

# FOUNDATIONS OF GAUGE AND PERSPECTIVE DUALITY\*

A.Y. ARAVKIN<sup>†</sup>, J.V. BURKE<sup>‡</sup>, D. DRUSVYATSKIY<sup>§</sup>, M.P. FRIEDLANDER<sup>¶</sup>, AND  
K. MACPHEE<sup>||</sup>

**Abstract.** Common numerical methods for constrained convex optimization are predicated on efficiently computing nearest points to the feasible region. The presence of a design matrix in the constraints yields feasible regions with more complex geometries. When the functional components are gauges, there is an equivalent optimization problem—the gauge dual—where the matrix appears only in the objective function and the corresponding feasible region is easy to project onto. We revisit the foundations of gauge duality and show that the paradigm arises from an elementary perturbation perspective. We therefore put gauge duality and Fenchel duality on an equal footing, explain gauge dual variables as sensitivity measures, and show how to recover primal solutions from those of the gauge dual. In particular, we prove that optimal solutions of the Fenchel dual of the gauge dual are precisely the primal solutions rescaled by the optimal value. The gauge duality framework is extended beyond gauges to the setting when the functional components are general nonnegative convex functions, including problems with piecewise linear quadratic functions and constraints that arise from generalized linear models used in regression.

**Key words.** convex optimization, gauge duality, nonsmooth optimization, perspective function

**AMS subject classifications.** 90C15, 90C25

**1. Introduction.** This work revolves around optimization problems of the form

$$(G_p) \quad \underset{x}{\text{minimize}} \quad \kappa(x) \quad \text{subject to} \quad \rho(b - Ax) \leq \sigma,$$

where  $A: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is a linear map,  $b \in \mathbb{R}^m$  is an  $m$ -vector, and  $\kappa$  and  $\rho$  are closed gauges – nonnegative, sublinear functions that vanish at the origin. In statistical and machine learning applications,  $\kappa$  is often a structure-inducing regularizer, such as the elastic net for group detection [23]. The function  $\rho$  may be interpreted as a penalty that measures the misfit between measurements  $b$  and the prediction  $Ax$ . For example,  $\rho$  can be the 2-norm or the Huber [15] function in the case of regression, or the logistic loss, used for classification problems [1, 16]. In high-dimensional applications, the number of measurements  $m$  is often much smaller than the dimension  $n$  of the predictor  $x$ , and the matrix  $A$  is only available through matrix-vector products  $Ax$  and  $A^T y$ .

The formulation  $(G_p)$  gives rise to two different “dual” problems:

$$(L_d) \quad \underset{y}{\text{maximize}} \quad \langle b, y \rangle - \sigma \rho^\circ(y) \quad \text{subject to} \quad \kappa^\circ(A^T y) \leq 1,$$

$$(G_d) \quad \underset{y}{\text{minimize}} \quad \kappa^\circ(A^T y) \quad \text{subject to} \quad \langle b, y \rangle - \sigma \rho^\circ(y) \geq 1.$$

---

\*February 28, 2017

<sup>†</sup>Department of Applied Mathematics, University of Washington, Seattle (sasha.aravkin@gmail.com). Research supported by the Washington Research Foundation Data Science Professorship.

<sup>‡</sup>Department of Mathematics, University of Washington, Seattle (burke@uw.edu). Research supported in part by NSF award DMS-1514559.

<sup>§</sup>Department of Mathematics, University of Washington, Seattle (ddrusv@uw.edu). Research partially supported by AFOSR YIP award FA9550-15-1-0237.

<sup>¶</sup>Departments of Computer Science and Mathematics, University of British Columbia, Vancouver, BC, Canada (mpf@cs.ubc.ca). Research supported by the ONR award N00014-16-1-2242.

<sup>||</sup>Department of Mathematics, University of Washington, Seattle (kmacphee@uw.edu).

Here  $\rho^\circ$  and  $\kappa^\circ$  are the polar gauges; see [section 1.2](#) for a precise definition. The first formulation  $(\mathbf{L}_d)$  is the classical Lagrangian (or Fenchel) dual, routinely used in the design and analysis of algorithms. Under mild interiority conditions, equality

$$\text{val}(\mathbf{G}_p) = \text{val}(\mathbf{L}_d)$$

holds and the optimal value of  $(\mathbf{L}_d)$  is attained. The second formulation  $(\mathbf{G}_d)$  is called the *gauge dual* and is less well-known. Gauge duality was introduced by Freund [11] for minimizing nonnegative sublinear functions over convex sets, and subsequently examined by Friedlander, Macédo, and Pong [13]. Under standard interiority conditions, equality

$$1 = \text{val}(\mathbf{G}_p) \cdot \text{val}(\mathbf{G}_d)$$

holds and the optimal value of  $(\mathbf{G}_d)$  is attained.

The gauge dual  $(\mathbf{G}_d)$  can be preferable for computation to the the primal  $(\mathbf{G}_p)$  and the Lagrangian dual  $(\mathbf{L}_d)$ . Indeed, numerous convex optimization algorithms rely on being able to project onto the feasible region easily. The appearance of the matrix  $A$  in the constraints of both  $(\mathbf{G}_p)$  and  $(\mathbf{L}_d)$  precludes such methods from being directly applicable. In contrast, the design matrix  $A$  appears in the gauge dual  $(\mathbf{G}_d)$  only in the objective. Moreover, typical applications occur in the regime  $m \ll n$ . For example,  $m$  is often logarithmic in  $n$  [6, 7, 10, 21]. Since the decision variables  $y$  of  $(\mathbf{G}_d)$  lie in the small dimensional space  $\mathbb{R}^m$ , projections onto the feasible region can be computed efficiently, for example by interior-point methods. Friedlander and Macédo [12] use gauge duality to derive an effective algorithm for an important class of spectral optimization problems that arise in signal-recovery applications, including phase recovery and blind deconvolution.

**1.1. A roadmap.** Broadly speaking, our goals are two-fold. First, we revisit the foundations of gauge duality in [section 3](#), reformulating these as in the modern approach to duality through a “perturbation framework”. That is, following Rockafellar and Wets [20, 11.H], consider an arbitrary convex function  $F$  on  $\mathbb{R}^n \times \mathbb{R}^m$  and define the *value functions*

$$(1.1) \quad p(y) := \inf_x F(x, y) \quad \text{and} \quad q(x) := \inf_y F^*(x, y).$$

This set-up immediately yields the primal-dual pair

$$(1.2) \quad p(0) = \inf_x F(x, 0) \quad \text{and} \quad p^{**}(0) = \sup_y -F^*(0, y).$$

Fenchel duality is a standard example that follows from an appropriate choice of  $F$ . We show that gauge duality fits equally well into this framework under a judicious choice of the perturbation function  $F$ , thereby putting Fenchel and gauge dualities on an equal footing. Strong duality, primal-dual optimality conditions, and an interpretation of the gauge dual solutions as sensitivity measures—i.e., subgradients of the value function—quickly follow ([section 3.2](#)). These results, in particular, answer the main open question posed by Freund in his original work [11] on an interpretation of gauge dual variables as sensitivity measures.

We also prove a striking relationship between optimal solutions of the Lagrangian dual of the gauge dual and the primal problem: the two coincide up to scaling by the optimal value function ([section 3.4](#)). Consequently Lagrangian primal-dual methods applied to the gauge dual can always be trivially translated to methods on the original

primal problem. We explore this viewpoint in [section 3.4](#) and illustrate its application to Chambolle and Pock’s primal-dual algorithm [\[8\]](#) in [section 7](#).

Our second aim is to extend the applicability of the gauge duality paradigm beyond gauges to capture more general convex problems. [Section 4](#) extends gauge duality to problems that involve convex functions that are merely nonnegative. The approach is based on using the perspective transform

$$f^p(x, \lambda) := \lambda f(\lambda^{-1}x), \quad \text{for } \lambda > 0$$

of a convex function  $f$  to reduce to the gauge setting. We call the resulting dual problem the *perspective dual*. The perspective-polar transformation, needed to derive the perspective dual problem, is developed in [section 4](#). We provide concrete illustrations of perspective duality for the logistic loss and the family of piecewise linear-quadratic functions in [section 5](#), which are used often in data-fitting applications. Numerical illustrations for several case-studies of perspective duals appear in [section 7](#).

**1.2. Notation.** The derivation of our results relies mainly on standard notions from convex analysis [\[18\]](#). We define these briefly below.

Throughout the paper,  $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$  denotes the extended-real-line, while  $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  and  $g: \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$  denote general closed convex functions. We routinely use the symbols  $b$  and  $A$  for an  $m$ -vector and a  $m \times n$  matrix, respectively. The domain and epigraph of  $f$  are the sets

$$\text{dom } f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\} \quad \text{and} \quad \text{epi } f := \{(x, \mu) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \leq \mu\}.$$

A function is *proper* if it has nonempty domain and is never  $-\infty$ , and it is called *closed* if its epigraph is closed, which corresponds to lower semi-continuity [\[18, Theorem 7.1\]](#). The closure of  $f$ , denoted  $\text{cl } f$ , is the function whose epigraph is the closure of  $\text{epi } f$ . The indicator function of a set  $\mathcal{Q} \subseteq \mathbb{R}^n$  is denoted by

$$\delta_{\mathcal{Q}}(x) = \begin{cases} 0 & \text{if } x \in \mathcal{Q} \\ +\infty & \text{otherwise.} \end{cases}$$

We use the symbol  $\text{ri}(\mathcal{Q})$  to denote the interior of  $\mathcal{Q}$  relative to its affine span. It is a standard fact that if a convex function  $f$  is finite at a point  $x \in \text{ri}(\text{dom } f)$ , then  $f$  is proper. We will use this observation implicitly in what follows.

The *conjugate* of a proper convex function  $f$  is  $f^*(y) := \sup_x \{\langle x, y \rangle - f(x)\}$ , which is a proper closed convex function [\[18, Theorem 12.2\]](#). In particular, for any convex set  $\mathcal{Q}$ , the conjugate  $\delta_{\mathcal{Q}}^*(y) = \sup_{x \in \mathcal{Q}} \langle x, y \rangle$  is called the *support function* of  $\mathcal{Q}$ . For any  $x \in \text{dom } f$ , the *subdifferential* of  $f$  at  $x$  is the set  $\partial f(x) := \{v \mid f(y) \geq f(x) + \langle v, y - x \rangle, \forall y\}$ . For any convex cone  $\mathcal{K}$ , the *polar cone* is the set

$$\mathcal{K}^\circ := \{y \mid \langle y, x \rangle \leq 0, \forall x \in \mathcal{K}\}.$$

Observe the equality  $\delta_{\mathcal{K}}^* = \delta_{\mathcal{K}^\circ}$  for any convex cone  $\mathcal{K}$ .

For any convex function  $f: \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ , its *perspective*  $f^p: \mathbb{R}^{n+1} \rightarrow \overline{\mathbb{R}}$  is the function whose epigraph  $\text{epi } f^p \subset \mathbb{R}^{n+1} \times \mathbb{R}$  is the cone generated by the set  $(\text{epi } f) \times \{1\}$ . Equivalently, we may write

$$f^p(x, \lambda) = \begin{cases} \lambda f(\lambda^{-1}x) & \text{if } \lambda > 0 \\ \delta_{\{0\}}(x) & \text{if } \lambda = 0 \\ +\infty & \text{if } \lambda < 0. \end{cases}$$

Though  $f^p$  may not be closed, the closure of  $f_p$  admits the convenient description

$$f^\pi(x, \lambda) := \begin{cases} \lambda f(\lambda^{-1}x) & \text{if } \lambda > 0 \\ f^\infty(x) & \text{if } \lambda = 0 \\ +\infty & \text{if } \lambda < 0, \end{cases}$$

where  $f^\infty(x)$  is the *recession function* of  $f$  [18, Theorem 8.5]. Importantly, when  $f$  is a proper convex function,  $f^\pi$  is positively homogeneous. A calculus for the perspective transform  $f \mapsto f^\pi$  is described by Aravkin, Burke, and Friedlander [2, Section 3.3]. We often apply more than one transformation to a function, and in those cases, the multiple transformations are applied in the order that they appear; e.g.,  $f^{\circ*} \equiv (f^\circ)^*$ .

**2. Gauge optimization and duality.** In this section, we review the main elements of gauge duality. The original description is due to Freund [11], but here we summarize the more recent treatment given by Friedlander, Macêdo, and Pong [13].

A convex function  $\kappa : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is called a *gauge* if it is nonnegative, positively homogeneous, and vanishes at the origin. The symbols  $\kappa : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  and  $\rho : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$  will always denote closed gauges. The polar of a gauge  $\kappa$  is the function  $\kappa^\circ$  defined by

$$(2.1) \quad \kappa^\circ(y) := \inf \{ \mu > 0 \mid \langle x, y \rangle \leq \mu \kappa(x), \forall x \},$$

which is also a gauge. For example, if  $\kappa$  is a norm than  $\kappa^\circ$  is the corresponding dual norm. Note the equality

$$\text{epi } \kappa^\circ = \{ (y, -\lambda) \mid (y, \lambda) \in (\text{epi } \kappa)^\circ \}.$$

It follows directly from the definition and positive homogeneity of  $\kappa$  that the polar can be characterized as the support function to the unit level set:

$$(2.2) \quad \kappa^\circ(y) = \sup \{ \langle y, x \rangle \mid \kappa(x) \leq 1 \}.$$

Moreover,  $\kappa$  and its polar  $\kappa^\circ$  satisfy a Hölder-like inequality

$$(2.3) \quad \langle x, y \rangle \leq \kappa(x) \cdot \kappa^\circ(y) \quad \forall x \in \text{dom } \kappa, \forall y \in \text{dom } \kappa^\circ,$$

which we refer to as the *polar-gauge inequality*.

