Journal of Machine Learning Research 14 (2013) 2689-2728 Submitted 2/12; Revised 1/13; Published 9/13

Sparse/Robust Estimation and Kalman Smoothing with Nonsmooth Log-Concave Densities: Modeling, Computation, and Theory*

Aleksandr Y. Aravkin SARAVKIN@US.IBM.COM
IBM T.J. Watson Research Center
1101 Kitchawan Rd
Yorktown Heights, NY 10598

James V. Burke JVBURKE@UW.EDU
Department of Mathematics
Box 354350
University of Washington
Seattle, WA, 98195-4350 USA

Gianluigi Pillonetto GIAPI@DEI.UNIPD.IT
Department of Information Engineering
Via Gradenigo 6/A
University of Padova
Padova, Italy

Editor: Aapo Hyvarinen

Abstract

We introduce a new class of quadratic support (QS) functions, many of which already play a crucial role in a variety of applications, including machine learning, robust statistical inference, sparsity promotion, and inverse problems such as Kalman smoothing. Well known examples of QS penalties include the ℓ_2, Huber, ℓ_1 and Vapnik losses. We build on a dual representation for QS functions, using it to characterize conditions necessary to interpret these functions as negative logs of true probability densities. This interpretation establishes the foundation for statistical modeling with both known and new QS loss functions, and enables construction of non-smooth multivariate distributions with specified means and variances from simple scalar building blocks.

The main contribution of this paper is a flexible statistical modeling framework for a variety of learning applications, together with a toolbox of efficient numerical methods for estimation. In particular, a broad subclass of QS loss functions known as piecewise linear quadratic (PLQ) penalties has a dual representation that can be exploited to design interior point (IP) methods. IP methods solve nonsmooth optimization problems by working directly with smooth systems of equations characterizing their optimality. We provide several numerical examples, along with a code that can be used to solve general PLQ problems.

The efficiency of the IP approach depends on the structure of particular applications. We consider the class of dynamic inverse problems using Kalman smoothing. This class comprises a wide variety of applications, where the aim is to reconstruct the state of a dynamical system with known process and measurement models starting from noisy output samples. In the classical case, Gaus-

*. The authors would like to thank Bradley M. Bell for insightful discussions and helpful suggestions. The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement no 257462 HYCON2 Network of excellence, by the MIUR FIRB project RBFR12M3AC - Learning meets time: a new computational approach to learning in dynamic systems.

sian errors are assumed both in the process and measurement models for such problems. We show that the extended framework allows arbitrary PLQ densities to be used, and that the proposed IP approach solves the generalized Kalman smoothing problem while maintaining the linear complexity in the size of the time series, just as in the Gaussian case. This extends the computational efficiency of the Mayne-Fraser and Rauch-Tung-Striebel algorithms to a much broader nonsmooth setting, and includes many recently proposed robust and sparse smoothers as special cases.

Keywords: statistical modeling, convex analysis, nonsmooth optimization, robust inference, sparsity optimization, Kalman smoothing, interior point methods

1. Introduction

Consider the classical problem of Bayesian parametric regression (MacKay, 1992; Roweis and Ghahramani, 1999) where the unknown $x \in \mathbb{R}^n$ is a random vector,[1] with a prior distribution specified using a known invertible matrix $G \in \mathbb{R}^{n\times n}$ and known vector $\mu \in \mathbb{R}^n$ via

$$\mu = Gx + w\,, \tag{1}$$

where w is a zero mean vector with covariance Q. Let z denote a linear transformation of x contaminated with additive zero mean measurement noise v with covariance R,

$$z = Hx + v\,, \tag{2}$$

where $H \in \mathbb{R}^{\ell\times n}$ is a known matrix, while v and w are independent. It is well known that the (unconditional) minimum variance linear estimator of x, as a function of z, is the solution to the following optimization problem:

$$\min_x \quad (z-Hx)^{\mathrm{T}}R^{-1}(z-Hx) + (\mu - Gx)^{\mathrm{T}}Q^{-1}(\mu - Gx)\,. \tag{3}$$

As we will show, (3) includes estimation problems arising in discrete-time dynamic linear systems which admit a state space representation (Anderson and Moore, 1979; Brockett, 1970). In this context, x is partitioned into N subvectors $\{x_k\}$, where each x_k represents the hidden system state at time instant k. For known data z, the classical Kalman smoother exploits the special structure of the matrices H, G, Q and R to compute the solution of (3) in $O(N)$ operations (Gelb, 1974). This procedure returns the minimum variance estimate of the state sequence $\{x_k\}$ when the additive noise in the system is assumed to be Gaussian.

