SIAM J. OPTIM.
Vol. 23, No. 3, pp. 1689–1717

VARIATIONAL PROPERTIES OF VALUE FUNCTIONS*

ALEKSANDR Y. ARAVKIN†, JAMES V. BURKE‡, AND MICHAEL P. FRIEDLANDER§

Abstract. Regularization plays a key role in a variety of optimization formulations of inverse problems. A recurring theme in regularization approaches is the selection of regularization parameters and their effect on the solution and on the optimal value of the optimization problem. The sensitivity of the value function to the regularization parameter can be linked directly to the Lagrange multipliers. This paper characterizes the variational properties of the value functions for a broad class of convex formulations, which are not all covered by standard Lagrange multiplier theory. An inverse function theorem is given that links the value functions of different regularization formulations (not necessarily convex). These results have implications for the selection of regularization parameters, and the development of specialized algorithms. Numerical examples illustrate the theoretical results.

Key words. convex optimization, optimal value functions, Lagrange multipliers, inverse problems

AMS subject classifications. 65K05, 65K10, 90C25

DOI. 10.1137/120899157

1. Introduction.

It is well known that there is a close connection between the sensitivity of the optimal value of a parametric optimization problem and its Lagrange multipliers. Consider the family of feasible convex optimization problems

$$\mathcal{P}(b,\tau) \qquad \underset{r,x}{\text{minimize}} \quad \rho(r) \quad \text{subject to} \quad Ax + r = b, \quad \phi(x) \le \tau,$$

where $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, and the functions $\phi : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$ and $\rho : \mathbb{R}^m \to \overline{\mathbb{R}}$ are closed, proper, and convex and continuous relative to their domains. The value function

$$v(b,r) := \inf_{r,x} \{\rho(r) \,|Ax + r = b,\ \phi(x) \le \tau\}$$

gives the optimal objective value of problem $\mathcal{P}(b,\tau)$ for fixed parameters b and τ. If $\mathcal{P}(b,\tau)$ is a feasible *ordinary convex program* [34, section 28], then under standard hypotheses the subdifferential of v is the set of pairs (u,μ), where $u \in \mathbb{R}^m$ and $\mu \in \mathbb{R}$ are the Lagrange multipliers of $\mathcal{P}(b,\tau)$ corresponding to the equality and inequality constraints, respectively. This connection is extensively explored in Rockafeller's 1993 survey paper [35].

If we allow ϕ to take on infinite values on the domain of the objective—which can occur, for example, if ϕ is an arbitrary gauge—then $\mathcal{P}(b,\tau)$ is no longer an ordinary convex program, and so the standard Lagrange multiplier theory does not apply.

*Received by the editors November 16, 2012; accepted for publication (in revised form) May 24, 2013; published electronically August 22, 2013.
http://www.siam.org/journals/siopt/23-3/89915.html

†IBM T.J. Watson Research Center, Yorktown Heights, NY 10598 (saravkin@us.ibm.com). The work of this author was partially supported by an NSERC CRD grant through the University of British Columbia.

‡Department of Mathematics, University of Washington, Seattle, WA 98195 (burke@math.washington.edu).

§Department of Computer Science, University of British Columbia, Vancouver B.C., V6T 1Z4, Canada (mpf@cs.ubc.ca). The work of this author was partially supported by NSERC Discovery grant 312104.

Multiplier theories that do apply to more general contexts can be found in [21, 16, 45, 8]. Remarkably, even in this general setting, it is possible to obtain explicit formulas of the subdifferential of the value function v useful in many applications.

1.1. Examples. We give two simple examples that illustrate the need for the extended Lagrange multiplier theory. Both are of the form

$$\underset{x}{\text{minimize}} \quad \tfrac{1}{2}\|Ax-b\|_2^2 \quad \text{subject to} \quad \gamma(x \mid U) \leq 1, \tag{1.1}$$

where

$$\gamma(x \mid U) := \inf\{\lambda \geq 0 \,|\, x \in \lambda U\}$$

is the *gauge function* for the nonempty closed convex set $U \subset \mathbb{R}^n$, which contains 0. Let $A = I$ and $b = (0,-1)^T$. Then the solution to (1.1) is just the 2-norm projection onto the set $\{x \,|\, \gamma(x \mid U) \leq 1\} = U$.

For our first example, we consider the set

$$U = \left\{ x \in \mathbb{R}^2 \,\middle|\, \tfrac{1}{2}x_1^2 \leq x_2 \right\},$$

defined in [34, section 10]. The gauge for this set is an example of a closed, proper, and convex function that is not locally bounded and therefore not continuous at a point in its effective domain. It is straightforward to show that

$$\gamma(x \mid U) = \begin{cases} \frac{x_1^2}{2x_2}, & x_2 > 0, \\ 0, & x_1 = 0 = x_2, \\ +\infty, & \text{otherwise.} \end{cases}$$

The constraint region for (1.1) is the set U, and the unique global solution is the point $x = 0$. However, since $0 = \gamma(0 \mid U) < 1$, the classical Lagrange multiplier theory fails: the solution is on the boundary of the feasible region, and yet no classical Lagrange multiplier exists. The problem is that the constraint is active at the solution but not active in the functional sense, i.e., $\gamma(0 \mid U) < 1$. In contrast, the extended multiplier theory of [45, Theorem 2.9.3] succeeds with the multiplier choice of 0.

For the second example, take $U = \mathbb{B}_2 \cap K$, where $\mathbb{B}_2$ is the unit ball associated with the Euclidean norm on $\mathbb{R}^2$. Then $\gamma(x \mid \mathbb{B}_2 \cap K) = \|x\|_2 + \delta(x \mid K)$, and the constraint region for (1.1) is the set $\mathbb{B}_2 \cap K$. Set $K = \{(x_1, x_2) \,|\, x_2 \geq 0\}$. Again, the origin is the unique global solution to this optimization problem, and no classical Lagrange multiplier for this problem exists.

In both of these examples, the multiplier theory in [45] can be applied to obtain a Lagrange multiplier theorem. In Theorem 5.2, we extend this theory and provide a characterization of these Lagrange multipliers that is useful in computation.

1.2. Formulations. Appropriate definitions of the functions ρ and ϕ can be used to represent a range of practical problems. Choosing ρ to be the 2-norm and ϕ to be any norm yields the canonical regularized least-squares problem

$$\underset{x,\,r}{\text{minimize}} \quad \|r\|_2 \quad \text{subject to} \quad Ax + r = b,\ \|x\| \leq \tau, \tag{1.2}$$

which optimizes the misfit between the data b and the forward model Ax, subject to keeping x appropriately bounded in some norm. The 2-norm constraint on x yields a Tikhonov regularization, popular in many inversion applications. A 1-norm constraint on x yields the Lasso problem [41], often used in sparse recovery and model-selection applications. Interestingly, when the optimal residual r is nonzero, the value function for this family of problems is always differentiable in both b and τ with

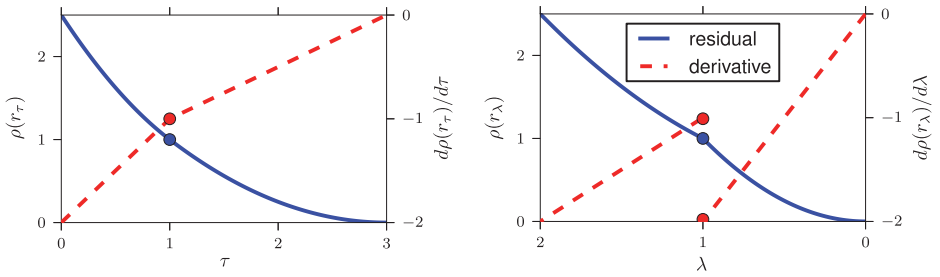

FIG. 1.1. *The misfit $\rho(r)$ (solid line) and its derivative (dashed line) as a function of the regularization parameter for a 1-norm regularized example. The left panel shows the constrained formulation $\mathcal{P}(b,\tau)$, which varies smoothly with τ; the right panel shows that the penalized formulation does not vary smoothly with λ (note the reversed axis).*

$$\nabla v(b,\tau) = \left(\frac{r}{\|r\|_2}, \ \frac{\|A^T r\|_*}{\|r\|_2} \right),$$

where $\|\cdot\|_*$ is the norm dual to $\|\cdot\|$. This gradient is derived by van den Berg and Friedlander [10, Theorem 2.2]. The analysis of the sensitivity in τ of the value function for the Lasso problem led to the development of the SPGL1 solver [9], currently used in a variety of sparse inverse problems, with particular success in large-scale sparse inverse problems [27]. A subsequent analysis [12] that allows $\phi(x)$ to be a gauge paved the way for other applications, such as group-sparsity promotion [11].

An alternative to $\mathcal{P}(b,\tau)$ is the class of penalized formulations

$$\mathcal{P}_{\text{L}}(b,\lambda) \qquad \underset{x}{\text{minimize}} \quad \rho(b - Ax) + \lambda\phi(x).$$

(The subscript "L" in the label reminds us that it can be interpreted as a Lagrangian of the original problem.) The nonnegative regularization parameter λ is used to control the tradeoff between the data misfit ρ and the regularization term ϕ. For example, talking $\rho(r) = \|r\|_2$ and $\phi(x) = \|x\|$ yields a formulation analogous to (1.2). This penalized formulation is commonly used in applications of Bayesian parametric regression [31, 37, 30, 42, 44], inference problems on dynamic linear systems [1, 15], feature selection, selective shrinkage, and compressed sensing [25, 20, 19], robust formulations [29, 24, 2, 23], support-vector regression [43, 26], classification [22, 33, 39], and functional reconstruction [6, 38, 17].

From an algorithmic point of view, the unconstrained formulation $\mathcal{P}_{\text{L}}(b,\lambda)$ may be preferable. However, the constrained formulation $\mathcal{P}(b,\tau)$ has the distinction that its value function $v(b,\tau)$ is jointly convex in its parameters; see section 1.3. In contrast, the optimal value of the penalized formulation $\mathcal{P}_{\text{L}}(b,\lambda)$ is not in general a convex function of its parameters. The following simple example

$$\rho(r) = \tfrac{1}{2}\|r\|_2^2, \quad \phi(x) = \|x\|_1, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

illustrates this situation. The optimal values of ρ in the formulations $\mathcal{P}(b,\tau)$ and $\mathcal{P}_{\text{L}}(b,\lambda)$, as functions of τ and λ, respectively, are given by

$$\rho_\tau = \begin{cases} \frac{1}{2} + \frac{1}{2}(\tau - 2)^2 & \text{for } \tau \in [0,1), \\ \frac{1}{4}(\tau - 3)^2 & \text{for } \tau \in [1,3), \\ 0 & \text{otherwise;} \end{cases} \qquad \rho_\lambda = \begin{cases} \lambda^2 & \text{for } \lambda \in [0,1), \\ \frac{1}{2} + \frac{1}{2}\lambda^2 & \text{for } \lambda \in [1,2), \\ 5/2 & \text{otherwise.} \end{cases}$$

The optimal values and their derivatives are shown in Figure 1.1, where it is clear that ρ_τ is convex (and in this case also smooth) in τ, but ρ_λ is not convex in λ.

The admissibility of variational analysis and convexity of the value function may convince some practitioners to explore formulations of type $\mathcal{P}(b,\tau)$ rather than $\mathcal{P}_{\text{L}}(b,\lambda)$. In fact, we give an example (in section 7) of how this variational information can be used for algorithm design in the context of large-scale inverse problems.

1.3. Approach. For many practical inverse problems, the formulation of primary interest is the residual-constrained formulation

$$\mathcal{P}_{\text{R}}(b,\sigma) \qquad \underset{x}{\text{minimize}} \quad \phi(x) \quad \text{subject to} \quad \rho(b-Ax) \le \sigma,$$

(the subscript "R" reminds us that this formulation reverses the objective and constraint functions from that of $\mathcal{P}(b,\tau)$) in part because estimates of a tolerance level σ on fitting the error $\rho(b-Ax)$ are more easily available than estimates of a bound on the penalty parameter on the regularization ϕ; cf. $\mathcal{P}_{\text{L}}(b,\lambda)$. However, the formulation $\mathcal{P}(b,\tau)$ can sometimes be easier to solve. The underlying numerical theme is to develop methods for solving $\mathcal{P}_{\text{R}}(b,\sigma)$ that use a sequence of solutions to the possibly easier problem $\mathcal{P}(b,\tau)$.

In section 2, we present an inverse function theorem for value functions that characterizes the relationship between $\mathcal{P}(b,\tau)$ and $\mathcal{P}_{\text{R}}(b,\sigma)$ and applies more generally to nonconvex problems. Pairs of problems of this type are classical, though typically paired in a max-min fashion. For example, the isoperimetric inequality and Queen Dido's problem are of this type; the greatest area surrounded by a curve of given length is related to the problem of finding the curve of least arc length surrounding a given area. (See [40] for a modern survey.) The Markowitz mean-variance portfolio theory is also based on such a pairing; minimizing volatility subject to a lower bound on expected return is related to maximizing expected return subject to an upper bound on volatility [32].

The application motivating our investigation is establishing conditions under which it is possible to implement a root-finding approach for the nonlinear equation

$$\text{find } \tau \text{ such that } v(b,\tau) = \sigma, \tag{1.3}$$

where $\mathcal{P}_{\text{R}}(b,\sigma)$ can be solved via a sequence of approximate solutions of $\mathcal{P}(b,\tau)$. This generalizes the approach used by van den Berg and Friedlander [10, 12] for large-scale sparse optimization applications. The convex case is especially convenient, because both value functions are decreasing and convex. When the value function is differentiable, Newton's method is globally monotonic and locally quadratic. In section 5 we establish the variational properties (including conditions necessary for differentiability) of $\mathcal{P}(b,\tau)$.

In section 4 we derive dual representations of $\mathcal{P}(b,\tau)$ and their optimality conditions. These are used in section 5 to characterize the variational properties of the value function v. The conjugate, horizon, and perspective functions arise naturally as part of the analysis, and we present a calculus (section 3) for these functions that allows explicit computation of the subdifferential of v for large classes of misfit functions ρ and regularization functions ϕ (see section 6).

One of the motivating problems for the general analysis and methods we present is the treatment of a robust misfit function ρ (such as the popular Huber penalty) in the context of sparsity promotion, which typically involves a nonsmooth regularizer ϕ. In

section 7 we demonstrate that the sensitivity analysis can be applied to solve a sparse nonnegative denoising problem with convex and nonconvex robust misfit measures.

The proofs of all of the results are relegated to the appendix (section 8).

1.4. Notation. For a matrix $A \in \mathbb{R}^{m\times n}$, the image and inverse image of the sets E and F, respectively, are given by the sets

$$AE = \{y \,|y = Ax,\, x \in E\} \quad \text{and} \quad A^{-1}F = \{x \,|Ax \in F\}.$$

For a function $p : \mathbb{R}^n \to \overline{\mathbb{R}}$, its epigraph is denoted epi $p = \{(x,\mu)\,|p(x) \le \mu\}$, and its level set is denoted $\text{lev}_p(\tau) = \{x\,|p(x) \le \tau\}$. The function p is said to be proper if $\text{dom}\, p \neq \emptyset$ and closed if epi p is a closed set. The function $\delta(x \mid X)$ is the indicator to a set X, i.e., $\delta(x \mid X) = 0$ if $x \in X$ and $\delta(x \mid X) = +\infty$ if $x \notin X$.

