variable and the regression vector from data via optimization. Another important example in statistics [15], signal recovery [5], and physics [16] is the Fisher information of a function  $x: \mathbb{R}^N \rightarrow ]0, +\infty[$ , namely

$$\int_{\mathbb{R}^N} \frac{\|\nabla x(t)\|_2^2}{x(t)} dt, \quad (1.2)$$

which hinges on the perspective function of the squared Euclidean norm (see [7] for further discussion).

In the literature, problems involving perspective functions are typically solved with a wide range of ad-hoc methods. Despite the ubiquity of perspective functions, no systematic structuring framework has been available to approach these problems. The goal of this paper is to fill this gap by showing that they are amenable to solution by proximal methods, which offer a broad array of splitting algorithms to solve complex nonsmooth problems with attractive convergence guarantees [1, 8, 11, 14]. The central element in the successful implementation of a proximal algorithm is the ability to compute the proximity operator of the functions present in the optimization problem. We therefore propose a systematic investigation of proximity operators for perspective functions and show that the proximal framework can efficiently solve perspective-function based problems, unveiling in particular new applications in high-dimensional statistics.

In Section 2, we introduce basic concepts from convex analysis and review essential properties of perspective function. We then study the proximity operator of perspective functions in Sections 3. We establish a characterization of the proximity operator and then provide examples of computation for concrete instances. Section 4 unveils new applications of perspective functions in high-dimensional statistics and demonstrates the flexibility and potency of the proposed framework to both model and solve complex problems in statistical data analysis.

## 2 Notation and background

### 2.1 Notation and elements of convex analysis

Throughout,  $\mathcal{H}$ ,  $\mathcal{G}$ , and  $\mathcal{K}$  are real Hilbert spaces and  $\mathcal{H} \oplus \mathcal{G}$  denotes their Hilbert direct sum. The symbol  $\|\cdot\|$  denotes the norm of a Hilbert space and  $\langle \cdot | \cdot \rangle$  the associated scalar product. The closed ball with center  $x \in \mathcal{K}$  and radius  $\rho \in ]0, +\infty[$  is denoted by  $B(x; \rho)$ .

A function  $f: \mathcal{K} \rightarrow ]-\infty, +\infty]$  is proper if  $\text{dom } f = \{x \in \mathcal{K} \mid f(x) < +\infty\} \neq \emptyset$ , coercive if  $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$ , and supercoercive if  $\lim_{\|x\| \rightarrow +\infty} f(x)/\|x\| = +\infty$ . Denote by  $\Gamma_0(\mathcal{K})$  the class of proper lower semicontinuous convex functions from  $\mathcal{K}$  to  $]-\infty, +\infty]$ , and let  $f \in \Gamma_0(\mathcal{K})$ . The conjugate of  $f$  is the function

$$f^*: \mathcal{K} \rightarrow [-\infty, +\infty] : u \mapsto \left( \sup_{x \in \mathcal{K}} \langle x | u \rangle - f(x) \right). \quad (2.1)$$

It also belongs to  $\Gamma_0(\mathcal{K})$  and  $f^{**} = f$ . The subdifferential of  $f$  is the set-valued operator

$$\partial f: \mathcal{K} \rightarrow 2^{\mathcal{K}} : x \mapsto \{u \in \mathcal{K} \mid (\forall y \in \text{dom } f) \langle y - x | u \rangle + f(x) \leq f(y)\}. \quad (2.2)$$

We have

$$(\forall x \in \mathcal{K})(\forall u \in \mathcal{K}) \quad u \in \partial f(x) \quad \Leftrightarrow \quad x \in \partial f^*(u). \quad (2.3)$$

Moreover,

$$(\forall x \in \mathcal{K})(\forall u \in \mathcal{K}) \quad f(x) + f^*(u) \geq \langle x | u \rangle \quad (2.4)$$

and

$$(\forall x \in \mathcal{K})(\forall u \in \mathcal{K}) \quad u \in \partial f(x) \quad \Leftrightarrow \quad f(x) + f^*(u) = \langle x | u \rangle. \quad (2.5)$$

If  $f$  is Gâteaux differentiable at  $x \in \text{dom } f$  with gradient  $\nabla f(x)$ , then

$$\partial f(x) = \{\nabla f(x)\}. \quad (2.6)$$

Let  $z \in \text{dom } f$ . The recession function of  $f$  is

$$(\forall y \in \mathcal{K}) \quad (\text{rec } f)(y) = \sup_{x \in \text{dom } f} (f(x + y) - f(x)) = \lim_{\alpha \rightarrow +\infty} \frac{f(z + \alpha y) - f(z)}{\alpha}. \quad (2.7)$$

The infimal convolution operation is denoted by  $\square$ . Now let  $C$  be a subset of  $\mathcal{K}$ . Then

$$\iota_C: \mathcal{K} \rightarrow \{0, +\infty\} : x \mapsto \begin{cases} 0, & \text{if } x \in C; \\ +\infty, & \text{if } x \notin C \end{cases} \quad (2.8)$$

is the indicator function of  $C$ ,

$$d_C: \mathcal{K} \rightarrow [0, +\infty] : x \mapsto \inf \|C - x\| \quad (2.9)$$

is the distance function to  $C$ , and

$$\sigma_C = \iota_C^*: \mathcal{K} \rightarrow [-\infty, +\infty] : u \mapsto \sup_{x \in C} \langle x | u \rangle \quad (2.10)$$

is the support function of  $C$ . If  $C$  is nonempty, closed, and convex then, for every  $x \in \mathcal{K}$ , there exists a unique point  $P_C x \in C$ , called the projection of  $x$  onto  $C$ , such that  $\|x - P_C x\| = d_C(x)$ . We have

$$(\forall x \in \mathcal{K})(\forall p \in \mathcal{K}) \quad p = P_C x \quad \Leftrightarrow \quad [p \in C \quad \text{and} \quad (\forall y \in C) \quad \langle y - p | x - p \rangle \leq 0]. \quad (2.11)$$

The normal cone to  $C$  is

$$N_C = \partial \iota_C: \mathcal{K} \rightarrow 2^{\mathcal{K}}: x \mapsto \begin{cases} \{u \in \mathcal{K} \mid \sup \langle C - x | u \rangle \leq 0\}, & \text{if } x \in C; \\ \emptyset, & \text{otherwise.} \end{cases} \quad (2.12)$$

For further background on convex analysis, see [1, 24].

## 2.2 Proximity operators

The proximity operator of  $f \in \Gamma_0(\mathcal{K})$  is

$$\text{prox}_f: \mathcal{K} \rightarrow \mathcal{K}: x \mapsto \underset{y \in \mathcal{K}}{\text{argmin}} \left( f(y) + \frac{1}{2} \|x - y\|^2 \right). \quad (2.13)$$

This operator was introduced by Moreau in 1962 [20] to model problems in unilateral mechanics. In [12], it was shown to play an important role in the investigation of various data processing problems, and it has become increasingly prominent in the general area of data analysis [10, 25]. We review basic properties and refer the reader to [1] for a more complete account.

Let  $f \in \Gamma_0(\mathcal{K})$ . Then

$$(\forall x \in \mathcal{K})(\forall p \in \mathcal{K}) \quad p = \text{prox}_f x \quad \Leftrightarrow \quad x - p \in \partial f(p). \quad (2.14)$$

If  $C$  is a nonempty closed convex subset of  $\mathcal{K}$ , then

$$\text{prox}_f = P_C. \quad (2.15)$$

Let  $\gamma \in ]0, +\infty[$ . The Moreau decomposition of  $x \in \mathcal{K}$  is

$$x = \text{prox}_{\gamma f} x + \gamma \text{prox}_{f^*/\gamma}(x/\gamma). \quad (2.16)$$

The following facts will also be needed.

**Lemma 2.1** *Let  $(\Omega, \mathcal{F}, \mu)$  be a complete  $\sigma$ -finite measure space, let  $\mathbb{K}$  be a separable real Hilbert space, and let  $\psi \in \Gamma_0(\mathbb{K})$ . Suppose that  $\mathcal{K} = L^2((\Omega, \mathcal{F}, \mu); \mathbb{K})$  and that  $\mu(\Omega) < +\infty$  or  $\psi \geq \psi(0) = 0$ . Set*

$$\begin{aligned} \Phi: \mathcal{K} &\rightarrow ]-\infty, +\infty] \\ x &\mapsto \begin{cases} \int_{\Omega} \psi(x(\omega)) \mu(d\omega), & \text{if } \psi \circ x \in L^1((\Omega, \mathcal{F}, \mu); \mathbb{R}); \\ +\infty, & \text{otherwise.} \end{cases} \end{aligned} \quad (2.17)$$

Let  $x \in \mathcal{K}$  and define, for  $\mu$ -almost every  $\omega \in \Omega$ ,  $p(\omega) = \text{prox}_{\psi} x(\omega)$ . Then  $p = \text{prox}_{\Phi} x$ .

*Proof.* By [1, Proposition 9.32],  $\Phi \in \Gamma_0(\mathcal{K})$ . Now take  $x$  and  $p$  in  $\mathcal{K}$ . Then it follows from (2.14) and [1, Proposition 16.50] that  $p(\omega) = \text{prox}_{\Phi}x(\omega)$   $\mu$ -a.e.  $\Leftrightarrow x(\omega) - p(\omega) \in \partial\psi(p(\omega))$   $\mu$ -a.e.  $\Leftrightarrow x - p \in \partial\Phi(p)$ .  $\Leftrightarrow p = \text{prox}_{\Phi}x$ .  $\square$

**Lemma 2.2** *Let  $D \neq \{0\}$  be a nonempty closed convex subset of  $\mathcal{K}$ , let  $x \in \mathcal{K}$ , and let  $\gamma \in ]0, +\infty[$ . Set  $f = \|\cdot\| + \sigma_D$  and  $C = \gamma D$ . Then*

$$\text{prox}_{\gamma f}x = \begin{cases} 0, & \text{if } d_C(x) \leq \gamma; \\ \left(1 - \frac{\gamma}{d_C(x)}\right)(x - P_Cx), & \text{if } d_C(x) > \gamma. \end{cases} \quad (2.18)$$

*If, in addition,  $D$  is a cone and  $K$  denotes its polar cone, then  $f = \|\cdot\| + \iota_K$  and*

$$\text{prox}_{\gamma f}x = \begin{cases} 0, & \text{if } \|P_Kx\| \leq \gamma; \\ \left(1 - \frac{\gamma}{\|P_Kx\|}\right)P_Kx, & \text{if } \|P_Kx\| > \gamma. \end{cases} \quad (2.19)$$

*Proof.* Using elementary convex analysis, we obtain

$$f = \iota_{B(0;1)}^* + \iota_D^* = (\iota_{B(0;1)} \square \iota_D)^* = \iota_{B(0;1)+D}^* = \sigma_{B(0;1)+D}. \quad (2.20)$$

Hence, it follows from (2.16) and (2.15) that

$$\text{prox}_{\gamma f}x = x - \gamma \text{prox}_{f^*/\gamma}(x/\gamma) = x - \gamma P_{B(0;1)+D}(x/\gamma). \quad (2.21)$$

However by [1, Propositions 28.1(ii) and 28.10],

$$\gamma P_{B(0;1)+D}(x/\gamma) = P_{B(0;\gamma)+C}x = \begin{cases} x, & \text{if } d_C(x) \leq \gamma; \\ P_Cx + \gamma \frac{x - P_Cx}{d_C(x)}, & \text{if } d_C(x) > \gamma. \end{cases} \quad (2.22)$$

Upon combining (2.21) and (2.22), we arrive at (2.18). Now suppose that, in addition,  $D$  is a cone. Then  $C = D$ ,  $\sigma_D = \iota_K$ , and (2.16) yields  $\text{Id} - P_D = P_K$ . Altogether, (2.18) reduces to (2.19).  $\square$

### 2.3 Perspective functions

We review here some essential properties of perspective functions.