In many circumstances, the estimator (3) performs poorly; put another way, quadratic penalization on model deviation is a bad model in many situations. For instance, it is not robust with respect to the presence of outliers in the data (Huber, 1981; Gao, 2008; Aravkin et al., 2011a; Farahmand et al., 2011) and may have difficulties in reconstructing fast system dynamics, for example, jumps in the state values (Ohlsson et al., 2012). In addition, sparsity-promoting regularization is often used in order to extract a small subset from a large measurement or parameter vector which has greatest impact on the predictive capability of the estimate for future data. This sparsity principle permeates many well known techniques in machine learning and signal processing, including feature selection, selective shrinkage, and compressed sensing (Hastie and Tibshirani, 1990; Efron et al., 2004; Donoho, 2006). In these cases, (3) is often replaced by a more general formulation

$$\min_x \quad V\left(Hx - z; R\right) + W\left(Gx - \mu; Q\right) \tag{4}$$

1. All vectors are column vectors, unless otherwise specified.

where the loss V may be the ℓ_2-norm, the Huber penalty (Huber, 1981), Vapnik's ε-insensitive loss, used in support vector regression (Vapnik, 1998; Hastie et al., 2001), or the hinge loss, leading to support vector classifiers (Evgeniou et al., 2000; Pontil and Verri, 1998; Schölkopf et al., 2000). The regularizer W may be the ℓ_2-norm, the ℓ_1-norm, as in the LASSO (Tibshirani, 1996), or a weighted combination of the two, yielding the elastic net procedure (Zou and Hastie, 2005). Many learning algorithms using infinite-dimensional reproducing kernel Hilbert spaces as hypothesis spaces (Aronszajn, 1950; Saitoh, 1988; Cucker and Smale, 2001) boil down to solving finite-dimensional problems of the form (4) by virtue of the representer theorem (Wahba, 1998; Schölkopf et al., 2001).

These robust and sparse approaches can often be interpreted as placing non-Gaussian priors on w (or directly on x) and on the measurement noise v. The Bayesian interpretation of (4) has been extensively studied in the statistical and machine learning literature in recent years, and probabilistic approaches used in the analysis of estimation and learning algorithms have been studied (Mackay, 1994; Tipping, 2001; Wipf et al., 2011). Non-Gaussian model errors and priors leading to a great variety of loss and penalty functions are also reviewed by Palmer et al. (2006) using convex-type representations, and integral-type variational representations related to Gaussian scale mixtures.

In contrast to the above approaches, in the first part of the paper, we consider a wide class of quadratic support (QS) functions and exploit their dual representation. This class of functions generalizes the notion of piecewise linear quadratic (PLQ) penalties (Rockafellar and Wets, 1998). The dual representation is the key to identifying which QS loss functions can be associated with a density, which in turn allows us to interpret the solution to the problem (4) as a MAP estimator when the loss functions V and W come from this subclass of QS penalties. This viewpoint allows statistical modeling using non-smooth penalties, such as the ℓ_1, hinge, Huber and Vapnik losses, which are all PLQ penalties. Identifying a statistical interpretation for this class of problems gives us several advantages, including a systematic constructive approach to prescribe mean and variance parameters for the corresponding model; a property that is particularly important for Kalman smoothing.

In addition, the dual representation provides the foundation for efficient numerical methods in estimation based on interior point optimization technology. In the second part of the paper, we derive the Karush-Kuhn-Tucker (KKT) equations for problem (4), and introduce interior point (IP) methods, which are iterative methods to solve the KKT equations using smooth approximations. This is essentially a smoothing approach to many (non-smooth) robust and sparse problems of interest to practitioners. Furthermore, we provide conditions under which the IP methods solve (4) when V and W come from PLQ densities, and describe implementation details for the entire class.

A concerted research effort has recently focused on the solution of regularized large-scale inverse and learning problems, where computational costs and memory limitations are critical. This class of problems includes the popular kernel-based methods (Rasmussen and Williams, 2006; Schölkopf and Smola, 2001; Smola and Schölkopf, 2003), coordinate descent methods (Tseng and Yun, 2008; Lucidi et al., 2007; Dinuzzo, 2011) and decomposition techniques (Joachims, 1998; Lin, 2001; Lucidi et al., 2007), one of which is the widely used sequential minimal optimization algorithm for support vector machines (Platt, 1998). Other techniques are based on kernel approximations, for example, using incomplete Cholesky factorization (Fine and Scheinberg, 2001), approximate eigen-decomposition (Zhang and Kwok, 2010) or truncated spectral representations (Pillonetto and Bell, 2007). Efficient interior point methods have been developed for ℓ_1-regularized problems (Kim et al., 2007), and for support vector machines (Ferris and Munson, 2003).