2. An inverse function theorem for optimal value functions. Let $\psi_i : X \subseteq \mathbb{R}^n \to \overline{\mathbb{R}}$, $i \in \{1,2\}$, be arbitrary scalar-valued functions, and consider the following pair of related problems and their associated value functions:

$$\mathcal{P}_{1,2}(\tau) \qquad v_1(\tau) := \inf_{x\in X}\ \psi_1(x) + \delta\left((x,\tau) \mid \text{epi}\ \psi_2\right),$$

$$\mathcal{P}_{2,1}(\sigma) \qquad v_2(\sigma) := \inf_{x\in X}\ \psi_2(x) + \delta\left((x,\sigma) \mid \text{epi}\ \psi_1\right).$$

This pair corresponds to the problems $\mathcal{P}(b,\tau)$ and $\mathcal{P}_{\text{R}}(b,\sigma)$, defined in section 1, with the identifications

$$\psi_1(x) = \rho(b - Ax) \quad \text{and} \quad \psi_2(x) = \phi(x).$$

Our goal in this section is to establish general conditions under which the value functions v_1 and v_2 satisfy the inverse-function relationship

$$v_1 \circ v_2 = \text{id},$$

and for which the the pair of problems $\mathcal{P}_{1,2}(\tau)$ and $\mathcal{P}_{2,1}(\sigma)$ have the same solution sets. The pair of problems $\mathcal{P}(b,\tau)$ and $\mathcal{P}_{\text{R}}(b,\sigma)$ always satisfy the conditions of the next theorem, which applies to functions that are not necessarily convex.

THEOREM 2.1. *Let $\psi_i : X \subseteq \mathbb{R}^n \to \overline{\mathbb{R}}$, $i \in \{1,2\}$, be as defined in $\mathcal{P}_{1,2}(\tau)$, and define*

$$\mathcal{S}_{1,2} := \left\{\tau \in \overline{\mathbb{R}}\,|\emptyset \neq \arg\min \mathcal{P}_{1,2}(\tau) \subset \{x \in X\,|\psi_2(x) = \tau\}\right\}.$$

Let $\mathcal{S}_{2,1}$ be defined symmetrically to $\mathcal{S}_{1,2}$ by interchanging the roles of the indices. Then, for every $\tau \in \mathcal{S}_{1,2}$,

(a) $v_2(v_1(\tau)) = \tau$ *and*

(b) $\arg\min \mathcal{P}_{1,2}(\tau) = \arg\min \mathcal{P}_{2,1}(v_1(\tau)) \subset \{x \in X\,||\}\,\psi_1(x) = v_1(\tau)$.

Moreover, $\mathcal{S}_{2,1} = \{v_1(\tau)\,|\tau \in \mathcal{S}_{1,2}\}$, and so

$$\left\{\left(\tau, v_1(\tau)\right)|\tau \in \mathcal{S}_{1,2}\right\} \;=\; \left\{\left(v_2(\sigma),\sigma\right)|\sigma \in \mathcal{S}_{2,1}\right\}.$$

3. Convex analysis. In order to present the duality results of section 4, we require a few basic tools from convex analysis. There are many excellent references for the necessary background material, with several appearing within the past 10 years. In this study we make use of Rockafellar [34] and Rockafellar and Wets [36], although similar results can be found elsewhere [8, 13, 14, 21, 28, 45]. We review the necessary results here.

3.1. Functional operations. The proper convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ generates the following convex functions:

1. *Legendre–Fenchel conjugate* of h:
$$h^*(y) := \sup_x \left[\langle y,\, x\rangle - h(x)\right].$$
2. *Horizon function* of h:
$$h^\infty(z) := \sup_{x \in \operatorname{dom} h} \left[h(x+z) - h(x)\right].$$
3. *Perspective function* of h:
$$\tilde{h}(z,\lambda) := \begin{cases} \lambda h(\lambda^{-1} z) & \text{if} \quad \lambda > 0, \\ \delta\left(z \mid 0\right) & \text{if} \quad \lambda = 0, \\ +\infty & \text{if} \quad \lambda < 0. \end{cases}$$
4. *Closure of the perspective function* of h:
$$h^\pi(z,\lambda) := \begin{cases} \lambda h(\lambda^{-1} z) & \text{if} \quad \lambda > 0, \\ h^\infty(z) & \text{if} \quad \lambda = 0, \\ +\infty & \text{if} \quad \lambda < 0. \end{cases}$$

Each of these functions can also be defined by considering the epigraphical perspective and properties of convex sets. Indeed, the horizon function h^∞ is usually defined to be the function whose epigraph is the horizon cone of the epigraph of h (see section 3.2 below). The definition given above is a consequence of [34, Theorem 8.5].

The perspective function of h is the positively homogeneous function generated by the convex function $\hat{h}(x,\lambda) := h(x) + \delta\left(\lambda \mid \{1\}\right)$ [34, pp. 35 and 67]. If h is additionally closed and proper, then so are h^* (Theorem 12.2), h^∞ (Theorem 8.5), and h^π (Corollary 8.5.2), where these results are from Rockafellar [34].

Note that for every closed, proper, and convex function h, the associated horizon and perspective function, h^∞ and h^π, are positively homogeneous and so can be represented as the support functional for some convex set [34, Theorem 13.2]. Moreover, if h is a support function, then $h^\infty = h^\pi = h$.

3.2. Cones. We associate the following cones with a convex set C and a convex function h:

1. *Polar cone:* The polar cone of C is denoted by
$$C^\circ := \left\{x^* \,|\langle x^*,\, x\rangle \le 0 \;\; \forall x \in C\right\}.$$
2. *Recession cone:* The recession cone of C is denoted by
$$C^\infty := \left\{x \,|C + x \subset C\right\} = \left\{x \,|y + \lambda x \in C \;\; \forall \lambda \ge 0, \forall y \in C\right\}.$$
3. *Barrier cone:* The barrier cone of C is denoted by
$$\operatorname{bar}(C) := \left\{x^* \,|\text{for some } \beta \in \mathbb{R}, \langle x,\, x^*\rangle \le \beta \;\; \forall x \in C\right\}.$$
4. *Horizon cone of h:* The horizon cone [34, Theorem 8.7] of h is denoted by
$$\operatorname{hzn}(h) := \left\{y \,|h^\infty(y) \le 0\right\} = [\operatorname{lev}_h(\tau)]^\infty \quad \forall\, \tau > \inf h.$$

A further excellent reference for horizon cones and functions is [7], where they are referred to as *asymptotic* cones and functions.

3.3. Calculus rules. The conjugate, horizon, and perspective transformations defined in section 3.1 posses a rich calculus. We use this calculus to obtain explicit expressions for the functions $\rho^*,\ \phi^*, (\phi^*)^\infty$, and $(\phi^*)^\pi$, which play a crucial role in the applications of section 6. The calculus for conjugates and horizons is developed in many references (e.g., [8, 13, 14, 21, 28, 45]); specific citations from [34] are provided. In order to establish the perspective calculus rules for affine composition and the inverse linear image, we note that addition is a special case of affine composition, and that infimal convolution is a special case of inverse linear image. Hence, we need only establish the perspective calculus formulas for affine composition and the inverse linear image: the formula for affine-composition follows from [34, Theorem 9.5] and the definition of the perspective transformation; the formula for inverse linear image is established in section 8.

Affine composition. Let $p : \mathbb{R}^m \to \overline{\mathbb{R}}$ be a closed proper convex function, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m$, such that $(\mathrm{Ran}\,(A) - b) \cap \mathrm{ri}\,(\mathrm{dom}\, p) \neq \emptyset$. Let

$$h(x) := p(Ax - b).$$

Then

$$\begin{aligned} h^*(y) &= \inf_{A^T u = y} [\langle b,\, u\rangle + p^*(u)] && \text{[34, Theorem 16.3]},\\ h^\infty(z) &= p^\infty(Az) && \text{[34, Theorem 9.5]},\\ h^\pi(x,\lambda) &= p^\pi(Ax - \lambda b, \lambda), \end{aligned}$$

where, for $\lambda = 0$,

$$h^\pi(x, 0) = p^\pi(Ax, 0) = p^\infty(Ax).$$

All three functions are closed, proper, and convex. The derivation of h^* also makes use of the observation that

$$\text{if} \quad g(x) := h(x - b), \quad \text{then} \quad g^*(v) = h^*(v) + \langle v,\, b\rangle\,. \tag{3.1}$$

Inverse linear image. Let $p : \mathbb{R}^n \to \overline{\mathbb{R}}$ be closed, proper, and convex, and let $A \in \mathbb{R}^{m\times n}$. Let

$$h(w) := \inf_{Ax = w} p(x)$$

be the inverse linear image of p under A. Then

$$h^*(y) = p^*(A^T y) \qquad \text{[34, Theorem 16.3]}.$$

If $(A^T)^{-1}\mathrm{ri}\,(\mathrm{dom}\, p^*) \neq \emptyset$, then

$$\begin{aligned} h^\infty(z) &= \inf_{Ax = z} p^\infty(x) && \text{[34, Theorem 9.2]},\\ h^\pi(w,\lambda) &= \inf_{Ax = w} p^\pi(x,\lambda) && \text{(Proof in section 8)}, \end{aligned}$$

where all of the functions $h,\ h^*,\ h^\infty$, and h^π are closed, proper, and convex.

Addition. Let $h_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, for $i = 1, \ldots, m$, be closed proper convex functions. If $h := h_1 + \cdots + h_m$ is not identically $+\infty$, then

$$\begin{aligned} h^\infty &= h_1^\infty + \cdots + h_m^\infty \qquad \text{[34, Theorem 9.2]}, \\ h^\pi &= h_1^\pi + \cdots + h_m^\pi, \end{aligned}$$

where both are closed, proper, and convex. Moreover, if $\bigcap_{i=1}^m \operatorname{ri}(\operatorname{dom} h_i) \neq \emptyset$, then

$$h^* = h_1^* \triangledown \cdots \triangledown h_m^* \qquad \text{[34, Theorem 16.4]}$$

is closed, proper, and convex, where $\triangledown$ denotes infimal convolution.

Infimal convolution. Let $h_i : \mathbb{R}^n \to \overline{\mathbb{R}}$, for $i = 1, \ldots, m$, be closed, proper, and convex functions. Let $h := h_1 \triangledown \cdots \triangledown h_m$. Then $h^* = h_1^* + \cdots + h_m^*$, and

$$\text{if} \quad \bigcap_{i=1}^m \operatorname{ri}(\operatorname{dom} h_i^*) \neq \emptyset, \quad \text{then} \quad h^\infty = h_1^\infty \triangledown \cdots \triangledown h_m^\infty \qquad \text{[34, Corollary 9.2.1]},$$

and

$$h^\pi(x, \lambda) = \inf_{\sum_{i=1}^m x_i = x} \left[h_1^\pi(x_1, \lambda) + \cdots + h_m^\pi(x_m, \lambda) \right].$$

All three functions are closed, proper, and convex.

4. The dual problem. For our analysis, it is convenient to consider the (equivalent) epigraphical formulation

$$(\mathcal{P}) \qquad v(b, \tau) = \operatorname*{minimize}_x \ f(x, b, \tau)$$

of $\mathcal{P}(b, \tau)$, where

$$f(x, b, \tau) := \rho(b - Ax) + \delta\left((x, \tau) \mid \operatorname{epi} \phi\right).$$

Because the functions ρ and ϕ are convex, it immediately follows that f is also convex. This fact gives the convexity of the value function v, since it is the inf-projection of the objective function in x [34, Theorem 5.3].

We use a duality framework derived from the one described in Rockafellar and Wets [36, Chapter 11, section H] and associate with $\mathcal{P}$ its dual problem and corresponding dual value function:

$$(\mathcal{D}) \qquad \hat{v}(b, \tau) := \operatorname*{maximize}_{u, \mu} \ \langle b, u \rangle + \tau\mu - f^*(0, u, \mu).$$

To derive this dual from [36, Theorem 11.39], define

$$f_{(b,\tau)}(x, \Delta b, \Delta\tau) := f(x, b + \Delta b, \tau + \Delta\tau) \ .$$

Then, by (3.1), $f^*_{(b,\tau)}(v, u, \mu) = f^*(v, u, \mu) - \langle b, u \rangle - \tau\mu$. Substituting this expression into [36, Theorem 11.39] gives $\mathcal{D}$.

The dual $\mathcal{D}$ is the key to understanding the variational behavior of the value function. To access these results we must compute the conjugate of f. For this it is useful to have an alternative representation for the support function of the epigraph, which is the conjugate of the indicator function appearing in f.

4.1. Reduced dual problem. In Theorem 4.2, we derive an equivalent representation of the dual problem $\mathcal{D}$ in terms of u alone. This is the *reduced* dual problem for $\mathcal{P}$. We first present a result about conjugates for epigraphs and lower level sets.

LEMMA 4.1 (Conjugates for epigraphs and lower level sets). *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be closed, proper, and convex. Then*

$$\delta^*((y,\mu) \,|\, \text{epi } h) = (h^*)^\pi(y,-\mu), \tag{4.1a}$$

$$\delta^*(y \,|\, \text{lev}_h(\tau)) = \text{cl}\left(\inf_{\mu\geq 0}\left[\tau\mu + (h^*)^\pi(y,\mu)\right]\right). \tag{4.1b}$$

Expressions (4.1b) and (4.1a) are easily derived from the case where $\tau = 0$ which is established in [34, Theorem 13.5] and [34, Corollary 13.5.1], respectively. In [34], it is shown that (4.1a) is a consequence of (4.1b). In section 8 we provide a different proof of Lemma 4.1, where it is shown that (4.1b) follows from (4.1a). The arguments provided in the proof are instructive for later computations.

The conjugate $f^*(y,u,\mu)$ of the perturbation function $f(x,b,\tau)$ defined in $\mathcal{P}$ is now easily computed:

$$\begin{aligned} f^*(y,u,\mu) &= \sup_{x,b,\tau}\left[\langle y,\, x\rangle + \langle u,\, b\rangle + \mu\tau - \rho(b-Ax) - \delta\left((x,\tau) \mid \text{epi } \phi\right)\right] \\ &= \sup_{x,w,\tau}\left[\langle y,\, x\rangle + \langle u,\, w + Ax\rangle + \mu\tau - \rho(w) - \delta\left((x,\tau) \mid \text{epi } \phi\right)\right] \\ &= \sup_{x,\tau}\left[\left\langle y + A^Tu,\, x\right\rangle + \mu\tau - \delta\left((x,\tau) \mid \text{epi } \phi\right)\right] + \sup_{w}\left[\langle u,\, w\rangle - \rho(w)\right] \\ &= (\phi^*)^\pi(y + A^Tu, -\mu) + \rho^*(u), \end{aligned} \tag{4.2}$$

where the final equality follows from (4.1a). With this representation of the conjugate of f, we obtain the following equivalent representations for the dual problem $\mathcal{D}$. The representation labeled $\mathcal{D}_r$ is of particular importance to our discussion. We refer to $\mathcal{D}_r$ as the *reduced dual*.