**Lemma 2.3** [7] *Let  $\varphi \in \Gamma_0(\mathcal{G})$ . Then the following hold:*

- (i)  $\tilde{\varphi}$  is a positively homogeneous function in  $\Gamma_0(\mathbb{R} \oplus \mathcal{G})$ .
- (ii) Let  $C = \{(\mu, u) \in \mathbb{R} \times \mathcal{G} \mid \mu + \varphi^*(u) \leq 0\}$ . Then  $(\tilde{\varphi})^* = \iota_C$  and  $\tilde{\varphi} = \sigma_C$ .
- (iii) Let  $\eta \in \mathbb{R}$  and  $y \in \mathcal{G}$ . Then

$$\partial\tilde{\varphi}(\eta, y) = \begin{cases} \{(\varphi(y/\eta) - \langle y \mid u \rangle/\eta, u) \mid u \in \partial\varphi(y/\eta)\}, & \text{if } \eta > 0; \\ \{(\mu, u) \in C \mid \sigma_{\text{dom } \varphi^*}(y) = \langle u \mid y \rangle\}, & \text{if } \eta = 0 \text{ and } y \neq 0; \\ C, & \text{if } \eta = 0 \text{ and } y = 0; \\ \emptyset, & \text{if } \eta < 0. \end{cases} \quad (2.23)$$

(iv) Suppose that  $\text{dom } \varphi^*$  is open or that  $\varphi$  is supercoercive, let  $\eta \in \mathbb{R}$ , and let  $y \in \mathcal{G}$ . Then

$$\partial\tilde{\varphi}(\eta, y) = \begin{cases} \{(\varphi(y/\eta) - \langle y | u \rangle/\eta, u) \mid u \in \partial\varphi(y/\eta)\}, & \text{if } \eta > 0; \\ C, & \text{if } \eta = 0 \text{ and } y = 0; \\ \emptyset, & \text{otherwise.} \end{cases} \quad (2.24)$$

We refer to the companion paper [7] for further properties of perspective functions as well as examples. Here are two important instances of (composite) perspective functions that will play a central role in Section 4.

**Lemma 2.4** Let  $L: \mathcal{H} \rightarrow \mathcal{G}$  be linear and bounded, let  $r \in \mathcal{G}$ , let  $u \in \mathcal{H}$ , let  $\alpha \in ]0, +\infty[$ , let  $\rho \in \mathbb{R}$ , and let  $q \in ]1, +\infty[$ . Set

$$f: \mathcal{H} \rightarrow ]-\infty, +\infty]: x \mapsto \begin{cases} \frac{\|Lx - r\|^q}{\alpha|\langle x | u \rangle - \rho|^{q-1}}, & \text{if } \langle x | u \rangle > \rho; \\ 0, & \text{if } Lx = r \text{ and } \langle x | u \rangle = \rho; \\ +\infty, & \text{otherwise} \end{cases} \quad (2.25)$$

and  $A: \mathcal{H} \rightarrow \mathbb{R} \oplus \mathcal{G}: x \mapsto (\langle x | u \rangle - \rho, Lx - r)$ . Then  $f = [\|\cdot\|^q/\alpha]^\sim \circ A \in \Gamma_0(\mathcal{H})$ .

*Proof.* This is a special case of [7, Example 4.2].  $\square$

**Lemma 2.5** [7, Example 3.6] Let  $\phi \in \Gamma_0(\mathbb{R})$  be an even function, let  $v \in \mathcal{G}$ , let  $\delta \in \mathbb{R}$ , and set

$$g: \mathbb{R} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \eta\phi(\|y\|/\eta) + \langle y | v \rangle + \delta\eta, & \text{if } \eta > 0; \\ (\text{rec } \phi)(\|y\|) + \langle y | v \rangle, & \text{if } \eta = 0; \\ +\infty, & \text{if } \eta < 0. \end{cases} \quad (2.26)$$

Then  $g = [\phi \circ \|\cdot\| + \delta\langle \cdot | v \rangle]^\sim \in \Gamma_0(\mathbb{R} \oplus \mathcal{G})$ .

### 3 Proximity operator of a perspective function

#### 3.1 Main result

We start with a characterization of the proximity operator of a perspective function when  $\text{dom } \varphi^*$  is open.

**Theorem 3.1** Let  $\varphi \in \Gamma_0(\mathcal{G})$ , let  $\gamma \in ]0, +\infty[$ , let  $\eta \in \mathbb{R}$ , and let  $y \in \mathcal{G}$ . Then the following hold:

- (i) Suppose that  $\eta + \gamma\varphi^*(y/\gamma) \leq 0$ . Then  $\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (0, 0)$ .
- (ii) Suppose that  $\text{dom } \varphi^*$  is open and that  $\eta + \gamma\varphi^*(y/\gamma) > 0$ . Then

$$\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (\eta + \gamma\varphi^*(p), y - \gamma p), \quad (3.1)$$

where  $p$  is the unique solution to the inclusion

$$y \in \gamma p + (\eta + \gamma\varphi^*(p))\partial\varphi^*(p). \quad (3.2)$$

If  $\varphi^*$  is differentiable at  $p$ , then  $p$  is characterized by  $y = \gamma p + (\eta + \gamma\varphi^*(p))\nabla\varphi^*(p)$ .

*Proof.* It follows from Lemma 2.3(ii) that

$$\tilde{\varphi} = \sigma_C, \quad \text{where } C = \{(\mu, u) \in \mathbb{R} \oplus \mathcal{G} \mid \mu + \varphi^*(u) \leq 0\}. \quad (3.3)$$

Since  $\varphi \in \Gamma_0(\mathcal{G})$ , we have  $\varphi^* \in \Gamma_0(\mathcal{G})$ . Therefore,  $C$  is a nonempty closed convex set. In turn, we derive from [9, Proposition 3.2] that  $\text{prox}_{\gamma\tilde{\varphi}} = \text{prox}_{\sigma_{\gamma C}}$  is a proximal thresholder on  $\gamma C$  in the sense that

$$(\forall \eta \in \mathbb{R})(\forall y \in \mathcal{G}) \quad \text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (0, 0) \Leftrightarrow (\eta, y) \in \gamma C. \quad (3.4)$$

(i): By (3.3) and (3.4),  $(\forall \eta \in \mathbb{R})(\forall y \in \mathcal{G}) \text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (0, 0) \Leftrightarrow \eta + \gamma\varphi^*(y/\gamma) \leq 0$ .

(ii): Set  $(\chi, q) = \text{prox}_{\gamma\tilde{\varphi}}(\eta, y)$  and  $p = (y - q)/\gamma$ . It follows from (2.14) that  $(\chi, q) \in \text{dom}(\gamma\partial\tilde{\varphi})$  and from (3.4) that  $(\chi, q) \neq (0, 0)$ . Hence, we deduce from Lemma 2.3(iv) that  $\chi > 0$ . Furthermore, we derive from (2.14) and Lemma 2.3(iii) that  $(\chi, q)$  is characterized by

$$\eta - \chi = \gamma\varphi(q/\chi) - \langle q/\chi \mid y - q \rangle \quad \text{and} \quad y - q \in \gamma\partial\varphi(q/\chi), \quad (3.5)$$

i.e.,

$$(\eta - \chi)/\gamma = \varphi(q/\chi) - \langle q/\chi \mid p \rangle \quad \text{and} \quad p \in \partial\varphi(q/\chi). \quad (3.6)$$

However, (2.5) asserts that

$$p \in \partial\varphi(q/\chi) \Leftrightarrow \varphi(q/\chi) + \varphi^*(p) = \langle q/\chi \mid p \rangle. \quad (3.7)$$

Hence, we derive from (3.6) that  $\varphi^*(p) = (\chi - \eta)/\gamma$ , i.e.,

$$\chi = \eta + \gamma\varphi^*(p). \quad (3.8)$$

Hence, by (2.3),

$$p \in \partial\varphi(q/\chi) \Leftrightarrow q \in \chi\partial\varphi^*(p) \Leftrightarrow y \in \gamma p + (\eta + \gamma\varphi^*(p))\partial\varphi^*(p). \quad (3.9)$$

Altogether, we have established the characterization (3.1)–(3.2), while the assertion concerning the differentiable case follows from (2.6).  $\square$

**Remark 3.2** Here is an alternative proof of Theorem 3.1. It follows from Lemma 2.3(ii) that

$$(\tilde{\varphi})^* = \iota_C, \quad \text{where } C = \{(\mu, u) \in \mathbb{R} \oplus \mathcal{G} \mid \mu + \varphi^*(u) \leq 0\} \quad (3.10)$$

is a nonempty closed convex set. Hence, using (2.16) and (2.15), we obtain

$$\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (\eta, y) - \gamma\text{prox}_{\gamma^{-1}\tilde{\varphi}^*}(\eta/\gamma, y/\gamma) = (\eta, y) - \gamma P_C(\eta/\gamma, y/\gamma) = (\eta, y) - P_{\gamma C}(\eta, y). \quad (3.11)$$

Now set  $(\pi, p) = P_C(\eta/\gamma, y/\gamma)$ . We deduce from (2.15), (2.16), and (2.12) that  $(\pi, p)$  is characterized by

$$(\eta/\gamma - \pi, y/\gamma - p) \in N_C(\pi, p). \quad (3.12)$$

(i): We have  $(\eta/\gamma, y/\gamma) \in C$ . Hence,  $(\pi, p) = (\eta/\gamma, y/\gamma)$  and (3.11) yields  $\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (0, 0)$ .

(ii): Set  $h: \mathbb{R} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty]: (\mu, u) \mapsto \mu + \varphi^*(u)$ . Then  $C = \text{lev}_{\leq 0} h$  and  $\text{dom } h = \mathbb{R} \times \text{dom } \varphi^*$  is open. It therefore follows from [1, Proposition 6.43(ii)] that

$$N_{\text{dom } h}(\pi, p) = \{(0, 0)\}. \quad (3.13)$$

Now let  $z \in \text{dom } \varphi^*$  and let  $\zeta \in ]-\infty, -\varphi^*(z)[$ . Then  $h(\zeta, z) < 0$ . Therefore, we derive from [1, Lemma 26.17 and Proposition 16.8] and (3.13) that

$$N_C(\pi, p) = \begin{cases} N_{\text{dom } h}(\pi, p) \cup \text{cone } \partial h(\pi, p), & \text{if } \pi + \varphi^*(p) = 0; \\ N_{\text{dom } h}(\pi, p), & \text{if } \pi + \varphi^*(p) < 0 \end{cases} \quad (3.14)$$

$$\begin{aligned} &= \begin{cases} \text{cone } \partial h(\pi, p), & \text{if } \pi + \varphi^*(p) = 0; \\ \{(0, 0)\}, & \text{if } \pi + \varphi^*(p) < 0 \end{cases} \\ &= \begin{cases} \text{cone } (\{1\} \times \partial\varphi^*(p)), & \text{if } \pi = -\varphi^*(p); \\ \{(0, 0)\}, & \text{if } \pi < -\varphi^*(p). \end{cases} \end{aligned} \quad (3.15)$$

Hence, if  $\pi < -\varphi^*(p)$ , then (3.12) yields  $(\eta/\gamma - \pi, y/\gamma - p) = (0, 0)$  and therefore  $(\eta/\gamma, y/\gamma) = (\pi, p) \in C$ , which is impossible since  $(\eta/\gamma, y/\gamma) \notin C$ . Thus, the characterization (3.12) becomes

$$\pi = -\varphi^*(p) \quad \text{and} \quad (\exists \nu \in ]0, +\infty[)(\exists w \in \partial\varphi^*(p)) \quad (\eta/\gamma + \varphi^*(p), y/\gamma - p) = \nu(1, w) \quad (3.16)$$

that is,  $y \in \gamma p + (\eta + \gamma\varphi^*(p))\partial\varphi^*(p)$ .