Define the following primal and dual feasible sets:

$$(2.4) \quad \mathcal{F}_p := \{ x \mid \rho(b - Ax) \leq \sigma \} \quad \text{and} \quad \mathcal{F}_d := \{ y \mid \langle b, y \rangle - \sigma \rho^\circ(y) \geq 1 \}.$$

The gauge primal ( $\mathbf{G}_p$ ) and dual ( $\mathbf{G}_d$ ) problems are said to be feasible, respectively, if the following intersections are nonempty:

$$\mathcal{F}_p \cap (\text{dom } \kappa) \quad \text{and} \quad A^T \mathcal{F}_d \cap (\text{dom } \kappa^\circ).$$

The primal and dual problems are *relatively strictly feasible*, respectively, if the following intersections are nonempty:

$$\text{ri } \mathcal{F}_p \cap (\text{ri dom } \kappa) \quad \text{and} \quad A^T \text{ri } \mathcal{F}_d \cap (\text{ri dom } \kappa^\circ).$$

If the intersections above are nonempty, with interior replacing relative interior, then we say that the problems are *strictly feasible*, respectively.

Assume throughout that  $\rho(b) > \sigma$ . Otherwise,  $\mathcal{F}_p$  contains the origin, which is a trivial solution of ( $\mathbf{G}_p$ ). We generally assume that  $\sigma$  is positive, though in certain cases, it is useful to allow  $\sigma = 0$  and then assume that  $\rho^{-1}(0) = \{0\}$ ; this allows us to extend many of the following results to problems where the feasible set  $\mathcal{F}_p$  is affine.

LEMMA 2.1 (Primal-dual constraint activity). *If the primal optimal value is attained at  $x^*$  with  $\kappa(x^*) > 0$ , then  $\rho(b - Ax^*) = \sigma$ . Similarly, if the dual optimal value is attained at  $y^*$  with  $\kappa^\circ(A^T y^*) > 0$ , then  $\langle b, y^* \rangle - \sigma \rho^\circ(y^*) = 1$ .*

*Proof.* We prove the contrapositive. If  $\rho(b - Ax^*) < \sigma$ , then by lower-semicontinuity of  $\rho$ , we have  $\rho(b - A(\lambda x^*)) < \sigma$  for all  $\lambda > 0$  close to 1. Consequently, we deduce  $\lambda \kappa(x^*) = \kappa(\lambda x^*) \geq \kappa(x^*)$ , with strict inequality unless  $\kappa(x^*) = 0$ . Since  $x^*$  is optimal, we conclude  $\kappa(x^*) = 0$ . The proof of the dual statement is similar.  $\square$

The duality relations in the gauge framework follow analogous principles to Lagrange duality, except that instead of an additive relationship between the primal and dual optimal values  $\nu_p$  and  $\nu_d$ , the relationship is multiplicative. The next result summarizes weak and strong duality for gauge optimization.

THEOREM 2.2 (Gauge duality [13]). *The following relationships hold for the gauge primal-dual pair  $(\mathbf{G}_p)$  and  $(\mathbf{G}_d)$ .*

(a) (Weak duality) *If  $x$  and  $y$  are primal and dual feasible, then*

$$1 \leq v_p v_d \leq \kappa(x) \cdot \kappa^\circ(A^T y).$$

(b) (Strong duality) *If the dual (resp. primal) is feasible and the primal (resp. dual) is relatively strictly feasible, then  $\nu_p \nu_d = 1$  and the gauge dual (resp. primal) attains its optimal value.*

**3. Perturbation analysis for gauge duality.** Modern treatment of duality in convex optimization is based on an interpretation of multipliers as giving sensitivity information relative to perturbations in the problem data. No such analysis, however, has existed for gauge duality. In this section we show that for a particular kind of perturbation, the gauge dual  $(\mathbf{G}_d)$  can in fact be derived via such an approach. This resolves a question posed by Friedlander, Macêdo, and Pong [13].

**3.1. The perturbation framework.** In this section we review the perturbation argument for deriving duality. Our summary follows the discussion in Rockafellar and Wets [20, 11.H]. Fix an arbitrary convex function  $F: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$  and consider the value functions defined by (1.1)–(1.2). Observe the equality  $q(0) = -p^{**}(0)$ . Rockafellar-Fenchel duality for the problem

$$\underset{x}{\text{minimize}} \quad h(Ax + y) + g(x),$$

where  $h$  and  $g$  are closed and convex, is obtained by setting  $F(x, y) = h(Ax + y) + g(x)$ . In that case, the primal-dual pair takes the familiar form

$$p(0) = \inf_x \{ h(Ax) + g(x) \} \quad \text{and} \quad p^{**}(0) = \sup_y \{ -h^*(-y) - g(A^T y) \}.$$

Under certain conditions, described in the following theorem, strong duality holds, i.e.,  $p(0) = p^{**}(0)$ , and the optimal value is attained.

THEOREM 3.1 (Multipliers and sensitivity [20, Theorem 11.39]). *Consider the primal-dual pair (1.2), where  $F: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$  is proper, closed, and convex.*

- (a) *The inequality  $p(0) \geq -q(0)$  always holds.*
- (b) *If  $0 \in \text{ri}(\text{dom } p)$ , then equality  $p(0) = -q(0)$  holds and the infimum  $q(0)$  is attained, if finite. Similarly, if  $0 \in \text{ri}(\text{dom } q)$ , then equality  $p(0) = -q(0)$  holds and the infimum  $p(0)$  is attained, if finite.*

- (c) The set  $\operatorname{argmax}_y -F^*(0, y)$  is nonempty and bounded if and only if  $0 \in \operatorname{int}(\operatorname{dom} p)$  and  $p(0)$  is finite, in which case  $\partial p(0) = \operatorname{argmax}_y -F^*(0, y)$ .
- (d) The set  $\operatorname{argmin}_x F(x, 0)$  is nonempty and bounded if and only if  $0 \in \operatorname{int}(\operatorname{dom} q)$  and  $q(0)$  is finite, in which case  $\partial q(0) = \operatorname{argmin}_x F(x, 0)$ .
- (e) Optimal solutions are characterized jointly through the conditions

$$\left. \begin{array}{l} \bar{x} \in \operatorname{argmin}_x F(x, 0) \\ \bar{y} \in \operatorname{argmax}_y -F^*(0, y) \\ F(\bar{x}, 0) = -F^*(0, \bar{y}) \end{array} \right\} \iff (0, \bar{y}) \in \partial F(\bar{x}, 0) \iff (\bar{x}, 0) \in \partial F^*(0, \bar{y}).$$

*Remark 3.2.* Part (b) of [Theorem 3.1](#) is stated in [20, Theorem 11.39] with interior in place of relative interior. We give a quick argument here for the claimed result with relative interiors. Suppose  $0 \in \operatorname{ri}(\operatorname{dom} p)$ . If  $p(0) = -\infty$ , then equality  $p(0) = -q(0)$  follows by (a). Hence we can suppose that  $p(0)$  is finite, and therefore  $p$  is proper. Thus there exists a subgradient  $\phi \in \partial p(0)$  [18, Theorem 23.4]. By the subgradient inequality, the following holds for any  $y$ :

$$p(0) \leq p(y) - \langle \phi, y \rangle = \inf_x \left\{ F(x, y) - \left\langle \begin{pmatrix} 0 \\ \phi \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \right\rangle \right\}.$$

Taking the infimum over  $y$ , we recognize the right-hand side as  $-F^*(0, \phi)$ . We deduce  $p(0) \leq -F^*(0, \phi) \leq -q(0)$ . Combining this with Part (a) of [Theorem 3.1](#) yields  $p(0) = -q(0)$  and we see that  $\phi$  attains  $\inf_y F^*(0, y)$ . The symmetric argument for the case  $0 \in \operatorname{ri}(\operatorname{dom} q)$  is analogous.

**3.2. Derivation of gauge duality as a perturbation.** We now show that the problems  $(G_p)$  and  $(G_d)$  constitute a primal-dual pair under the framework set out by [Theorem 3.1](#). The key is to postulate the correct pairing function  $F(\cdot, \cdot)$ .

**3.2.1. The perturbation function.** Our starting point is the primal perturbation scheme:

$$v_p(y) := \inf_{\mu > 0, x} \{ \mu \mid \rho(b - Ax + \mu y) \leq \sigma, \kappa(x) \leq \mu \}.$$

Note that  $v_p(0)$  is equal to the optimal value of the primal  $(G_p)$ . Since  $y$  and  $\mu$  multiply each other in the description above, it is convenient to reparametrize the problem by setting  $\lambda := 1/\mu$  and  $w := x/\mu$ . By positive homogeneity of  $\kappa$  and  $\rho$ , this yields the equivalent description

$$v_p(y) = \inf_{\lambda > 0, w} \{ 1/\lambda \mid \rho(\lambda b - Aw + y) \leq \sigma \lambda, w \in \mathcal{U}_\kappa \},$$

where  $\mathcal{U}_\kappa := \{ w \mid \kappa(w) \leq 1 \}$  is the unit level set for  $\kappa$ . In particular, this reparameterization shows that  $v_p$  is convex because it is the infimal projection of a convex function; it is proper when the primal  $(G_p)$  is feasible. Note that minimizing  $1/\lambda$  is equivalent to minimizing  $-\lambda$ . With this in mind, define the convex function  $F: \mathbb{R}^{n+1} \times \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$ :

$$(3.1) \quad F(w, \lambda, y) := -\lambda + \delta_{(\operatorname{epi} \rho) \times \mathcal{U}_\kappa} \left( W \begin{pmatrix} w \\ \lambda \\ y \end{pmatrix} \right), \quad W := \begin{pmatrix} -A & b & I \\ 0 & \sigma & 0 \\ I & 0 & 0 \end{pmatrix}.$$

The function  $F$  is proper and closed because  $0 \in \operatorname{dom} F$ , and  $\kappa$  and  $\rho$  are closed. The associated infimal projection

$$p(y) := \inf_{w, \lambda} F(w, \lambda, y)$$

is essentially the negative reciprocal of  $v_p$ . We formalize this in the following lemma. We omit the proof since it is immediate.

LEMMA 3.3. *Equality  $v_p(y) = -1/p(y)$  holds provided that  $v_p(y)$  is nonzero and finite. Moreover,  $v_p(y) = 0$  implies  $p(y) = -\infty$ , and  $p(y) = 0$  implies  $v_p(y) = +\infty$ .*

We now compute the conjugate of  $F$ , which is needed to derive the dual value function  $q$ . By Rockafellar and Wets [20, Theorem 11.23(b)],

$$F^*(w, \lambda, y) = \text{cl} \inf_{z, \beta, r} \left\{ \delta_{(\text{epi } \rho) \times \mathcal{U}_\kappa}^* \begin{pmatrix} z \\ \beta \\ r \end{pmatrix} \mid W^T \begin{pmatrix} z \\ \beta \\ r \end{pmatrix} = \begin{pmatrix} w \\ \lambda \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\},$$

where the closure operation  $\text{cl}$  is applied to the function on the right-hand-side with respect to the argument  $(w, \lambda, y)$ . Because  $W$  is nonsingular, there is a unique vector  $(z, \beta, r)$  that satisfies the constraints in the description of  $F^*$ . The closure operation  $\text{cl}$  therefore turns out to be superfluous, and we can further simplify the description to

$$\begin{aligned} F^*(w, \lambda, y) &= \delta_{(\text{epi } \rho) \times \mathcal{U}_\kappa}^* \begin{pmatrix} y \\ \sigma^{-1}(1 + \lambda - \langle b, y \rangle) \\ w + A^T y \end{pmatrix} \\ &= \delta_{\text{epi } \rho}^* \begin{pmatrix} y \\ \sigma^{-1}(1 + \lambda - \langle b, y \rangle) \end{pmatrix} + \delta_{\mathcal{U}_\kappa}^*(w + A^T y). \end{aligned}$$

Taking into account the equalities  $\delta_{\text{epi } \rho}^*(z_1, z_2) = \delta_{\text{epi } \rho^\circ}(z_1, -z_2)$  and  $\delta_{\mathcal{U}_\kappa}^* = \kappa^\circ$ , this expression transforms to

$$F^*(w, \lambda, y) = \delta_{\text{epi } \rho^\circ} \begin{pmatrix} y \\ -\sigma^{-1}(1 + \lambda - \langle b, y \rangle) \end{pmatrix} + \kappa^\circ(w + A^T y).$$

In particular, we conclude

$$(3.2) \quad F^*(0, 0, y) = \begin{cases} \kappa^\circ(A^T y) & \text{if } \langle b, y \rangle - \sigma \rho^\circ(y) \geq 1 \\ +\infty & \text{otherwise,} \end{cases}$$

Thus the dual problem  $\sup_y -F^*(0, 0, y)$  recovers, up to a sign change, the required gauge dual problem.

We are thus justified in defining the dual perturbation function  $v_d(w, \lambda) := \inf_y F^*(w, \lambda, y)$  or equivalently

$$v_d(w, \lambda) = \inf_y \{ \kappa^\circ(A^T y + w) \mid \langle b, y \rangle - \sigma \rho^\circ(y) \geq 1 + \lambda \}.$$

Note that  $v_d(0, 0)$  is the optimal value of  $(\mathbf{G}_d)$ . In summary,  $(-1/v_p)$  and  $v_d$ , respectively, play the roles of  $p$  and  $q$  as defined in (1.1) and used in Theorem 3.1.

**3.2.2. Proof of gauge duality (Theorem 2.2).** We now use the perturbation framework to prove the gauge duality result given by Theorem 2.2. The following auxiliary result ties the feasibility of the gauge pair  $(\mathbf{G}_p)$  and  $(\mathbf{G}_d)$  to the domain of the value function. The proof of this result, which is largely an application of the calculus of relative interiors, is deferred to Appendix A.

LEMMA 3.4 (Feasibility and domain of the value function). *If the primal  $(\mathbf{G}_p)$  is relatively strictly feasible, then  $0 \in \text{ri}(\text{dom } p)$ . If the dual  $(\mathbf{G}_d)$  is relatively strictly feasible, then  $0 \in \text{ri}(\text{dom } v_d)$ . The analogous implications, where the  $\text{ri}$  operator is replaced by the  $\text{int}$  operator, hold under strict feasibility (not relative).*

As in the hypotheses of [Theorem 2.2](#), in this subsection we denote the optimal primal and dual values by  $v_p$  and  $v_d$  (without arguments), i.e.,  $v_p \equiv v_p(0)$  and  $v_d \equiv v_d(0, 0)$ . Similarly, we let  $p \equiv p(0)$ .