In contrast, general and efficient solvers for state space estimation problems of the form (4) are missing in the literature. The last part of this paper provides a contribution to fill this gap, spe-

cializing the general results to the dynamic case, and recovering the classical efficiency results of the least-squares formulation. In particular, we design new Kalman smoothers tailored for systems subject to noises coming from PLQ densities. Amazingly, it turns out that the IP method studied by Aravkin et al. (2011a) generalizes perfectly to the entire class of PLQ densities under a simple verifiable non-degeneracy condition. In practice, IP methods converge in a small number of iterations, and the effort per iteration depends on the structure of the underlying problem. We show that the IP iterations for all PLQ Kalman smoothing problems can be computed with a number of operations that scales linearly in N, as in the quadratic case. This theoretical foundation generalizes the results recently obtained by Aravkin et al. (2011a), Aravkin et al. (2011b), Farahmand et al. (2011) and Ohlsson et al. (2012), framing them as particular cases of the general framework presented here.

The paper is organized as follows. In Section 2 we introduce the class of QS convex functions, and give sufficient conditions that allow us to interpret these functions as the negative logs of associated probability densities. In Section 3 we show how to construct QS penalties and densities having a desired structure from basic components, and in particular how multivariate densities can be endowed with prescribed means and variances using scalar building blocks. To illustrates this procedure, further details are provided for the Huber and Vapnik penalties. In Section 4, we focus on PLQ penalties, derive the associated KKT system, and present a theorem that guarantees convergence of IP methods under appropriate hypotheses. In Section 5, we present a few simple well-known problems, and compare a basic IP implementation for these problems with an ADMM implementation (all code is available online). In Section 6, we present the Kalman smoothing dynamic model, formulate Kalman smoothing with PLQ penalties, present the KKT system for the dynamic case, and show that IP iterations for PLQ smoothing preserve the classical computational efficiency known for the Gaussian case. We present numerical examples using both simulated and real data in Section 7, and make some concluding remarks in Section 8. Section A serves as an appendix where supporting mathematical results and proofs are presented.

2. Quadratic Support Functions And Densities

In this section, we introduce the class of Quadratic Support (QS) functions, characterize some of their properties, and show that many commonly used penalties fall into this class. We also give a statistical interpretation to QS penalties by interpreting them as negative log likelihoods of probability densities; this relationship allows prescribing means and variances along with the general quality of the error model, an essential requirement of the Kalman smoothing framework and many other areas.

2.1 Preliminaries

We recall a few definitions from convex analysis, required to specify the domains of QS penalties. The reader is referred to Rockafellar (1970) and Rockafellar and Wets (1998) for more detailed reading.

- (Affine hull) Define the affine hull of any set $C \subset \mathbb{R}^n$, denoted by $\mathrm{aff}(C)$, as the smallest affine set (translated subspace) that contains C.
- (Cone) For any set $C \subset \mathbb{R}^n$, denote by cone C the set $\{tr | r \in C, t \in \mathbb{R}_+\}$.
- (Domain) For $f(x) : \mathbb{R}^n \to \overline{\mathbb{R}} = \{\mathbb{R} \cup \infty\}$, $\mathrm{dom}(f) = \{x : f(x) < \infty\}$.

- (Polars of convex sets) For any convex set $C \subset \mathbb{R}^m$, the polar of C is defined to be

$$C^\circ := \{r | \langle r, d\rangle \leq 1 \; \forall \, d \in C\},$$

  and if C is a convex cone, this representation is equivalent to

$$C^\circ := \{r | \langle r, d\rangle \leq 0 \; \forall \, d \in C\}.$$

- (Horizon cone). Let $C \subset \mathbb{R}^n$ be a nonempty convex set. The horizon cone C^∞ is the convex cone of 'unbounded directions' for C, that is, $d \in C^\infty$ if $C + d \subset C$.

- (Barrier cone). The barrier cone of a convex set C is denoted by $\mathrm{bar}(C)$:

$$\mathrm{bar}(C) := \{x^* | \text{for some } \beta \in \mathbb{R}, \langle x, x^*\rangle \leq \beta \;\; \forall x \in C\}\,.$$

- (Support function). The support function for a set C is denoted by $\delta^*(x\,|\,C)$:

$$\delta^*(x\,|\,C) := \sup_{c \in C} \langle x, c\rangle \;\; .$$

2.2 QS Functions And Densities

We now introduce the QS functions and associated densities that are the focus of this paper. We begin with the dual representation, which is crucial to both establishing a statistical interpretation and to the development of a computational framework.

Definition 1 (Quadratic Support functions and penalties) *A QS function is any function* $\rho(U, M, b, B; \cdot) : \mathbb{R}^n \to \overline{\mathbb{R}}$ *having representation*

$$\rho(U, M, b, B; y) = \sup_{u \in U} \left\{ \langle u, b + By\rangle - \tfrac{1}{2}\langle u, Mu\rangle \right\} \;, \tag{5}$$

where $U \subset \mathbb{R}^m$ is a nonempty convex set, $M \in \mathcal{S}^n_+$ the set of real symmetric positive semidefinite matrices, and $b + By$ is an injective affine transformation in y, with $B \in \mathbb{R}^{m \times n}$, so, in particular, $m \leq n$ and $\mathrm{null}(B) = \{0\}$.