**Remark 3.3** Let  $\varphi \in \Gamma_0(\mathcal{G})$  be such that  $\text{dom } \varphi^*$  is open, let  $\gamma \in ]0, +\infty[$ , let  $\eta \in \mathbb{R}$ , and let  $y \in \mathcal{G}$  be such that  $\eta + \gamma\varphi^*(y/\gamma) > 0$ . We derive from (3.5) that  $y/\chi - q/\chi \in \partial(\gamma\varphi/\chi)(q/\chi)$  and then from (2.14) that  $q = \chi \text{prox}_{\gamma\varphi/\chi}(y/\chi)$ . Using (2.16), we can also write  $q = y - \text{prox}_{\chi\gamma\varphi^*(\cdot/\gamma)}y$ . Hence, we deduce from Theorem 3.1 the implicit relation

$$\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = \chi \left( 1, \text{prox}_{\gamma\varphi/\chi}(y/\chi) \right), \quad \text{where} \quad \chi = \eta + \gamma\varphi^* \left( \frac{\text{prox}_{\chi\gamma\varphi^*(\cdot/\gamma)}y}{\gamma} \right). \quad (3.17)$$

The next example is based on distance functions.

**Example 3.4** Let  $\varphi = \phi \circ d_D$ , where  $D = B(0; 1) \subset \mathcal{G}$  and  $\phi \in \Gamma_0(\mathbb{R})$  is an even function such that  $\phi(0) = 0$  and  $\phi^*$  is differentiable on  $\mathbb{R}$ . It follows from [1, Examples 13.3(iv) and 13.23] that  $\varphi^* = \|\cdot\| + \phi^* \circ \|\cdot\|$ . Note that, since  $\varphi$  and  $\phi$  are even and satisfy  $\varphi(0) = 0$  and  $\phi(0) = 0$ ,  $\varphi^*$  and  $\phi^*$  are even and satisfy  $\varphi^*(0) = 0$  and  $\phi^*(0) = 0$  as well by [1, Propositions 13.18 and 13.19]. In turn,  $\phi^{*\prime}(0) = 0$  and we therefore derive from [1, Corollary 16.38(iii) and Example 16.25] that

$$(\forall u \in \mathcal{G}) \quad \partial\varphi^*(u) = \begin{cases} \left\{ \frac{1 + \phi^{*\prime}(\|u\|)}{\|u\|} u \right\}, & \text{if } u \neq 0; \\ B(0; 1), & \text{if } u = 0. \end{cases} \quad (3.18)$$



We have  $\text{dom } \varphi^* = \mathcal{G}$  and, in view of Theorem 3.1(ii), we need only assume that  $\eta + \gamma\varphi^*(y/\gamma) > 0$ , i.e.,

$$\eta + \|y\| + \gamma\phi^*(\|y\|/\gamma) > 0. \quad (3.19)$$

Then (3.2) and (3.18) yield

$$\begin{cases} y = \gamma p + \left( \eta + \gamma(\|p\| + \phi^*(\|p\|)) \right) \frac{1 + \phi^{*\prime}(\|p\|)}{\|p\|} p, & \text{if } p \neq 0; \\ \|y\| \leq \eta, & \text{if } p = 0. \end{cases} \quad (3.20)$$

In view of Remark 3.2, the normal cone to the set  $C$  of (3.10) at  $(0, 0)$  is

$$K = \{(\eta, y) \in [0, +\infty[ \times \mathcal{G} \mid \|y\| \leq \eta\}. \quad (3.21)$$

So, for every  $(\eta, y) \in K$ ,  $P_C(\eta/\gamma, y/\gamma) = (0, 0)$  and  $\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (\eta, y)$ . Now suppose that  $(\eta, y) \notin K$ . Then  $p \neq 0$  and, taking the norm in the upper line of (3.20), we obtain

$$\gamma\|p\| + \left( \eta + \gamma(\|p\| + \phi^*(\|p\|)) \right) (1 + \phi^{*\prime}(\|p\|)) = \|y\|. \quad (3.22)$$

Set

$$\psi: s \mapsto s + \left( \frac{\eta}{\gamma} + s + \phi^*(s) \right) (1 + \phi^{*\prime}(s)) - \frac{\|y\|}{\gamma} \quad (3.23)$$

and define

$$\theta: s \mapsto \frac{1}{2} \left( \left( \frac{\eta}{\gamma} + s + \phi^*(s) \right)^2 + s^2 \right) - \frac{\|y\|s}{\gamma}. \quad (3.24)$$

Since  $\phi^*$  is convex,  $\theta$  is strongly convex and it therefore admits a unique minimizer  $t$ . Therefore  $\psi(t) = \theta'(t) = 0$  and  $\|p\| = t = \psi^{-1}(\|y\|/\gamma)$  is the unique solution to (3.22). In turn, (3.20) yields

$$p = \frac{t}{\|y\| + \gamma\psi(t)} y, \quad (3.25)$$

and we obtain  $\text{prox}_{\gamma\tilde{\varphi}}(\eta, y)$  via (3.1).

Next, we compute the proximity operator of a special case of the perspective function introduced in Lemma 2.5.

**Corollary 3.5** *Let  $v \in \mathcal{G}$ , let  $\delta \in \mathbb{R}$ , and let  $\phi \in \Gamma_0(\mathbb{R})$  be an even function such that  $\phi(0) = 0$  and  $\phi^*$  is differentiable on  $\mathbb{R}$ . Define*

$$g: \mathbb{R} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \eta\phi(\|y\|/\eta) + \delta\eta + \langle y \mid v \rangle, & \text{if } \eta > 0; \\ 0, & \text{if } y = 0 \text{ and } \eta = 0; \\ +\infty, & \text{otherwise.} \end{cases} \quad (3.26)$$

Let  $\gamma \in ]0, +\infty[$ , let  $\eta \in \mathbb{R}$ , let  $y \in \mathcal{G}$ , and set

$$\psi: s \mapsto \left( \phi^*(s) + \frac{\eta}{\gamma} - \delta \right) \phi^{*\prime}(s) + s. \quad (3.27)$$

Then  $\psi$  is invertible. Moreover, if  $\eta + \gamma\phi^*(\|y/\gamma - v\|) > \gamma\delta$ , set

$$t = \psi^{-1}(\|y/\gamma - v\|) \quad \text{and} \quad p = \begin{cases} v + \frac{t}{\|y - \gamma v\|}(y - \gamma v), & \text{if } y \neq \gamma v; \\ v, & \text{if } y = \gamma v. \end{cases} \quad (3.28)$$

Then

$$\text{prox}_{\gamma g}(\eta, y) = \begin{cases} (\eta + \gamma(\phi^*(t) - \delta), y - \gamma p), & \text{if } \eta + \gamma\phi^*(\|y/\gamma - v\|) > \gamma\delta; \\ (0, 0), & \text{if } \eta + \gamma\phi^*(\|y/\gamma - v\|) \leq \gamma\delta. \end{cases} \quad (3.29)$$

*Proof.* This is a special case of Theorem 3.1 with  $\varphi = \phi \circ \|\cdot\| + \delta + \langle \cdot, v \rangle$ . Indeed, as shown in [7, Example 3.6], (3.26) is a special case of (2.26). Hence, we derive from Lemma 2.5 that  $g = \tilde{\varphi} \in \Gamma_0(\mathbb{R} \oplus \mathcal{G})$ . Next, we obtain from [1, Example 13.7 and Proposition 13.20(iii)] that

$$\varphi^* = \phi^* \circ \|\cdot - v\| - \delta \quad (3.30)$$

and therefore that

$$\nabla\varphi^*: \mathcal{G} \rightarrow \mathcal{G}: z \mapsto \begin{cases} \frac{\phi^*(\|z - v\|)}{\|z - v\|}(z - v), & \text{if } z \neq v; \\ 0, & \text{if } z = v. \end{cases} \quad (3.31)$$

In view of Theorem 3.1, it remains to assume that  $\eta + \gamma\varphi^*(y/\gamma) > 0$ , i.e.,  $\eta + \phi^*(\|y/\gamma - v\|) > \gamma\delta$ , and to show that the point  $(t, p)$  provided by (3.28) satisfies

$$t = \|p - v\| \quad \text{and} \quad y = \gamma p + (\eta + \gamma\varphi^*(p))\nabla\varphi^*(p). \quad (3.32)$$

We consider two cases:

- $y = \gamma v$ : Since  $\phi$  is an even convex function such that  $\phi(0) = 0$ ,  $\phi^*$  has the same properties by [1, Propositions 13.18 and 13.19]. Hence, going back to Remark 3.2, since  $\phi^*$  is differentiable, the points that have  $(\pi, p) = (\delta, v)$  as a projection onto  $C = \{(\mu, u) \in \mathbb{R} \oplus \mathcal{G} \mid \mu + \phi^*(\|u - v\|) \leq \delta\}$  are the points on the ray  $\{(\delta + \lambda, v) \mid \lambda \in [0, +\infty[ \}$ . Thus, we derive from (3.11) that

$$y = \gamma v \Leftrightarrow P_C(\eta/\gamma, y/\gamma) = (\pi, p) = (\delta, v) \Leftrightarrow p = v \Leftrightarrow t = 0 \Leftrightarrow \text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (\eta, y) - \gamma(\delta, v) = (\eta - \gamma\delta, y - \gamma p). \quad (3.33)$$

Since  $\phi^*(0) = 0$ , we recover (3.29).

- $y \neq \gamma v$ : As seen in (3.33),  $p \neq v$ . Using (3.30) and (3.31), (3.32) can be rewritten as

$$t = \|p - v\| \quad \text{and} \quad y - \gamma v = \gamma(p - v) + \frac{(\eta + \gamma\phi^*(\|p - v\|) - \gamma\delta)\phi^*(\|p - v\|)}{\|p - v\|}(p - v), \quad (3.34)$$

that is,

$$t = \|p - v\| \quad \text{and} \quad y/\gamma - v = \frac{\|p - v\| + (\eta/\gamma - \delta + \phi^*(\|p - v\|))\phi^*(\|p - v\|)}{\|p - v\|}(p - v). \quad (3.35)$$

In view of (3.27), this is equivalent to

$$t = \|p - v\| \quad \text{and} \quad y/\gamma - v = \frac{\psi(\|p - v\|)}{\|p - v\|}(p - v). \quad (3.36)$$

Upon taking the norm on both sides of the second equality, we obtain

$$\psi(t) = \psi(\|p - v\|) = \|y/\gamma - v\|. \quad (3.37)$$

We note that, since  $\phi^*$  is convex,  $\psi$  is the derivative of the strongly convex function

$$\theta: s \mapsto \frac{1}{2}(\phi^{*2}(s) + s^2) + \left(\frac{\eta}{\gamma} - \delta\right)\phi^*(s). \quad (3.38)$$

Consequently,  $\psi$  is strictly increasing [1, Proposition 17.13], hence invertible. It follows that  $t = \psi^{-1}(\|y/\gamma - v\|)$ . In turn, (3.36) yields (3.28).