*Proof of [Theorem 2.2](#). Part (a):* We proceed by proving that the two inequalities (i)  $(1/v_p) \leq v_d$  and (ii)  $(1/v_d) \leq v_p$  hold always. This in particular will imply that the assumptions of part (a) guarantee  $v_p$  and  $v_d$  are nonzero and finite. Hence the conclusion of (a) follows. We begin with (i). [Theorem 3.1](#) guarantees the inequality

$$(3.3) \quad p \geq -\inf_y F^*(0, 0, y) = -v_d.$$

By [Lemma 3.3](#), whenever  $v_p$  is nonzero and finite, equality  $p = -1/v_p$  holds, which together with (3.3) yields (i). If, on the other hand,  $v_p = +\infty$ , then (i) is trivial. Finally, if  $v_p = 0$ , [Lemma 3.3](#) yields  $p = -\infty$ , and hence (3.3) implies  $v_d = +\infty$ , and (i) again holds. Thus, (i) holds always. To establish (ii), it suffices to consider the case  $v_d = 0$ . From (3.3) we conclude  $p \geq 0$ , that is either  $p = 0$  or  $p = +\infty$ . By [Lemma 3.3](#), the first case  $p = 0$  implies  $v_p = +\infty$  and therefore (ii) holds. The second case  $p = +\infty$  implies that the primal problem is infeasible, that is  $v_p = +\infty$ , and again (ii) holds. Thus (ii) holds always, as required.

*Part (b):* Suppose the dual is feasible and the primal is relatively strictly feasible. Part (a) implies that both  $v_p$  and  $v_d$  are nonzero and finite and hence  $1 \leq v_p \cdot v_d = -v_d/p$ . On the other hand, by [Lemma 3.4](#) the assumption that the primal is relatively strictly feasible implies  $0 \in \text{ri}(\text{dom } p)$ . This last inequality thus implies  $p$  is finite, and hence  $p(\cdot)$  is proper. Hence by [Theorem 3.1](#), equality  $p = -v_d$  holds and the infimum in the dual  $v_d$  is attained. Thus we deduce  $1 = v_p \cdot v_d$ , as claimed.

Conversely, suppose that the primal is feasible and the dual is relatively strictly feasible. The first assumption implies  $0 \in \text{ri}(\text{dom } q)$  by [Lemma 3.4](#). This in turn implies  $p = -v_d$  and that the infimum in  $p$  is attained. Since the primal is feasible, by [Lemma 3.3](#),  $p$  is nonzero, and hence  $1 = v_p \cdot v_d$  and the infimum in the primal is attained.  $\square$

**3.3. Optimality conditions.** The perturbation framework can be harnessed to develop optimality conditions for the gauge pair that relate the primal-dual solutions to subgradients of the corresponding value function. This yields a version of parts (c) and (d) in [Theorem 3.1](#) specialized to gauge duality.

**THEOREM 3.5** (Gauge multipliers and sensitivity). *The following relationships hold for the gauge primal-dual pair  $(G_p)$  and  $(G_d)$ .*

(a) *If the primal is strictly feasible and the dual is feasible, then the set of optimal solutions for the dual is nonempty and bounded, and coincides with*

$$\partial p(0) = \partial(-1/v_p)(0).$$

(b) *If the dual is strictly feasible and the primal is feasible, then the set of optimal solutions for the primal is nonempty and bounded with solutions given by  $x^* = w^*/\lambda^*$ , where*

$$(w^*, \lambda^*) \in \partial v_d(0, 0) \quad \text{and} \quad \lambda^* > 0.$$

*Proof.* Part (a). Because the primal problem is strictly feasible, it follows from [Lemma 3.4](#) that  $0 \in \text{int}(\text{dom } p)$ , and because the dual is feasible,  $p(0)$  is finite. [Theorem 3.1](#) and [Lemma 3.3](#) then imply the conclusion of Part (a).

Part (b). Because the dual problem is strictly feasible, it follows from [Lemma 3.4](#) that  $0 \in \text{int}(\text{dom } v_d)$ , and because the primal is feasible,  $v_d(0)$  is finite. [Theorem 3.1](#) then implies that the optimal primal set is nonempty and bounded, and

$\operatorname{argmin}_{w,\lambda} F(w, \lambda, 0) = \partial v_d(0, 0)$ . Because the primal problem is feasible, any pair  $(w^*, \lambda^*) \in \operatorname{argmin}_{w,\lambda} F(w, \lambda, 0)$  must satisfy  $\lambda^* > 0$  by [Lemma 3.3](#). Thus, this inclusion is equivalent to  $x^* = w^*/\lambda^*$  being optimal for the primal problem, with optimal value  $1/\lambda^*$ . This proves Part (b).  $\square$

We next use the sensitivity interpretation given by [Theorem 3.5](#) to develop a set of explicit necessary and sufficient optimality conditions that mirror the more familiar KKT conditions from Lagrange duality.

**THEOREM 3.6** (Optimality conditions). *Suppose both the gauge primal and gauge dual problems are strictly feasible. Then the pair  $(x^*, y^*)$  is primal-dual optimal if and only if it satisfies the conditions*

$$(3.4a) \quad \rho(b - Ax^*) = \sigma \quad (\text{primal activity})$$

$$(3.4b) \quad \langle b, y^* \rangle - \sigma \rho^\circ(y^*) = 1 \quad (\text{dual activity})$$

$$(3.4c) \quad \langle x^*, A^T y^* \rangle = \kappa(x^*) \cdot \kappa^\circ(A^T y^*) \quad (\text{objective alignment})$$

$$(3.4d) \quad \langle b - Ax^*, y^* \rangle = \rho(b - Ax^*) \cdot \rho^\circ(y^*). \quad (\text{constraint alignment})$$

*Proof.* Suppose  $(x^*, y^*)$  is primal-dual optimal. Then by strong duality,  $\kappa(x^*)$  and  $\kappa^\circ(A^T y^*)$  are both nonzero and finite, and [Lemma 2.1](#) tells us that the constraints  $\rho(b - Ax^*) = \sigma$  and  $\langle b, y^* \rangle - \sigma \rho^\circ(y^*) = 1$  are active. Hence, (3.4a)-(3.4b) hold.

Now define  $\lambda^* := 1/\kappa(x^*)$  and  $w^* := \lambda^* x^*$ . Note that  $\kappa(w^*) = 1$ . By [Theorem 3.1\(e\)](#) and [Theorem 3.5\(b\)](#),  $(x^*, y^*)$  is primal-dual optimal if and only if  $(0, 0, y^*) \in \partial F(w^*, \lambda^*, 0)$ . By [[20](#), Theorem 6.14] and [[18](#), Theorem 23], whenever  $\rho(\lambda b - Aw) = \lambda \sigma > 0$  and  $\kappa(w) \leq 1$  we have

$$\begin{aligned} \partial F(w, \lambda, 0) &= - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + W^T N_{(\operatorname{epi} \rho) \times \mathcal{U}_k} \begin{pmatrix} \lambda b - Aw \\ \sigma \lambda \\ w \end{pmatrix} \\ &= - \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + W^T (\operatorname{cone}(\partial \rho(\lambda b - Aw) \times \{-1\})) \times \{v \mid \kappa^\circ(v) \leq \langle v, w \rangle\} \\ &= \left\{ \begin{pmatrix} v - \mu A^T z \\ \mu(\langle b, z \rangle - \sigma) - 1 \\ \mu z \end{pmatrix} \mid \mu \geq 0, z \in \partial \rho(\lambda b - Aw), \kappa^\circ(v) \leq \langle v, w \rangle \right\}. \end{aligned}$$

In particular, this subdifferential formula holds for  $(w, \lambda) = (w^*, \lambda^*)$ . We deduce existence of  $z^* \in \partial \rho(\lambda^* b - Aw^*)$  and  $\mu^* \geq 0$  such that the following hold:

$$(3.5a) \quad \mu^* (\langle b, z^* \rangle - \sigma) = 1$$

$$(3.5b) \quad y^* = \mu^* z^*$$

$$(3.5c) \quad \kappa^\circ(\mu^* A^T z^*) \leq \langle \mu^* A^T z^*, w^* \rangle.$$

Notice that  $\mu^* = 0$  cannot satisfy (3.5a), so (3.5c) together with the polar-gauge inequality implies

$$\kappa^\circ(A^T y^*) \kappa(w^*) = \kappa^\circ(A^T y^*) \leq \langle A^T y^*, w^* \rangle \leq \kappa^\circ(A^T y^*) \kappa(w^*).$$

Therefore equality holds throughout, and dividing through by  $\lambda^*$  we see that (3.4c) is satisfied. Finally, recall that from the characterization (2.2) of the polar,

$$(3.6) \quad \mu \rho(u) = \mu \sup_y \{ \langle y, u \rangle \mid \rho^\circ(y) \leq 1 \} = \sup_y \{ \langle y, u \rangle \mid \rho^\circ(y) \leq \mu \}$$

which implies

$$(3.7) \quad \mu \partial \rho(u) = \operatorname{argmax}_y \{ \langle y, u \rangle \mid \rho^\circ(y) \leq \mu \}.$$

In particular, if  $y \in \mu \partial \rho(u)$  then  $\langle y, u \rangle \geq \langle 0, u \rangle = 0$ . If  $\rho(\lambda^* b - Aw^*) = 0$ , then (3.7) implies that  $y^* = \mu^* z^* \in \mu^* \partial \rho(\lambda^* b - Aw^*)$  must satisfy

$$0 \leq \langle y^*, \lambda^* b - Aw^* \rangle \leq \rho(\lambda^* b - Aw^*) \cdot \rho^\circ(y^*) = 0,$$

which gives condition (3.4d) after dividing through by  $\lambda^*$ . On the other hand, if  $\rho(u) > 0$  then the set (3.7) is given by  $\{y \mid \mu \rho(u) = \langle y, u \rangle = \rho^\circ(y) \rho(u)\}$ . Thus we again have  $\langle y^*, \lambda^* b - Aw^* \rangle = \rho^\circ(y^*) \cdot \rho(\lambda^* b - Aw^*)$ , and dividing through by  $\lambda^*$  gives (3.4d). This finishes one direction of the proof.

For the reverse implication, suppose that  $(\bar{x}, \bar{y})$  satisfies (3.4a)-(3.4d). Then clearly  $\bar{x}$  satisfies the primal constraint, and  $\bar{y}$  satisfies the dual constraint. By weak duality, to show that  $(\bar{x}, \bar{y})$  is primal-dual optimal it is sufficient to show that  $\kappa(\bar{x}) \cdot \kappa^\circ(A^T \bar{y}) = 1$ . Adding (3.4c) and (3.4d), we obtain

$$\langle b, \bar{y} \rangle = \kappa(\bar{x}) \cdot \kappa^\circ(A^T \bar{y}) + \rho(b - A\bar{x}) \cdot \rho^\circ(\bar{y}).$$

Plug in  $\rho(b - A\bar{x}) = \sigma$  and then use (3.4b) to get  $\kappa(\bar{x}) \cdot \kappa^\circ(A^T \bar{y}) = 1$ , as desired.  $\square$

The following corollary describes a variation of the optimality conditions outlined by Theorem 3.6. These conditions assume that a solution  $y^*$  of the dual problem is available, and gives conditions that can be used to determine a corresponding solution of the primal problem. A application of the following result appears in section 6.

**COROLLARY 3.7** (Gauge primal-dual recovery). *Suppose that the primal  $(G_p)$  and dual  $(G_d)$  are strictly feasible. If  $y$  is optimal for  $(G_d)$ , then for any  $x \in \mathbb{R}^n$  the following conditions are equivalent:*

- (a)  $x$  is optimal for  $(G_p)$ ;
- (b)  $\langle x, A^T y \rangle = \kappa(x) \cdot \kappa^\circ(A^T y)$  and  $b - Ax \in \sigma \partial \rho^\circ(y)$ ;
- (c)  $A^T y \in \kappa^\circ(A^T y) \cdot \partial \kappa(x)$  and  $b - Ax \in \sigma \partial \rho^\circ(y)$ .

*Proof.* We use the optimality conditions given in Theorem 3.6. Note that by Lemma 2.1 we have equality (3.4b) in the dual constraint.

We first show that (b) implies (a). Suppose (b) holds. Then (3.4c) holds automatically. From the characterization (2.2) of the polar, we have

$$(3.8) \quad \sigma \rho^\circ(y) = \sigma \sup_{\rho(z) \leq 1} \langle y, z \rangle = \sup_{\rho(z) \leq \sigma} \langle y, z \rangle,$$

and thus  $\sigma \partial \rho^\circ(y) = \partial(\sigma \rho^\circ)(y)$  is the set of maximizing elements in this supremum. Because  $b - Ax \in \sigma \partial \rho^\circ(y)$ , it therefore holds that  $\rho(b - Ax) \leq \sigma$ . If we additionally use the polar-gauge inequality, we deduce that

$$\sigma \rho^\circ(y) = \langle y, b - Ax \rangle \leq \rho(b - Ax) \cdot \rho^\circ(y) \leq \sigma \rho^\circ(y),$$

and therefore the above inequalities are all tight. Thus conditions (3.4a) and (3.4d) hold, and by Theorem 3.6,  $(x, y)$  is a primal-dual optimal pair.

We next show that (a) implies (b). Suppose that  $x$  is optimal for  $(G_p)$ . Then the first condition of (b) holds by (3.4c), and (3.4a) and (3.4d) combine to give us

$$\sigma \rho^\circ(y) = \rho(b - Ax) \cdot \rho^\circ(y) = \langle b - Ax, y \rangle.$$

This implies that  $z := b - Ax$  is a maximizing element of the supremum in (3.8), and thus  $b - Ax \in \sigma \partial \rho^\circ(y)$ .