THEOREM 4.2 (Dual representations). *For problem $\mathcal{P}$ define the functions*

$$\begin{aligned} g_\tau(u) &:= \rho^*(u) + \delta^*\left(A^Tu \,\middle|\, \text{lev}_\phi(\tau)\right), \\ p_\tau(s,\mu) &:= \tau\mu + (\phi^*)^\pi(s,\mu). \end{aligned}$$

Then the value function for $\mathcal{D}$ has the following equivalent characterizations:

$$\begin{aligned} \hat{v}(b,\tau) &= \sup_u\left[\langle b,\, u\rangle - \rho^*(u) - \inf_{\mu\geq 0}\; p_\tau(A^Tu,\mu)\right] \\ &= \sup_u\left[\langle b,\, u\rangle - \rho^*(u) - \delta^*\left(A^Tu \,\middle|\, \text{lev}_\phi(\tau)\right)\right] && (\mathcal{D}_r) \\ &= g_\tau^*(b) && (4.3a) \\ &= \text{cl}\left(v(\cdot,\tau)\right)(b), && (4.3b) \end{aligned}$$

where the closure operation in the (4.3b) *refers to the lower semicontinuous hull of the convex function $b \mapsto v(b,\tau)$. In particular, this implies the weak duality inequality $\hat{v}(b,\tau) \leq v(b,\tau)$. Moreover, if the function ρ is differentiable, the solution u to $\mathcal{D}_r$ is unique.*

In the large-scale setting, the primal problem $\mathcal{P}(b,\tau)$ is usually solved using a primal method that does not give direct access to the multiplier $\overline{\mu}$ for the inequality constraint $\phi(x) \leq \tau$. For example, $\mathcal{P}(b,\tau)$ may be solved using a variant of the

gradient projection algorithm. However, one can still obtain an approximation to the optimal dual variable $\overline{u}$ in $\mathcal{D}_r$, typically through the residual corresponding to the current iterate. For this reason, one needs a way to obtain an approximation to $\overline{\mu}$ from an approximation to $\overline{u}$ (i.e., given $\overline{u}$, compute $\overline{\mu}$). Lemma 4.1 and Theorem 4.2 show that this can be done by solving the problem $\inf_{\mu\geq 0}\ p_\tau(A^T\overline{u},\mu)$ for $\overline{\mu}$. Indeed, in the sequel we show that in many important cases there is a closed form expression for the solution $\overline{\mu}$. The following lemma serves to establish a precise relationship between the solution $\overline{u}$ to the reduced dual $\mathcal{D}_r$ and the solution pair $(\overline{u},\overline{\mu})$ to the dual $\mathcal{D}$.

LEMMA 4.3. *Let ϕ be as in $\mathcal{P}$ with $\tau > \inf\phi$ and $\overline{x}\in \mathrm{lev}_\phi(\tau)$.*

1. *For every s, we have*

$$\delta^*(s\,|\,\mathrm{lev}_\phi(\tau)\,) \leq \inf_{\mu\geq 0} p_\tau(s,\mu). \tag{4.4}$$

2. *Let $(\overline{x},\overline{s})$ satisfy $\overline{s}\in N\,(\overline{x}\,|\,\mathrm{lev}_\phi(\tau)\,)$ and define*

$$S_1 = \operatorname*{arg\,min}_{\mu\geq 0} p_\tau(\overline{s},\mu) \quad \textit{and} \quad S_2 = \left\{ \bar{\mu}\geq 0 \;\middle|\; \begin{array}{l} \overline{s}\in \overline{\mu}^+\partial\phi(\overline{x}) \\ 0 = \overline{\mu}(\phi(\overline{x})-\tau) \end{array} \right\},$$

where, for $x\in\mathrm{dom}\,\phi$,

$$\mu^+\partial\phi(x) := \begin{cases} \{\mu z\,|\,z\in\partial\phi(x)\,\} & \text{if } \mu>0 \text{ and } x\in\mathrm{dom}\,\partial\phi, \\ N\,(x\mid \mathrm{dom}\,\phi\,) & \text{if } \mu=0 \text{ or } \partial\phi(x)=\emptyset. \end{cases}$$

If either S_1 or S_2 is nonempty, then $S_1=S_2$ and equality holds in (4.4).

In Zălinescu the object $\mu^+\partial\phi(x)$ is denoted as $\partial(\mu\phi)(x)$ [45, p. 141], where

$$(\mu\phi)(x) := \begin{cases} \mu\phi(x) & \text{if } \lambda>0 \text{ and} \\ \delta\,(x\mid\mathrm{dom}\,\phi) & \text{if } \lambda=0. \end{cases}$$

We choose the notation $\mu^+\partial\phi(x)$ to emphasize that there is an underlying limiting operation at play, e.g., see [36, Definition 8.3 and Proposition 8.12].

The final lemma of this section concerns conditions under which solutions to $\mathcal{P}$ and $\mathcal{D}_r$ exist. This is closely tied to the horizon behavior of these problems and the notion of coercivity.

DEFINITION 4.4. *A function $h:\mathbb{R}^n\to\overline{\mathbb{R}}$ is said to be α-coercive if*

$$\lim_{\|x\|\to\infty} \frac{f(x)}{\|x\|^\alpha} = +\infty.$$

In particular, h is said to be 0-coercive, or simply coercive, if $\lim_{\|x\|\to\infty} f(x)=\infty$.

LEMMA 4.5 (Coercivity of primal and dual objectives).

1. *The objective function $f(\cdot,b,\tau)$ of $\mathcal{P}$ is coercive if and only if*

$$\mathrm{hzn}\,(\phi)\cap[-A^{-1}\mathrm{hzn}\,(\rho)] = \{0\}. \tag{4.5a}$$

2. *The objective function of the reduced dual $\mathcal{D}_r$ is coercive if and only if*

$$b\in \mathrm{int}\,(\mathrm{dom}\,\rho + A\mathrm{lev}_\phi(\tau))\,. \tag{4.5b}$$

5. Variational properties of the value function. Using $\mathcal{D}$ and representation of the conjugate of the objective of $\mathcal{P}$ (cf. (4.2)), we can specialize [36, Theorem 11.39] to obtain a characterization of the subdifferential of the value function, as well as sufficient conditions for strong duality.

THEOREM 5.1 (Strong duality and subgradient of the value function). *Let v and $\hat{v}$ be as in $\mathcal{P}$ and $\mathcal{D}$, respectively. It is always the case that*

$$v(b,\tau) \geq \hat{v}(b,\tau) \qquad \textit{(weak duality).}$$

If $(b,\tau) \in \operatorname{int}(\operatorname{dom} v)$, then

$$v(b,\tau) = \hat{v}(b,\tau) \qquad \textit{(strong duality)}$$

and $\partial v(b,\tau) \neq \emptyset$ with

$$\partial v(b,\tau) := \underset{u,\ \mu \geq 0}{\arg\max} \left[\langle b,\, u \rangle - \rho^*(u) - p_\tau(A^T u, -\mu) \right].$$

Furthermore, for fixed $(b,\tau) \in \mathbb{R}^m \times \mathbb{R}$,

$$\operatorname{dom} f(\cdot, b, \tau) \neq \emptyset \iff b \in \operatorname{dom} \rho + A(\operatorname{lev}_\phi(\tau)).$$

In particular, this implies that

$$(b,\tau) \in \operatorname{int}(\operatorname{dom} v) \iff b \in \operatorname{int}(\operatorname{dom} \rho + A(\operatorname{lev}_\phi(\tau))).$$

We now derive a characterization of the subdifferential $\partial v(b,\tau)$ based on the solutions of the reduced dual $\mathcal{D}_r$.

THEOREM 5.2 (Value function subdifferential). *Suppose that*

$$b \in \operatorname{ri}(\operatorname{dom} \rho) + A \operatorname{ri}(\operatorname{lev}_\phi(\tau)) \quad \textit{and} \tag{5.1a}$$

$$\operatorname{ri}(\operatorname{dom} \rho^*) \cap [A^{-T} \operatorname{ri}(\operatorname{bar}(\operatorname{lev}_\phi(\tau)))] \neq \emptyset. \tag{5.1b}$$

1. *If the pair $(\overline{x}, \overline{u})$ satisfies*

$$\overline{x} \in \operatorname{lev}_\phi(\tau),\ \overline{u} \in \partial\rho(b - A\overline{x}),\ \textit{and}\ A^T\overline{u} \in N\left(\overline{x} \,|\, \operatorname{lev}_\phi(\tau)\right), \tag{5.1c}$$

   then $\overline{x}$ solves $\mathcal{P}$ and $\overline{u}$ solves $\mathcal{D}_r$.
2. *If $\overline{x}$ solves $\mathcal{P}$ and (5.1a) holds, there exists $\overline{u}$ such that $(\overline{x}, \overline{u})$ satisfies (5.1c).*
3. *If $\overline{u}$ solves $\mathcal{D}_r$ and (5.1b) holds, there exists $\overline{x}$ such that $(\overline{x}, \overline{u})$ satisfies (5.1c).*
4. *If either (4.5a) and (5.1a) holds, or (4.5b) and (5.1b) holds, then $\partial v(b,\tau) \neq \emptyset$ and $\arg\min_{\mu \geq 0}\ p_\tau(A^T\overline{u}, \mu) \neq \emptyset$ for all $(\overline{x}, \overline{u}) \in \mathbb{R}^n \times \mathbb{R}^m$ satisfying (5.1c) with*

$$\partial v(b,\tau) = \left\{ \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} (\overline{x}, \overline{u}) \in \mathbb{R}^n \times \mathbb{R}^m \textit{ satisfy (5.1c) and} \\ \overline{\mu} \in \arg\min_{\mu \geq 0}\ p_\tau(A^T\overline{u}, \mu) \end{array} \right\} \tag{5.1d}$$

$$= \left\{ \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} \exists\, \overline{x} \in \operatorname{lev}_\phi(\tau) \textit{ s.t. } 0 \in -A^T\overline{u} + \overline{\mu}^+ \partial\phi(\overline{x}), \\ \textit{where } \overline{u} \in \partial\rho(b - A\overline{x}), \\ \overline{\mu} \geq 0,\ \textit{and}\ \overline{\mu}(\phi(\overline{x}) - \tau) = 0 \end{array} \right\}. \tag{5.1e}$$

The representation (5.1e) expresses the elements of $\partial v(b,\tau)$ in terms of classical Lagrange multipliers when $\overline{\mu} > 0$, and extends the classical theory when $\overline{\mu} = 0$. (See Lemma 4.3 for the definition of $\mu^+\partial\phi(x)$.) Because v is convex, it is subdifferentially regular, and so for fixed b, we can obtain the subdifferential of v with respect to τ alone [36, Corollary 10.11], i.e.,

$$\partial_\tau v(b,\tau) = \left\{ \omega \,\middle|\, \begin{pmatrix} u \\ \omega \end{pmatrix} \in \partial v(b,\tau) \right\}.$$

6. Applications. In this section we apply the calculus rules of section 3.3 in conjunction with Theorem 5.2 to evaluate the subdifferential of the value function in three important special cases: where ϕ is a gauge-plus-indicator (section 6.1), a quadratic support function (section 6.2), and an affine composition with a quadratic support function (section 6.3). In all cases we allow ρ to be an arbitrary convex function.

6.1. Gauge-plus-indicator. The case where ρ is a linear least-squares objective and ϕ is a gauge function is studied in [12]. We generalize this case by allowing the convex function ρ to be possibly nonsmooth and non-finite-valued, and take

$$\phi(x) := \gamma\left(x \mid U\right) + \delta\left(x \mid X\right), \tag{6.1}$$

where U is a nonempty closed convex set containing the origin. Here, $\gamma\left(x \mid U\right)$ is the gauge function defined in (1.1). It is evident from the definition of a gauge that ϕ is also a gauge if and only if X is a convex cone. Since $0 \in U$, it follows from [34, Theorem 14.5] that $\gamma\left(\cdot \mid U\right) = \delta^*(\cdot \,|\, U^\circ)$, where

$$U^\circ = \{ v \,|\, \langle v,\, u\rangle \leq 1 \;\forall\, u \in U \}$$

is the polar of the set U.

Observe that the requirement $x \in X$ is unaffected by varying τ in the constraint $\phi(x) \leq \tau$. Indeed, the problem $\mathcal{P}$ is unchanged if we replace ρ and ϕ by

$$\hat{\rho}(y,x) := \rho(y) + \delta\left(x \mid X\right) \quad \text{and} \quad \hat{\phi}(x) := \gamma\left(x \mid U\right) \tag{6.2}$$

with A and b replaced by

$$\hat{b} := \begin{pmatrix} b \\ 0 \end{pmatrix} \quad \text{and} \quad \hat{A} := \begin{bmatrix} A \\ -I \end{bmatrix}.$$

Hence, the generalization of [12] discussed here only concerns the application to more general convex functions ρ.

There are two ways one can proceed with this application. One can use ϕ as given in (6.1) or use $\hat{\rho}$ and $\hat{\phi}$ as defined in (6.2). We choose the former in order to highlight the presence of the abstract constraint $x \in X$. But we emphasize—regardless of the formulation chosen—that the end result is the same.

Lemma 6.1. *Let ϕ be as given in* (6.1). *The following formulas hold:*

$$\gamma\left(\cdot \mid U\right) = \delta^*(\cdot \,|\, U^\circ), \tag{6.3a}$$

$$\operatorname{dom}\gamma\left(\cdot \mid U\right) = \operatorname{cone}\left(U\right) = \operatorname{bar}\left(U^\circ\right), \tag{6.3b}$$

$$\operatorname{dom}\phi = \operatorname{cone}\left(U\right) \cap X, \tag{6.3c}$$

$$\operatorname{lev}_\phi(\tau) = (\tau U) \cap X, \tag{6.3d}$$

$$\operatorname{hzn}\left(\phi\right) = U^\infty \cap X^\infty, \text{ and} \tag{6.3e}$$

$$\operatorname{cl}\left(\operatorname{bar}\left(\operatorname{lev}_\phi(\tau)\right)\right) = \operatorname{cl}\left(\operatorname{bar}\left(U\right) + \operatorname{bar}\left(X\right)\right). \tag{6.3f}$$

If it is further assumed that

$$\operatorname{ri}(\tau U) \cap \operatorname{ri}(X) \neq \emptyset, \tag{6.4}$$

and then we also have

$$\phi^*(z) = \min_s [\delta^*(z - s \,|\, X) + \delta(s \mid U^\circ)], \tag{6.5a}$$

$$(\phi^*)^\pi(z, \mu) = \min_s [\delta^*(z - s \,|\, X) + \delta(s \mid \mu U^\circ)], \tag{6.5b}$$

$$\delta^*(z \,|\, \operatorname{lev}_\phi(\tau)) = \min_{\mu \geq 0} [\tau\mu + (\phi^*)^\pi(z, \mu)] \tag{6.5c}$$

$$= \min_s [\delta^*(z - s \,|\, X) + \tau\gamma(s \mid U^\circ)], \text{ and} \tag{6.5d}$$

$$N(x \,|\, \operatorname{lev}_\phi(\tau)) = N(x \,|\, \tau U) + N(x \,|\, X). \tag{6.5e}$$

If $\overline{s}$ minimizes (6.5d)*, then $\overline{\mu} := \gamma(\overline{s} \mid U^\circ)$ minimizes* (6.5c)*.*

By Theorem 5.1, the subdifferential of $v(b, \tau)$ is obtained by solving the dual problem (8.4) or the reduced dual $\mathcal{D}_r$. When ϕ is given by (6.1), the results of Lemma 6.1 show that the dual and the reduced dual take the form

$$\begin{aligned} &\sup_{u,\mu} \left[\langle b, u\rangle + \tau\mu - (\phi^*)^\pi(A^T u, -\mu) - \rho^*(u)\right] \\ &\quad = \sup_u \left[\langle b, u\rangle - \rho^*(u) - \delta^*\left(A^T u \,\middle|\, \operatorname{lev}_\phi(\tau)\right)\right] \\ &\quad = \sup_u \left[\langle b, u\rangle - \rho^*(u) - \min_s [\delta^*\left(A^T u - s \,\middle|\, X\right) + \tau\gamma(s \mid U^\circ)]\right] \end{aligned} \tag{6.6}$$

$$= \sup_{u,s} \left[\langle b, u\rangle - \rho^*(u) - \delta^*\left(A^T u - s \,\middle|\, X\right) - \delta^*(s \,|\, \tau U)\right]. \tag{6.7}$$

Moreover, if $(\overline{u}, \overline{s})$ solves (6.7), then $(\overline{u}, \overline{\mu})$ solves (6.6) with $\overline{\mu} = -\gamma(\overline{s} \mid U^\circ)$, and

$$(\overline{u}, -\gamma(\overline{s} \mid U^\circ)) \in \partial v(b, \tau).$$

We have the following version of Theorem 5.2 when ϕ is given by (6.1).