□

**Example 3.6** Define

$$g: \mathbb{R} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty] : (\eta, y) \mapsto \begin{cases} -\sqrt{\eta^2 - \|y\|^2}, & \text{if } \eta > 0 \text{ and } \|y\| \leq \eta; \\ 0, & \text{if } y = 0 \text{ and } \eta = 0; \\ +\infty, & \text{otherwise,} \end{cases} \quad (3.39)$$

let  $\gamma \in ]0, +\infty[$ , let  $\eta \in \mathbb{R}$ , let  $y \in \mathcal{G}$ , and define

$$\psi: s \mapsto \left(2 + \frac{\eta}{\gamma\sqrt{1+s^2}}\right)s. \quad (3.40)$$

If  $\eta + \sqrt{\gamma^2 + \|y\|^2} > 0$ , set

$$p = \begin{cases} \frac{t}{\|y\|}y, & \text{if } y \neq 0; \\ 0, & \text{if } y = 0, \end{cases} \quad \text{where } t = \psi^{-1}\left(\frac{\|y\|}{\gamma}\right). \quad (3.41)$$

Then

$$\text{prox}_{\gamma g}(\eta, y) = \begin{cases} \left(\eta + \gamma\sqrt{1+t^2}, y - \gamma p\right), & \text{if } \eta + \sqrt{\gamma^2 + \|y\|^2} > 0; \\ (0, 0), & \text{if } \eta + \sqrt{\gamma^2 + \|y\|^2} \leq 0. \end{cases} \quad (3.42)$$

*Proof.* This is a special case of Corollary 3.5 with  $\delta = 0$ ,  $v = 0$ , and

$$\phi: s \mapsto \begin{cases} -\sqrt{1-s^2}, & \text{if } |s| \leq 1; \\ +\infty, & \text{otherwise.} \end{cases} \quad (3.43)$$

It follows from [1, Example 13.2(vi) and Corollary 13.33] that  $\phi^*: s \mapsto \sqrt{1+s^2}$ . Hence,  $\phi^{*'}: s \mapsto s/\sqrt{1+s^2}$  and we derive (3.42) from (3.29). □

**Example 3.7** Let  $v \in \mathcal{G}$ , let  $\delta \in \mathbb{R}$ , let  $\alpha \in ]0, +\infty[$ , let  $q \in ]1, +\infty[$ , and consider the function

$$g: \mathbb{R} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \frac{\|y\|^q}{\alpha\eta^{q-1}} + \delta\eta + \langle y | v \rangle, & \text{if } \eta > 0; \\ 0, & \text{if } y = 0 \text{ and } \eta = 0; \\ +\infty, & \text{otherwise.} \end{cases} \quad (3.44)$$

Let  $\gamma \in ]0, +\infty[$ , set  $q^* = q/(q-1)$ , set  $\varrho = (\alpha(1-1/q^*))^{q^*-1}$ , and take  $\eta \in \mathbb{R}$  and  $y \in \mathcal{G}$ . If  $q^*\gamma^{q^*-1}\eta + \varrho\|y\|^{q^*} > \gamma\delta$  and  $y \neq \gamma v$ , let  $t$  be the unique solution in  $]0, +\infty[$  to the equation

$$s^{2q^*-1} + \frac{q^*(\eta - \gamma\delta)}{\gamma\varrho} s^{q^*-1} + \frac{q^*}{\varrho^2} s - \frac{q^*\|y - \gamma v\|}{\gamma\varrho^2} = 0 \quad (3.45)$$

and set

$$p = \begin{cases} v + \frac{t}{\|y - \gamma v\|} (y - \gamma v), & \text{if } y \neq \gamma v; \\ v, & \text{if } y = \gamma v. \end{cases} \quad (3.46)$$

Then

$$\text{prox}_{\gamma g}(\eta, y) = \begin{cases} (\eta + \gamma(\varrho t^{q^*} - \delta)/q^*, y - \gamma p), & \text{if } q^*\gamma^{q^*-1}\eta + \varrho\|y\|^{q^*} > \gamma\delta; \\ (0, 0), & \text{if } q^*\gamma^{q^*-1}\eta + \varrho\|y\|^{q^*} \leq \gamma\delta. \end{cases} \quad (3.47)$$

*Proof.* This is a special case of Corollary 3.5 with  $\phi = |\cdot|^q/\alpha$ . Indeed, we derive from [1, Example 13.2(i) and Proposition 13.20(i)] that  $\phi^* = \varrho|\cdot|^{q^*}/q^*$ , which implies that (3.46)–(3.47) follow from (3.29).  $\square$

**Example 3.8** Let  $v \in \mathcal{G}$ , let  $\alpha \in ]0, +\infty[$ , let  $\delta \in \mathbb{R}$ , and consider the function

$$g: \mathbb{R} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \frac{\|y\|^2}{\alpha\eta} + \delta\eta + \langle y | v \rangle, & \text{if } \eta > 0; \\ 0, & \text{if } y = 0 \text{ and } \eta = 0; \\ +\infty, & \text{otherwise.} \end{cases} \quad (3.48)$$

We obtain a special case of Example 3.7 with  $q = q^* = 2$ . Now let  $\gamma \in ]0, +\infty[$ , and take  $\eta \in \mathbb{R}$  and  $y \in \mathcal{G}$ . If  $4\gamma\eta + \alpha\|y\|^2 \leq 2\gamma\delta$ , then  $\text{prox}_{\gamma g}(\eta, y) = (0, 0)$ . Suppose that  $4\gamma\eta + \alpha\|y\|^2 > 2\gamma\delta$ . First, if  $y = \gamma v$ , then  $\text{prox}_{\gamma g}(\eta, y) = (\eta - \gamma\delta/2, 0)$ . Next, suppose that  $y \neq \gamma v$  and let  $t$  be the unique solution in  $]0, +\infty[$  to the depressed cubic equation

$$s^3 + \frac{4\alpha(\eta - \gamma\delta) + 8\gamma}{\alpha^2\gamma} s - \frac{8\|y - \gamma v\|}{\alpha^2\gamma} = 0. \quad (3.49)$$

Then we derive from (3.46)–(3.47) that

$$\text{prox}_{\gamma g}(\eta, y) = \left( \eta + \frac{\gamma}{2} \left( \frac{\alpha t^2}{2} - \delta \right), \left( 1 - \frac{\gamma t}{\|y - \gamma v\|} \right) (y - \gamma v) \right). \quad (3.50)$$

Note that (3.49) can be solved explicitly via Cardano's formula [4, Chapter 4] to obtain  $t$ .

We conclude this subsection by investigating integral functions constructed from integrands that are perspective functions.

**Proposition 3.9** *Let  $(\Omega, \mathcal{F}, \mu)$  be a measure space, let  $\mathbb{G}$  be a separable real Hilbert space, and let  $\varphi \in \Gamma_0(\mathbb{G})$ . Set  $\mathcal{H} = L^2((\Omega, \mathcal{F}, \mu); \mathbb{R})$  and  $\mathcal{G} = L^2((\Omega, \mathcal{F}, \mu); \mathbb{G})$ , and suppose that  $\mu(\Omega) < +\infty$  or  $\varphi \geq \varphi(0) = 0$ . For every  $x \in \mathcal{H}$ , set  $\Omega_0(x) = \{\omega \in \Omega \mid x(\omega) = 0\}$  and  $\Omega_+(x) = \{\omega \in \Omega \mid x(\omega) > 0\}$ . Define*

$$\Phi: \mathcal{H} \oplus \mathcal{G} \rightarrow ]-\infty, +\infty]: (x, y) \mapsto \begin{cases} \int_{\Omega_0(x)} (\text{rec } \varphi)(y(\omega)) \mu(d\omega) + \int_{\Omega_+(x)} x(\omega) \varphi\left(\frac{y(\omega)}{x(\omega)}\right) \mu(d\omega), \\ +\infty, \end{cases} \quad \text{if } \begin{cases} x \geq 0 \text{ } \mu\text{-a.e.} \\ (\text{rec } \varphi)(y) 1_{\Omega_0(x)} + x\varphi(y/x) 1_{\Omega_+(x)} \in L^1((\Omega, \mathcal{F}, \mu); \mathbb{R}); \\ \text{otherwise.} \end{cases} \quad (3.51)$$

Now let  $x \in \mathcal{H}$  and  $y \in \mathcal{G}$ , and set, for  $\mu$ -almost every  $\omega \in \Omega$ ,  $(p(\omega), q(\omega)) = \text{prox}_{\tilde{\varphi}}(x(\omega), y(\omega))$ . Then  $\text{prox}_{\Phi}(x, y) = (p, q)$ .

*Proof.* Set  $z = (x, y)$ . It follows from Lemma 2.3(i) that  $\tilde{\varphi} \in \Gamma_0(\mathbb{R} \oplus \mathbb{G})$ , and [7, Proposition 5.1] asserts that  $\Phi$  is a well-defined function in  $\Gamma_0(\mathbb{R} \oplus \mathcal{G})$  with

$$\Phi(z) = \int_{\Omega} \tilde{\varphi}(z(\omega)) \mu(d\omega). \quad (3.52)$$

Therefore, the result is obtained by applying Lemma 2.1 with  $\mathbb{K} = \mathbb{R} \oplus \mathbb{G}$  and  $\mathcal{K} = \mathcal{H} \oplus \mathcal{G}$ .  $\square$

**Remark 3.10** Proposition 3.9 provides a general setting for computing the proximity operators of abstract integral functionals by reducing it to the computation of the proximity operator of the integrand. In particular, by suitably choosing the underlying measure space and the integrand, it provides a framework for computing the proximity operators of the integral function based on perspective functions discussed in [7], which include general divergences. For instance, discrete  $N$ -dimensional divergences are obtained by setting  $\Omega = \{1, \dots, N\}$  and  $\mathcal{F} = 2^\Omega$ , and letting  $\mu$  be the counting measure (hence  $\mathcal{H} = \mathcal{G} = \mathbb{R}^N$ ) and  $\mathbb{G} = \mathbb{R}$ . While completing the present paper, it has come to our attention that the computation of the proximity operators of discrete divergences has also been recently addressed in [13].

## 3.2 Further results

A convenient assumption in Theorem 3.1(ii) is that  $\text{dom } \varphi^*$  is open, as it allowed us to rule out the case when

$$\text{prox}_{\gamma \tilde{\varphi}}(\eta, y) = (0, q) \quad \text{and} \quad q \neq 0, \quad (3.53)$$

and to reduce (3.14) to (3.15) using (3.13). In general, (3.13) has the form  $N_{\text{dom } h}(\pi, p) = \{0\} \times N_{\text{dom } \varphi^*} p$  and, if  $\text{dom } \varphi^*$  is simple enough, explicit expressions can still be obtained. To shed more light

on the case (3.53), consider the scenario in which  $q \neq 0$  and  $\text{dom } \varphi^*$  is closed, and set  $p = (y - q)/\gamma$ . Then, in view of (2.14), (3.53) yields  $(\eta/\gamma, p) \in \partial\tilde{\varphi}(0, q)$ . In turn, we derive from (2.23) that

$$\varphi^*(p) \leq -\eta/\gamma \quad \text{and} \quad \sigma_{\text{dom } \varphi^*}(q) = \langle p \mid q \rangle. \quad (3.54)$$

Thus,

$$p \in \text{dom } \varphi^* \quad \text{and} \quad (\forall z \in \text{dom } \varphi^*) \quad \langle z - p \mid y/\gamma - p \rangle \leq 0, \quad (3.55)$$

and we infer from (2.11) that  $p = P_{\text{dom } \varphi^*}(y/\eta)$ . Therefore,

$$\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = (0, y - \gamma P_{\text{dom } \varphi^*}(y/\eta)) = (0, y - P_{\gamma\text{dom } \varphi^*}y) \quad (3.56)$$

and we note that the condition  $q \neq 0$  means that  $y \notin \gamma \text{dom } \varphi^*$ . We provide below examples in which  $\text{dom } \varphi^*$  is a simple proper closed subset of  $\mathcal{G}$  and the proximity operator of the perspective function of  $\varphi$  can be computed explicitly.