Finally, to show the equivalence of (b) and (c), note that by the polar-gauge inequality,  $\langle x, A^T y \rangle = \kappa(x) \cdot \kappa^\circ(A^T y)$  if and only if  $x$  minimizes the convex function  $\kappa^\circ(A^T y) \kappa(\cdot) - \langle \cdot, A^T y \rangle$ . This, in turn, is true if and only if  $0 \in \kappa^\circ(A^T y) \partial \kappa(x) - A^T y$ , or equivalently  $A^T y \in \kappa^\circ(A^T y) \cdot \partial \kappa(x)$ .  $\square$

**3.4. The relationship between Lagrange and gauge multipliers.** We now use the perturbation framework for duality to establish a relationship between gauge dual and Lagrange dual variables. We begin with an auxiliary result that characterizes the subdifferential of the perspective function.

LEMMA 3.8 (Subdifferential of perspective function). *Let  $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  be a closed proper convex function. Then for  $(x, \mu) \in \text{dom } g^\pi$ , equality holds:*

$$\partial g^\pi(x, \mu) = \begin{cases} \{ (z, -g^*(z)) \mid z \in \partial g(x/\mu) \} & \text{if } \mu > 0 \\ \{ (z, -\gamma) \mid (z, \gamma) \in \text{epi } g^*, z \in \partial g^\infty(x) \} & \text{if } \mu = 0. \end{cases}$$

*Proof.* Recall that for any closed proper convex function  $f$ , we have

$$(3.9) \quad \partial f(x) = \underset{z}{\text{argmax}} \{ \langle z, x \rangle - f^*(z) \},$$

and in particular if  $\mathcal{C}$  is a nonempty closed convex set, then  $\partial \delta_{\mathcal{C}}^*(x) = \text{argmax}_{z \in \mathcal{C}} \langle z, x \rangle$  [18, Theorem 23.5 and Corollary 23.5.3]. By [18, Corollary 13.5.1], we have  $g^\pi = \delta_{\mathcal{C}}^*$  where  $\mathcal{C} = \{ (z, \gamma) \mid g^*(z) \leq -\gamma \}$  is a closed convex set. If  $(x, \mu) \in \text{dom } g^\pi$ , then  $\mathcal{C}$  is nonempty and

$$(3.10) \quad \partial g^\pi(x, \mu) = \underset{(z, \gamma) \in \mathcal{C}}{\text{argmax}} \{ \langle (x, \mu), (z, \gamma) \rangle \} = \underset{(z, \gamma) \in \mathcal{C}}{\text{argmax}} \{ \langle z, x \rangle + \mu \gamma \}.$$

Suppose now that  $\mu > 0$ . Then

$$\sup_{(z, \gamma) \in \mathcal{C}} \{ \langle z, x \rangle + \mu \gamma \} = \sup_{z \in \text{dom } g^*} \{ \langle z, x \rangle - \mu g^*(z) \} = \mu \cdot \sup_{z \in \text{dom } g^*} \{ \langle z, x/\mu \rangle - g^*(z) \}.$$

Hence, again by (3.9),  $(z, \gamma) \in \text{argmax}_{(z, \gamma) \in \mathcal{C}} \{ \langle z, x \rangle + \mu \gamma \}$  if and only if  $z \in \partial g(x/\mu)$  and  $-\gamma = g^*(z)$ . On the other hand, if  $\mu = 0$  then

$$\sup_{(z, \gamma) \in \mathcal{C}} \langle z, x \rangle = \sup_{z \in \text{dom } g^*} \langle z, x \rangle = \delta_{\text{dom } g^*}^*(x) = g^\infty(x).$$

Hence, again by (3.9),  $(z, \gamma) \in \text{argmax}_{(z, \gamma) \in \mathcal{C}} \langle z, x \rangle$  if and only if  $z \in \partial g^\infty(x)$  and  $(z, -\gamma) \in \text{epi } g^*$ .  $\square$

We now state the main result relating the optimal solutions of  $(\mathbf{G}_p)$  to the optimal solutions of the Lagrange dual of  $(\mathbf{G}_d)$ .

THEOREM 3.9. *Suppose that the gauge dual  $(\mathbf{G}_d)$  is strictly feasible and the primal  $(\mathbf{G}_p)$  is feasible. Let  $(L)$  denote the Lagrange dual of  $(\mathbf{G}_d)$ , and let  $\nu_L$  denote its optimal value. Then*

$$z^* \text{ is optimal for } (L) \iff z^*/\nu_L \text{ is optimal for } (\mathbf{G}_p).$$

*Proof.* We first note that (L) can be derived via the framework of [Theorem 3.1](#) through the Lagrangian value function

$$h(w) = \inf_y \left\{ \kappa^\circ(A^T y + w) + \delta_{\langle b, \cdot \rangle - \sigma \rho^\circ(\cdot) \geq 1}(y) \right\}.$$

Here  $h$  plays the role of  $p$  in [Theorem 3.1](#); cf. [20, Example 11.41]. The feasibility of  $(\mathbf{G}_d)$  guarantees that  $h(0)$  is finite, and by [Lemma 3.4](#) we have

$$(\mathbf{G}_d) \text{ strictly feasible} \implies 0 \in \text{int}(\text{dom } v_d) \implies 0 \in \text{int}(\text{dom } h).$$

Thus by [Theorem 3.1](#) the optimal points  $z^*$  for (L) are characterized by  $z^* \in \partial h(0)$ . Note also  $h(0) = \nu_L$ .

On the other hand, by [Theorem 3.5\(b\)](#) the solutions to  $(\mathbf{G}_p)$  are precisely the points  $w^*/\lambda^*$  such that  $(w^*, \lambda^*) \in \partial v_d(0, 0)$ . Thus to relate the solution sets of (L) and  $(\mathbf{G}_p)$ , we must relate  $\partial h(0)$  and  $\partial v_d(0, 0)$ .

For  $\lambda$  in a neighborhood of zero, by positive homogeneity of  $\kappa^\circ$  and  $\rho^\circ$  we have

$$v_d(w, \lambda) = (1 + \lambda)h\left(\frac{w}{1 + \lambda}\right) = \inf_y \left\{ (1 + \lambda)\kappa^\circ\left(A^T y + \frac{w}{1 + \lambda}\right) + \delta_{\langle b, \cdot \rangle - \sigma \rho^\circ(\cdot) \geq 1}(y) \right\}.$$

Thus by [Lemma 3.8](#),  $\partial v_d(0, 0) = \{(z, -h^*(z)) \mid z \in \partial h(0)\}$ . However, for  $z \in \partial h(0)$  the Fenchel-Young equality gives us

$$0 = \langle 0, z \rangle = h^*(z) + h(0) = h^*(z) + \nu_L.$$

Thus we obtain the convenient description

$$\partial v_d(0, 0) = \partial h(0) \times \{h(0)\} = \partial h(0) \times \{\nu_L\}$$

and the set of optimal solutions for  $(\mathbf{G}_p)$  is precisely  $\frac{1}{\nu_L}\partial h(0)$ .  $\square$

**4. Perspective duality.** We now transition away from gauge functions with the aim of extending the gauge duality framework to include functions that are not necessarily sublinear. Thus, for the remainder of the paper, we consider functions  $f : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}_+$  and  $g : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}_+$ , which are closed, convex and nonnegative over their domains. The aim of this section is to derive and analyze the *perspective-dual* pair

$$(\mathbf{N}_p) \quad \underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad g(b - Ax) \leq \sigma,$$

$$(\mathbf{N}_d) \quad \underset{y, \alpha, \mu}{\text{minimize}} \quad f^\sharp(A^T y, \alpha) \quad \text{subject to} \quad \langle b, y \rangle - \sigma \cdot g^\sharp(y, \mu) \geq 1 - (\alpha + \mu).$$

The functions  $f^\sharp$  and  $g^\sharp$  are the polars of the perspective transforms of  $f$  and  $g$ . This perspective-polar transform is key to deriving this duality notion, which allows the application of the gauge-duality framework discussed in previous sections. In the next section we describe properties of that transformation and its application to the derivation of the perspective-dual pair.

**4.1. Perspective-polar transform.** Define the perspective-polar transform by  $f^\sharp := (f^\pi)^\circ$ , which is the polar gauge to  $f^\pi$ . Note that  $f^\sharp$  also coincides with  $(f^p)^\circ$ .

An explicit characterization of the perspective-polar transform is given by

$$(4.1) \quad f^\sharp(z, -\xi) = \inf \{ \mu > 0 \mid \langle z, x \rangle \leq \xi + \mu f(x), \forall x \}.$$

This representation can be obtained by applying the definition of the gauge polar (2.1) to the perspective transform as follows:

$$\begin{aligned} f^\sharp(x, -\xi) &= \inf \{ \mu > 0 \mid \langle z, x \rangle - \xi \lambda \leq \mu f^p(x, \lambda), \forall x, \forall \lambda \} \\ &= \inf \{ \mu > 0 \mid \langle z, x \rangle - \xi \lambda \leq \mu \lambda f(x/\lambda), \forall x, \forall \lambda > 0 \} \\ &= \inf \{ \mu > 0 \mid \langle z, \lambda x \rangle - \xi \lambda \leq \mu \lambda f(x), \forall x, \forall \lambda > 0 \}, \end{aligned}$$

which yields (4.1) after dividing through by  $\lambda$ . Rockafellar's extension [18, p.136] of the polar gauge transform to nonnegative convex functions that vanish at the origin coincides with  $f^\sharp(x, -1)$ .

The following theorem provides an alternative characterization of the perspective-polar transform in terms of the more familiar Fenchel conjugate  $f^*$ . It also provides an expression for the perspective-polar of  $f$  in terms of the Minkowski function generated by the epigraph of the conjugate of  $f$ , i.e.,

$$\gamma_{\text{epi } f^*}(x, \tau) := \inf \{ \lambda > 0 \mid (x, \tau) \in \lambda \text{epi } f^* \},$$

which is a gauge. Nonnegativity of  $f$  is not required, and so the assumption is dropped for this result.

**THEOREM 4.1.** *For any closed proper convex function  $f$  with  $0 \in \text{dom } f$ , we have  $f^{\pi^*}(z, -\xi) = \delta_{\text{epi } f^*}(z, \xi)$ . If in addition  $f$  is nonnegative,  $f^\sharp(z, -\xi) = \gamma_{\text{epi } f^*}(z, \xi)$ .*

*Proof.* Because of the assumptions on  $f$ , we have  $f^\pi(x, 0) = \lim_{\lambda \rightarrow 0^+} f^\pi(x, \lambda)$  for each  $x \in \mathbb{R}^n$  [18, Corollary 8.5.2]. Thus we have the following chain of equalities:

$$\begin{aligned} f^{\pi^*}(z, -\xi) &= \sup \{ \langle z, x \rangle - \lambda \xi - f^\pi(x, \lambda) \mid x \in \mathbb{R}^n, \lambda \in \mathbb{R} \} \\ &= \sup \{ \langle z, x \rangle - \lambda \xi - \lambda f(\lambda^{-1}x) \mid x \in \mathbb{R}^n, \lambda > 0 \} \\ &= \sup \{ \langle z, \lambda y \rangle - \lambda \xi - \lambda f(y) \mid y \in \mathbb{R}^n, \lambda > 0 \} \\ &= \sup \{ \lambda \cdot \sup_y \{ \langle z, y \rangle - \xi - f(y) \} \mid \lambda > 0 \} \\ &= \sup \{ \lambda \cdot (f^*(z) - \xi) \mid \lambda > 0 \} \\ &= \delta_{\text{epi } f^*}(z, \xi). \end{aligned}$$

This proves the first statement. Now additionally suppose that  $f$  is nonnegative. Because  $f^\pi$  is closed, it is identical to its biconjugate, and so  $f^\pi(x, \lambda) = \delta_{\text{epi } f^*}^*(x, -\lambda)$ . Also,  $\text{epi } f^*$  is closed and convex, and contains the origin because  $f$  is nonnegative. Therefore, it follows from Rockafellar [18, Corollary 15.1.2] that

$$f^\sharp(z, -\xi) \equiv f^{\pi^*}(z, -\xi) = \delta_{\text{epi } f^*}^{\star\circ}(z, \xi) = \gamma_{\text{epi } f^*}(z, \xi),$$

as claimed. □

The following result relates the level sets of the perspective-polar transform to the level sets of the conjugate perspective. This result is useful in deriving the constraint sets for certain perspective-dual problems for which there is no closed form for the perspective polar; cf. [Example 5.4](#).

**THEOREM 4.2 (Level-set equivalence).** *For any nonnegative, closed proper convex function  $f$  with  $0 \in \text{dom } f$ , and for any scalar  $\mu$ ,*

$$f^\sharp(z, \xi) \leq \mu \iff [0 \leq \mu \text{ and } f^{\star\pi}(z, \mu) \leq -\xi].$$

*Proof.* The following chain of equivalences follows from Theorem 4.1:

$$\begin{aligned}
(4.2) \quad f^\sharp(z, \xi) \leq \mu &\iff \gamma_{\text{epi } f^*}(z, -\xi) \leq \mu \\
&\iff \inf \{ \lambda > 0 \mid (z, -\xi) \in \lambda \text{epi } f^* \} \leq \mu \\
&\iff \inf \{ \lambda > 0 \mid f^*(z/\lambda) \leq -\xi/\lambda \} \leq \mu \\
&\iff \inf \{ \lambda > 0 \mid f^{*\pi}(z, \lambda) \leq -\xi \} \leq \mu.
\end{aligned}$$

Define  $\alpha = \inf \{ \lambda > 0 \mid f^{*\pi}(z, \lambda) \leq -\xi \}$ .

First suppose that  $\alpha \leq \mu$ . Then clearly  $0 \leq \mu$ . If  $\alpha < \mu$ , there exists  $\lambda$  with  $0 < \lambda < \mu$  such that  $f^{*\pi}(z, \lambda) \leq -\xi$ . Since  $f$  is nonnegative, we then have  $\lambda f \leq \mu f$ , and thus  $(\lambda f)^* \geq (\mu f)^*$ . In particular,

$$f^{*\pi}(z, \mu) = (\mu f)^*(z) \leq (\lambda f)^*(z) = f^{*\pi}(z, \lambda) \leq -\xi.$$

On the other hand, if  $\alpha = \mu$ , there exists a sequence  $\lambda_k \rightarrow \mu^+$  such that  $f^{*\pi}(z, \lambda_k) \leq -\xi$  for each  $n$ . Now by the lower semi-continuity of  $f^{*\pi}$ , we obtain

$$f^{*\pi}(z, \mu) \leq \lim_{k \rightarrow \infty} f^{*\pi}(z, \lambda_k) \leq -\xi.$$

This shows the forward implication of the theorem.