THEOREM 6.2. *Let ϕ be given by* (6.1) *under the assumption that* (6.4) *holds, and consider the following two conditions:*

$$b \in \operatorname{ri}(\operatorname{dom}\rho + A[\tau U \cap X]) = \operatorname{ri}(\operatorname{dom}\rho) + A[\operatorname{ri}(\tau U) \cap \operatorname{ri}(X)] \tag{6.8}$$

and

$$\exists\ \hat{u} \in \operatorname{ri}(\operatorname{dom}\rho^*) \quad \text{such that} \quad A^T\hat{u} \in \operatorname{ri}(\operatorname{bar}(U)) + \operatorname{ri}(\operatorname{bar}(X)). \tag{6.9}$$

1. *If the triple $(\overline{x}, \overline{u}, \overline{s})$ satisfies*

$$\overline{u} \in \partial\rho(b - A\overline{x}), \qquad \overline{x} \in X \cap (\tau U), \tag{6.10a}$$

$$\overline{s} \in N(\overline{x} \,|\, \tau U), \quad \text{and} \quad A^T\overline{u} - \overline{s} \in N(\overline{x} \,|\, X), \tag{6.10b}$$

   then $\overline{x}$ solves $\mathcal{P}(b, \tau)$ and $(\overline{u}, \overline{s})$ solves (6.7)*.*
2. *If $\overline{x}$ solves $\mathcal{P}(b, \tau)$ and* (6.8) *holds, then there exists a pair $(\overline{u}, \overline{s})$ such that $(\overline{x}, \overline{u}, \overline{s})$ satisfies* (6.10)*.*
3. *If $(\overline{u}, \overline{s})$ solves* (6.7) *and* (6.9) *holds, then there exists $\overline{x}$ such that $(\overline{x}, \overline{u}, \overline{s})$ satisfies* (6.10)*.*

4. *If either*

$$U^\infty \cap X^\infty \cap [-A^{-1}\mathrm{hzn}\,(\rho)] = \{0\} \quad \textit{and (6.8) holds} \tag{6.11}$$

or

$$b \in \mathrm{int}\,(\mathrm{dom}\,\rho + A[\tau U \cap X]) \textit{ and (6.9) holds,} \tag{6.12}$$

then $\partial v(b,\tau) \neq \emptyset$ *and is given by*

(6.13)

$$\partial v(b,\tau) = \left\{ \begin{pmatrix} \overline{u} \\ -\gamma\,(\overline{s} \mid U^\circ) \end{pmatrix} \middle| \, (\overline{x},\overline{u},\overline{s}) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \textit{ satisfy (6.10)} \right\}$$

$$= \left\{ \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix} \middle| \begin{array}{c} \exists\ \overline{x} \in X \cap (\tau U) \textit{ s.t.} \\ 0 \in -A^T\overline{u} + N\,(\overline{x}\,|\,X\,) + \overline{\mu}^+ \partial\gamma\,(\overline{x} \mid U)\,, \textit{ where} \\ \overline{u} \in \partial\rho(b - A\overline{x}),\ 0 \leq \overline{\mu} \textit{ and } \overline{\mu}(\gamma\,(\overline{x} \mid U) - \tau) = 0 \end{array} \right\}.$$

6.1.1. Gauge penalties. In [12], the authors study the case where ρ is a linear least-squares objective, ϕ is a gauge functional, and $X = \mathbb{R}^n$. In this case, [12, Lemma 2.1] and [12, Theorem 2.2(b)] can be deduced from (6.7) and (6.13), respectively. Another application is to the case where ρ is finite-valued and smooth, ϕ is a norm, and X is a generalized box. In this case, all of the conditions of Theorems 5.1 and 6.2 are satisfied, solutions to both $\mathcal{P}(b,\tau)$ and (6.7) exist, and v is differentiable. In particular, consider the nonnegative 1-norm-constrained inversion, where

$$\phi(x) = \|x\|_1 + \delta\left(x \mid \mathbb{R}^n_+\right),$$

and ρ is any differentiable convex function. The subdifferential characterization given in Theorem 5.1 can be explicitly computed via Theorem 6.2. In the notation of (6.1),

$$U = \{x|\,\|x\|_1 \leq 1\,|=\} : \mathbb{B}_1,$$

and X in (6.1) is $\mathbb{R}^n_+$. Since the function ρ is differentiable, the solution $\overline{u}$ to the dual (6.7) is unique [34, Theorem 26.3]. Therefore, Theorem 6.2 gives the existence of a *unique* gradient

$$\nabla_b v(b,\tau) = -A^T \nabla\rho(b - A\bar{x}),$$

where $\bar{x}$ is any solution that achieves the optimal value. The derivative with respect to τ is immediately given by Theorem 6.2 as

$$\nabla_\tau v(b,\tau) = -\gamma\left(A^T\nabla\rho(b - A\bar{x}) \mid U^\circ\right) = -\|A^T\nabla\rho(b - A\bar{x})\|_\infty. \tag{6.14}$$

Note that (6.14) has the same algebraic form when x is unconstrained. The nonnegativity constraint on x is reflected in the derivative only through its effect on the optimal point $\bar{x}$.

6.2. Quadratic support functions. We now consider the case

$$\phi(x) := \sup_{w \in U} \left[\langle x,\, w\rangle - \tfrac{1}{2}\,\langle w,\, Bw\rangle\right], \tag{6.15}$$

where $U \subset \mathbb{R}^n$ is nonempty, closed, and convex with $0 \in U$, and $B \in \mathbb{R}^{n\times n}$ is positive semidefinite. We call this class of functions quadratic support (QS) functions. This surprisingly rich and useful class is found in many applications. A deeper study of its properties and uses can be found in [4]. Note that the conjugate of ϕ is given by

$$\phi^*(w) = \tfrac{1}{2}\langle w,\, Bw\rangle + \delta\,(w \mid U)\,. \tag{6.16}$$

If the set U is polyhedral convex, then the function ϕ is called a piecewise linear-quadratic (PLQ) penalty function [36, Example 11.18]. Since B is positive semidefinite there is a matrix $L \in \mathbb{R}^{n\times k}$ such that $B = LL^T$, where k is the rank of B. Using L, the calculus rules in section 3.3 give the following alternative representation for ϕ:

$$\begin{aligned}
\phi(x) &= \sup_{w\in U}\left[\langle w,\, x\rangle - \tfrac{1}{2}\|L^T w\|_2^2 - \delta\,(w \mid U)\right]\\
&= \inf_{x_1+x_2=x}\left[\delta^*(x_1 \,|\, U\,) + \inf_{Ls=x_2}\tfrac{1}{2}\,\|s\|_2^2\right]\\
&= \inf_s\left[\tfrac{1}{2}\,\|s\|_2^2 + \delta^*(x - Ls \,|\, U\,)\right]\\
&= \inf_s\left[\tfrac{1}{2}\,\|s\|_2^2 + \gamma\,(x - Ls \mid U^\circ)\right], \qquad (6.17)
\end{aligned}$$

where the final equality follows from [34, Theorem 14.5] since $0 \in U$. Note that the function class (6.15) includes all gauge functionals for sets containing the origin. By (6.16), it easily follows that

$$(\phi^*)^\pi(w,\mu) = \begin{cases} \frac{1}{2\mu}\,\|w\|_B^2 + \delta\,(w \mid \mu U) & \text{if } \mu > 0,\\ \delta\,(w \mid U^\infty \cap \mathrm{Nul}\,(B)) & \text{if } \mu = 0,\\ +\infty & \text{if } \mu < 0, \end{cases}$$

where $\|\cdot\|_B$ denotes the seminorm induced by B, i.e.,

$$\|w\|_B := \sqrt{w^T Bw}.$$

The next result catalogues important properties of the function ϕ given in (6.15).

LEMMA 6.3. *Let ϕ be given by* (6.15) *with $\tau > 0$. Then*

$$\begin{aligned}
\mathrm{dom}\,\phi &= \mathrm{cone}\,(U^\circ) + \mathrm{Ran}\,(B) \qquad \textit{and}\\
\mathrm{hzn}\,(\phi) &= \mathrm{cone}\,(U)^\circ,
\end{aligned}$$

and in particular, ϕ is coercive if and only if $0 \in \mathrm{int}\,(U)$. Moreover,

$$\delta^*(w \,|\, \mathrm{lev}_\phi(\tau)\,) = \min_{\lambda\ge 0}\left[\tau\lambda + (\phi^*)^\pi(w,\lambda)\right] \tag{6.18}$$

$$= \begin{cases} \tau\gamma\,(w \mid U) + \frac{\|w\|_B^2}{2\gamma(w|U)} & \textit{if } \gamma\,(w \mid U) > \|w\|_B/\sqrt{2\tau},\\ \sqrt{2\tau}\|w\|_B & \textit{if } \gamma\,(w \mid U) \le \|w\|_B/\sqrt{2\tau}, \end{cases} \tag{6.19}$$

where the minimizing λ in (6.18) *is given by*

$$\lambda = \max\left\{\gamma\,(w \mid U)\,,\ \frac{\|w\|_B}{\sqrt{2\tau}}\right\}. \tag{6.20}$$

In particular, the formula (6.19) *implies that*

$$\mathrm{bar}\,(\mathrm{lev}_\phi(\tau)) = \mathrm{dom}\,(\delta^*(\cdot \,|\, \mathrm{lev}_\phi(\tau)\,)) = \mathrm{dom}(\gamma\,(\cdot \mid U)) = \mathrm{cone}\,(U)\,.$$

We now apply Theorem 5.2 to the case where ϕ is given by (6.15).

THEOREM 6.4. *Let ϕ be given by* (6.15) *and consider the following two conditions:*

$$\exists\ \hat{x} \in \mathrm{ri}\,(\mathrm{dom}\,\phi) \quad \textit{such that} \quad \phi(\hat{x}) < \tau \ \textit{and}\ b - A\hat{x} \in \mathrm{ri}\,(\mathrm{dom}\,\rho) \tag{6.21}$$

and

$$\exists\ \hat{u} \in \mathrm{ri}\,(\mathrm{dom}\,\rho^*) \quad \textit{such that} \quad A^T\hat{u} \in \mathrm{ri}\,(\mathrm{cone}\,(U))\,. \tag{6.22}$$

1. *If the pair $(\overline{x}, \overline{u})$ satisfy*

$$\overline{x} \in \mathrm{lev}_\phi(\tau),\ \overline{u} \in \partial\rho(b - A\overline{x}), \quad \textit{and} \quad A^T\overline{u} \in N\left(\overline{x} \,|\, \mathrm{lev}_\phi(\tau)\right), \tag{6.23}$$

   then $\overline{x}$ solves $\mathcal{P}(b,\tau)$ and $\overline{u}$ solves $\mathcal{D}_r$.
2. *If $\overline{x}$ solves $\mathcal{P}(b,\tau)$ and* (6.21) *holds, then there exists $\overline{u}$ such that* (6.23) *holds.*
3. *If $\overline{u}$ solves $\mathcal{D}_r$ and* (6.22) *holds, then there exists $\overline{x}$ such that* (6.23) *holds.*
4. *If either*

$$\mathrm{cone}\,(U)^\circ \cap [-A^{-1}\mathrm{hzn}\,(\rho)] = \{0\} \ \textit{and}\ (6.21)\ \textit{holds} \tag{6.24}$$

   or

$$b \in \mathrm{int}\,(\mathrm{dom}\,\rho + A\mathrm{lev}_\phi(\tau)) \ \textit{and}\ (6.22)\ \textit{holds,} \tag{6.25}$$

   then $\partial v(b,\tau) \neq \emptyset$ and is given by

$$\begin{aligned}\partial v(b,\tau) &= \left\{ \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} \exists\ \overline{x}\ \textit{s.t.}\ (\overline{x},\overline{u})\ \textit{satisfy}\ (6.23)\ \textit{and} \\ \overline{\mu} = \max\left\{\gamma\left(A^T\overline{u} \mid U\right), \|A^T\overline{u}\|_B/\sqrt{2\tau}\right\} \end{array} \right\} \\ &= \left\{ \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} \exists\ \overline{x} \in \mathrm{lev}_\phi(\tau)\ \textit{s.t.}\ 0 \in -A^T\overline{u} + \overline{\mu}^+\partial\phi(\overline{x}),\ \textit{where} \\ \overline{u} \in \partial\rho(b - A\overline{x}),\ 0 \leq \overline{\mu},\ \textit{and}\ \overline{\mu}(\phi(\overline{x}) - \tau) = 0 \end{array} \right\}. \end{aligned} \tag{6.26}$$

In the following corollary we exploit the structure of ϕ to refine the multiplier description of the $\partial v(b,\tau)$ given in (6.26).

COROLLARY 6.5. *Consider the problem $\mathcal{P}(b,\tau)$ with ϕ given by* (6.15). *A pair $(\overline{x},\overline{u})$ satisfies* (6.23) *if and only if $\overline{x} \in \mathrm{lev}_\phi(\tau)$, $\overline{u} \in \partial\rho(b - A\overline{x})$, and either*

$$A^T\overline{u} \in N\left(\overline{x} \,|\, \mathrm{dom}\,\phi\right) \quad \textit{or} \tag{6.27a}$$

$$\exists\ \overline{\mu} > 0,\ \overline{w} \in U \quad \textit{such that} \quad \overline{x} \in B\overline{w} + N\left(\overline{w} \,|\, U\right) \ \textit{and}\ A^T\overline{u} = \overline{\mu w}. \tag{6.27b}$$

6.2.1. Huber penalty. A popular function in the PLQ class is the Huber penalty [29]:

$$\phi(x) = \sup_{w \in [-\kappa,\kappa]^n} \left[\langle x,\, w\rangle - \tfrac{1}{2}\left\|w\right\|_2^2\right] = \sum_{i=1}^n \phi_i(x_i); \quad \phi_i(x_i) := \begin{cases} \frac{1}{2}x_i^2 & \text{if } |x_i| \leq \kappa, \\ \kappa|x_i| - \kappa^2/2 & \text{otherwise.} \end{cases}$$

The Huber function is of form (6.15) with $B = I$ and $U = [-\kappa,\kappa]^n$. In this case, $U^\infty \cap \mathrm{Nul}\,(B) = \{0\}$ so that the conditions of Corollary 6.5 hold.