When $0 \in U$, we refer to the associated QS function as a penalty, since it is necessarily non-negative.

Remark 2 *When U is polyhedral, $0 \in U$, $b = 0$ and $B = I$, we recover the basic piecewise linear-quadratic penalties characterized in Rockafellar and Wets (1998, Example 11.18).*

Theorem 3 *Let U, M, B, b be as in Definition 1, and set $K = U^\infty \cap \mathrm{null}(M)$. Then*

$$B^{-1}[\mathrm{bar}(U) + \mathrm{Ran}\,(M) - b] \subset \mathrm{dom}[\rho(U, M, B, b; \cdot)] \subset B^{-1}[K^\circ - b]\,,$$

with equality throughout when $\mathrm{bar}(U) + \mathrm{Ran}\,(M)$ is closed, where $\mathrm{bar}(U) = \mathrm{dom}\,(\delta^*(\cdot\,|\,U))$ is the barrier cone of U. In particular, equality always holds when U is polyhedral.

We now show that many commonly used penalties are special cases of QS (and indeed, of the PLQ) class.

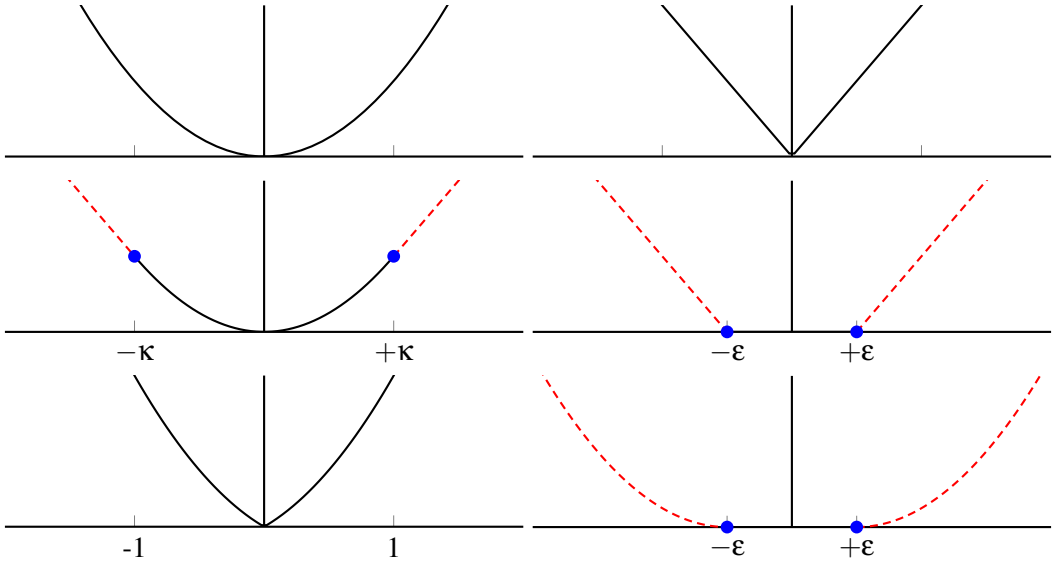

Figure 1: Scalar ℓ_2 (top left), ℓ_1 (top right), Huber (middle left), Vapnik (middle right), elastic net (bottom left) and smooth insensitive loss (bottom right) penalties

Remark 4 (scalar examples) *ℓ_2, ℓ_1, elastic net, Huber, hinge, and Vapnik penalties are all representable using the notation of Definition 1.*

1. *ℓ_2: Take $U = \mathbb{R}$, $M = 1$, $b = 0$, and $B = 1$. We obtain*

$$\rho(y) = \sup_{u \in \mathbb{R}} \left\{ uy - u^2/2 \right\} .$$

The function inside the sup *is maximized at $u = y$, hence $\rho(y) = \frac{1}{2}y^2$, see top left panel of Figure 1.*

2. *ℓ_1: Take $U = [-1,1]$, $M = 0$, $b = 0$, and $B = 1$. We obtain*

$$\rho(y) = \sup_{u \in [-1,1]} \{uy\} .$$

The function inside the sup *is maximized by taking $u = \text{sign}(y)$, hence $\rho(y) = |y|$, see top right panel of Figure 1.*

3. *Elastic net: $\ell_2 + \lambda\ell_1$. Take*

$$U = \mathbb{R} \times [-\lambda, \lambda],\ b = \begin{bmatrix} 0 \\ 0 \end{bmatrix},\ M = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},\ B = \begin{bmatrix} 1 \\ 1 \end{bmatrix} .$$

This construction reveals the general calculus of PLQ addition, see Remark 5. See bottom right panel of Figure 1.