**Example 3.11** Suppose that  $D \neq \{0\}$  is a nonempty closed convex cone in  $\mathcal{G}$  and define

$$\varphi = \vartheta + \iota_D, \quad \text{where} \quad \vartheta = \sqrt{1 + \|\cdot\|_{\mathcal{G}}^2}. \quad (3.57)$$

Since  $\text{dom } \vartheta = \mathcal{G}$ , we have  $\varphi^* = (\vartheta + \iota_D)^* = \vartheta^* \square \iota_{D^\ominus}$ , where  $D^\ominus$  is the polar cone of  $D$  and (combine [1, Examples 13.2(vi) and 13.7])

$$\vartheta^*: \mathcal{G} \rightarrow ]-\infty, +\infty] : u \mapsto \begin{cases} -\sqrt{1 - \|u\|_{\mathcal{G}}^2}, & \text{if } \|u\|_{\mathcal{G}} \leq 1; \\ +\infty, & \text{if } \|u\|_{\mathcal{G}} > 1. \end{cases} \quad (3.58)$$

Thus,  $\text{dom } \varphi^* = \text{dom } (\vartheta^* \square \iota_{D^\ominus}) = \text{dom } \vartheta^* + \text{dom } \iota_{D^\ominus} = B(0; 1) + D^\ominus$  is closed as the sum of two closed convex sets, one of which is bounded. As a result, since  $D^\ominus \neq \mathcal{G}$ ,

$$\text{dom } \varphi^* \text{ is a proper closed subset of } \mathcal{G}. \quad (3.59)$$

Now set  $\mathcal{K} = \mathbb{R} \oplus \mathcal{G}$  and  $K = [0, +\infty[ \times D$ , and let  $\gamma \in ]0, +\infty[$ ,  $\eta \in \mathbb{R}$ , and  $y \in \mathcal{G}$ . Then  $\|(\eta, y)\|_{\mathcal{K}} = \sqrt{|\eta|^2 + \|y\|_{\mathcal{G}}^2}$  and, as shown in [7, Example 3.5],

$$\tilde{\varphi} = \|\cdot\|_{\mathcal{K}} + \iota_K. \quad (3.60)$$

Hence, we derive from (2.19) that

$$\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = \begin{cases} (0, 0), & \text{if } \|P_K(\eta, y)\|_{\mathcal{K}} \leq \gamma; \\ \left(1 - \frac{\gamma}{\|P_K(\eta, y)\|_{\mathcal{K}}}\right) P_K(\eta, y), & \text{if } \|P_K(\eta, y)\|_{\mathcal{K}} > \gamma. \end{cases} \quad (3.61)$$

We thus obtain an explicit expression as soon as  $P_K$  is explicit although  $\text{dom } \varphi^*$  is not open. As an illustration, let  $N \geq 2$  be an integer, set  $\mathcal{G} = \mathbb{R}^{N-1}$ , let  $D = [0, +\infty[^{N-1}$ , and denote by  $\|\cdot\|_N$  the usual  $N$ -dimensional Euclidean norm. Then  $\varphi = \sqrt{1 + \|\cdot\|_{N-1}^2} + \iota_D$ ,  $K = [0, +\infty[^N$ , and (3.61) becomes

$$\text{prox}_{\gamma\tilde{\varphi}}(\eta, y) = \begin{cases} (0, 0), & \text{if } \|(\eta_+, y_+)\|_N \leq \gamma; \\ \left(1 - \frac{\gamma}{\|(\eta_+, y_+)\|_N}\right) (\eta_+, y_+), & \text{if } \|(\eta_+, y_+)\|_N > \gamma, \end{cases} \quad (3.62)$$

where  $\eta_+ = \max\{0, \eta\}$  and  $y_+$  is defined likewise componentwise.

The second example provides the proximity operator of the perspective function of the Huber function.

**Example 3.12 (perspective of the Huber function)** Following [7, Example 3.2], let  $\rho \in ]0, +\infty[$  and consider the perspective function

$$\tilde{\varphi}: \mathbb{R}^2 \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \rho|y| - \frac{\eta\rho^2}{2}, & \text{if } |y| > \eta\rho \text{ and } \eta > 0; \\ \frac{|y|^2}{2\eta}, & \text{if } |y| \leq \eta\rho \text{ and } \eta > 0; \\ \rho|y|, & \text{if } \eta = 0; \\ +\infty, & \text{if } \eta < 0 \end{cases} \quad (3.63)$$

of the Huber function

$$\varphi: \mathbb{R} \rightarrow ]-\infty, +\infty]: y \mapsto \begin{cases} \rho|y| - \frac{\rho^2}{2}, & \text{if } |y| > \rho; \\ \frac{|y|^2}{2}, & \text{if } |y| \leq \rho. \end{cases} \quad (3.64)$$

Then  $\varphi^* = |\cdot|^2/2 + \iota_{[-\rho, \rho]}$  and  $\text{dom } \varphi^*$  is therefore a proper closed subset of  $\mathbb{R}$ . In addition, (3.10) yields

$$C = \{(\mu, u) \in ]-\infty, 0] \times [-\rho, \rho] \mid \mu + |u|^2/2 \leq 0\}. \quad (3.65)$$

Now let  $\eta \in \mathbb{R}$ , let  $y \in \mathbb{R}$ , and set  $(\chi, q) = \text{prox}_{\gamma\tilde{\varphi}}(\eta, y)$ . Then the following hold:

- (i) If  $\eta + |y|^2/(2\gamma) \leq 0$  and  $|y| \leq \gamma\rho$ , then Theorem 3.1(i) yields  $(\chi, q) = (0, 0)$ .
- (ii) We have  $\chi = 0 \Leftrightarrow \eta/\gamma \leq -\rho^2/2$ . Hence, if  $\eta \leq -\gamma\rho^2/2$  and  $|y| > \gamma\rho$ , (3.56) yields  $(\chi, q) = (0, y - P_{[-\gamma\rho, \gamma\rho]}y) = (0, y - \gamma\rho \text{sign}(y))$ .
- (iii) If  $\eta > -\gamma\rho^2/2$  and  $|y| > \rho\eta + \gamma\rho(1 + \rho^2/2)$ , then  $(\eta/\gamma, y/\gamma) \in (-\rho^2/2, \rho \text{sign}(y)) + N_C(-\rho^2/2, \rho \text{sign}(y))$  and therefore  $P_C(\eta/\gamma, y/\gamma) = (-\rho^2/2, \rho \text{sign}(y))$ . Hence, (3.11) yields  $(\chi, q) = (\eta + \gamma\rho^2/2, y - \gamma\rho \text{sign}(y))$ .
- (iv) If  $\eta > -\gamma\rho^2/2$  and  $|y| \leq \rho\eta + \gamma\rho(1 + \rho^2/2)$ , then  $(\chi, q) = \text{prox}_{\gamma[|\cdot|^2/2]^\sim}(\eta, y)$  is obtained by setting  $v = 0$ ,  $\delta = 0$ , and  $\alpha = 2$  in Example 3.8.

The last example concerns the Vapnik loss function.

**Example 3.13 (perspective of the Vapnik function)** Following [7, Example 3.4], let  $\varepsilon \in ]0, +\infty[$  and consider the perspective function

$$\tilde{\varphi}: \mathbb{R}^2 \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} d_{[-\varepsilon\eta, \varepsilon\eta]}(y), & \text{if } \eta \geq 0; \\ +\infty, & \text{if } \eta < 0 \end{cases} \quad (3.66)$$

of the Vapnik  $\varepsilon$ -insensitive loss function [28]

$$\varphi = \max\{|\cdot| - \varepsilon, 0\}. \quad (3.67)$$

We have  $\varphi = d_{[-\varepsilon, \varepsilon]} = \iota_{[-\varepsilon, \varepsilon]} \square |\cdot|$  and therefore  $\varphi^* = \varepsilon |\cdot| + \iota_{[-1, 1]}$ . Furthermore, (3.10) becomes

$$C = \{(\mu, u) \in ]-\infty, 0] \times [-1, 1] \mid \mu + \varepsilon|u| \leq 0\}. \quad (3.68)$$

Now let  $\eta \in \mathbb{R}$ , let  $y \in \mathbb{R}$ , and set  $(\chi, q) = \text{prox}_{\gamma\varphi}(\eta, y)$ . Then the following hold:

- (i) If  $\eta + \varepsilon|y| \leq 0$  and  $|y| \leq \gamma$ , then Theorem 3.1(i) yields  $(\chi, q) = (0, 0)$ .
- (ii) We have  $\chi = 0 \Leftrightarrow \eta/\gamma \leq -\varepsilon$ . Hence, if  $\eta \leq -\gamma\varepsilon$  and  $|y| > \gamma$ , (3.56) yields  $(\chi, q) = (0, y - P_{[-\gamma, \gamma]}y) = (0, y - \gamma \text{sign}(y))$ .
- (iii) If  $\eta > -\gamma\varepsilon$  and  $|y| > \varepsilon\eta + \gamma(1 + \varepsilon^2)$ , then  $(\eta/\gamma, y/\gamma) \in (-\varepsilon, \text{sign}(y)) + N_C(-\varepsilon, \text{sign}(y))$  and therefore  $P_C(\eta/\gamma, y/\gamma) = (-\varepsilon, \text{sign}(y))$ . Hence, (3.11) yields  $(\chi, q) = (\eta + \gamma\varepsilon, y - \gamma \text{sign}(y))$ .
- (iv) If  $|y| > -\eta/\varepsilon$  and  $\varepsilon\eta \leq |y| \leq \varepsilon\eta + \gamma(1 + \varepsilon^2)$ , then  $P_C(\eta/\gamma, y/\gamma)$  coincides with the projection of  $(\eta/\gamma, y/\gamma)$  onto the half-space with outer normal vector  $(1, \varepsilon \text{sign}(y))$  and which has the origin on its boundary. As a result, (3.11) yields  $(\chi, q) = ((\eta + \varepsilon|y|)/(1 + \varepsilon^2), \varepsilon(\eta + \varepsilon|y|)\text{sign}(y)/(1 + \varepsilon^2))$ .
- (v) If  $\eta \geq 0$  and  $|y| \leq \varepsilon\eta$ , then  $P_C(\eta/\gamma, y/\gamma) = (0, 0)$  and (3.11) yields  $(\chi, q) = (\eta, y)$ .

## 4 Applications in high-dimensional statistics

Sections 2 and 3 provide a unifying framework to model a variety of problems around the notion of a perspective function. By applying the results of Section 3 in existing proximal algorithms, we obtain efficient methods to solve complex problems. To illustrate this point, we focus on a specific application area: high-dimensional regression in the statistical linear model.

### 4.1 Penalized linear regression

We consider the standard statistical linear model

$$z = Xb + \sigma e, \quad (4.1)$$

where  $z = (\zeta_i)_{1 \leq i \leq n} \in \mathbb{R}^n$  is the response,  $X \in \mathbb{R}^{n \times p}$  a design (or feature) matrix,  $b = (\beta_j)_{1 \leq j \leq p} \in \mathbb{R}^p$  a vector of regression coefficients,  $\sigma \in ]0, +\infty[$ , and  $e = (\varepsilon_i)_{1 \leq i \leq n}$  the noise vector; each  $\varepsilon_i$  is the realization of a random variable with mean zero and variance 1. Henceforth, we denote by  $X_{i \cdot}$  the  $i$ th row of  $X$  and by  $X_{\cdot j}$  the  $j$ th column of  $X$ . In the high-dimensional setting where  $p > n$ , a typical assumption about the regression vector  $b$  is sparsity. In this scenario, the Lasso [27] has become a fundamental tool for variable selection and predictive modeling. It is based on solving the penalized least-squares problem

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2n} \|Xb - z\|_2^2 + \lambda \|b\|_1, \quad (4.2)$$

where  $\lambda \in [0, +\infty[$  is a regularization parameter that aims at controlling the sparsity of the solution. The Lasso has strong performance guarantees in terms of support recovery, estimation, and predictive performance if one takes  $\lambda \propto \sigma \|X^\top e\|_\infty$ . In the high-dimensional setting, two shortcomings of the



Lasso are the introduction of bias in the final estimates due to the  $\ell^1$  norm and lack of knowledge about the quantity  $\sigma$  which necessitates proper tuning of  $\lambda$  via model selection strategies that is dependent on  $\sigma$ . Bias reduction can be achieved by using a properly weighted  $\ell^1$  norm, resulting in the adaptive Lasso [30] formulation

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2n} \|Xb - z\|_2^2 + \lambda \sum_{j=1}^p w_j |\beta_j|, \quad (4.3)$$

where the fixed weights  $w_j \in ]0, +\infty[$  are estimated from data. In [30], it was shown that, for suitable choices of  $w_j$ , the adaptive Lasso produces (asymptotically) unbiased estimates of  $b$ . One of the first methods to alleviate the  $\sigma$ -dependency of the Lasso has been the Sqrt-Lasso [2]. The Sqrt-Lasso problem is based on the formulation

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|Xb - z\|_2 + \lambda \|b\|_1. \quad (4.4)$$

This optimization problem can be cast as second order cone program (SOCP) [2]. The modification of the objective function can be interpreted as an (implicit) scaling of the Lasso objective function by an estimate  $\|Xb - z\|_2 / \sqrt{n}$  of  $\sigma$  [19], leading to

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2\sqrt{n}} \frac{\|Xb - z\|_2^2}{\frac{1}{\sqrt{n}} \|Xb - z\|_2} + \lambda \|b\|_1. \quad (4.5)$$

In [2], it was shown that the tuning parameter  $\lambda$  does not depend on  $\sigma$  in Sqrt-Lasso.