A graph of the scalar component function ϕ_i is shown in Figure 6.1. The Huber penalty is robust to outliers, since it increases linearly rather than quadratically outside the threshold defined by κ. For any misfit function ρ, Theorem 6.4 can be

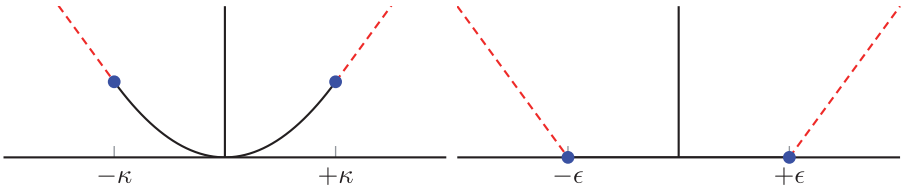

FIG. 6.1. *Huber (left) and Vapnik (right) penalties.*

used to easily compute the subgradient $\partial v(b,\tau)$ of the value function. If the regularity condition (6.21) is satisfied (e.g., if ρ is finite valued), then Theorem 6.4 implies that

$$\partial v(b,\tau) = \left\{ \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} (\overline{x},\overline{u}) \text{ satisfy (6.23) and} \\ \overline{\mu} = \max\left\{\kappa\|A^T\overline{u}\|_\infty, \|A^T\overline{u}\|_2/\sqrt{2\tau}\right\} \end{array} \right\}.$$

In particular, if ρ is differentiable finite-valued, $\overline{u} = \nabla\rho(b - A\overline{x})$ is unique and

$$\nabla v(b,\tau) = \begin{pmatrix} \overline{u} \\ -\overline{\mu} \end{pmatrix}.$$

6.3. Affine composition with QS functions. Next consider the case where ϕ takes the form

$$\phi(x) := \psi(Hx + c), \quad \text{where} \quad \psi(y) := \sup_{w\in U}[\langle y, w\rangle - \tfrac{1}{2}\langle w, Bw\rangle], \tag{6.28}$$

$H \in \mathbb{R}^{\nu\times n}$ is injective, $c \in \mathbb{R}^\nu$, $U \subset \mathbb{R}^\nu$ is nonempty, closed, and convex with $0 \in U$, and $B \in \mathbb{R}^{\nu\times\nu}$ is symmetric and positive semidefinite. We assume that

$$\exists\, \hat{x} \text{ such that } H\hat{x} + c \in \operatorname{ri}(\operatorname{dom}\psi),$$

where $\operatorname{dom}\psi = \operatorname{cone}(U^\circ) + \operatorname{Ran}(B)$ (Lemma 6.3). We show that the function ϕ in (6.28) is an instance of the quadratic support functions considered in section 6.2. To see this we make the following definitions:

$$\tilde{x} = \begin{pmatrix} x \\ s \end{pmatrix}, \; \tilde{y} = \begin{pmatrix} y \\ z \end{pmatrix}, \; \tilde{w} = \begin{pmatrix} v \\ w \end{pmatrix}, \; \tilde{U} = \{0\}\times U,$$

$$\tilde{b} = \begin{pmatrix} b \\ c \end{pmatrix}, \; \tilde{A} = \begin{bmatrix} A & 0 \\ -H & I \end{bmatrix}, \; \tilde{B} = \begin{bmatrix} 0 & 0 \\ 0 & B \end{bmatrix}, \; \tilde{\rho}\begin{pmatrix} y \\ z \end{pmatrix} = \rho(y) + \delta(z \mid \{0\}), \text{ and}$$

$$\tilde{\phi}\begin{pmatrix} x \\ s \end{pmatrix} = \sup_{\binom{v}{w}\in\tilde{U}} \left[\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, \begin{pmatrix} x \\ s \end{pmatrix}\right\rangle - \tfrac{1}{2}\left\langle \begin{pmatrix} v \\ w \end{pmatrix}, \tilde{B}\begin{pmatrix} v \\ w \end{pmatrix}\right\rangle\right] = \delta^*(x \mid \{0\}) + \psi(s).$$

With these definitions, the two problems $\mathcal{P}(b,\tau)$ and

$$\text{minimize} \quad \tilde{\rho}(\tilde{b} - \tilde{A}\tilde{x}) \quad \text{subject to} \quad \tilde{\phi}(\tilde{x}) \le \tau$$

are equivalent. In addition, we have the relationships

$$\tilde{\rho}^*\begin{pmatrix} u \\ r \end{pmatrix} = \rho^*(u) + \delta^*(r \mid \{0\}), \quad \tilde{\phi}^*\begin{pmatrix} v \\ w \end{pmatrix} = \delta(v \mid \{0\}) + \psi^*(w),$$

$$\gamma\left(\begin{pmatrix} v \\ w \end{pmatrix} \mid \tilde{U}\right) = \delta(v \mid \{0\}) + \gamma(w \mid U), \quad \text{and} \quad \left\|\begin{pmatrix} v \\ w \end{pmatrix}\right\|_{\tilde{B}} = \delta^*(v \mid \{0\}) + \|w\|_B.$$

Moreover, the reduced dual $\mathcal{D}_r$ becomes

$$\sup_{H^T r = A^T u} \left[\langle b,\, u \rangle + \langle c,\, r \rangle - \rho^*(u) - \delta^*(r \,|\, \mathrm{lev}_\psi(\tau)) \right]. \tag{6.29}$$

Using standard methods of convex analysis, we obtain the following result as a direct consequence of Theorem 6.4 and [36, Corollary 10.11].

THEOREM 6.6. *Let ϕ be given by* (6.28)*, and consider the following two conditions:*

$$\exists\ \hat{x} \text{ such that } H\hat{x} + c \in \mathrm{ri}\,(\mathrm{dom}\,\psi)\,,\ \psi(H\hat{x}+c) < \tau,\ \text{and } b - A\hat{x} \in \mathrm{ri}\,(\mathrm{dom}\,\rho) \tag{6.30}$$

and

$$\exists\ \hat{u} \in \mathrm{ri}\,(\mathrm{dom}\,\rho^*) \text{ and } \hat{r} \in \mathrm{ri}\,(\mathrm{cone}\,(U)) \text{ such that } \begin{pmatrix} \hat{u} \\ \hat{r} \end{pmatrix} \in \mathrm{Nul}\left(\begin{bmatrix} A \\ -H \end{bmatrix}^T \right). \tag{6.31}$$

1. *If the triple $(\overline{x}, \overline{u}, \overline{r})$ satisfies*
   (6.32)
   $$\overline{x} \in \mathrm{lev}_\phi(\tau),\ \overline{u} \in \partial\rho(b - A\overline{x}),\ \overline{r} \in N\left(H\overline{x} + c \,|\, \mathrm{lev}_\psi(\tau)\right),\ \text{and } A^T\overline{u} = H^T\overline{r},$$
   then $\overline{x}$ solves $\mathcal{P}(b,\tau)$ and $(\overline{u},\overline{r})$ solves (6.29)*.*
2. *If $\overline{x}$ solves $\mathcal{P}(b,\tau)$ and* (6.30) *holds, there exists $(\overline{u},\overline{r})$ such that* (6.32) *holds.*
3. *If $(\overline{u},\overline{r})$ solves* (6.29) *and* (6.31) *holds, there exists $\overline{x}$ such that* (6.32) *holds.*
4. *If either*
   $$H^{-1}[\mathrm{cone}\,(U)^\circ] \cap [-A^{-1}\mathrm{hzn}\,(\rho)] = \{0\} \text{ and } (6.30) \text{ holds}$$
   or
   $$\begin{pmatrix} b \\ c \end{pmatrix} \in \mathrm{int}\left(\mathrm{dom}\,\rho \times \mathrm{lev}_\psi(\tau) + \mathrm{Ran}\left(\begin{bmatrix} A \\ -H \end{bmatrix} \right) \right) \text{ and } (6.31) \text{ holds,}$$
   then $\partial v(b,\tau) \neq \emptyset$ and is given by
   $$\begin{aligned} \partial v(b,c,\tau) &= \left\{ \begin{pmatrix} \overline{u} \\ \overline{r} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} \exists\ \overline{x} \in \mathbb{R}^n \text{ s.t. } (\overline{x},\overline{u},\overline{r}) \text{ satisfy } (6.32) \text{ and} \\ \overline{\mu} = \max\left\{ \gamma\left(\overline{r} \mid U\right), \|\overline{r}\|_B / \sqrt{2\tau} \right\} \end{array} \right\} \\ &= \left\{ \begin{pmatrix} \overline{u} \\ \overline{r} \\ -\overline{\mu} \end{pmatrix} \;\middle|\; \begin{array}{c} \exists\ \overline{x} \in \mathbb{R}^n \text{ s.t. } c + H\overline{x} \in \mathrm{lev}_\psi(\tau), \\ \overline{u} \in \partial\rho(b - A\overline{x}),\ \overline{r} \in \overline{\mu}^+ \partial\psi(c + H\overline{x}),\ \overline{\mu} \geq 0, \\ \overline{\mu}(\psi(c + H\overline{x}) - \tau) = 0,\ \text{and } A^T\overline{u} = H^T\overline{r} \end{array} \right\}. \end{aligned}$$

COROLLARY 6.7. *Consider the problem $\mathcal{P}(b,\tau)$ with ϕ given by* (6.28)*. Then $(\overline{x},\overline{u},\overline{r})$ satisfies* (6.32) *if and only if*

$$\begin{gathered} H\overline{x} + c \in \mathrm{lev}_\psi(\tau), \quad \overline{u} \in \partial\rho(b - A\overline{x}), \quad A^T\overline{u} = H^T\overline{r}, \\ \text{and either} \quad \overline{r} \in N\left(H\overline{x} + c \,|\, \mathrm{dom}\,\psi\right) \quad \text{or} \\ \exists\ \overline{\mu} \geq 0,\ \overline{w} \in U \quad \text{such that} \quad Hx + c \in B\overline{w} + N\left(\overline{w} \,|\, U\right) \text{ and } \overline{r} = \overline{\mu w}. \end{gathered}$$

6.3.1. Vapnik penalty. The Vapnik penalty

$$\rho(r) = \sup_{u \in [0,1]^{2n}} \left\{ \left\langle \begin{bmatrix} r - \epsilon \\ -r - \epsilon \end{bmatrix}, u \right\rangle \right\} = (r - \epsilon)_+ + (-r - \epsilon)_+$$

is an important example in the PLQ class which is most easily represented as the composition of an affine transformation with a PLQ function. The scalar version is shown in the right panel of Figure 6.1. In this case,

$$H = \begin{bmatrix} I \\ -I \end{bmatrix}, \quad c = -\begin{bmatrix} \epsilon \mathbf{1} \\ \epsilon \mathbf{1} \end{bmatrix}, \quad B = \mathbf{0} \in \mathbb{R}^{2n \times 2n}, \quad \text{and} \quad U = [0,1]^{2n}.$$

In order to satisfy (6.32), we need to find a triple $(\overline{x}, \overline{u}, \overline{w})$ with $\overline{w} = [\overline{w}_1\ \overline{w}_2]^T \in [0,1]^{2n}$ so that $\overline{u} \in \partial\rho(b - A\overline{x})$ and $A^T\overline{u} = H^T\overline{w} = \overline{w}_1 - \overline{w}_2$. We claim that either $\overline{w}_1(i) = 0$ or $\overline{w}_2(i) = 0$ for all i. To see this, observe that $\overline{w} \in N\left(H\overline{x} + c \,|\, \mathrm{lev}_\psi(\tau)\right)$, so

$$\left\langle \overline{w}, y - \begin{bmatrix} \overline{x} - \epsilon \\ -\overline{x} - \epsilon \end{bmatrix} \right\rangle \leq 0$$

whenever $\psi(y) \leq \tau$. Taking y first with $-\epsilon$ as the only nonzero in the ith coordinate, and then with $-\epsilon$ in the only nonzero in the $(n+i)$th coordinate, we get

$$\overline{w}_1(i)(-\overline{x}(i)) \leq 0 \quad \text{and} \quad \overline{w}_2(i)(\overline{x}(i)) \leq 0.$$

If $x(i) < 0$, from the first equation we get $\overline{w}_1(i) = 0$, while if $x(i) > 0$, we get $\overline{w}_2(i) = 0$ from the second equation. If $x(i) = 0$, then taking $y = 0$ gives

$$\overline{w}_1(i)\epsilon \leq 0 \quad \text{and} \quad \overline{w}_2(i)\epsilon \leq 0,$$

so $\overline{w}_1(i) = \overline{w}_2(i) = 0$. Since $A^T\overline{u} = \overline{w}_1 - \overline{w}_2$ and $\overline{w}_1(i)$ or $\overline{w}_2(i)$ is 0 for each i, we get $\mu = \gamma\left(\overline{w} \mid [0,1]^{2n}\right) = \|A^T\overline{u}\|_\infty$. Hence, the subdifferential ∂v is computed in precisely the same way for the Vapnik regularization as for the 1-norm.

7. Numerical example: Robust nonnegative basis pursuit . In this example, we recover a nonnegative undersampled sparse signal from a set of very noisy measurements using several formulations of $\mathcal{P}$. We compare the performance of three different penalty functions ρ: least-squares, Huber (see section 6.2.1), and a nonconvex penalty arising from the student's t distribution (see, e.g., [5, 3]). The regularizing function ϕ in all of the examples is the sum of the 1-norm and the indicator of the positive orthant (see section 6.1.1).

The formulations using Huber and Student's t misfits are robust alternatives to the nonnegative basis pursuit problem [18]. The Huber misfit agrees with the quadratic penalty for small residuals but is relatively insensitive to larger residuals. The student's t misfit is the negative likelihood of the student's t distribution,

$$\rho_s(x) = \log(1 + x^2/\nu), \tag{7.1}$$

where ν is the degrees of freedom parameter.

For each penalty ρ, our aim is to solve the problem

$$\underset{x \geq 0}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \rho(b - Ax) \leq \sigma$$

via a series of approximate solutions of $\mathcal{P}$. The 1-norm regularizer on x encourages a sparse solution. In particular, we solve the nonlinear equation (1.3), where v is the value function of $\mathcal{P}$. This is the approach used by the SPGL1 software package [12]; the underlying theory, however, does not cover the Huber function. Also, ϕ is not everywhere finite valued, which violates [12, Assumption 3.1]. Finally, the student's t misfit (7.1) is nonconvex; however, the inverse function relationship (cf. Theorem 2.1) still holds, so we can achieve our goal, provided we can solve the root-finding problem.