To show the reverse implication, suppose  $0 \leq \mu$  and  $f^{*\pi}(z, \mu) \leq -\xi$ . If  $0 < \mu$ , then by (4.2) we immediately have  $f^\sharp(z, \xi) \leq \mu$ . Suppose then  $\mu = 0$ . We want to show  $f^\sharp(z, \xi) \leq 0$ . However, by (4.1) this is equivalent to the following:

$$(4.3) \quad \inf \{ \lambda > 0 \mid \langle z, w \rangle + \xi \leq \lambda f(w), \forall w \in \mathbb{R}^n \} = 0.$$

If  $z = 0$ , this simply says that  $\xi \leq 0$ . However, if  $z \neq 0$ , then  $\xi \leq 0$  is implied by our assumption because

$$0 = f^{*\pi}(0, 0) = f^{*\pi}(z, \mu) \leq -\xi.$$

On the other hand, if  $z \neq 0$  then

$$(z, 0, -\xi) \in \text{epi } f^{*\pi} = \text{cl}(\text{epi } f^{*p}).$$

Thus there exists a sequence  $(z_k, \mu_k, r_k)$  with  $\lim_{k \rightarrow \infty} (z_k, \mu_k, r_k) = (z, 0, -\xi)$  and  $f^{*p}(z_k, \mu_k) \leq r_k$  for all  $n$ . Since  $z \neq 0$ , we can assume without loss of generality that  $z_k \neq 0$  and  $\mu_k > 0$  for all  $k$ . Then for each  $k$ , we have  $\mu_k f^*(z_k/\mu_k) \leq r_k$ , which is equivalent to the following:

$$\begin{aligned}
\mu_k f^*(z_k/\mu_k) \leq r_k &\iff \sup_w \{ \langle w, z_k \rangle - \mu_k f(w) \} \leq r_k \\
&\iff \langle w, z_k \rangle - \mu_k f(w) \leq r_k, \forall w \in \mathbb{R}^n \\
&\iff \langle w, z_k \rangle - r_k \leq \mu_k f(w), \forall w \in \mathbb{R}^n \\
&\iff \mu_k \geq \inf \{ \lambda > 0 \mid \langle w, z_k \rangle - r_k \leq \lambda f(w), \forall w \in \mathbb{R}^n \} \geq 0.
\end{aligned}$$

Take limits as  $k \rightarrow \infty$  to deduce (4.3), as required.  $\square$

**4.1.1. Examples.** We now apply the perspective-polar transform to two important special cases.

*Example 4.3* (Gauge functions). Suppose that  $f$  is a closed proper gauge. Then

$$(4.4) \quad f^\sharp(z, \xi) = f^\circ(z) + \delta_{\mathbb{R}_-}(\xi).$$

Use expression (4.1) for this derivation. When  $\xi > 0$ , take  $x = 0$  in the infimum in (4.1) to deduce that  $f^\sharp(z, \xi) = +\infty$ . On the other hand, when  $\xi \leq 0$ , the positive homogeneity of  $f$  implies that  $f^\sharp(z, \xi) = f^\circ(z)$ ; we leave the details to the reader. Note that if  $f$  vanishes at the origin, then  $f^\sharp(z, \xi) = +\infty$  for all  $\xi > 0$ .

*Example 4.4* (Separable sums). The previous example considers univariate penalties. The formulas, however, easily extend to separable sums of such univariate penalties. Consider the sum

$$f(x) := \sum_{i=1}^n f_i(x_i),$$

where each convex function  $f_i : \mathbb{R}^{n_i} \rightarrow \overline{\mathbb{R}}_+$  is nonnegative. Then a straightforward computation shows that  $f^p(x, \lambda) = \sum_{i=1}^n f_i^p(x_i, \lambda)$ . Furthermore, taking into account [13, Proposition 2.4], which expresses the polar of a separable sum of gauges, we deduce

$$f^\sharp(z, \xi) = \max_{i=1, \dots, n} f_i^\sharp(z_i, \xi).$$

**4.2. Derivation of the perspective dual via a lift.** We next derive the relationship between the primal and dual problems  $(\mathbf{N}_p)$  and  $(\mathbf{N}_d)$  by lifting  $(\mathbf{N}_p)$  to an equivalent gauge optimization problem, and then recognizing  $(\mathbf{N}_d)$  as its gauge dual.

**THEOREM 4.5** (Gauge lifting of the primal). *A point  $x^*$  is optimal for  $(\mathbf{N}_p)$  if and only if  $(x^*, 1)$  is optimal for the gauge problem*

$$(4.5) \quad \underset{x, \lambda}{\text{minimize}} \quad f^\pi(x, \lambda) \quad \text{subject to} \quad \rho \left( \begin{bmatrix} b \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} A & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} \right) \leq \sigma,$$

where  $\rho$  is the gauge  $\rho(z, \mu, \tau) := g^\pi(z, \tau) + \delta_{\{0\}}(\mu)$ .

*Proof.* By definition of  $f^\pi$ ,  $x^*$  is optimal for  $(\mathbf{N}_p)$  if and only if the pair  $(x^*, 1)$  is optimal for

$$\underset{x, \lambda}{\text{minimize}} \quad f^\pi(x, \lambda) \quad \text{subject to} \quad \lambda = 1, \quad g^\pi(b - Ax, \lambda) \leq \sigma.$$

The following equivalence follows from the definition of  $\rho$ :

$$[\lambda = 1 \quad \text{and} \quad g^\pi(b - Ax, \lambda) \leq \sigma] \quad \iff \quad \rho(b - Ax, 1 - \lambda, \lambda) \leq \sigma.$$

Thus we arrive at the constraint region in (4.5).  $\square$

**COROLLARY 4.6** (Gauge dual of the lift). *Problem  $(\mathbf{N}_d)$  is the gauge dual of (4.5).*

*Proof.* It follows from the form of the gauge pair  $(\mathbf{G}_p)$  and  $(\mathbf{G}_d)$  that the gauge dual of (4.5) is

$$(4.6) \quad \underset{y, \beta, \alpha}{\text{minimize}} \quad f^{\pi^\circ} \left( \begin{bmatrix} A^T & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} y \\ \alpha \\ \mu \end{bmatrix} \right) \\ \text{subject to} \quad \langle (y, \alpha, \mu), (b, 1, 1) \rangle - \sigma^\circ(y, \alpha, \mu) \geq 1.$$

Note that  $\rho$  is separable in  $(z, \mu)$  and  $\beta$ , so we can appeal to [13, Proposition 2.4] to obtain the expression

$$\rho^\circ(y, \alpha, \mu) = \max \{ g^{\pi^\circ}(y, \mu), \delta_{\{0\}}^\circ(\alpha) \}.$$

Since  $\delta_{\{0\}}^\circ(\alpha)$  is identically zero, the result follows.  $\square$

The next result generalizes the gauge duality result of [Theorem 2.2](#) to the case where  $f$  and  $g$  are nonnegative but not necessarily gauges. We parallel the construction in (2.4), and for this section only redefine the following feasible sets:

$$\begin{aligned} \mathcal{F}_p &:= \{ x \mid g(b - Ax) \leq \sigma \} \\ \mathcal{F}_d &:= \{ (y, \alpha, \mu) \mid \langle b, y \rangle - \sigma g^\sharp(y, \mu) \geq 1 - (\alpha + \mu) \}. \end{aligned}$$

Thus,  $(\mathbf{N}_p)$  is *relatively strictly feasible* if

$$\text{ri } \mathcal{F}_p \cap \text{ri dom } f \neq \emptyset.$$

Similarly,  $(\mathbf{N}_d)$  is *relatively strictly feasible* if there exists a triple  $(y, \alpha, \mu)$  such that

$$(A^T y, \alpha) \in \text{ri dom } f^\sharp \quad \text{and} \quad \langle b, y \rangle - \sigma g^\sharp(y, \mu) > 1 - (\alpha + \mu).$$

*Strict feasibility* follows the same definitions, except that the relative interior operation replaced by the interior operation. Finally, let  $\nu_p$  and  $\nu_d$  denote the optimal values of the primal-dual pair  $(\mathbf{N}_p)$  and  $(\mathbf{N}_d)$ .

**THEOREM 4.7** (Perspective duality). *The following relationships hold for the perspective dual pair  $(\mathbf{N}_p)$  and  $(\mathbf{N}_d)$ .*

(a) (*Weak duality*) *If  $x$  and  $(y, \alpha, \mu)$  are primal and dual feasible, then*

$$(4.7) \quad 1 \leq \nu_p \nu_d \leq f(x) \cdot f^\sharp(A^T y, \alpha).$$

(b) (*Strong duality*) *If the dual (resp. primal) is feasible and the primal (resp. dual) is relatively strictly feasible, then  $\nu_p \nu_d = 1$  and the perspective dual (resp. primal) attains its optimal value.*

*Proof.* The weak duality inequality (4.7) follows immediately from the analogous result in [Theorem 2.2](#), together with [Theorem 4.5](#) and [Corollary 4.6](#).

Next we demonstrate that  $(\mathbf{N}_p)$  is relatively strictly feasible if and only if (4.5) is relatively strictly feasible. By definition, (4.5) is relatively strictly feasible if and only if there exists a point  $(x, 1) \in \text{ri dom } f^\pi$  such that

$$(b - Ax, 0, 1) \in \text{ri dom } \rho \quad \text{and} \quad \rho(b - Ax, 0, 1) = g(b - Ax) < \sigma.$$

We now seek a description of  $\text{ri dom } f^\pi$ . We have

$$\begin{aligned} \text{dom } f^\pi &= \{ (x, \mu) \mid f^\pi(x, \mu) < \infty \} \\ &= \text{cl}(\{0\} \cup \{ (x, \mu) \mid \mu > 0, f(x/\mu) < \infty \}) = \text{cl cone}((\text{dom } f) \times \{1\}). \end{aligned}$$

By [18, Corollary 6.8.1], the above description yields

$$\text{ri dom } f^\pi = \{ (x, \mu) \mid \mu > 0, x \in \mu \text{ ri dom } f \}.$$

Thus  $(x, 1) \in \text{ri dom } f^\pi$  if and only if  $x \in \text{ri dom } f$ . Thus,

$$\text{dom } \rho = \{ (y, 0, \mu) \mid (y, \mu) \in \text{dom } g^\pi \},$$

and so

$$\text{ri dom } \rho = \{ (y, 0, \mu) \mid (y, \mu) \in \text{ri dom } g^\pi \} = \{ (y, \mu, 0) \mid \mu > 0, y \in \mu \text{ ri dom } g \}.$$

In particular, the condition  $(b - Ax, 0, 1) \in \text{ri dom } \rho$  is equivalent to  $b - Ax \in \text{ri dom } g$ . Thus the conditions for relative strict feasibility of (4.5) and  $(\mathbf{N}_p)$  are the same.

It is also straightforward to verify that  $(\mathbf{N}_d)$  is relatively strictly feasible if and only if  $(\mathbf{N}_p)$  is relatively strictly feasible. Strong duality then follows from relative interiority, Corollary 4.6, Theorem 4.5, and the analogous strong-duality result for gauges in Theorem 2.2.  $\square$

**4.3. Optimality conditions.** The following result generalizes Theorem 3.5 to include the perspective-dual pair.

**THEOREM 4.8** (Perspective optimality). *Suppose  $(\mathbf{N}_p)$  is strictly feasible. Then the tuple  $(x, y, \alpha, \mu)$  is perspective primal-dual optimal if and only if*

$$\begin{aligned} (4.8a) \quad & g(b - Ax) = \sigma && \text{(primal activity)} \\ (4.8b) \quad & \langle b, y \rangle - \sigma g^\sharp(y, \mu) = 1 - (\alpha + \mu) && \text{(dual activity)} \\ (4.8c) \quad & \langle x, A^T y \rangle + \alpha = f(x) \cdot f^\sharp(A^T y, \alpha) && \text{(objective alignment)} \\ (4.8d) \quad & \langle b - Ax, y \rangle + \mu = g(b - Ax) \cdot g^\sharp(y, \mu). && \text{(constraint alignment)} \end{aligned}$$

*Proof.* By construction,  $x$  is optimal for  $(\mathbf{N}_p)$  if and only if  $(x, 1)$  is optimal for its gauge reformulation (4.5). Apply Theorem 3.5 to (4.5) and the corresponding gauge dual  $(\mathbf{N}_d)$  to obtain the required conditions.  $\square$

The following result mirrors Corollary 3.7 for the perspective-duality case.

**COROLLARY 4.9** (Perspective primal-dual recovery). *Suppose that the primal  $(\mathbf{N}_p)$  is strictly feasible. If  $(y, \alpha, \mu)$  is optimal for  $(\mathbf{N}_d)$ , then for any  $x \in \mathbb{R}^n$ , the following conditions are equivalent:*

- (a)  $x$  is optimal for  $(\mathbf{N}_p)$ ;
- (b)  $\langle x, A^T y \rangle + \alpha = f(x) \cdot f^\sharp(A^T y, \alpha)$  and  $(b - Ax, 1) \in \sigma \partial g^\sharp(y, \mu)$ ;
- (c)  $A^T y \in f^\sharp(A^T y, \alpha) \cdot \partial f(x)$  and  $(b - Ax, 1) \in \sigma \partial g^\sharp(y, \mu)$ .