Alternative approaches rely on the idea of simultaneously and explicitly estimating  $b$  and  $\sigma$  from the data. The scaled Lasso [26], a robust hybrid of ridge and Lasso regression [23], and the TREX [19] are important instances. In the following, we will show that these estimators are based on perspective functions under the unifying statistical framework of concomitant estimation. We will introduce a novel family of estimators and show how the corresponding optimization problems can be solved using proximal algorithms. In particular, we will derive novel proximal algorithms for solving both the standard TREX and a novel generalized version of the TREX which includes the Sqrt-Lasso as special case.

## 4.2 Penalized concomitant M-estimators

In statistics, the task of simultaneously estimating a regression vector  $b$  and an additional model parameter is referred to as concomitant estimation. In [17], Huber introduced a generic method for formulating “maximum likelihood-type” estimators (or M-estimators) with a concomitant parameter from a convex criterion. Using our perspective function framework, we can extend this framework and introduce the class of penalized concomitant M-estimators defined through the convex optimization problem

$$\underset{\sigma \in \mathbb{R}, \tau \in \mathbb{R}, b \in \mathbb{R}^p}{\text{minimize}} \quad \sum_{i=1}^n \tilde{\varphi}_i(\sigma, X_i \cdot b - \zeta_i) + \sum_{j=1}^p \tilde{\psi}_j(\tau, a_j^\top b), \quad (4.6)$$

with concomitant variables  $\sigma$  and  $\tau$  under the assumptions outlined in Theorem 3.1 and in Section 3.2. Here,  $\varphi_i \in \Gamma_0(\mathbb{R})$ ,  $\psi_j \in \Gamma_0(\mathbb{R})$ , and  $a_j \in \mathbb{R}^p$ . The terms  $\tilde{\varphi}_i$  are data fitting terms and  $\tilde{\psi}_j$  are penalty terms. A prominent instance of this family of estimators is the scaled Lasso [26] formulation

$$\underset{b \in \mathbb{R}^p, \sigma \in ]0, +\infty[}{\text{minimize}} \quad \frac{1}{2n} \frac{\|Xb - z\|_2^2}{\sigma} + \frac{\sigma}{2} + \lambda \|b\|_1, \quad (4.7)$$

which yields estimates equivalent to the Sqrt-Lasso. Here, setting  $\varphi_i = |\cdot|^2/(2n) + 1/2$  and  $\psi_j = \lambda|\cdot|$  leads to the scaled (or concomitant) Lasso formulation (see Lemma 2.5, Corollary 3.5, and [21]). Other function choices result in well-known estimators. For instance, taking each  $\varphi_i$  to be the Huber function (see Example 3.12) and each  $\psi_j$  to be the Berhu (reversed Huber) function recovers the robust Lasso variant, introduced and discussed in [23]. Setting each  $\psi_j = \lambda|w_j \cdot|$  to be a weighted  $\ell^1$  component results in the ‘‘Huber + adaptive Lasso’’ estimator, analyzed theoretically in [18]. Note that for the latter two approaches, no dedicated optimization algorithms exist that can solve the corresponding optimization problem with provable convergence guarantees. Combining the proximity operators introduced here with proximal algorithms enables us to design such algorithms. To exemplify this powerful framework we focus next on a particular instance of a penalized concomitant M-estimator, the TREX estimator, and derive proximity operators and proximal algorithms.

### 4.3 Proximal algorithms for the TREX

The TREX [19] extends Sqrt-Lasso and scaled Lasso by taking into account the unknown noise distribution of  $e$ . Recalling that a theoretically desirable tuning parameter for the Lasso is  $\lambda \propto \sigma \|X^\top e\|_\infty$ , the TREX scales the Lasso objective by an estimate of this quantity, namely,

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad \frac{\|Xb - z\|_2^2}{\|X^\top(Xb - z)\|_\infty} + \alpha \|b\|_1. \quad (4.8)$$

The parameter  $\alpha > 0$  can be set to a constant value ( $\alpha = 1/2$  being the default choice). In [19], promising statistical results were reported where an approximate version of the TREX, with no tuning of  $\alpha$ , has been shown to be a valid alternative to the Lasso. A major technical challenge in the TREX formulation is the non-convexity of the optimization problem. In [3], this difficulty is overcome by showing that the TREX problem, although non-convex, can be solved by observing that problem (4.8) can be equivalently expressed as finding the best solution to  $2p$  convex problems of the form

$$\underset{\substack{b \in \mathbb{R}^p \\ x_j^\top(Xb - z) > 0}}{\text{minimize}} \quad \frac{\|Xb - z\|_2^2}{\alpha x_j^\top(Xb - z)} + \|b\|_1, \quad \text{where } x_j = sX_{:,j}, \quad \text{with } s \in \{-1, 1\}. \quad (4.9)$$

Each subproblem can be reformulated as a standard SOCP and numerically solved using generic SOCP solvers [3]. Next we show how our perspective function approach allows us to derive proximal algorithms for not only the TREX subproblems and but also for novel generalized versions of the TREX. The proximal algorithms construct a sequence  $(b_k)_{k \in \mathbb{N}}$  that is guaranteed to converge to a solution to (4.9).

#### 4.3.1 Proximal operators for the TREX subproblem

We first note that the data fitting term of the TREX subproblem (4.9) is the special case of (2.25) where  $\mathcal{H} = \mathbb{R}^p$ ,  $\mathcal{G} = \mathbb{R}^n$ ,  $q = 2$ ,  $L = X$ ,  $r = z$ ,  $u = X^\top x_j$ , and  $\rho = x_j^\top z$ . Given  $\alpha \in ]0, +\infty[$ , the data

fitting term of the TREX subproblem thus assumes the form

$$f_j: \mathbb{R}^p \rightarrow ]-\infty, +\infty]: b \mapsto \begin{cases} \frac{\|Xb - z\|_2^2}{\alpha x_j^\top (Xb - z)}, & \text{if } x_j^\top (Xb - z) > 0; \\ 0, & \text{if } Xb = z; \\ +\infty, & \text{otherwise,} \end{cases} \quad (4.10)$$

and the corresponding TREX subproblem is to

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad f_j(b) + \|b\|_1. \quad (4.11)$$

Now consider the linear transformation

$$M_j: \mathbb{R}^p \rightarrow \mathbb{R} \times \mathbb{R}^n: b \mapsto (x_j^\top Xb, Xb) \quad (4.12)$$

and introduce

$$g_j: \mathbb{R} \times \mathbb{R}^n \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \frac{\|y - z\|_2^2}{\alpha(\eta - x_j^\top z)}, & \text{if } \eta > x_j^\top z; \\ 0, & \text{if } y = z \text{ and } \eta = x_j^\top z; \\ +\infty, & \text{otherwise.} \end{cases} \quad (4.13)$$

Then  $f_j = g_j \circ M_j$ . Upon setting  $h = \|\cdot\|_1$ , we see that (4.11) is of the form

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \quad g_j(M_j b) + h(b). \quad (4.14)$$

Next, we determine the proximity operators  $\text{prox}_{g_j}$  and  $\text{prox}_h$ , as only those are needed in modern proximal splitting methods [8, 11] to solve (4.14). The proximity operator  $\text{prox}_h$  is the standard soft thresholding operator. A formula for  $\text{prox}_{g_j}$  is provided by Example 3.8 up to a shift by  $(x_j^\top z, z)$ . Let  $\gamma \in ]0, +\infty[$  and let  $g$  be as in (3.48). Combining Example 3.8 and [1, Proposition 23.29(ii)], we obtain, for every  $\eta \in \mathbb{R}$  and every  $y \in \mathbb{R}^n$ ,

$$\begin{aligned} \text{prox}_{\gamma g_j}(\eta, y) &= (x_j^\top z, z) + \text{prox}_{\gamma g}(\eta - x_j^\top z, y - z) \\ &= \begin{cases} (\eta + \alpha\gamma\|p\|_2^2/4, y - \gamma p), & \text{if } 4\gamma(\eta - x_j^\top z) + \alpha\|y - z\|_2^2 > 0; \\ (x_j^\top z, z), & \text{if } 4\gamma(\eta - x_j^\top z) + \alpha\|y - z\|_2^2 \leq 0, \end{cases} \end{aligned} \quad (4.15)$$

where

$$p = \begin{cases} \frac{t}{\|y - z\|} (y - z), & \text{if } y \neq z; \\ 0, & \text{if } y = z, \end{cases} \quad (4.16)$$

and where  $t$  is the unique solution in  $]0, +\infty[$  to the depressed cubic equation

$$s^3 + \frac{4\alpha(\eta - x_j^\top z) + 8\gamma}{\alpha^2\gamma} s - \frac{8\|y - z\|}{\alpha^2\gamma} = 0. \quad (4.17)$$

### 4.3.2 Proximal operators for generalized TREX estimators

Thus far, we have shown that the data-fitting function in the TREX subproblem (4.9) is a special case of (2.25). However, the full potential of (2.25) is revealed by taking a general  $q \in ]1, +\infty[$ , leading to the composite perspective function

$$f_{j,q}: \mathbb{R}^p \rightarrow ]-\infty, +\infty]: b \mapsto \begin{cases} \frac{\|Xb - z\|_2^q}{\alpha |x_j^\top (Xb - z)|^{q-1}}, & \text{if } x_j^\top (Xb - z) > 0; \\ 0, & \text{if } Xb = z; \\ +\infty, & \text{otherwise.} \end{cases} \quad (4.18)$$

This function is the data fitting term of a generalized TREX subproblem for the corresponding global generalized TREX objective

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} \frac{\|Xb - z\|_2^q}{\alpha \|X^\top (Xb - z)\|_\infty^{q-1}} + \|b\|_1. \quad (4.19)$$

This objective function provides a novel family of generalized TREX estimators, parameterized by  $q$ . The first important observation is that, in the limiting case  $q \rightarrow 1$ , the generalized TREX estimator collapses to the Sqrt-Lasso (4.4). Secondly, particular choices of  $q$  allow very efficient computation of proximity operators for the generalized TREX subproblems. Considering the linear transformation  $M_j: \mathbb{R}^p \rightarrow \mathbb{R} \times \mathbb{R}^n: b \mapsto (x_j^\top Xb, Xb)$  and introducing

$$g_{j,q}: \mathbb{R} \times \mathbb{R}^n \rightarrow ]-\infty, +\infty]: (\eta, y) \mapsto \begin{cases} \frac{\|y - z\|_2^q}{\alpha |\eta - x_j^\top z|^{q-1}}, & \text{if } \eta > x_j^\top z; \\ 0, & \text{if } y = z \text{ and } \eta = x_j^\top z; \\ +\infty, & \text{otherwise} \end{cases} \quad (4.20)$$

we arrive at  $f_{j,q} = g_{j,q} \circ M_j$ . Setting  $h = \|\cdot\|_1$  the corresponding problem is to

$$\underset{b \in \mathbb{R}^p}{\text{minimize}} g_{j,q}(M_j b) + h(b). \quad (4.21)$$