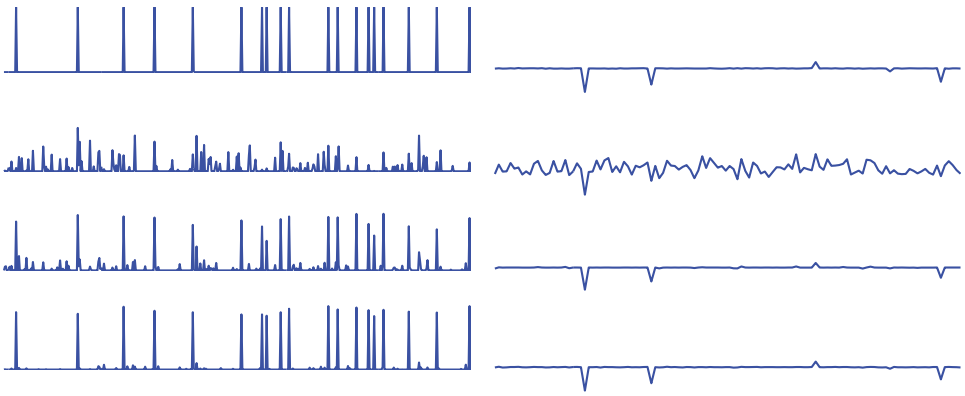

FIG. 7.1. *Left, top to bottom: true signal, and reconstructions via least-squares, Huber, and student's t. Right, top to bottom: true errors, and least-squares, Huber, and student's t residuals.*

Formula (6.14) computes the derivative of the value function associated with $\mathcal{P}(b,\tau)$ for any convex differentiable ρ. The derivative requires $\nabla\rho$, evaluated at the optimal residual associated with $\mathcal{P}(b,\tau)$. For the Huber case, this is given by

$$(\nabla\rho(b - A\bar{x}))_i = \operatorname{sign}(b_i - A_i\bar{x}) \cdot \min(|b_i - A_i\bar{x}|, \kappa).$$

The student's t misfit is also smooth, but nonconvex. Therefore, the formula (6.14) may still be applied—with the caveat that there is no guarantee of success. However, in all of the numerical experiments, we are able to find the root of (1.3).

We consider a common compressive sensing example: we want to recover a 20-sparse vector in $\mathbb{R}^{512}_+$ from 120 measurements. We use a Gaussian measurement matrix $A \in \mathbb{R}^{100\times 1024}$, where each entry is sampled from the distribution $N(0, 1/10)$. We generate measurements to test the BPDN formulation according to

$$b = Ax + w + \zeta,$$

where $w \sim N(0, 0.005^2)$ is small Gaussian error, while ζ contains five randomly placed large outliers sampled from $N(0, 4)$. For each penalty ρ, the σ parameter is the true measure of the error in that penalty, i.e., $\sigma_\rho = \rho(\zeta)$. This allows a fair comparison between the penalties.

We expect the Huber function to out-perform the least squares penalty by budgeting the error level σ to allow a few large outliers, which will never happen with the quadratic. We expect the student's t penalty to work even better, because it is nonconvex and grows sublinearly as outliers increase. The results in Figure 7.1 demonstrate that this is indeed the case. In many instances the Huber function is able to do just as well as the student's t; however, often the student's t does better (and never worse). Both robust penalties always do better than the least squares fit. The code is implemented in and extended version of SPGL1 and can be downloaded from https://github.com/saravkin/spgl1. The particular experiment presented here can be found in tests/spgl1TestNN.m.

8. Appendix: Proofs of results.

Proof of Theorem 2.1. Let $\tau \in \mathcal{S}_{1,2}$ and set $\sigma_\tau = v_1(\tau)$. By assumption, $\arg\min \mathcal{P}_{1,2}(\tau) \neq \emptyset$. Let $x_\tau \in \arg\min \mathcal{P}_{1,2}(\tau)$, so that $\psi_1(x_\tau) = \sigma_\tau$ and $\psi_2(x_\tau) = \tau$.

In particular, x_τ is feasible for $\mathcal{P}_{2,1}(\sigma_\tau)$. Let $\hat{x}$ be any other feasible point for $\mathcal{P}_{2,1}(\sigma_\tau)$ so that $\psi_1(\hat{x}) \le \sigma_\tau = v_1(\tau) = \psi_1(x_\tau)$. If $\psi_1(\hat{x}) < \sigma_\tau = v_1(\tau)$, then $\psi_2(\hat{x}) > \tau$ since otherwise we contradict the definition of $v_1(\tau)$. If $\psi_1(\hat{x}) = \sigma_\tau$, then we claim that $\psi_2(\hat{x}) \ge \tau$. Indeed, if $\psi_2(\hat{x}) < \tau$, then $\hat{x} \in \arg\min \mathcal{P}_{1,2}(\tau)$ but $\psi_2(\hat{x}) < \tau$, which contradicts the fact that $\tau \in \mathcal{S}_{1,2}$. Hence, every feasible point for $\mathcal{P}_{2,1}(\sigma_\tau)$ has $\psi_2(\hat{x}) \ge \tau$ with equality only if $\psi_1(\hat{x}) = \sigma_\tau$. But x_τ is feasible for $\mathcal{P}_{2,1}(\sigma_\tau)$ with $\psi_2(x_\tau) = \tau$. Therefore, $x_\tau \in \arg\min \mathcal{P}_{2,1}(\sigma_\tau) \subset \{x \in X \,|\psi_1(x) = \sigma_\tau\}$. Consequently, $v_2(v_1(\tau)) = \tau$ and

$$\emptyset \neq \arg\min \mathcal{P}_{1,2}(\tau) \subset \arg\min \mathcal{P}_{2,1}(\sigma_\tau) \subset \{x \in X \,|\psi_1(x) = \sigma_\tau\}. \tag{8.1}$$

We now show that $\arg\min \mathcal{P}_{2,1}(\sigma_\tau) \subset \arg\min \mathcal{P}_{1,2}(\tau)$. Let $\hat{x} \in \arg\min \mathcal{P}_{2,1}(\sigma_\tau)$. In particular, $\hat{x}$ is feasible for $\mathcal{P}_{2,1}(\sigma_\tau)$, so, by what we have already shown, $\psi_2(\hat{x}) \ge \tau$ with equality only if $\psi_1(\hat{x}) = \sigma_\tau$. But, by our choice of $\hat{x}$, $\psi_2(\hat{x}) = v_2(v_1(\tau)) = \tau$, so $\psi_1(\hat{x}) = \sigma_\tau$, i.e., $\hat{x} \in \arg\min \mathcal{P}_{1,2}(\tau)$.

It remains to establish the final statement of the theorem. By (8.1), we already have that $\{v_1(\tau) \,|\tau \in \mathcal{S}_{1,2}\} \subset \mathcal{S}_{2,1}$, so we need only establish the reverse inclusion. For this, let $\sigma \in \mathcal{S}_{2,1}$ and set $\tau_\sigma = v_2(\sigma)$. By interchanging the indices and applying the first part of the theorem, we have from (8.1) that

$$\emptyset \neq \arg\min \mathcal{P}_{2,1}(\sigma) \subset \arg\min \mathcal{P}_{1,2}(\tau_\sigma) \subset \{x \in X \,|\psi_2(x) = \tau_\sigma\}.$$

That is, $\tau_\sigma \in \mathcal{S}_{1,2}$, and, by (a), $\sigma = v_1(v_2(\sigma)) = v_1(\tau_\sigma)$.

Proof of the inverse linear image (section 3.3). For $\lambda > 0$, observe that

$$\begin{aligned} h^\pi(w,\lambda) &= \lambda \inf_{Ax = \lambda^{-1}w} p(x) \\ &= \lambda \inf_{As = w} p(\lambda^{-1}s) \qquad (s := \lambda x) \\ &= \inf_{As=w} p^\pi(s,\lambda) \qquad (8.2) \\ &= \inf\left\{ p^\pi(s,\zeta) \,\middle|\, \hat{A}\begin{pmatrix} s \\ \zeta \end{pmatrix} = \begin{pmatrix} w \\ \lambda \end{pmatrix} \right\}, \qquad (8.3) \end{aligned}$$

where

$$\hat{A} = \begin{bmatrix} A & 0 \\ 0 & 1 \end{bmatrix}.$$

Again by [34, Theorem 9.2] in conjunction with [34, Corollary 16.2.1], the function in (8.3) is closed if $(\hat{A}^T)^{-1} \operatorname{dom}(p^\pi)^* \neq \emptyset$. Since, by [34, Corollary 13.5.1], $\operatorname{dom}(p^\pi)^* = \{(u,\eta) \,|p^*(u) \le -\eta\}$, we have

$$(\hat{A}^T)^{-1} \operatorname{dom}(p^\pi)^* \neq \emptyset \quad \text{if and only if} \quad (A^T)^{-1} \operatorname{dom} p^* \neq \emptyset.$$

Hence, by assumption, the function in (8.3) is closed, proper, and convex and equals $h^\pi(w,\lambda)$ on the relative interior of its domain. Since $h^\pi(w,\lambda)$ is closed, (8.2) implies that these functions must coincide.

Proof of Lemma 4.1. We first prove (4.1a). The conjugate of $\delta\left((x,\tau) \mid \text{epi } h\right)$ is obtained as follows:

$$\begin{aligned}
\delta^*((y,\mu)\,|\,\text{epi } h) &= \sup_{\tau,x}\left[\langle y,\, x\rangle + \mu\tau - \delta\left((x,\tau) \mid \text{epi } h\right)\right] \\
&= \sup_{\tau,x\in\operatorname{dom} h}\left[\langle y,\, x\rangle + \mu\tau - \delta\left(h(x)-\tau \mid \mathbb{R}_-\right)\right] \\
&= \sup_{\omega,x\in\operatorname{dom} h}\left[\,\langle y,\, x\rangle + \mu(h(x)-\omega) - \delta\left(\omega \mid \mathbb{R}_-\right)\right] \qquad (\omega := h(x)-\tau) \\
&= \sup_{x\in\operatorname{dom} h}\left[\,\langle y,\, x\rangle + \mu h(x) + \sup_{\omega}\left[-\mu\omega - \delta\left(\omega \mid \mathbb{R}_-\right)\right]\right] \\
&= \sup_{x\in\operatorname{dom} h}\left[\langle y,\, x\rangle + \mu h(x) + \delta\left(\mu \mid \mathbb{R}_-\right)\right].
\end{aligned}$$

For $\mu < 0$, we obtain

$$\delta^*((y,\mu)\,|\,\text{epi } h) = -\mu \sup_x\left[\left\langle -\mu^{-1}y,\, x\right\rangle - h(x)\right] = -\mu h^*(-\mu^{-1}y).$$

Since h^* is necessarily a closed proper convex function, we obtain the result.

To see (4.1b), first note that the function

$$q(y) := \inf_{\mu>0}[\tau\mu + \mu h^*(y/\mu)] = \inf_{\mu\geq 0}[\tau\mu + (h^*)^\pi(y,\mu)]$$

is the positively homogeneous function generated by the function $y \mapsto \tau + h(y)$ [34, p. 35], and so it is convex in y. Next observe that the conjugate of q is given by

$$\begin{aligned}
q^*(x) &= \sup_y\left[\langle x,\, y\rangle - \inf_{\mu\geq 0}[\tau\mu + (h^*)^\pi(y,\mu)]\right] \\
&= \sup_{y,\mu\geq 0}\left[\langle x,\, y\rangle + \tau(-\mu) - (h^*)^\pi(y,\mu)\right] \\
&= \sup_{(y,\mu)}\left[\langle x,\, y\rangle + \tau\mu - (h^*)^\pi(y,-\mu)\right] \quad \text{(exchange } -\mu \text{ for } \mu) \\
&= \sup_{(y,\mu)}\left[\langle x,\, y\rangle + \tau\mu - \delta^*((y,\mu)\,|\,\text{epi } h)\right] \quad \text{(by (4.1a))} \\
&= \delta\left((x,\tau) \mid \text{epi } h\right) = \delta\left(x \mid \text{lev}_h(\tau)\right).
\end{aligned}$$

The result now follows from the biconjugate theorem [34, Theorem 12.2].

Proof of Theorem 4.2. Combining $\mathcal{D}$ with (4.1b) and (4.2) gives

$$\begin{aligned}
\hat{v}(b,\tau) &:= \sup_{u,\mu}\left[\langle b,\, u\rangle + \tau\mu - (\phi^*)^\pi(A^Tu,-\mu) - \rho^*(u)\right] \\
&= \sup_u\left[\langle b,\, u\rangle - \rho^*(u) - \inf_{\mu\leq 0}\left[\tau(-\mu) + (\phi^*)^\pi(A^Tu,-\mu)\right]\right] \\
&= \sup_u\left[\langle b,\, u\rangle - \rho^*(u) - \inf_{\mu\geq 0}\left[\tau(\mu) + (\phi^*)^\pi(A^Tu,\mu)\right]\right] \\
&= \sup_u\left[\langle b,\, u\rangle - \rho^*(u) - \delta^*\big(A^Tu \,\big|\, \text{lev}_\phi(\tau)\big)\right],
\end{aligned} \tag{8.4}$$

where the final equality follows from (4.1b). The equivalence (4.3a) follows from the definition of the conjugate, and the equivalence (4.3b) follows from [34, Theorems 16.3 and 16.4], which tell us that

$$
\begin{aligned}
g_\tau^*(b) &= \operatorname{cl}\left(\rho\triangledown\left[\delta^*\left(A^T\cdot\,\middle|\,\operatorname{lev}_\phi(\tau)\right)\right]^*\right)(b)\\
&= \operatorname{cl}\left(\inf_{w_1+w_2=\cdot}\left[\rho(w_1)+\inf_{Ax=w_2}\delta\left(x\mid\operatorname{lev}_\phi(\tau)\right)\right]\right)(b)\\
&= \operatorname{cl}\left(\inf_{\phi(x)\le\tau}\rho(\cdot-Ax)\right)(b)\\
&= \operatorname{cl}\left(v(\cdot,\tau)\right)(b)\,.
\end{aligned}
$$

The uniqueness of u when ρ is differentiable follows from the essential strict convexity of ρ^* [34, Theorem 26.3].

Proof of Lemma 4.3.

Part 1. The inequality follows immediately from (4.1b). But it is also easily derived from the observation that if $\mu > 0$ and $x \in \operatorname{lev}_\phi(\tau)$, then

$$
\begin{aligned}
\tau\mu + \mu\phi^*(s/\mu) &\ge \tau\mu + \mu[\langle x,\, s/\mu\rangle - \phi(x)] \quad \text{(Fenchel–Young inequality)}\\
&\ge \phi(x)\mu + \langle x,\, s\rangle - \mu\phi(x)\\
&= \langle x,\, s\rangle\,.
\end{aligned}
$$

Taking the sup over $x \in \operatorname{lev}_\phi(\tau)$ gives the result.

Part 2. The proof uses the following three key facts:

(i) By [34, Theorems 23.5 and 23.7], for any nonempty closed convex set U and $\overline{u} \in U$,

$$
\overline{v} \in N\left(\overline{u}\,|\,U\right) \iff \overline{u} \in \partial\delta^*(\overline{v}\,|\,U) = \operatorname*{arg\,max}_{u\in U}\langle\overline{v},\, u\rangle\,. \tag{8.5}
$$

(ii) The Fenchel–Young inequality tells us that

$$
\tau + \phi^*(\overline{s}) \ge \phi(\overline{x}) + \phi^*(\overline{s}) \ge \langle\overline{s},\, \overline{x}\rangle\,. \tag{8.6}
$$

(iii) see [8, Lemma 26.17] or [45, Corollary 2.9.5]. Let $g : \overline{\mathbb{R}} \to$ be a convex function and $\tau \in \mathbb{R}$ be such that $\tau > \inf g$. Then for every $x \in \operatorname{lev}_g(\tau)$

$$
N\left(x\,|\,\operatorname{lev}_g(\tau)\right) = \begin{cases} N\left(x\,|\,\operatorname{dom} g\right) \cup \operatorname{cone}\left(\partial g(x)\right) & \text{if } g(x) = \tau,\\ N\left(x\,|\,\operatorname{dom} g\right) & \text{if } g(x) < \tau. \end{cases} \tag{8.7}
$$

We divide the proof into two parts: (A) if $S_1 \ne \emptyset$, show $S_1 \subset S_2$, and (B) if $S_2 \ne \emptyset$, show $S_2 \subset S_1$ and equality holds in (4.4). Combined, these implications establish Part 2 of the lemma.