4. *Huber: Take* $U = [-\kappa, \kappa]$, $M = 1$, $b = 0$, *and* $B = 1$. *We obtain*

$$\rho(y) = \sup_{u \in [-\kappa,\kappa]} \left\{ uy - u^2/2 \right\} ,$$

with three explicit cases:

   (a) *If* $y < -\kappa$, *take* $u = -\kappa$ *to obtain* $-\kappa y - \frac{1}{2}\kappa^2$.
   (b) *If* $-\kappa \leq y \leq \kappa$, *take* $u = y$ *to obtain* $\frac{1}{2}y^2$.
   (c) *If* $y > \kappa$, *take* $u = \kappa$ *to obtain a contribution of* $\kappa y - \frac{1}{2}\kappa^2$.

This is the Huber penalty, shown in the middle left panel of Figure 1.

5. *Hinge loss: Taking* $B = 1$, $b = -\varepsilon$, $M = 0$ *and* $U = [0,1]$ *we have*

$$\rho(y) = \sup_{u \in U} \{(y - \varepsilon)u\} = (y - \varepsilon)_+ .$$

To verify this, just note that if $y < \varepsilon$, $u^* = 0$; *otherwise* $u^* = 1$.

6. *Vapnik loss is given by* $(y-\varepsilon)_+ + (-y-\varepsilon)_+$. *We immediately obtain its PLQ representation by taking*

$$B = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \; b = -\begin{bmatrix} \varepsilon \\ \varepsilon \end{bmatrix}, \; M = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \; U = [0,1] \times [0,1]$$

to yield

$$\rho(y) = \sup_{u \in U} \left\{ \left\langle \begin{bmatrix} y - \varepsilon \\ -y - \varepsilon \end{bmatrix}, u \right\rangle \right\} = (y-\varepsilon)_+ + (-y-\varepsilon)_+ .$$

The Vapnik penalty is shown in the middle right panel of Figure 1.

7. *Soft hinge loss function (Chu et al., 2001). Combining ideas from examples 4 and 5, we can construct a 'soft' hinge loss; that is, the function*

$$\rho(y) = \begin{cases} 0 & \text{if} \quad y < \varepsilon \\ \frac{1}{2}(y-\varepsilon)^2 & \text{if} \quad \varepsilon < y < \varepsilon + \kappa \\ \kappa(y-\varepsilon) - \frac{1}{2}(\kappa)^2 & \text{if} \quad \varepsilon + \kappa < y \, . \end{cases}$$

that has a smooth (quadratic) transition rather than a kink at ε : *Taking* $B = 1$, $b = -\varepsilon$, $M = 1$ *and* $U = [0, \kappa]$ *we have*

$$\rho(y) = \sup_{u \in [0,\kappa]} \{(y-\varepsilon)u\} - \tfrac{1}{2}u^2 \, .$$

To verify this function has the explicit representation given above, note that if $y < \varepsilon$, $u^* = 0$; *if* $\varepsilon < y < \kappa + \varepsilon$, *we have* $u^* = (y-\varepsilon)_+$, *and if* $\kappa + \varepsilon < y$, *we have* $u^* = \kappa$.

8. *Soft insensitive loss function (Chu et al., 2001). Using example 7, we can create a symmetric soft insensitive loss function (which one might term the Hubnik) by adding together to soft hinge loss functions:*

$$\begin{aligned} \rho(y) &= \sup_{u \in [0,\kappa]} \{(y-\varepsilon)u\} - \tfrac{1}{2}u^2 + \sup_{u \in [0,\kappa]} \{(-y-\varepsilon)u\} - \tfrac{1}{2}u^2 \\ &= \sup_{u \in [0,\kappa]^2} \left\{ \left\langle \begin{bmatrix} y - \varepsilon \\ -y - \varepsilon \end{bmatrix}, u \right\rangle \right\} - \tfrac{1}{2} u^T \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} u \, . \end{aligned}$$

See bottom bottom right panel of Figure 1.

Note that the affine generalization (Definition 1) is needed to form the elastic net, the Vapnik penalty, and the SILF function, as all of these are sums of simpler QS penalties. These sum constructions are examples of a general calculus which allows the modeler to build up a QS density having a desired structure. This calculus is described in the following remark.

Remark 5 *Let $\rho_1(y)$ and $\rho_2(y)$ be two QS penalties specified by U_i, M_i, b_i, B_i, for $i = 1,2$. Then the sum $\rho(y) = \rho_1(y) + \rho_2(y)$ is also a QS penalty, with*

$$U = U_1 \times U_2,\ M = \begin{bmatrix} M_1 & 0 \\ 0 & M_2 \end{bmatrix},\ b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},\ B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}\ .$$

Notwithstanding the catalogue of scalar QS functions in Remark 4 and the gluing procedure described in Remark 5, the supremum in Definition 1 appears to be a significant roadblock to understanding and designing a QS function having specific properties. However, with some practice the design of QS penalties is not as daunting a task as it first appears. A key tool in understanding the structure of QS functions are Euclidean norm projections onto convex sets.