*Proof.* By construction,  $x$  is optimal for  $(\mathbf{N}_p)$  if and only if  $(x, 1)$  is optimal for its gauge reformulation (4.5). Apply Corollary 3.7 to (4.5) and its gauge dual  $(\mathbf{N}_d)$  to obtain the equivalence of (a) and (b). To show the equivalence of (b) and (c), note that by the polar-gauge inequality,  $\langle (x, 1), (A^T y, \alpha) \rangle \leq f^\pi(x, 1) \cdot f^\sharp(A^T y, \alpha)$  for all  $x$ , or equivalently,

$$(4.9) \quad \langle x, A^T y \rangle + \alpha \leq f(x) \cdot f^\sharp(A^T y, \alpha), \quad \forall x.$$

The inequality is tight for a fixed  $x$  if and only if  $x$  minimizes the function  $h = f^\sharp(A^T y, \alpha) f(\cdot) - \langle \cdot, A^T y \rangle - \alpha$ . This in turn is equivalent to  $0 \in \partial h(x)$ , or

$$A^T y \in f^\sharp(A^T y, \alpha) \cdot \partial f(x).$$

This shows the equivalence of (b) and (c) and completes the proof.  $\square$

Section 6 gives applications of Corollary 4.9 for recovering primal optimal solutions from perspective-dual optimal solutions.

**4.4. Reformulations of the perspective dual.** An important simplification occurs when the functions involved in  $(\mathbf{N}_p)$  are gauges.

**COROLLARY 4.10** (Simplification for gauges). *If  $f$  is a gauge, then a triple  $(y, \alpha, \mu)$  is optimal for  $(\mathbf{N}_d)$  if and only if  $(y, \mu)$  is optimal for*

$$\underset{y, \alpha}{\text{minimize}} \quad f^\circ(A^T y) \quad \text{subject to} \quad \langle b, y \rangle - \sigma g^\sharp(y, \mu) \geq 1 - \mu$$

*If in addition  $g$  is a gauge, then a triple  $(y, \alpha, \mu)$  is optimal for  $(\mathbf{N}_d)$  if and only if  $y$  solves  $(\mathbf{G}_d)$ .*

*Proof.* These follow from the formulas for  $f^\sharp$  and  $g^\sharp$  established in [section 4.1.1](#).  $\square$

[Theorem 4.2](#) gives an expression for the level sets of  $g$  in terms of its conjugate polar.

**COROLLARY 4.11.** *The point  $(y, \alpha, \mu)$  is optimal for  $(\mathbf{N}_d)$  if and only if there exists a scalar  $\xi$  such that  $(y, \alpha, \mu, \xi)$  is optimal for the problem*

$$\begin{aligned} \underset{y, \alpha, \mu, \xi}{\text{minimize}} \quad & f^\sharp(A^T y, \alpha) \\ \text{subject to} \quad & \langle b, y \rangle - \sigma \xi = 1 - (\alpha + \mu) \\ & g^{*\pi}(y, \xi) \leq -\mu, \quad \xi \geq 0. \end{aligned}$$

*Proof.* Immediate from [Theorem 4.2](#) after introducing a new variable  $\xi := (\langle b, y \rangle + \alpha + \mu - 1)/\sigma$  in  $(\mathbf{N}_d)$ .  $\square$

**5. Examples: piecewise linear-quadratic and GLM constraints.** From a computational standpoint, the perspective-dual formulation may be an attractive alternative to the original problem. The efficiency of this approach requires that the dual constraints are in some sense more tractable than those of the primal. For example, we may consider the dual feasible set “easy” if it admits an efficient procedure for projecting onto that set. We examine two special cases that admit tractable dual problems in this sense. The first case is the family of piecewise linear quadratic (PLQ) functions, introduced by Rockafellar [19] and subsequently examined by Rockafellar and Wets [20, p.440], and Aravkin, Burke, and Pillonetto [3]. The second case is the logistic loss. For this section only, we assume for the sake of simplicity that the objective  $f$  is a gauge, so that the perspective dual in each of these cases simplifies as in [Corollary 4.10](#). The more general case still applies.

**5.1. PLQ constraints.** The family of PLQ functions is a large class of convex functions that includes such commonly used penalties as the Huber function, the Vapnik  $\epsilon$ -loss, and the hinge loss. The last two are used in support-vector regression and classification [3]. PLQ functions take the form

$$(5.1) \quad g(y) = \sup_{u \in U} \left\{ \langle u, By + b \rangle - \frac{1}{2} \|Lu\|_2^2 \right\}, \quad U := \{u \in \mathbb{R}^l \mid Wu \leq w\},$$

where  $g$  is defined by linear operators  $L : \mathbb{R}^l \rightarrow \mathbb{R}^l$  and  $W \in \mathbb{R}^{k \times l}$ , a vector  $w \in \mathbb{R}^k$ , and an injective affine transformation  $B(\cdot) + b$  from  $\mathbb{R}^k$  to  $\mathbb{R}^l$ . We may assume without loss of generality that  $B(\cdot) + b$  is the identity transformation, since the primal problem  $(\mathbf{N}_p)$  already allows for composition of the constraint function  $g$  with an affine transformation. We also assume that  $U$  contains the origin, which implies that  $g$  is nonnegative and thus can be interpreted as a penalty function. Aravkin, Burke, and Pillonetto [3] describe a range of PLQ functions that often appear in applications.

The conjugate representation of  $g$ , given by

$$(5.2) \quad g^*(y) = \delta_U(y) + \frac{1}{2}\|Ly\|^2,$$

is useful for deriving its polar perspective  $g^\sharp$ . In the following discussion, it is convenient to interpret the quadratic function  $-(1/2\mu)\|Ly\|^2$  as a closed convex function of  $\mu \in \mathbb{R}_-$ , and thus when  $\mu = 0$ , we make the definition  $-(1/2\mu)\|Ly\|^2 = \delta_{\{0\}}(y)$ .

**THEOREM 5.1.** *If  $g$  is a PLQ function, then*

$$\begin{aligned} g^\sharp(y, \mu) &= \delta_{\mathbb{R}_-}(\mu) + \max \left\{ \gamma_U(y), -(1/2\mu)\|Ly\|^2 \right\} \\ &= \delta_{\mathbb{R}_-}(\mu) + \max \left\{ -(1/2\mu)\|Ly\|^2, \max_{i=1, \dots, k} \{W_i^T y / w_i\} \right\}, \end{aligned}$$

where  $W_1^T, \dots, W_k^T$  are the rows of  $W$  that define  $U$  in (5.1).

*Proof.* First observe that when  $g$  is PLQ,  $\text{epi } g^* = \{(y, \tau) \mid y \in U, \frac{1}{2}\|Ly\|^2 \leq \tau\}$ . Apply [Theorem 4.1](#) and simplify to obtain the chain of equalities

$$\begin{aligned} g^\sharp(y, \mu) &= \gamma_{\text{epi } g^*}(y, -\mu) = \inf \{ \lambda > 0 \mid (y, -\mu) \in \lambda \text{epi } g^* \} \\ &= \inf \{ \lambda > 0 \mid y/\lambda \in U, \frac{1}{2\lambda^2}\|Ly\|^2 \leq -\mu/\lambda \} \\ &= \delta_{\mathbb{R}_-}(\mu) + \max \left\{ \gamma_U(y), -\frac{1}{2\mu}\|Ly\|^2 \right\}. \end{aligned}$$

Because  $U$  is polyhedral, we can make the explicit description

$$\begin{aligned} \gamma_U(y) &= \inf \{ \lambda > 0 \mid y \in \lambda U \} \\ &= \inf \{ \lambda > 0 \mid W(y/\lambda) \leq w \} = \max \left\{ 0, \max_{i=1, \dots, k} \{W_i^T y / w_i\} \right\}. \end{aligned}$$

This follows from considering cases on the signs of the  $W_i^T y$ , and noting that  $w \geq 0$  because  $U$  contains the origin.  $\square$

The next example illustrates how [Theorem 5.1](#) can be applied to compute the perspective-polar transform of the Huber function.

*Example 5.2* (Huber function). The Huber function

$$h_\eta(x) := \begin{cases} \frac{1}{2\eta}x^2 & \text{if } |x| \leq \eta \\ |x| - \frac{\eta}{2} & \text{if } |x| > \eta, \end{cases}$$

can be considered as a smooth approximation to the absolute value function, where the positive parameter  $\eta$  controls the degree of smoothness. This function can equivalently be stated in conjugate form

$$h_\eta(x) = \sup_{u \in [-\eta, \eta]} \{ ux - (\eta/2)u^2 \} = \sup_u \{ ux - [\delta_{[-\eta, \eta]}(u) + (\eta/2)u^2] \},$$

which reveals  $h_\eta^*(y) = \delta_{[-\eta, \eta]}(y) + (\eta/2)y^2$ . We then apply [Theorem 4.1](#) to obtain

$$\begin{aligned} h_\eta^\sharp(z, \xi) &= \gamma_{h_\eta^*}(z, -\xi) \\ &= \inf \{ \lambda > 0 \mid (z, -\xi) \in \lambda \text{epi } h_\eta^* \} \\ &= \inf \{ \lambda > 0 \mid |z|/\lambda \leq \eta, (\eta/2\lambda)z^2 \leq -\xi \} \\ &= \delta_{\mathbb{R}_-}(\xi) + \max \left\{ |z|/\eta, -(\eta/2\xi)z^2 \right\}. \end{aligned}$$

Note that this can easily be extended beyond the univariate case to a separable sum by applying Example 4.4.

We can now write down an explicit formulation of the perspective dual problem  $(\mathbf{N}_d)$  when the primal problem  $(\mathbf{N}_p)$  has a PLQ-constrained feasible region (i.e.,  $g$  is PLQ) and a gauge objective (i.e.,  $f$  is a closed gauge). The constraint set of  $(\mathbf{N}_d)$  simplifies significantly so that, for example, a first-order projection method might be applied to solve the problem. Apply Theorem 5.1 and introduce a scalar variable  $\xi$  to rephrase the dual problem  $(\mathbf{N}_d)$  as

$$(5.3) \quad \begin{aligned} & \underset{y, \mu, \xi}{\text{minimize}} && f^\circ(A^T y) \\ & \text{subject to} && \langle b, y \rangle + \mu - \sigma \xi = 1 \\ & && W y \leq \xi w, \quad -\frac{1}{2\mu} \|Ly\|^2 \leq \xi \\ & && \mu \leq 0, \quad \xi \geq 0. \end{aligned}$$

We can further simplify the constraint set using the fact that

$$(5.4) \quad [ \|Ly\|^2 \leq -2\mu\xi \text{ and } \mu \leq 0, \xi \geq 0 ] \iff \left\| \begin{bmatrix} 2Ly \\ \xi + 2\mu \end{bmatrix} \right\|_2 \leq \xi - 2\mu,$$

Thus, projecting a point  $\bar{y}$  onto the feasible set of (5.3) can be done by solving a second-order cone (SOCP) program. For many important cases, the operator  $L$  is typically extremely sparse. For example, when  $g$  is a sum of separable Huber functions,  $L = \sqrt{\eta}I$ . Hence in many practical cases, this projection problem could be solved efficiently using SOCP solvers that take advantage of sparsity, e.g., Gurobi [14].

**5.2. Generalized linear models and the Bregman divergence.** The family of exponentially-distributed random variables has density functions of the form

$$(5.5) \quad p(y | \theta) = \exp[\langle \theta, y \rangle - \phi^*(\theta) - p_0(y)],$$

where  $\theta$  is an  $n$ -vector of parameters, and the conjugate of  $\phi$  is the cumulant generating function of the distribution; the function  $p_0$  serves to normalize the distribution. The maximum likelihood estimate (MLE) can be obtained as the minimizer of the negative log-likelihood function

$$(5.6) \quad h(\theta) := -\log p(y | \theta) = \phi^*(\theta) - \langle \theta, y \rangle + p_0(y).$$

In applications that impose an *a priori* distribution on the parameters, the goal is to find an approximation to the MLE estimate that penalizes a regularization function  $f$  (a surrogate for the prior). In other words, we look for parameters  $\theta$  that satisfy

$$(5.7) \quad h(\theta) - \inf h \leq \sigma \quad \text{and} \quad f(\theta) \text{ is small}$$

for some small positive value  $\sigma$ . The typical approach for computing such  $\theta$  in practice is to solve the regularized optimization problem

$$\underset{\theta}{\text{minimize}} \quad \lambda h(\theta) + f(\theta),$$

where  $\lambda$  is calibrated to obtain the required approximation of the MLE estimate.

Distribution	$\phi(\theta)$	$\phi^*(y)$	$d_{\phi^*}(\theta; \nabla\phi(y))$
Gaussian ( $\sigma^2 = 1$ )	$\frac{1}{2}\theta^2$	$\frac{1}{2}y^2$	$\frac{1}{2}(y - \theta)^2$
Poisson	$\theta \log \theta - \theta$	$e^y$	$e^\theta - y\theta + \phi(y)$
Bernoulli	$\theta \log \theta + (1 - \theta) \log(1 - \theta)$	$\log(1 + e^y)$	$\log(1 + e^\theta) - y\theta + \phi(y)$

Table 5.1: The seed functions associated with common exponential distributions, and the corresponding Bregman divergence constraint shown in (5.9) [9]. Domains of functions  $\phi$  and  $\phi^*$  are not given. In higher dimensions,  $\phi$ ,  $\phi^*$ , and  $d_{\phi^*}$  will be separable sums of the given one-dimensional terms, assuming independent data.

The calibration procedure for  $\lambda$  can be avoided entirely by instead imposing the conditions (5.7) directly as part of the formulation, and solving the constrained problem

$$(5.8) \quad \underset{\theta}{\text{minimize}} \quad f(\theta) \quad \text{subject to} \quad h(\theta) - \inf h \leq \sigma.$$

We now show how to eliminate  $(\inf h)$  from the formulation through the *Bregman divergence* function

$$d_\phi(v; w) := \phi(v) - \phi(w) - \langle \nabla\phi(w), v - w \rangle.$$

We assume that  $\phi$  is a closed convex function of the Legendre type: a function for which  $C := \text{int}(\text{dom } \phi)$  is not empty,  $\phi$  is strictly convex on  $C$ , and  $\phi$  is essentially smooth [18, p.258]. The key implication of this assumption is that the gradient map is a one-to-one correspondence between  $C$  and  $C^* := \text{int}(\text{dom } \phi^*)$ , and  $\nabla\phi^* = (\nabla\phi)^{-1}$ . This relationship allows us to assert that

$$\theta = \nabla\phi(y) \quad \text{if and only if} \quad \theta = \text{argmin } h(\theta).$$

Substitute this expression for  $\theta$  into the constraint of (5.8), and again use the inverse relationship between the gradients of  $\phi$  and its conjugate, to determine that

$$(5.9) \quad h(\theta) - \inf h \leq \sigma \quad \text{if and only if} \quad d_{\phi^*}(\theta; \nabla\phi(y)) \leq \sigma.$$

Furthermore, the Bregman divergence exhibits two properties that make it relevant within the context of the formulation  $(N_p)$  considered in this paper: it is everywhere positive except when its arguments coincide (and in that case its value is zero), and it is strictly convex in its first argument. Banerjee et al. [4] detail the close connection between the family of exponential distributions and the Bregman divergence defined by the cumulant generating function. Table 5.1 gives several examples of functions  $\phi$  that give rise to common divergences; more examples are given by Banerjee et al. [4, Table 1] and Dhillon and Tropp [9, Tables 2.1–2.2].