The proximity operator  $\text{prox}_{g_{j,q}}$  is provided by Example 3.7, where  $\delta = 0$  and  $v = 0$ , up to a shift by  $(x_j^\top z, z)$ . Let  $g$  be the function in (3.44) and let  $\gamma \in ]0, +\infty[$ . Set  $q^* = q/(q-1)$ , set  $\varrho = (\alpha(1-1/q^*))^{q^*-1}$ , and take  $(\eta, y) \in \mathbb{R} \times \mathcal{G}$ . If  $q^* \gamma^{q^*-1} (\eta - x_j^\top z) + \varrho \|y - z\|_2^{q^*} > 0$  and  $y \neq z$ , let  $t \in ]0, +\infty[$  be the unique solution to the polynomial equation

$$s^{2q^*-1} + \frac{q^*(\eta - x_j^\top z)}{\gamma \varrho} s^{q^*-1} + \frac{q^*}{\varrho^2} s - \frac{q^* \|y - z\|}{\gamma \varrho^2} = 0. \quad (4.22)$$

Set

$$p = \begin{cases} \frac{t}{\|y - z\|} (y - z), & \text{if } y \neq z; \\ 0, & \text{if } y = z. \end{cases} \quad (4.23)$$

Then we derive from Example 3.7 that

$$\text{prox}_{\gamma g_{j,q}}(\eta, y) = \begin{cases} (\eta + \gamma \varrho t^{q^*} / q^*, y - \gamma p), & \text{if } q^* \gamma^{q^*-1} (\eta - x_j^\top z) + \varrho \|y - z\|_2^{q^*} > 0; \\ (x_j^\top z, z), & \text{if } q^* \gamma^{q^*-1} (\eta - x_j^\top z) + \varrho \|y - z\|_2^{q^*} \leq 0. \end{cases} \quad (4.24)$$

The key step in the calculation of the proximity operator is to solve (4.22) efficiently. The solution is explicit for  $q = 2$ , as discussed in Example 3.8. For  $q = 3$ , we obtain a quartic equation that can also be solved explicitly. For  $q \in \{(i + 1)/i \mid i \in \mathbb{N}, i \geq 2\}$  (4.22) is a polynomial with integer exponents and is thus amenable to efficient root finding algorithms. For a general  $q$ , a one-dimensional line search for convex functions on a bounded interval needs to be performed.

### 4.3.3 Douglas-Rachford for generalized TREX subproblems

Problem (4.14) is a standard composite problem and can be solved via several proximal splitting methods that require only the ability to compute  $\text{prox}_{g_j}$  and  $\text{prox}_h$ ; see [6] and references therein. For large scale problems, one could also employ recent algorithms that benefit from block-coordinate [11] or asynchronous block-iterative implementations [8], while still guaranteeing the convergence of their sequence  $(b_k)_{k \in \mathbb{N}}$  of iterates to a solution to the problem. In this section, we focus on a simple implementation based on the Douglas-Rachford splitting method [1] in the context of the generalized TREX estimation to illustrate the applicability and versatility of the tools presented in Sections 2 and 3.

Define  $F: (b, c) \mapsto h(b) + g_{j,q}(c)$  and  $G = \iota_V$ , where  $V$  is the graph of  $M_j$ , i.e.,  $V = \{(b, c) \in \mathbb{R}^p \times \mathbb{R}^{n+1} \mid M_j b = c\}$ . Then we can rewrite (4.14) as

$$\underset{\mathbf{x}=(b,c) \in \mathbb{R}^p \times \mathbb{R}^{n+1}}{\text{minimize}} \quad F(\mathbf{x}) + G(\mathbf{x}) \quad (4.25)$$

Let  $\gamma \in ]0, +\infty[$ , let  $\mathbf{y}_0 \in \mathbb{R}^{p+n+1}$ , and let  $(\mu_k)_{k \in \mathbb{N}}$  be a sequence in  $]0, 2[$  such that  $\inf_{k \in \mathbb{N}} \mu_k > 0$  and  $\sup_{k \in \mathbb{N}} \mu_k < 2$ . The Douglas-Rachford algorithm is

$$\begin{aligned} & \text{for } k = 0, 1, \dots \\ & \left[ \begin{array}{l} \mathbf{x}_k = \text{prox}_{\gamma G} \mathbf{y}_k \\ \mathbf{z}_k = \text{prox}_{\gamma F}(2\mathbf{x}_k - \mathbf{y}_k) \\ \mathbf{y}_{k+1} = \mathbf{y}_k + \mu_k(\mathbf{z}_k - \mathbf{x}_k). \end{array} \right. \end{aligned} \quad (4.26)$$

The sequence  $(\mathbf{x}_k)_{k \in \mathbb{N}}$  is guaranteed to converge to a solution to (4.25) [1, Corollary 27.4]. Note that

$$\text{prox}_F: (b, c) \mapsto (\text{prox}_h b, \text{prox}_{g_{j,q}} c) \quad (4.27)$$

and, in view of (2.15),

$$\text{prox}_G: (b, c) \mapsto (v, M_j v), \quad \text{where } v = b - M_j^\top (\text{Id} + M_j M_j^\top)^{-1} (M_j b - c) \quad (4.28)$$

is the projection operator onto  $V$ . Hence, upon setting  $R_j = M_j^\top (\text{Id} + M_j M_j^\top)^{-1}$ ,  $\mathbf{x}_k = (b_k, c_k) \in \mathbb{R}^p \times \mathbb{R}^{n+1}$ ,  $\mathbf{y}_k = (x_k, y_k) \in \mathbb{R}^p \times \mathbb{R}^{n+1}$ , and  $\mathbf{z}_k = (z_k, t_k) \in \mathbb{R}^p \times \mathbb{R}^{n+1}$ , we can rewrite (4.26) as

$$\begin{aligned} & \text{for } k = 0, 1, \dots \\ & \left[ \begin{array}{l} q_k = M_j x_k - y_k \\ b_k = x_k - R_j q_k \\ c_k = M_j b_k \\ z_k = \text{prox}_{\gamma h}(2b_k - x_k) \\ t_k = \text{prox}_{\gamma g_{j,q}}(2c_k - y_k) \\ x_{k+1} = x_k + \mu_k(z_k - b_k) \\ y_{k+1} = y_k + \mu_k(t_k - c_k). \end{array} \right. \end{aligned} \quad (4.29)$$

Then  $(b_k)_{k \in \mathbb{N}}$  converges to a solution  $b$  to (4.14) or (4.21). Note that the matrix  $R_j$  needs to be precomputed only once by inverting a positive definite symmetric matrix.

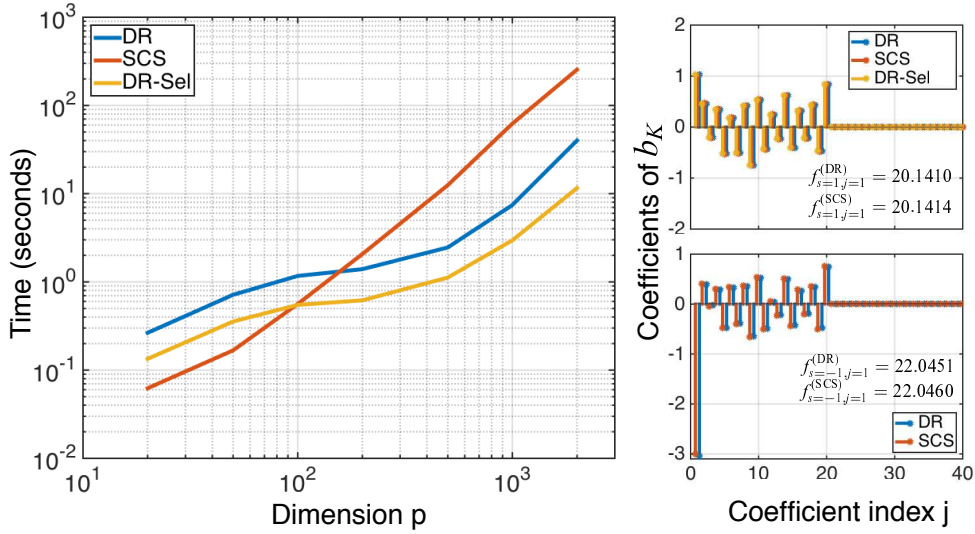


Figure 1: Left panel: Average wall-clock time (seconds) versus dimension  $p$  for solving the TREX subproblems with Douglas-Rachford (DR), SCS, and DR-Sel (Douglas-Rachford with online sign selection). Right panel: Both plots show the first 40 variables of a typical  $p = 2000$  TREX solution (top for  $s = +1$ ). The  $m = 20$  first indices are the non-zero indices in  $b^*$ . Insets show the TREX subproblem objective function values for  $s = \pm 1$  and  $X_{\cdot 1}$ , reached by Douglas-Rachford and SCS. DR-Sel selects the correct signed subproblem as verified a posteriori by the minimum function value ( $f_{s=1, j=1}^{(DR)} = 20.1410$  versus  $f_{s=-1, j=1}^{(DR)} = 22.0451$ ).

## 4.4 Numerical illustrations

We illustrate the convergence behavior of the Douglas-Rachford algorithm for TREX problems and the statistical performance of generalized TREX estimators using numerical experiments. All presented algorithms and experimental evaluations are implemented in MATLAB and are available at <http://github.com/muellsen/TREX>. All algorithms are run in MATLAB 2015a on a MacBook Pro with 2.8 GHz Intel Core i7 and 16 GB 1600 MHz DDR3 memory.

### 4.4.1 Evaluation of the Douglas-Rachford scheme on TREX subproblems

We first examine the scaling behavior of the Douglas-Rachford scheme for the TREX subproblem on linear regression tasks. We simulate synthetic data according to the linear model (4.1) with  $m = 20$  nonzero variables, regression vector  $b^* = [-1, 1, -1, \dots, 0_{p-m}^\top]^\top$ , and feature vectors  $X_{i\cdot} \sim N(0, \Sigma)$  with  $\Sigma_{ii} = 1$  and  $\Sigma_{ij} = 0.3$ , and Gaussian noise  $\varepsilon_i \sim N(0, \sigma^2)$  with  $\sigma = 1$ . Each column  $X_{\cdot j}$  is normalized to have norm  $\sqrt{n}$ . We fix the sample size  $n = 200$  and consider the dimension  $p \in \{20, 50, 100, 200, 500, 1000, 2000\}$ . We solve one standard TREX subproblem (for  $s \in \{-1, 1\}$ ,  $X_{\cdot 1}$ ,  $\alpha = 0.5$ ) over  $d = 20$  random realizations of  $X$  and  $e$ . For the TREX subproblem we consider the proximal Douglas-Rachford algorithm 4.29 with parameters  $\mu_k \equiv 1.95$  and  $\gamma = 70$ . We declare that the Douglas-Rachford algorithm has converged at iteration  $K$  if  $\min\{\|b_{K+1} - b_K\|, \|y_{K+1} - y_K\|\} \leq 10^{-10}$ , resulting in the final estimate  $b_K$ .

In practice, the Douglas-Rachford algorithm for the TREX subproblem can be enhanced by an

online sign selection rule (DR-Sel). When a TREX subproblem for fixed  $X_{:,j}$  is considered, we can solve the problem for  $s \in \{-1, 1\}$  concurrently for a small number  $k_0$  of iterations (standard setting  $k_0 = 50$ ) and select the signed optimization problem with best progress in terms of objective function value.

We compare the run time scaling and solution quality of Douglas-Rachford and DR-Sel with those of the state-of-the-art Splitting Conic Solver (SCS). SCS is a general-purpose first-order proximal method that provides numerical solutions to several standard classes of optimization problems, including SOCPs and Semidefinite Programs (SDPs). We use SCS in indirect mode [22] to solve the SOCP formulation of the TREX subproblem [3] with convergence tolerance  $10^{-4}$ .