Theorem 6 (Projection Theorem for Convex Sets) *[Zarantonello (1971)] Let $Q \in \mathbb{R}^{n\times n}$ be symmetric and positive definite and let $C \subset \mathbb{R}$ be non-empty, closed and convex. Then Q defines an inner product on $\mathbb{R}^n$ by $\langle x, t\rangle_Q = x^T Qy$ with associated Euclidean norm $\|x\|_Q = \sqrt{\langle x, x\rangle_Q}$. The projection of a point $y \in \mathbb{R}^n$ onto C in norm $\|\cdot\|_Q$ is the unique point $P_Q(y \mid C)$ solving the least distance problem*

$$\inf_{x\in C} \|y - x\|_Q, \tag{6}$$

and $z = P_Q(y\,|\,C)$ if and only if $z \in C$ and

$$\langle x - z, y - z\rangle_Q \leq 0 \quad \forall\, x \in C\ . \tag{7}$$

Note that the least distance problem (6) is equivalent to the problem

$$\inf_{x\in C} \tfrac{1}{2}\|y - x\|_Q^2\ .$$

In the following lemma we use projections as well as duality theory to provide alternative representations for QS penalties.

Theorem 7 *Let $M \in \mathbb{R}^{n\times n}$ be symmetric and positive semi-definite matrix, let $L \in \mathbb{R}^{n\times k}$ be any matrix satisfying $M = LL^T$ where $k = \mathrm{rank}(M)$, and let $U \subset \mathbb{R}^n$ be a non-empty, closed and convex set that contains the origin. Then the QS function $\rho := \rho(U, M, 0, I; \cdot)$ has the primal representations*

$$\rho(y) = \inf_{s\in\mathbb{R}^k} \left[\tfrac{1}{2}\|s\|_2^2 + \delta^*(y - Ls\,|\,U)\right] = \inf_{s\in\mathbb{R}^k} \left[\tfrac{1}{2}\|s\|_2^2 + \gamma(y - Ls\,|\,U^\circ)\right]\ , \tag{8}$$

where, for any convex set V,

$$\delta^*(z\,|\,V) := \sup_{v\in V} \langle z, v\rangle \quad \textit{and} \quad \gamma(z\,|\,V) := \inf\{t\,|\,t \geq 0,\ z \in tV\}$$

are the support and gauge functionals for V, respectively.

If it is further assumed that $M \in \mathcal{S}^n_{++}$ the set of positive definite matrices, then ρ has the representations

$$\begin{aligned}
\rho(y) &= \inf_{s\in\mathbb{R}^k} \left[\tfrac{1}{2}\|s\|^2_M + \gamma\left(M^{-1}y - s \,\middle|\, M^{-1}U^\circ\right)\right] && (9)\\
&= \tfrac{1}{2}\|P_M\left(M^{-1}y \,|\, U\right)\|^2_M + \gamma\left(M^{-1}y - P_M(M^{-1}y|\,U) \,\middle|\, M^{-1}U^\circ\right) && (10)\\
&= \inf_{s\in\mathbb{R}^k} \left[\tfrac{1}{2}\|s\|^2_{M^{-1}} + \gamma(y - s \,|\, U^\circ)\right] && (11)\\
&= \tfrac{1}{2}\|P_{M^{-1}}(y|\,MU)\|^2_{M^{-1}} + \gamma(y - P_{M^{-1}}(y|\,MU)\,|\,U^\circ) && (12)\\
&= \tfrac{1}{2}y^TM^{-1}y - \inf_{u\in U} \tfrac{1}{2}\|u - M^{-1}y\|^2_M && (13)\\
&= \tfrac{1}{2}\|P_M(M^{-1}y|\,U)\|^2_M + \left\langle M^{-1}y - P_M(M^{-1}y|\,U),\, P_M(M^{-1}y|\,U)\right\rangle_M && (14)\\
&= \tfrac{1}{2}y^TM^{-1}y - \inf_{v\in MU} \tfrac{1}{2}\|v - y\|^2_{M^{-1}} && (15)\\
&= \tfrac{1}{2}\|P_{M^{-1}}(y|\,MU)\|^2_{M^{-1}} + \langle y - P_{M^{-1}}(y|\,MU),\, P_{M^{-1}}(y|\,MU)\rangle_{M^{-1}}\ . && (16)
\end{aligned}$$

In particular, (15) *says $\rho(y) = \frac{1}{2}y^TM^{-1}y$ whenever $y \in MU$. Also note that, by* (8)*, one can replace the gauge functionals in* (9)-(12) *by the support functional of the appropriate set where $M^{-1}U^\circ = (MU)^\circ$.*

The formulas (9)-(16) show how one can build PLQ penalties having a wide range of desirable properties. We now give a short list of a few examples illustrating how to make use of these representations.