(A) Let $\overline{\mu} \in S_1$. We show that $\overline{\mu} \in S_2$. First suppose $\phi(\overline{x}) < \tau$. By (8.7), $N\left(\overline{x}\,|\,\operatorname{lev}_\phi(\tau)\right) = N\left(\overline{x}\,|\,\operatorname{dom}\phi\right)$. Hence, by (8.5), $\overline{s} \in N\left(\overline{x}\,|\,\operatorname{dom}\phi\right)$. Therefore, if $\overline{\mu} = 0$, we have $\overline{\mu} \in S_2$.

On the other hand, if $\overline{\mu} > 0$, by (8.5) and the fact that $N\left(\overline{x}\,|\,\operatorname{lev}_\phi(\tau)\right) = N\left(\overline{x}\,|\,\operatorname{dom}\phi\right)$, we have

$$
\begin{aligned}
\langle\overline{s},\, \overline{x}\rangle &= \delta^*(\overline{s}\,|\,\operatorname{dom}\phi)\\
&= (\phi^*)^\infty(\overline{s}) && \text{[34, Theorem 13.3]}\\
&= \tau 0 + (\phi^*)^\pi(\overline{s}, 0)\\
&\ge \tau\overline{\mu} + (\phi^*)^\pi(\overline{s}, \overline{\mu})\\
&> \overline{\mu}\phi(\overline{x}) + \overline{\mu}\phi^*(\overline{s}/\overline{\mu})\\
&\ge \overline{\mu}\,\langle\overline{s}/\overline{\mu},\, \overline{x}\rangle && \text{(Fenchel–Young inequality)}\\
&= \langle\overline{s},\, \overline{x}\rangle\,.
\end{aligned}
$$

Since this cannot occur, it must be the case that $\overline{\mu} = 0$ and $\overline{\mu} \in S_2$.

Now suppose that $\phi(\overline{x}) = \tau$ and $\overline{s} = 0$. Then, for $\mu > 0$, $p_\tau(\overline{s}, \mu) = (\tau + \phi^*(0))\mu \geq 0$ by (8.6), and, for $\mu = 0$, $p_\tau(\overline{s}, \mu) = (\phi^*)^\infty(0) = 0$. Therefore, $0 = \inf_{0 \leq \mu} p_\tau(\overline{s}, \mu)$ with $\mu = 0 \in S_1$. But, in this case, it is also clear that $0 \in S_2 \neq \emptyset$, since $\overline{s} = 0 \in N(\overline{x} \,|\, \operatorname{dom} \phi)$. Thus, if $\overline{\mu} = 0$, we have $\overline{\mu} \in S_2$. If $\overline{\mu} > 0$, then $0 = \tau + \phi^*(0)$ since $0 = p_\tau(0, \overline{\mu}) = (\tau + \phi^*(0))\overline{\mu}$. But then, by (8.6), $\phi(\overline{x}) + \phi^*(0) = \langle \overline{s}, \overline{x} \rangle = 0$ so that $\overline{s} = 0 \in \partial\phi(\overline{x})$. However, $\phi(\overline{x}) = \tau > \inf \phi$, so $0 \notin \partial\phi(\overline{x})$ [34, Theorem 23.5(b)]. This contradiction implies that if $\overline{s} = 0$, then we must also have $\overline{\mu} = 0$, and, in particular, we have $S_1 \subset S_2$.

Finally, suppose that $\phi(\overline{x}) = \tau$ and $\overline{s} \neq 0$. Then, by (8.7),

$$\text{either} \quad \text{(a) } \overline{s} \in \operatorname{cone}(\partial\phi(\overline{x})) \quad \text{or} \quad \text{(b) } \overline{s} \in N(\overline{x} \,|\, \operatorname{dom} \phi).$$

Let us first suppose that $\overline{s} \notin N(\overline{x} \,|\, \operatorname{dom} \phi)$ so, in particular, $\overline{s} \in \operatorname{cone}(\partial\phi(\overline{x}))$. As an immediate consequence, we have that $S_2 \neq \emptyset$ and the only values of μ for which $\overline{s} \in \mu^+ \partial\phi(\overline{x})$ have $\mu > 0$ since $\overline{s} \notin N(\overline{x} \,|\, \operatorname{dom} \phi)$. Let $0 < \hat{\mu} \in S_2$. If $\overline{\mu} = 0$, then

$$\begin{aligned}
\delta^*(\overline{s} \,|\, \operatorname{dom} \phi) &= (\phi^*)^\infty(\overline{s}) \\
&= \inf_{0 \leq \mu} p_\tau(\overline{s}, \mu) \\
&\leq \tau\hat{\mu} + \hat{\mu}\phi^*(\overline{s}/\hat{\mu}) \\
&= \hat{\mu}\phi(\overline{x}) + \hat{\mu}[\langle \overline{s}/\hat{\mu}, \overline{x} \rangle - \phi(\overline{x})] \quad \left[\begin{array}{c} \overline{s}/\hat{\mu} \in \partial\phi(\overline{x}) \text{ and} \\ \text{[34, Theorem 23.5(d)]} \end{array} \right] \\
&= \langle \overline{s}, \overline{x} \rangle \\
&\leq \delta^*(\overline{s} \,|\, \operatorname{dom} \phi)
\end{aligned}$$

so that $\langle \overline{s}, \overline{x} \rangle = \delta^*(\overline{s} \,|\, \operatorname{dom} \phi)$, or, equivalently, $\overline{s} \in N(\overline{x} \,|\, \operatorname{dom} \phi)$, contradicting the choice of $\overline{s}$. Hence, it must be the case that $\overline{\mu} > 0$. Again let $0 < \hat{\mu} \in S_2$. Then, by Part 1,

$$\begin{aligned}
\delta^*(\overline{s} \,|\, \operatorname{lev}_\phi(\tau)) &\leq p_\tau(\overline{s}, \overline{\mu}) \\
&= \inf_{0 \leq \mu} p_\tau(\overline{s}, \mu) \\
&\leq \tau\hat{\mu} + \hat{\mu}\phi^*(\overline{s}/\hat{\mu}) \\
&= \hat{\mu}\phi(\overline{x}) + \hat{\mu}[\langle \overline{s}/\hat{\mu}, \overline{x} \rangle - \phi(\overline{x})] \quad \left[\begin{array}{c} \overline{s}/\hat{\mu} \in \partial\phi(\overline{x}) \text{ and} \\ \text{[34, Theorem 23.5(d)]} \end{array} \right] \\
&= \langle \overline{s}, \overline{x} \rangle \\
&\leq \delta^*(\overline{s} \,|\, \operatorname{lev}_\phi(\tau))
\end{aligned}$$

so that $\langle \overline{s}, \overline{x} \rangle = \overline{\mu}[\phi(\overline{x}) + \phi^*(\overline{s}/\overline{\mu})]$, or, equivalently, $\overline{s} \in \overline{\mu}\partial\phi(\overline{x})$. Hence, $\overline{\mu} \in S_2$.

Finally, consider the case where $0 \neq \overline{s} \in N(\overline{x} \,|\, \operatorname{dom} \phi)$. Then

$$\begin{aligned}
\inf_{\mu \geq 0} p_\tau(\overline{s}, \mu) &\leq p_\tau(\overline{s}, 0) \\
&= (\phi^*)^\infty(\overline{s}) \\
&= \delta^*(\overline{s} \,|\, \operatorname{dom} \phi) && \text{[34, Theorem 13.3]} \\
&= \langle \overline{s}, \overline{x} \rangle && \text{(by (8.5))} \\
&= \delta^*(\overline{s} \,|\, \operatorname{lev}_\phi(\tau)) && \text{(again by (8.5))} \\
&\leq \inf_{\mu \geq 0} p_\tau(\overline{s}, \mu) && \text{(Part 1),}
\end{aligned}$$

so $0 \in S_1$ and $0 \in S_2$. If $\overline{\mu} > 0$, then this string of equivalences also implies that $\langle \overline{s}, \overline{x} \rangle = p_\tau(\overline{s}, \overline{\mu}) = \overline{\mu}[\phi(\overline{x}) + \phi^*(\overline{s}/\overline{\mu})]$, or, equivalently, $\overline{s} \in \overline{\mu}\partial\phi(\overline{x})$ so that $\overline{\mu} \in S_2$. Putting this all together, we get that $S_1 \subset S_2$.

(B) Let $\overline{\mu} \in S_2$. If $\overline{\mu} = 0$, then

$$\begin{aligned} p_\tau(\overline{s}, 0) &= (\phi^*)^\infty(\overline{s}) \\ &= \delta^*(\overline{s} \mid \operatorname{dom}\phi) \quad [34, \text{ Theorem } 13.3] \\ &= \langle \overline{s}, \overline{x} \rangle \\ &\le \delta^*(\overline{s} \mid \operatorname{lev}_\phi(\tau)) \\ &\le \inf_{\mu \ge 0} p_\tau(\overline{s}, \mu) \qquad \text{(Part 1)}. \end{aligned}$$

Therefore, $\overline{\mu} = 0 \in S_1$ and equality holds in (4.4).

On the other hand, if $\overline{\mu} > 0$, then $\overline{s}/\overline{\mu} \in \partial\phi(\overline{x})$, and so

$$\begin{aligned} \tau\bar{\mu} + (\phi^*)^\pi(\overline{s}, \bar{\mu}) &= \bar{\mu}[\phi(\overline{x}) + \phi^*(\overline{s}/\bar{\mu})] \\ &= \bar{\mu}\, \langle \overline{x}, \overline{s}/\bar{\mu} \rangle \qquad \left[\begin{array}{c} \overline{s}/\hat{\mu} \in \partial\phi(\overline{x}) \text{ and} \\ [34, \text{ Theorem } 23.5(\text{d})] \end{array} \right] \\ &= \langle \overline{x}, \overline{s} \rangle \\ &\le \delta^*(\overline{s} \mid \operatorname{lev}_\phi(\tau)) \\ &\le \inf_{\mu \ge 0} [\tau\mu + (\phi^*)^\pi(\overline{s}, \mu)] \qquad \text{(Part 1)}. \end{aligned}$$

Hence, $\overline{\mu} \in S_1$ and equality holds in (4.4).

Proof of Lemma 4.5.

Part 1. The primal coercivity equivalence follows from [34, Theorems 8.4 and 8.7] since $\operatorname{hzn}(f(\cdot, b, \tau)) = \operatorname{hzn}(\phi) \cap [-A^{-1}\operatorname{hzn}(\rho)]$.

Part 2. For the dual coercivity equivalence, let $\hat{g}(u) = g_\tau(u) - \langle b, u \rangle$, which is the objective of the reduced dual $\mathcal{D}_r$. By (4.3b), $\hat{g}^*(0) = g_\tau^*(b) = \operatorname{cl}(v(\cdot, \tau)) \le v(b, \tau)$. Therefore, the result follows from [34, Corollary 14.2.2] since by (8.8), $\operatorname{dom} v(\cdot, \tau) = \operatorname{dom}\rho + A\operatorname{dom}\phi$.

Proof of Theorem 5.1. The expression for f^* is derived in (4.2). The weak and strong duality relationships as well as the expression for ∂v follow immediately from [36, Theorem 11.39].

Next, note that

$$\operatorname{dom} f(\cdot, b, \tau) \neq \emptyset \iff \left[\begin{array}{c} \exists x \in \operatorname{lev}_\phi(\tau) \\ b - Ax \in \operatorname{dom}\rho \end{array} \right] \iff b \in \operatorname{dom}\rho + A\operatorname{lev}_\phi(\tau). \tag{8.8}$$

Now assume that $b \in \operatorname{int}(\operatorname{dom}\rho + A(\operatorname{lev}_\phi(\tau)))$. Recall from [34, Theorem 6.6 and Corollary 6.6.2] that

$$\operatorname{int}(\operatorname{dom}\rho + A(\operatorname{lev}_\phi(\tau))) = \operatorname{ri}(\operatorname{dom}\rho) + A(\operatorname{ri}(\operatorname{lev}_\phi(\tau))). \tag{8.9}$$

Moreover, by [34, Theorem 7.6], for any convex function p,

$$\operatorname{ri}(\operatorname{lev}_p(\tau)) = \{x \in \operatorname{ri}(\operatorname{dom} p) \,|\, p(x) < \tau\}. \tag{8.10}$$

Since $b \in \operatorname{int}(\operatorname{dom}\rho + A(\operatorname{lev}_\phi(\tau)))$, (8.9)–(8.10) imply the existence of $\overline{w} \in \operatorname{ri}(\operatorname{dom}\rho)$ and $\overline{x} \in \operatorname{ri}(\operatorname{dom}\phi)$ with $\phi(\overline{x}) < \tau$ such that $b = \overline{w} + A\overline{x}$. Since ϕ is relatively

continuous on the relative interior of its domain [34, Theorem 10.1], there exists $\delta > 0$ such that

$$\begin{aligned}(\overline{w} + \delta I\!B) \cap \operatorname{dom}\rho &\subset \operatorname{ri}(\operatorname{dom}\rho),\\ (\overline{x} + \delta I\!B) \cap \operatorname{dom}\phi &\subset \operatorname{ri}(\operatorname{dom}\phi),\\ \phi(x) < \tfrac{1}{2}(\phi(\overline{x}) + \tau) \;\forall\, x \in (\overline{x} + \delta I\!B) &\cap \operatorname{dom}\phi.\end{aligned}$$

Set $S_\rho = (\overline{w} + \delta I\!B) \cap \operatorname{dom}\rho$ and $S_\phi = (\overline{x} + \delta I\!B) \cap \operatorname{dom}\phi$. Since

$$\begin{aligned}\operatorname{cone}(S_\rho + AS_\phi - b) &= \operatorname{cone}(S_\rho - \overline{w}) + A\operatorname{cone}(S_\phi - \overline{x})\\ &= \operatorname{span}(\operatorname{dom}\rho - \overline{w}) + A\operatorname{span}(\operatorname{dom}\phi - \overline{x})\\ &= \operatorname{span}(\operatorname{dom}\rho + A\operatorname{dom}\phi - b)\\ &\supset \operatorname{cone}(\operatorname{dom}\rho + A\operatorname{dom}\phi - b)\\ &= \mathbb{R}^m \qquad (b \in \operatorname{int}(\operatorname{dom}\rho + A(\operatorname{lev}_\phi(\tau))))\,,\end{aligned}$$

we have $0 \in \operatorname{int}(S_\rho + AS_\phi - b)$. Therefore, there exits an $\epsilon > 0$ such that $b + \epsilon I\!B \subset S_\rho + AS_\phi$. Consequently, if $\hat{b} \in b + \epsilon I\!B$ and $|\hat{\tau} - \tau| < \frac{1}{2}(\tau - \phi(\overline{x}))$, then $\operatorname{dom} f(\cdot, \hat{b}, \hat{\tau}) \neq \emptyset$ and so $(\hat{b}, \hat{\tau}) \in \operatorname{dom} v$.