The run time scaling results are shown in Figure 1. We emphasize that the scaling experiments are not meant to measure absolute algorithmic performance but rather efficiency with respect to optimization formulations that are subsequently solved by proximal algorithms. We observe that SCS with the SOCP formulation of TREX compares favorably with Douglas-Rachford and DR-Sel in low dimensions while, for  $p > 200$ , both Douglas-Rachford variants perform better. DR-Sel outperforms Douglas-Rachford by a factor of 2 to 4 and always selects the correct signed subproblem (data not shown). The TREX solutions found by SCS and Douglas-Rachford are close in terms of  $\|b^{(DR)} - b^{(SCS)}\|$ , with DR typically reaching slightly lower function values than SCS. Values for the first 40 dimensions of a typical solution  $b_K$  in  $p = 2000$  dimensions are shown in Figure 1 (right panels).

#### 4.4.2 Behavior of generalized TREX estimators

We next study the effect of the exponent  $q$  on the statistical behavior of the generalized TREX estimator. We use the synthetic setting outlined in [29] to study the phase transition behavior of the different generalized TREX estimators. We generate data from the linear model (4.1) with  $p = 64$  and  $m = \lceil 0.4p^{3/4} \rceil$  nonzero variables, regression vector  $b^* = [-1, 1, -1, \dots, 0_{p-m}^\top]^\top$ , and feature vectors  $X_{i,:} \sim N(0, \Sigma)$  with  $\Sigma_{ii} = 1$  and  $\Sigma_{ij} = 0$  and Gaussian noise  $e$  with  $\sigma = 0.5$ . Each column  $X_{:,j}$  is normalized to have norm  $\sqrt{n}$ . We define the rescaled sample size according to  $\theta(n, p, m) = n/(2m \log(p - m))$  and consider  $\theta(n, p, m) \in \{0.2, 0.4, \dots, 1.6\}$ . At  $\theta(n, p, m) = 1$ , the probability of exact recovery of the support of  $b^*$  is 0.5 for the (Sqrt)-Lasso with oracle regularization parameter [29]. We consider the generalized TREX with different exponents  $q \in \{9/8, 7/6, 3/2, 2\}$  and the Sqrt-Lasso as limiting case  $q = 1$ . For all generalized TREX estimators we consider regularization parameters  $\alpha \in \{0.1, 0.15, \dots, 2\}$ . For Sqrt-Lasso we consider the standard regularization path setting outlined in [21]. We solve all generalized TREX problems with the Douglas-Rachford scheme using the previously described parameter and convergence settings. We measure the probability of exact support recovery and Hamming distance to the true support over  $d = 12$  repetitions. We threshold all “numerical zeros” in the generalized TREX solutions vectors at level 0.05. For all solutions closest to the true support in terms of Hamming distance, we also calculate estimation error  $\|b_K - b^*\|_2^2/n$  and prediction error  $\|Xb_K - Xb^*\|_2^2/n$ . Figure 2 shows average performance results across all repetitions. We observe several interesting phenomena for the family of generalized TREX estimators. In terms of exact recovery, the performance is slightly better than predicted by theory (see gray dashed line in Figure 2 top left panel), with decrease in performance for increasing  $q$ . This is also consistent with average Hamming distance measurements (top right panel). We observe that generalized TREX oracle solutions (according to the minimum Hamming distance criterion) show best performance in terms of estimation and prediction error for exponents  $q \in \{9/8, 7/6\}$ , followed by  $q \in \{3/2, 2\}$ .

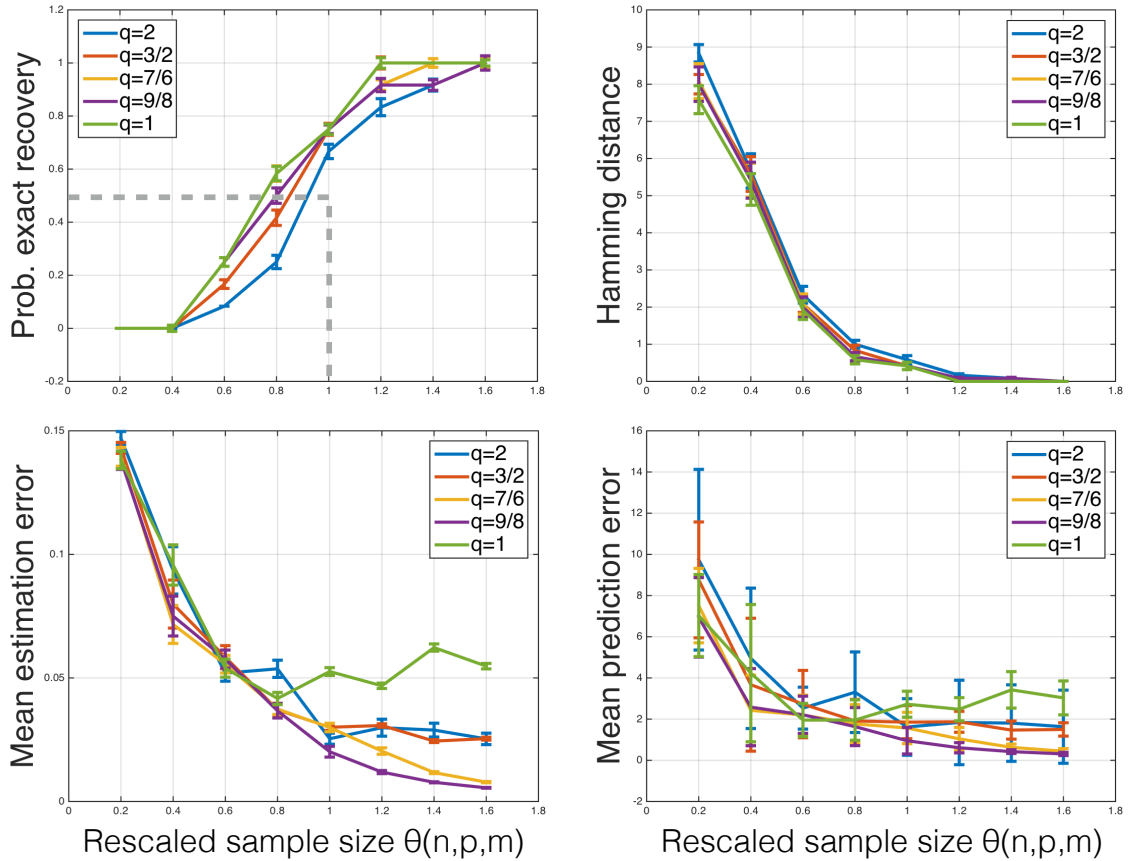


Figure 2: Top row: Probability (and standard error) of exact support recovery versus rescaled sample size  $\theta(n, p, m)$  for generalized TREX with  $q \in \{1, 9/8, 7/6, 3/2, 2\}$ ; top right panel: Average Hamming distance to true support. Bottom row: Mean estimation error  $\|b_K - b^*\|_2^2/n$  (left panel) and mean prediction error  $\|Xb_K - Xb^*\|_2^2/n$  (right panel).

The present numerical experiments highlight the usefulness of the family of generalized TREX estimators for sparse linear regression problems. Further theoretical research is needed to derive asymptotic properties of generalized TREX. A central prerequisite for establishing generalized TREX as statistical estimator is to solve the underlying optimization problem with provable guarantees. We have shown that our perspective function framework along with efficient computation of proximity operators enables this important task in a seamless way.

**Acknowledgement.** We thank Dr. Jacob Bien for valuable discussions. The Simons Foundation is acknowledged for partial financial support of this research. The work of P. L. Combettes was also partially supported by the CNRS MASTODONS project under grant 2016TABASCO.



## References

- [1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2011.
- [2] A. Belloni, V. Chernozhukov, and L. Wang, Square-root lasso: pivotal recovery of sparse signals via conic programming, *Biometrika*, vol. 98, pp. 791–806, 2011.
- [3] J. Bien, I. Gaynanova, J. Lederer, and C. L. Müller, Non-convex global minimization and false discovery rate control for the TREX, 2016. <http://arxiv.org/abs/1604.06815>
- [4] G. Birkhoff and S. Mac Lane, *A Survey of Modern Algebra*, 4th edition. Macmillan, New York, 1977.
- [5] J. M. Borwein, A. S. Lewis, and D. Noll, Maximum entropy reconstruction using derivative information, part 1: Fisher information and convex duality, *Math. Oper. Res.*, vol. 21, pp. 442–468, 1996.
- [6] P. L. Combettes, Systems of structured monotone inclusions: Duality, algorithms, and applications, *SIAM J. Optim.*, vol. 23, pp. 2420–2447, 2013.
- [7] P. L. Combettes, Perspective functions: Properties, constructions, and examples, 2016. <https://arxiv.org/abs/1610.01552>
- [8] P. L. Combettes and J. Eckstein, Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions, *Math. Programming*, published online 2016-07-05.
- [9] P. L. Combettes and J.-C. Pesquet, Proximal thresholding algorithm for minimization over orthonormal bases, *SIAM J. Optim.*, vol. 18, pp. 1351–1376, 2007.
- [10] P. L. Combettes and J.-C. Pesquet, Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, (H. H. Bauschke et al., eds), pp. 185–212. Springer, New York, 2011.
- [11] P. L. Combettes and J.-C. Pesquet, Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping, *SIAM J. Optim.*, vol. 25, pp. 1221–1248, 2015.
- [12] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.*, vol. 4, pp. 1168–1200, 2005.
- [13] M. El Gheche, G. Chierchia, and J.-C. Pesquet, Proximity operators of discrete information divergences, 2016. <https://arxiv.org/pdf/1606.09552v1.pdf>
- [14] F. Facchinei and J.-S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer-Verlag, New York, 2003.
- [15] R. A. Fisher, Theory of statistical estimation, *Proc. Cambridge. Philos. Soc.*, vol. 22, pp. 700–725, 1925.
- [16] B. R. Frieden and R. A. Gatenby (eds.), *Exploratory Data Analysis Using Fisher Information*. Springer, New York, 2007.
- [17] P. J. Huber, *Robust Statistics*, 1st ed. Wiley, New York, 1981.
- [18] S. Lambert-Lacroix and L. Zwald, Robust regression through the Huber’s criterion and adaptive lasso penalty, *Electron. J. Stat.*, vol. 5, pp. 1015–1053, 2011.
- [19] J. Lederer and C. L. Müller, Don’t fall for tuning parameters: Tuning-free variable selection in high dimensions with the TREX, *Proc. Twenty-Ninth AAAI Conf. Artif. Intell.*, pp. 2729–2735. AAAI Press, Austin, 2015.
- [20] J. J. Moreau, Fonctions convexes duales et points proximaux dans un espace hilbertien, *C. R. Acad. Sci. Paris Sér. A Math.*, vol. 255, pp. 2897–2899, 1962.
- [21] E. Ndiaye, O. Fercoq, A. Gramfort, V. Leclère, and J. Salmon, Efficient smoothed concomitant lasso estimation for high dimensional regression, 2016. <https://arxiv.org/pdf/1606.02702v1.pdf>

- [22] B. O’Donoghue, E. Chu, N. Parikh, and S. Boyd, Conic optimization via operator splitting and homogeneous self-dual embedding, *J. Optim. Theory Appl.*, vol. 169, pp. 1042–1068, 2016.
- [23] A. B. Owen, A robust hybrid of lasso and ridge regression, *Contemp. Math.*, vol. 443, pp. 59–71, 2007.
- [24] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
- [25] S. Sra, S. Nowozin, and S. J. Wright, *Optimization for Machine Learning*. MIT Press, Cambridge, MA, 2012.
- [26] T. Sun and C. Zhang, Scaled sparse linear regression, *Biometrika*, vol. 99, pp. 879–898, 2012.
- [27] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc.*, vol. B58, pp. 267–288, 1996.
- [28] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. Springer, New York, 2000.
- [29] M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using  $\ell_1$ -constrained quadratic programming (Lasso), *IEEE Trans. Inform. Theory*, vol. 55, pp. 2183–2202, 2009.
- [30] H. Zou, The adaptive lasso and its oracle properties, *J. Amer. Stat. Assoc.*, vol. 101, pp. 1418–1429, 2006.