Remark 8 (General examples) *In this remark we show how the representations in Lemma 7 can be used to build QS penalties with specific structure. In each example we specify the components U, M, b, and B for the QS function $\rho := \rho(U, M, b, B; \cdot)$.*

1. *Norms. Any norm $\|\cdot\|$ can be represented as a QS function by taking $M = 0$, $B = I$, $b = 0$, $U = \mathbb{B}^\circ$, where $\mathbb{B}$ is the unit ball of the desired norm. Then, by* (8)*, $\rho(y) = \|y\| = \gamma(y\,|\,\mathbb{B})$.*

2. *Gauges and support functions. Let U be any closed convex set containing the origin, and Take $M = 0, B = I, b = 0$. Then, by* (8)*, $\rho(y) = \gamma(y\,|\,U^\circ) = \delta^*(y\,|\,U)$.*

3. *Generalized Huber functions. Take any norm $\|\cdot\|$ having closed unit ball $\mathbb{B}$. Let $M \in \mathcal{S}^n_{++}$, $B = I$, $b = 0$, and $U = \mathbb{B}^\circ$. Then, by the representation* (12)*,*

$$\rho(y) = \tfrac{1}{2}P_{M^{-1}}(y|\,M\mathbb{B}^\circ)^TM^{-1}P_{M^{-1}}(y|\,M\mathbb{B}^\circ) + \|y - P_{M^{-1}}(y|\,M\mathbb{B}^\circ)\|\ .$$

   In particular, for $y \in M\mathbb{B}^\circ$, $\rho(y) = \frac{1}{2}y^TM^{-1}y$.

   If we take $M = I$ and $\|\cdot\| = \kappa^{-1}\|\cdot\|_1$ for $\kappa > 0$ (i.e., $U = \kappa\mathbb{B}_\infty$ and $U^\circ = \kappa^{-1}\mathbb{B}_1$), then ρ is the multivariate Huber function described in item 4 of Remark 4. In this way, Theorem 7 shows how to generalize the essence of the Huber norm to any choice of norm. For example, if we take $U = \kappa\mathbb{B}_M = \{\kappa u \,|\, \|u\|_M \le 1\}$, then, by (14)*,*

$$\rho(y) = \begin{cases} \frac{1}{2}\|y\|^2_{M^{-1}} & ,\text{if } \|y\|_{M^{-1}} \le \kappa \\ \kappa\|y\|_{M^{-1}} - \frac{\kappa^2}{2} & ,\text{if } \|y\|_{M^{-1}} > \kappa\ . \end{cases}$$

4. *Generalized hinge-loss functions. Let $\|\cdot\|$ be a norm with closed unit ball $\mathbb{B}$, let K be a non-empty closed convex cone in $\mathbb{R}^n$, and let $v \in \mathbb{R}^n$. Set $M = 0$, $b = -v$, $B = I$, and $U = -(\mathbb{B}^\circ \cap K^\circ) = \mathbb{B}^\circ \cap (-K)^\circ$. Then, by (Burke, 1987, Section 2),*

$$\rho(y) = \operatorname{dist}(y \,|\, v - K) = \inf_{u \in K} \|y - b + u\| .$$

If we consider the order structure "$\leq_K$" induced on $\mathbb{R}^n$ by

$$y \leq_K v \quad \Longleftrightarrow \quad v - y \in K ,$$

then $\rho(y) = 0$ if and only if $y \leq_K v$. By taking $\|\cdot\| = \|\cdot\|_1$, $K = \mathbb{R}^n_+$ so $(-K)^\circ = K$, and $v = \varepsilon\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones, we recover the multivariate hinge loss function in Remark 4.

5. *Order intervals and Vapnik loss functions. Let $\|\cdot\|$ be a norm with closed unit ball $\mathbb{B}$, let $K \subset \mathbb{R}^n$ be a non-empty symmetric convex cone in the sense that $K^\circ = -K$, and let $w <_K v$, or equivalently, $v - w \in \operatorname{intr}(K)$. Set*

$$U = (\mathbb{B}^\circ \cap K) \times (\mathbb{B}^\circ \cap K^\circ), \quad M = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad b = -\begin{pmatrix} v \\ w \end{pmatrix}, \quad \text{and} \quad B = \begin{bmatrix} I \\ I \end{bmatrix} .$$

Then

$$\rho(y) = \operatorname{dist}(y \,|\, v - K) + \operatorname{dist}(y \,|\, w + K) .$$

Observe that $\rho(y) = 0$ if and only if $w \leq_K y \leq_K v$. The set $\{y \,|\, w \leq_K y \leq_K v\}$ is an "order interval" (Schaefer, 1970). If we take $w = -v$, then $\{y \,|\, -v \leq_K y \leq_K v\}$ is a symmetric neighborhood of the origin. By taking $\|\cdot\| = \|\cdot\|_1$, $K = \mathbb{R}^n_+$, and $v = \varepsilon\mathbf{1}$=-$w$, we recover the multivariate Vapnik loss function in Remark 4. Further examples of symmetric cones are $\mathcal{S}^n_+$ and the Lorentz or ℓ_2 cone (Güler and Hauser, 2002).