Suppose that we are given a data set  $\{(a_i, b_i)\}_{i=1}^m \subset \mathbb{R}^n \times \mathbb{R}$ , where each vector  $a_i$  describes features associated with observations  $b_i$ . Assume that the vector  $b$  of observations is distributed according to an exponential density given by (5.5), where the parameters  $\theta := \theta(a_1, \dots, a_n)$  may depend on the feature vectors. In particular, this occurs when the observations  $b_i$  are independent and distributed as

$$\tilde{p}(y_i | \theta_i) = \exp[\theta_i y_i - \tilde{\phi}^*(\theta_i) - \tilde{p}_0(y_i)],$$

in which case  $\phi^*(\theta) := \sum_{i=1}^m \tilde{\phi}^*(\theta_i)$  and  $p_0(y) := \sum_{i=1}^m \tilde{p}_0(y_i)$  are separable sums. We assume a linear dependence between the parameters and feature vectors, and thus set  $\theta = Ax$ , where the matrix  $A$  has rows  $a_i^T$ . We may now obtain a regularized MLE estimate defined by (5.8) by solving

$$\underset{x}{\text{minimize}} \quad f(x) \quad \text{subject to} \quad d_{\phi^*}(Ax; \nabla\phi(b)) \leq \sigma.$$

We use [Corollary 4.11](#) to derive the perspective dual, which requires the computation of the conjugate of  $g(z) := d_{\phi^*}(z; \nabla\phi(b))$ :

$$\begin{aligned} g^*(y) &= \sup_z \{ \langle z, y \rangle - d_{\phi^*}(z; \nabla\phi(b)) \} \\ &= \sup_z \{ \langle z, y \rangle - \phi^*(z) + \phi^*(\nabla\phi(b)) + \langle b, z - \nabla\phi(b) \rangle \} \\ &= \phi^*(\nabla\phi(b)) - \langle b, \nabla\phi(b) \rangle + \phi(y + b), \end{aligned}$$

where we simplify the expression using the inverse relationship between the gradients of  $\phi$  and its conjugate. Assume for simplicity that  $f$  is a gauge, which is typical when it serves as a regularization function. In that case, the perspective dual reduces to

$$(5.10) \quad \begin{aligned} &\underset{y, \mu, \xi}{\text{minimize}} \quad f^\circ(A^T y) \\ &\text{subject to} \quad \phi^\pi(y + \xi b, \xi) \leq \xi[\langle b, \nabla\phi(b) \rangle - \phi^*(\nabla\phi(b)) - \sigma] - 1, \quad \xi \geq 0; \end{aligned}$$

cf. [Corollaries 4.10](#) and [4.11](#).

*Example 5.3* (Gaussian distribution). As a first example, consider the case where the  $b_i$  are distributed as independent Gaussian variables with unit variance. In this case,  $\phi := \frac{1}{2} \|\cdot\|^2$  and the above constraints specialize to

$$\frac{1}{2\xi} \|y\|^2 + \langle b, y \rangle \leq -(1 + \sigma\xi), \quad \xi \geq 0.$$

This is an example of a PLQ constraint, which falls into the category of problems described in [section 5.1](#).

*Example 5.4* (Poisson distribution). Consider the case where the observations  $b_i$  are independent Poisson observations. Straightforward calculations using the functions in [Table 5.1](#) show that the perspective dual constraints for the Poisson case reduce to

$$\sum_{i=1}^m z_i \log(z_i/\xi) \leq \beta\xi + \sum_{i=1}^m z_i - (1 + \sigma\xi), \quad z = y + \xi b, \quad \xi \geq 0,$$

where  $\beta = \sum_{i=1}^m (b_i + b_i \log b_i)$  is a constant. By introducing new variables, this can be further simplified to require only affine constraints and  $m$  relative-entropy constraints. To solve projection subproblems onto a constraint set of this form, we note that

$$F(x, y, r) = 400(-\log(x/y) - \log(\log(x/y) - r/y) - 4 \log(y))$$

is a self-concordant barrier for the set

$$\{ (x, y, r) \mid y > 0, y \log(y/x) \leq r \},$$

which is the epigraph of the relative entropy function; see [Nesterov and Nemirovskii \[17, Proposition 5.1.4\]](#) and [Boyd and Vandenberghe \[5, Example 9.8\]](#). Standard interior methods can therefore be used to project onto the constraint set.

*Example 5.5* (Bernoulli distribution). When the observations  $b_i$  are independent Bernoulli observations, the perspective dual constraints in (5.10) reduce to

$$\sum_{i=1}^m [z_i \log(z_i/\xi) + (\xi - z_i) \log((\xi - z_i)/\xi)] \leq \beta\xi - (1 + \sigma\xi), \quad z = y + \xi b, \quad \xi \geq 0,$$

where  $\beta = \sum_{i=1}^m (b_i \log b_i + (1 - b_i) \log(1 - b_i))$  is a constant. By introducing new variables, this can be rewritten with only affine constraints and  $2m$  relative-entropy constraints. Thus the projection subproblems can be solved as in the Poisson case.

**6. Examples: recovering primal solutions.** Once we have constructed and solved the gauge or perspective dual problems, we have two available approaches for recovering a corresponding primal optimal solution. If we applied a (Lagrange) primal-dual algorithm (e.g., the algorithm of Chambolle and Pock [8]) to solve the dual, then [Theorem 3.9](#) gives a direct recipe for constructing a primal solution from the output of the algorithm. On the other hand, if we applied a primal-only algorithm to solve the dual, then we must instead rely on [Corollary 3.7](#) or [Corollary 4.9](#) to recover a primal solution. Interestingly, the alignment conditions in these theorems can provide insight into the structure of the primal optimal solution, as illustrated by the following examples.

**6.1. Recovery for basis pursuit denoising.** The example in this section illustrates how [Corollary 3.7](#) can be used to recover primal optimal solutions from dual optimal solutions for a simple gauge problem. Consider the gauge dual pair

$$(6.1a) \quad \underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \|b - Ax\|_2 \leq \sigma$$

$$(6.1b) \quad \underset{y}{\text{minimize}} \quad \|A^T y\|_\infty \quad \text{subject to} \quad \langle b, y \rangle - \sigma \|y\|_2 \geq 1,$$

which corresponds to the basis pursuit denoising problem. The 1-norm in the primal objective encourages sparsity in  $x$  while the constraint enforces a maximum deviation between a forward model  $Ax$  and observations  $b$ .

Assume that  $y^*$  is optimal for the dual problem, and set  $z = A^T y^*$ . Define the active set

$$I(z) = \{i \mid |z_i| = \|z\|_\infty\}$$

as the set of indices of  $z$  that achieve the optimal objective value of the gauge dual. We use [Corollary 3.7](#) to determine properties of a primal solution  $x^*$ . In particular, the first part of [Corollary 3.7\(b\)](#) holds if and only if  $x_i^* = 0$  for all  $i \notin I(z)$ , and  $\text{sign}(x_i^*) = \text{sign}(z_i)$  for all  $i \in I(z)$ . Thus, the maximal-in-modulus elements of  $A^T y^*$  determine the support for any primal optimal solution  $x^*$ . The second condition in [Corollary 3.7\(b\)](#) holds if and only if  $b - Ax = \sigma y^* / \|y^*\|_2$ . In order to satisfy this last condition, we would solve the least-squares problem restricted to the support of the solution, i.e.,

$$\underset{x}{\text{minimize}} \quad \|b - Ax - \sigma y^* / \|y^*\|_2\|_2^2 \quad \text{subject to} \quad x_i = 0 \quad \forall i \notin I(z).$$

(Note that  $y^* \neq 0$ , since otherwise the primal problem is infeasible.) The efficiency of this least-squares solve depends on the number of elements in  $I(z)$ . For many applications of basis pursuit denoising, for example, we expect the support to be small relative to the length of  $x$ , and in that case, the least-squares recovery problem is expected to be a relatively inexpensive subproblem. We may interpret the role of the dual problem as that of determining the optimal support of the primal, and the role of the above least-squares problem is to recover the actual values of the support.

**6.2. Sparse recovery with Huber misfit.** Consider the problem

$$(6.2) \quad \underset{x}{\text{minimize}} \quad \|x\|_1 \quad \text{subject to} \quad \sum_{i=1}^m h_\eta(Ax - b) \leq \sigma,$$

where  $h_\eta$  is the Huber function defined in [Example 5.2](#). This problem corresponds to  $(\mathbf{N}_p)$  with  $f(x) = \|x\|_1$  and  $g(r) = \sum_{i=1}^m h_\eta(r_i)$ . Suppose that the tuple  $(y, \alpha, \mu)$  is optimal for the perspective dual with  $\mu < 0$ , and that the optimal value of  $(\mathbf{N}_p)$  is attained. Because  $f$  is a gauge, [Corollary 4.10](#) asserts that  $\alpha^* = 0$ , and hence part (b) of [Corollary 4.9](#) reduces to the conditions

$$(6.3a) \quad \langle x, A^T y \rangle = f(x) \cdot f^\circ(A^T y)$$

$$(6.3b) \quad (b - Ax, 1) \in \sigma \partial g^\sharp(y, \mu).$$

As we did for the basis pursuit example ([§6.1](#)), we use [\(6.3a\)](#) to deduce the support of the optimal primal solution. Next, it follows from [Theorem 5.1](#) that when  $g$  is PLQ,

$$g^\sharp(y, \mu) = \delta_{\mathbb{R}_-}(\mu) + \max \left( \max_{i=1, \dots, k} \{W_i^T y / w_i\}, -\frac{1}{2\mu} \|Ly\|^2 \right).$$

In particular, because  $g$  is a separable sum of Huber functions,  $W = [I \ -I]^T$ ,  $w$  is the constant vector of all ones, and  $L = \sqrt{\eta}I$ . Since  $\mu < 0$ , it follows that

$$\partial g^\sharp(y, \mu) = \partial \left( \max \left\{ \|y\|_\infty, -\frac{\eta}{2\mu} \|y\|^2 \right\} \right) (y, \mu).$$

Define the set

$$V(y, \mu) = \{v_1, \dots, v_{2m+1}\} := \left\{ y_1, \dots, y_m, -y_1, \dots, -y_m, -\frac{\eta}{2\mu} \|y\|^2 \right\},$$

and let  $J(y, \mu) := \{j \mid |v_j| = \max_{i=1, \dots, 2m+1} |v_i|\}$  be the set of maximizing indices for  $V(y, \mu)$ . Then

$$\partial g^\sharp(y, \mu) = \text{conv} \{ \nabla v_j \mid j \in J(y, \mu) \}.$$

More concretely, precisely the following terms are contained in the convex hull above:

- $\left( -\frac{\eta}{\mu} y, \frac{\eta}{2\mu^2} \|y\|^2 \right)$  if  $-\frac{\eta}{2\mu} \|y\|^2 \geq \|y\|_\infty$ ;
- $(\text{sign}(y_i) \cdot e_i, 0)$  if  $i \in [m]$  and  $|y_i| = \|y\|_\infty \geq -\frac{\eta}{2\mu} \|y\|^2$ .

(Here,  $e_i$  is the  $i$ th standard basis vector.) Note that if there is an optimal solution to  $(\mathbf{N}_p)$ , then [Theorem 4.9](#) tells us that  $\left( -\frac{\eta}{\mu} y, \frac{\eta}{2\mu^2} \|y\|^2 \right)$  must be included in this convex hull, otherwise it is impossible to have  $(b - Ax, 1) \in \partial g^\sharp(y, \mu)$ .

In summary, [Corollary 4.9](#) tells us that to find an optimal solution  $x$  for  $(\mathbf{N}_p)$ , we need to solve a linear program to ensure that  $(b - Ax, 1) \in \text{conv} \{ \nabla v_j \mid j \in J(y, \mu) \}$  subject to the optimal support of  $x$ , as determined by [\(6.3a\)](#). In cases where the size of the support is expected to be small (as might be expected with a 1-norm objective), this required linear program can be solved efficiently. Compare this result to the analogous example from the gauge case, outlined in [section 6.1](#).

**7. Numerical experiment: sparse robust regression.** To illustrate the usefulness of the primal-from-dual recovery procedure implied by [Theorem 3.9](#), we continue to examine the sparse robust regression problem [\(6.2\)](#), as considered by Aravkin et al. [\[2\]](#). The aim is to find a sparse signal (e.g., a spike train) from measurements contaminated by outliers. These experiments have been performed with the following data:  $m = 120$ ,  $n = 512$ ,  $\sigma = 0.2$ ,  $\eta = 1$ , and  $A$  is a Gaussian matrix. The true solution  $x_{\text{true}} \in \{-1, 0, 1\}$  is a spike train which has been constructed to have 20 nonzero entries, and the true noise  $b - Ax_{\text{true}}$  has been constructed to have 5 outliers.

We compare two approaches for solving problem [\(6.2\)](#). In both, we use Chambolle and Pock’s (CP) algorithm [\[8\]](#), which is primal-dual (in the sense of Lagrange duality) and can be adapted to solve both the primal problem [\(6.2\)](#) and its perspective dual [\(5.3\)](#). Other numerical methods could certainly be applied to either of these problems; we note, however, that applying a primal-only method to [\(5.3\)](#) would require us to use the methods of [section 6](#) rather than [Theorem 3.9](#) for the recovery of a primal solution.