On the other hand, if $(b, \tau) \in \operatorname{int}(\operatorname{dom} v)$, then $\operatorname{dom} f(\cdot, \hat{b}, \hat{\tau}) \neq \emptyset$ for all $(\hat{b}, \hat{\tau})$ near (b, τ) so that $\operatorname{dom} f(\cdot, \hat{b}, \tau) \neq \emptyset$ for all $\hat{b}$ near b. Hence, $b \in \operatorname{int}(\operatorname{dom}\rho + A(\operatorname{lev}_\phi(\tau)))$.

Proof of Theorem 5.2.

Part 1. First note that (5.1c) is equivalent to the optimality condition

$$0 \in -A^T \partial\rho(b - A\overline{x}) + \partial\delta\left(\overline{x} \mid \operatorname{lev}_\phi(\tau)\right) \tag{8.11}$$

for the problem $\mathcal{P}$, and hence by [34, Theorem 23.8], $\overline{x}$ solves $\mathcal{P}$. Moreover, by [34, Theorem 23.5], (5.1c) is equivalent to

$$b - A\overline{x} \in \partial\rho^*(\overline{u}), \quad \overline{x} \in \partial\delta^*\left(A^T\overline{u} \,\middle|\, \operatorname{lev}_\phi(\tau)\right),$$

or, equivalently,

$$b \in \partial\rho^*(\overline{u}) + A\partial\delta^*\left(A^T\overline{u} \,\middle|\, \operatorname{lev}_\phi(\tau)\right), \tag{8.12}$$

which by [34, Theorem 23.8] implies that $\overline{u}$ solves the reduced dual $\mathcal{D}_r$.

Part 2. If $\overline{x}$ solves $\mathcal{P}$, then

$$0 \in \partial\left[\rho\big(b - A(\cdot)\big) + \delta\left(\cdot \mid \operatorname{lev}_\phi(\tau)\right)\right](\overline{x}),$$

which by [34, Theorems 23.8, 23.9] and (5.1a) is equivalent to (8.11), which in turn is equivalent to (5.1c).

Part 3. If $\overline{u}$ solves $\mathcal{D}_r$, then

$$b \in \partial\left[\rho^*(\cdot) + \delta^*\left(A^T(\cdot) \,\middle|\, \operatorname{lev}_\phi(\tau)\right)\right](\overline{u}),$$

which by [34, Theorems 23.8, 23.9] and (5.1b) is equivalent to (8.12), which in turn is equivalent to (5.1c).

Part 4. The equivalence (5.1e) follows from (5.1d), Part 2 of Lemma 4.3, and the fact that $A^T\overline{u} \in N\left(\overline{x} \,|\, \mathrm{lev}_\phi(\tau)\right)$ if and only if $\overline{x} \in \partial\delta^*\left(A^T\overline{u} \,\middle|\, \mathrm{lev}_\phi(\tau)\right)$.

To see (5.1d), note that (4.5a), (5.1a), and Part 1 of Lemma 4.5 imply that the primal objective is coercive, so a solution $\overline{x}$ exists. Hence, by Part 2, there exists $\overline{u}$ so that $(\overline{x}, \overline{u})$ satisfies (5.1c).

Analogously, (4.5b), (5.1b), and Part 2 of Lemma 4.5 imply that the solution $\overline{u}$ to the dual exists, and so by Part 3, there exists $\overline{x}$ such that the pair $(\overline{x}, \overline{u})$ satisfies (5.1c). In either case, the subdifferential is nonempty and is given by (5.1d).

Proof of Lemma 6.1. Formula (6.3a) is just [34, Theorem 14.5]. The first equation in (6.3b) is obvious and the second follows from (6.3a) and the definition of the barrier cone. The formula (6.3c) is now obvious. Formulas (6.3d) and (6.3e) follow immediately from the definitions and [34, Corollary 8.3.3]. Formula (6.3f) follows from (6.3e), [34, Corollary 14.2.1], and [34, Corollary 16.4.2].

First note that (6.4) implies that $\mathrm{ri}\,(\mathrm{cone}\,(U)) \cap \mathrm{ri}\,(X) \neq \emptyset$. Hence, the formula (6.5a) follows from [34, Theorem 16.4] and (6.3c). To see (6.5b), observe that the expression on the RHS is again an infimal convolution for which inf = min for the same reason as for (6.5a). The equivalence with $(\phi^*)^\pi(z, \mu)$ follows from the calculus rules in section 3.3. For formula (6.5d), first note that

$$\begin{aligned}\inf_{\mu\geq 0}[\tau\mu + (\phi^*)^\pi(z,\mu)] &= \inf_{\mu\geq 0}\left[\tau\mu + \inf_s[\delta^*(z-s \,|\, X) + \delta\,(s \mid \mu U^\circ)]\right] \\ &= \inf_s\left[\delta^*(z-s \,|\, X) + \inf_{\mu\geq 0}[\tau\mu + \delta\,(s \mid \mu U^\circ)]\right] \\ &= \inf_s\left[\delta^*(z-s \,|\, X) + \tau\gamma\,(s \mid U^\circ)\right].\end{aligned}$$

Again, the final infimum in this derivation is an infimal convolution for which inf = min for the same reasons as in (6.5a) since, by (6.3c) and [34, Theorem 14.5],

$$\mathrm{dom}\,((\tau\gamma\,(\cdot \mid U^\circ))^*) = \mathrm{dom}\,((\delta^*(\cdot \,|\, \tau U))^*) = \mathrm{dom}\,\delta\,(\cdot \mid \tau U) = \tau U.$$

Therefore, an optimal $\overline{s}$ in this infimal convolution exists giving $\overline{\mu} = \gamma\,(\overline{s} \mid U^\circ)$ as the optimal solution to the first min in (6.5d).

Formula (6.5e) is an immediate consequence of (6.3d), (6.4), and [34, Corollary 23.8.1].

Proof of Theorem 6.2. By (6.3d) and the calculus rules for the relative interior [34, section 6], (5.1a) and (6.8) are equivalent. Similary, by (6.3f) and [34, Theorem 6.3], (5.1b) and (6.9) are equivalent.

Part 1. Since (6.4) holds, the formula (6.5e) holds and so (6.10) and (5.1c) are equivalent. Hence, the result follows from Part 1 of Theorem 5.2.

Part 2. Since (5.1a) and (6.8) are equivalent, the result follows from Part 2 of Theorem 5.2.

Part 3. Since (5.1b) and (6.9) are equivalent, the result follows from Part 3 of Theorem 5.2.

Part 4. By (6.3e), (6.11) is equivalent to (4.5a) and (5.1a), and, by (6.3c), (6.12) is equivalent to (4.5b) and (5.1b). Therefore, by Theorem 5.2, (6.13) is equivalent to (5.1d) since $\tau\gamma\,(s \mid U^\circ) = \inf_{\mu\geq 0}[\tau\mu + \delta\,(s \mid \mu U^\circ)$. The final equivalence is identical to that of Theorem 5.2.

Proof of Lemma 6.3. The formula for $\operatorname{dom}\phi$ follows from (6.17). Indeed, by (6.17), $x \in \operatorname{dom}\phi$ if and only if there exists $s \in \mathbb{R}^k$ such that $x - Ls \in \operatorname{dom}\gamma(\cdot \mid U^\circ) = \operatorname{cone}(U^\circ)$, or, equivalently, $x \in \operatorname{cone}(U^\circ) + \operatorname{Ran}(L) = \operatorname{cone}(U^\circ) + \operatorname{Ran}(B)$. The formula for $\operatorname{hzn}(\phi)$ follows immediately from [34, Theorem 14.2] and (6.16). In particular, ϕ is coercive if and only if $\{0\} = \operatorname{hzn}(\phi)$, or, equivalently, $\operatorname{cone}(U) = \mathbb{R}^n$, i.e., $0 \in \operatorname{int}(U)$.

Next we show that the λ given in (6.20) solves (6.18). First observe that the optimal λ must be greater than $\gamma(w \mid U)$, and from elementary calculus, the minimizer of the hyperbola $\frac{1}{2\lambda}\|w\|_B^2 + \tau\lambda$ for $\lambda \geq 0$ is given by $\|w\|_B/\sqrt{2\tau}$. Therefore, the minimizing λ is given by (6.20). Substituting this value into (6.18) gives (6.19).

It is now easily shown that the function in (6.19) is lower semicontinuous. Therefore, the equivalence in (6.18) follows from (4.1b).

Proof of Theorem 6.4. By [34, Theorem 7.6],

$$\operatorname{ri}(\operatorname{lev}_\phi(\tau)) = \{x \,|x \in \operatorname{ri}(\operatorname{dom}\phi)\,,\ \phi(x) < \tau\}\,.$$

Hence, by Lemma 6.3, the equivalence between (5.1) and (6.21), (6.22), (6.24), (6.25), respectively, is easily seen. Therefore, Parts 1–4 follow immediately from Theorem 5.2.

Proof of Corollary 6.5. Condition (6.27a) occurs when $\overline{\mu} = 0$ since $0^+\partial\phi(\overline{x}) = N(\overline{x} \mid \operatorname{dom}\phi)$. When $\overline{\mu} > 0$, by [34, Theorem 23.5], $\partial\phi(x) = \arg\max_{w\in U}[\langle x,\, w\rangle - \frac{1}{2}\langle w,\, Bw\rangle]$, so that $w \in \partial\phi(x)$ if and only if $x \in Bw + N(w \mid U)$.

Acknowledgments. The authors are grateful to two anonymous referees for their remarkably careful reading of an intricate paper. Their detailed list of comments and suggestions led to many fixes and improvements.

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N.J., 1979.

[2] A. Aravkin, B. Bell, J. V. Burke, and G. Pillonetto, *An ℓ_1-Laplace robust Kalman smoother*, IEEE Trans. Automat. Control, 56 (2011), pp. 2898–2911.

[3] A. Aravkin, J. Burke, and G. Pillonetto, *Robust and trend following kalman smoothers using Student's t*, in Proceedings of the 16th IFAC Symposium on System Identification, 2012.

[4] A. Aravkin, J. Burke, and G. Pillonetto, *Sparse/robust estimation and kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory*, J. Mach. Learn. Res., to appear.

[5] A. Aravkin, M. P. Friedlander, F. Herrmann, and T. van Leeuwen, *Robust inversion, dimensionality reduction, and randomized sampling*, Math. Program., 134 (2012), pp. 101–125.

[6] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[7] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer, New York, 2003.

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

[9] E. Vanden Berg and M. P. Friedlander, *SPGL1: A Solver for Large-Scale Sparse Reconstruction*. http://www.cs.ubc.ca/labs/scl/index.php/Main/Spgl1 (June 2007).

[10] E. Vanden Berg and M. P. Friedlander, *Probing the pareto frontier for basis pursuit solutions*, SIAM J. Sci. Comput., 31 (2008), pp. 890–912.

[11] E. v. Berg and M. P. Friedlander, *Theoretical and empirical results for recovery from multiple measurements*, IEEE Trans. Inform. Theory, 56 (2010), pp. 2516–2527.

[12] E. v. Berg and M. P. Friedlander, *Sparse optimization with least-squares constraints*, SIAM J. Optim., 21 (2011), pp. 1201–1229.

[13] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*, Springer, New York, 2000.
[14] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.
[15] R. Brockett, *Finite Dimensional Linear Systems*, Wiley, New York, 1970.
[16] C. Combari, M. Laghdir, and L. Thibault, *Sous-différentiels de fonctions convexes composées*, Ann. Sci. Math. Québec, 18 (1994), pp. 119–148.
[17] F. Cucker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc., 39 (2001), pp. 1–49.
[18] D. L. Dohono and J. Tanner, *Sparse nonnegative solution of underdetermined linear equations by linear programming*, Proc. Natl. Acad. Sci. USA, 102 (2005), pp. 9446–9451.
[19] D. Donoho, *Compressed sensing*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1289–1306.
[20] B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani, *Least angle regression*, Annals Statist., 32 (2004), pp. 407–499.
[21] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, Elsevier, New York, 1976.
[22] T. Evgeniou, M. Pontil, and T. Poggio, *Regularization networks and support vector machines*, Adv. Comput. Math., 13 (2000), pp. 1–150.
[23] S. Farahmand, G. Giannakis, and D. Angelosante, *Doubly robust smoothing of dynamical processes via outlier sparsity constraints*, IEEE Trans. Signal Process., 59 (2011), pp. 4529–4543.
[24] J. Gao, *Robust l1 principal component analysis and its Bayesian variational inference*, Neural Comput., 20 (2008), pp. 555–572.
[25] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Monogr. Statist. Appl. Probab. 43, Chapman and Hall, London, 1990.
[26] T. J. Hastie, R. J. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, 2001.
[27] F. J. Herrmann, M. P. Friedlander, and O. Yilmaz, *Fighting the curse of dimensionality: compressive sensing in exploration seismology*, IEEE Signal Proc. Magazine, 29 (2012), pp. 88–100.
[28] J. B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, New York, 2001.
[29] P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.
[30] D. Mackay, *Bayesian non-linear modelling for the prediction competition.*, ASHRAE Trans., 100 (1994), pp. 3704–3716.
[31] D. J. C. MacKay, *Bayesian interpolation*, Neural Comput., 4 (1992), pp. 415–447.
[32] H. M. Markowitz, *Mean-Variance Analysis in Portfolio Choice and Capital Markets*, Wiley, New York, 1987.
[33] M. Pontil and A. Verri, *Properties of support vector machines*, Neural Comput., 10 (1998), pp. 955–974.
[34] R. T. Rockafellar, *Convex Analysis*, Princeton Landmarks Math., Princeton University Press, Princeton, NJ, 1970.
[35] R. T. Rockafellar, *Lagrange multipliers and optimality*, SIAM Rev., 35 (1993), pp. 183–238.
[36] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer, New York, 1998.
[37] S. Roweis and Z. Ghahramani, *A unifying review of linear Gaussian models*, Neural Comput., 11 (1999), pp. 305–345.
[38] S. Saitoh, *Theory of Reproducing Kernels and its Applications*, Longman, Longman, 1988.
[39] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, *New support vector algorithms*, Neural Comput., 12 (2000), pp. 1207–1245.
[40] R. Tapia, *The Isoperimetric Problem Revisited: Extracting a Short Proof of Sufficiency From Euler's Approach to Necessity*, Technical report, Rice University, 2013.
[41] R. Tibshirani, *Regression shrinkage and selection via the Lasso*, J. Roy. Statist. Soc. Ser. B., 58 (1996), pp. 267–288.
[42] M. Tipping, *Sparse Bayesian learning and the relevance vector machine*, J. Mach. Learn. Res., 1 (2001), pp. 211–244.
[43] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998.
[44] D. Wipf, B. Rao, and S. Nagarajan, *Latent variable Bayesian models for promoting sparsity*, IEEE Trans. Inform. Theory, 57 (2011), pp. 6236–6255.
[45] C. Zălinescu, *Convex Analysis in General Vector Spaces*, World Scientific, River Edge, NJ, 2002.