The CP method applies to the convex optimization problem

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \quad F(Kw) + G(w)$$

where  $K : \mathbb{R}^d \rightarrow \mathbb{R}^l$  is a linear operator and  $F : \mathbb{R}^l \rightarrow \overline{\mathbb{R}}$  and  $G : \mathbb{R}^d \rightarrow \overline{\mathbb{R}}$  are closed convex functions. The CP iterations are given by

$$(7.1a) \quad w^{k+1} := \text{prox}_{\alpha_y F^*} (w^k + \alpha_w K(2z^k - z^{k-1}))$$

$$(7.1b) \quad z^{k+1} := \text{prox}_{\alpha_z G} (z^k - \alpha_z K^* w^{k+1}),$$

where  $\text{prox}_{\alpha f}(x) = \text{argmin}_y \{f(y) + \frac{1}{2\alpha} \|x - y\|_2^2\}$ . The positive scalars  $\alpha_w$  and  $\alpha_z$  are chosen to satisfy  $\alpha_w \alpha_z \|K\|^2 < 1$ . Setting  $F = \delta_{g(b-\cdot) \leq \sigma}$ ,  $K = A$ , and  $G = \|\cdot\|_1$  yields the primal problem [\(6.2\)](#). In this case, the proximal operators  $\text{prox}_{\alpha F^*}$  and  $\text{prox}_{\alpha G}$  can be computed using the Moreau identity, i.e.,

$$\text{prox}_{\alpha F^*}(w) = w - \text{prox}_{(\alpha F)^*}(w) = w - \alpha \Pi_F(w/\alpha)$$

$$\text{prox}_{\alpha G}(z) = z - \text{prox}_{(\alpha G)^*}(z) = z - \Pi_{\alpha \mathbb{B}_\infty}(z/\alpha)$$

where  $\Pi_F$  is the projection onto the sublevel set in the definition of  $F$ , and  $\Pi_{\alpha \mathbb{B}_\infty}$  is the projection onto the infinity-norm ball of radius  $\alpha$ . We implement  $\Pi_F$  using Convex [\[22\]](#) with the Gurobi solver.

On the other hand, to apply CP to the perspective dual problem [\(5.3\)](#), one instead takes  $F = f^\circ = \|\cdot\|_\infty$ ,  $K = A^T$ , and  $G = \delta_{\mathcal{Q}}$ , where  $\mathcal{Q}$  is the constraint set for [\(5.3\)](#). To compute  $\text{prox}_{\alpha_z G}$ , which is the projection onto  $\mathcal{Q}$ , we solve the SOCP [\(5.4\)](#) using Gurobi. To compute  $\text{prox}_{\alpha F^*}(w)$ , we again use the Moreau identity and project onto level sets of  $\|\cdot\|_1$ .

[Figure 7.1](#) compares the outcomes of running CP on the primal and perspective dual problems. This experiment exhibited similar behavior when run 500 times with different realizations of the random data, and so here we report on a single problem instance. Note that performing an iteration of CP on the perspective dual is significantly faster than performing an iteration of CP on the primal because  $\Pi_{\mathcal{Q}}$  can be computed much more efficiently than  $\Pi_F$  (see the discussion in [section 5.1](#)). This also appears to make convergence of CP on the perspective dual more stable, as seen in [Figure 7.1\(a\)](#). [Figure 7.1\(c\)-\(d\)](#) illustrate the sparsity patterns of the iterates  $x_k$  relative to those  $x_{\text{true}}$ ; notably, we recover the correct sparsity patterns using [Theorem 3.9](#). The recovery procedure outlined in [section 6.2](#) also recovers the correct sparsity pattern, when applied to the final perspective dual iterate.

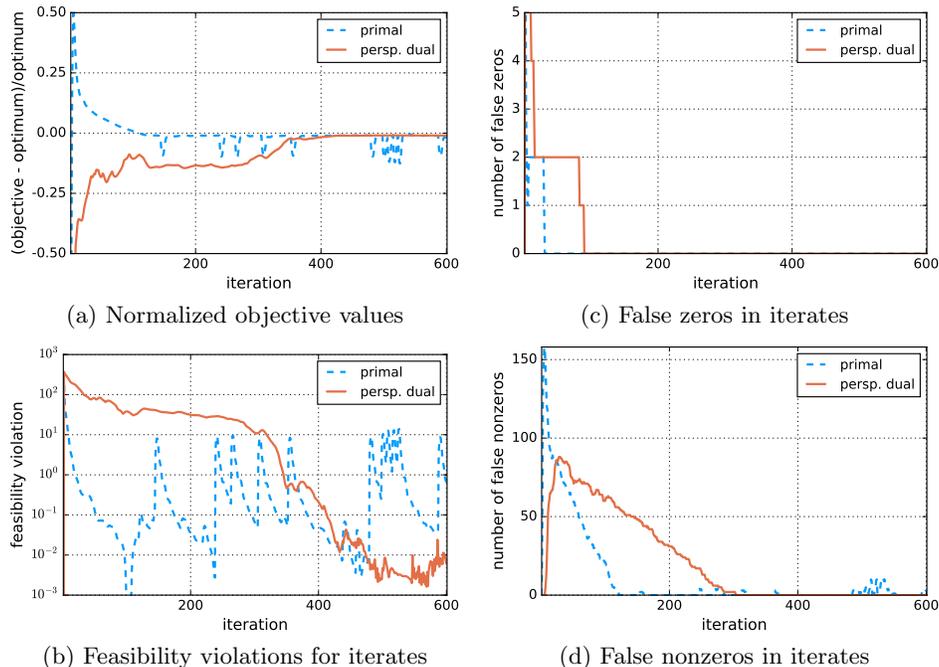


Fig. 7.1: The CP algorithm applied to sparse robust regression; see [section 7](#). Dashed lines indicate CP applied to the primal problem (6.2), and solid lines indicate CP applied to the perspective dual problem (5.3) where the primal solution is recovered via the method of [Theorem 3.9](#). Plots show (a) normalized deviation of objective value  $\|x^k\|_1$  from optimal value  $\|x_{true}\|_1$ ; (b) infeasibility measure  $\max(g(b - Ax^k) - \sigma, 0)$  for iterate  $x^k$ ; (c) number false zeros in iterate  $x^k$  relative to  $x_{true}$ ; (d) number of false nonzeros in iterate  $x^k$  relative to  $x_{true}$ .

**8. Discussion.** Gauge duality is fascinating in part because it shares many symmetric properties with Lagrange duality, and yet Freund’s 1987 development of the concept flows from an entirely different principle based on polarity of the sets that define the gauge functions. On the other hand, Lagrange duality proceeds from a perturbation argument, which yields as one of its hallmarks a sensitivity interpretation of the dual variables. The discussion in [section 3](#) reveals that both duality notions can be derived from the same Rockafellar-Fenchel perturbation framework. The derivation of gauge duality using this framework appears to be its first application to a perturbation that does not lead to Lagrange duality. This new link between gauge duality and the perturbation framework establishes a sensitivity interpretation for gauge dual variables, which has not been available until now.

One motivation for this work is to explore alternative formulations of optimization problems that might be computationally advantageous for certain problem classes. The phase-retrieval problem, based on an SDP formulation, was a first application of ideas from gauge duality for developing large-scale solvers [12]. That approach, however, was limited in its flexibility because it required gauge functions. The discussions of [section 4](#) pave the way to new extensions, such as different models of the measurement

process, as described in [section 5.2](#).

Another implication of this work is that it establishes the foundations for exploring a new breed of primal-dual algorithms. Our own application of a primal-dual algorithm (Chambolle and Pock’s method [8]) to the perspective-dual problem, together with a procedure for extracting a primal estimate, is a first exploratory step towards developing variations of such methods that directly work on the perspective-dual pair. Future directions of research include the development of such algorithms together with their attendant convergence properties, and to understand the classes of problems for which they are practicable.

## REFERENCES

- [1] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and S. Roy. Level-set methods for convex optimization. *arXiv preprint arXiv:1602.01506*, 2016.
- [2] A. Y. Aravkin, J. V. Burke, and M. P. Friedlander. Variational properties of value functions. *SIAM Journal on optimization*, 23(3):1689–1717, 2013.
- [3] A. Y. Aravkin, J. V. Burke, and G. Pillonetto. Linear system identification using stable spline kernels and PLQ penalties. In *52nd IEEE Conference on Decision and Control*, pages 5168–5173, Dec 2013.
- [4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.
- [5] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge university press, 2004.
- [6] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, Feb 2006.
- [7] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.
- [8] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.
- [9] I. S. Dhillon and J. A. Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2007.
- [10] D. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
- [11] R. M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.
- [12] M. P. Friedlander and I. Macêdo. Low-rank spectral optimization via gauge duality. *SIAM Journal on Scientific Computing*, 28(3):A1616–A1638, 2016.
- [13] M. P. Friedlander, I. Macêdo, and T. K. Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.
- [14] I. Gurobi Optimization. Gurobi optimizer reference manual, 2015.
- [15] P. Huber. *Robust Statistics*. Wiley, 1981.
- [16] J. A. Nelder and R. J. Baker. Generalized linear models. *Encyclopedia of statistical sciences*, 1972.
- [17] Y. Nesterov and A. Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.
- [18] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1972.
- [19] R. T. Rockafellar. First- and second-order epi-differentiability in nonlinear programming. *Trans. Amer. Math. Soc.*, 307(1):75–108, May 1988.
- [20] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.
- [21] J. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory*, 52(3):1030–1051, 2006.
- [22] M. Udell, K. Mohan, D. Zeng, J. Hong, S. Diamond, and S. Boyd. Convex optimization in Julia. *SC14 Workshop on High Performance Technical Computing in Dynamic Languages*, 2014.
- [23] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

**Appendix A. Proof of [Lemma 3.4](#).** Suppose first that the primal  $(G_p)$  is

relatively strictly feasible. A point  $y$  lies in the domain of  $p$  if and only if the system

$$\begin{bmatrix} y \\ 0 \\ 0 \end{bmatrix} \in M \begin{bmatrix} w \\ \lambda \end{bmatrix} + (\text{epi } \rho) \times \mathcal{U}_\kappa, \quad \text{where } M := \begin{pmatrix} A & -b \\ 0 & -\sigma \\ -I & 0 \end{pmatrix},$$

is solvable for  $(w, \lambda)$ . Thus the set  $(\text{dom } p) \times \{0\} \times \{0\}$  coincides with

$$(A.1) \quad L \cap (\text{range } M + (\text{epi } \rho) \times \mathcal{U}_\kappa),$$

where  $L$  is the linear subspace  $L := \{(a, b, c) \mid b = 0, c = 0\}$ . We aim to show  $(0, 0, 0)$  is in the relative interior of (A.1), which will show that  $0 \in \text{ri}(\text{dom } p)$ . Appealing to [18, Lemma 7.3] and [18, Theorem 7.6], we have the representations

$$\begin{aligned} \text{ri}(\text{epi } \rho) &= \{(z, r) \in \text{ri}(\text{dom } \rho) \times \mathbb{R} \mid \rho(z) < r\} \\ \text{ri}\mathcal{U}_\kappa &= \{x \in \text{ri}(\text{dom } \kappa) \mid \kappa(x) < 1\}. \end{aligned}$$

From relative strict feasibility of  $(G_p)$ , we deduce existence of an  $x \in \text{ri}(\text{dom } \kappa)$  with

$$b - Ax \in \text{ri}(\text{dom } \rho) \quad \text{and} \quad \rho(b - Ax) < \sigma.$$

Fix a constant  $r > \kappa(x)$  and define the pair  $(w, \lambda) := (x/r, 1/r)$ . Then we immediately have  $(b\lambda - Aw, \sigma\lambda) \in \text{ri}(\text{epi } \rho)$  and  $\kappa(w) < 1$ . It follows that the vector  $-M \begin{bmatrix} w \\ \lambda \end{bmatrix}$  lies in  $\text{ri}(\text{epi } \rho) \times \text{ri}\mathcal{U}_\kappa$ . Thus  $(0, 0, 0)$  lies in the intersection

$$(A.2) \quad (\text{ri } L) \cap (\text{range } M + (\text{ri}(\text{epi } \rho) \times \text{ri}\mathcal{U}_\kappa)).$$

Appealing to [18, Theorem 6.5, Corollary 6.6.2], we deduce that (A.2) is precisely the relative interior of the intersection (A.1). Thus  $y = 0$  lies in the relative interior of  $\text{dom } p$  as claimed.

Next, suppose that the gauge dual  $(G_d)$  is strictly feasible. By definition of  $f^*$ , the tuple  $(w, \lambda)$  lies in the domain of  $v_d$  if and only if

$$(w, 0, -\lambda) \in (\text{dom } \kappa^\circ \times \text{epi}(\sigma\rho^\circ - \langle b, \cdot \rangle + 1)) - \text{range } B, \quad \text{with } B := \begin{bmatrix} A^T \\ I \\ 0 \end{bmatrix}.$$

Thus  $\text{dom } v_d$  is linearly isomorphic to the intersection

$$(A.3) \quad L' \cap ((\text{dom } \kappa^\circ \times \text{epi}(\sigma\rho^\circ - \langle b, \cdot \rangle + 1)) - \text{range } B),$$

where  $L'$  is the linear subspace  $L' := \{(a, b, c) \mid b = 0\}$ . However, by [18, Lemma 7.3], relative strict feasibility of the dual  $(G_d)$  amounts to the inclusion

$$(0, 0, 0) \in (\text{ri}(\text{dom } \kappa^\circ) \times \text{ri}(\text{epi}(\sigma\rho^\circ - \langle b, \cdot \rangle + 1))) - \text{range } B.$$

Using [18, Corollary 6.5.1, Corollary 6.6.2], we deduce that strict feasibility of  $(G_d)$  implies  $(0, 0, 0)$  is in the relative interior of the intersection (A.3), and thus  $0 \in \text{ri}(\text{dom } v_d)$ , as claimed.

Finally, the exact same arguments, but with relative interiors replaced by interiors, will prove the claims relating strict feasibility and interiority.