The examples given above show that one can also construct generalized versions of the elastic net as well as the soft insensitive loss functions defined in Remark 4. In addition, cone constraints can also be added by using the identity $\delta^*(\cdot \,|\, K^\circ) = \delta(\cdot \,|\, K)$. These examples serve to illustrate the wide variety of penalty functions representable as QS functions. Computationally, one is only limited by the ability to compute projections described in Theorem 7. For more details on computational properties for QS functions, see Aravkin et al. (2013), Section 6.

In order to characterize QS functions as negative logs of density functions, we need to ensure the integrability of said density functions. The function $\rho(y)$ is said to be *coercive* if $\lim_{\|y\| \to \infty} \rho(y) = \infty$, and coercivity turns out to be the key property to ensure integrability. The proof of this fact and the characterization of coercivity for QS functions are the subject of the next two theorems (see Appendix for proofs).

Theorem 9 (QS integrability) *Suppose $\rho(y)$ is a coercive QS penalty. Then the function $\exp[-\rho(y)]$ is integrable on $\operatorname{aff}[\operatorname{dom}(\rho)]$ with respect to the $\dim(\operatorname{aff}[\operatorname{dom}(\rho)])$-dimensional Lebesgue measure.*

Theorem 10 *A QS function ρ is coercive if and only if $[B^{\mathrm{T}}\operatorname{cone}(U)]^\circ = \{0\}$.*

Theorem 10 can be used to show the coercivity of familiar penalties. In particular, note that if $B = I$, then the QS function is coercive if and only if U contains the origin in its interior.

Corollary 11 *The penalties ℓ_2, ℓ_1, elastic net, Vapnik, and Huber are all coercive.*

Proof We show that all of these penalties satisfy the hypothesis of Theorem 10.

ℓ_2: $U = \mathbb{R}$ and $B = 1$, so $\left[B^{\mathrm{T}}\mathrm{cone}(U)\right]^\circ = \mathbb{R}^\circ = \{0\}$.

ℓ_1: $U = [-1,1]$, so $\mathrm{cone}(U) = \mathbb{R}$, and $B = 1$.

Elastic Net: In this case, $\mathrm{cone}(U) = \mathbb{R}^2$ and $B = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]$.

Huber: $U = [-\kappa,\kappa]$, so $\mathrm{cone}(U) = \mathbb{R}$, and $B = 1$.

Vapnik: $U = [0,1]\times[0,1]$, so $\mathrm{cone}(U) = \mathbb{R}^2_+$. $B = \left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]$, so $B^{\mathrm{T}}\mathrm{cone}(U) = \mathbb{R}$.

■

One can also show the coercivity of the above examples using their primal representations. However, our main objective is to pave the way for a modeling framework where multi-dimensional penalties can be constructed from simple building blocks and then solved by a uniform approach using the dual representations alone.

We now define a family of distributions on $\mathbb{R}^n$ by interpreting piecewise linear quadratic functions ρ as negative logs of corresponding densities. Note that the support of the distributions is always contained in $\mathrm{dom}\,\rho$, which is characterized in Theorem 3.

Definition 12 (QS densities) *Let $\rho(U,M,B,b;y)$ be any coercive extended QS penalty on $\mathbb{R}^n$. Define $\mathbf{p}(y)$ to be the following density on $\mathbb{R}^n$:*

$$\mathbf{p}(y) = \begin{cases} c^{-1}\exp\left[-\rho(y)\right] & y \in \mathrm{dom}\,\rho \\ 0 & \text{else,} \end{cases}$$

where

$$c = \left(\int_{y\in\mathrm{dom}\,\rho} \exp\left[-\rho(y)\right] dy\right),$$

and the integral is with respect to the $\dim(\mathrm{dom}(\rho))$*-dimensional Lebesgue measure.*

QS densities are true densities on the affine hull of the domain of ρ. The proof of Theorem 9 can be easily adapted to show that they have moments of all orders.

3. Constructing QS Densities

In this section, we describe how to construct multivariate QS densities with prescribed means and variances. We show how to compute normalization constants to obtain scalar densities, and then extend to multivariate densities using linear transformations. Finally, we show how to obtain the data structures U,M,B,b corresponding to multivariate densities, since these are used by the optimization approach in Section 4.

We make use of the following definitions. Given a sequence of column vectors $\{r_k\} = \{r_1,\ldots,r_N\}$ and matrices $\{\Sigma_k\} = \{\Sigma_1,\ldots,\Sigma_N\}$, we use the notation

$$\mathrm{vec}(\{r_k\}) = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix}, \ \mathrm{diag}(\{\Sigma_k\}) = \begin{bmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \Sigma_N \end{bmatrix}.$$

In definition 12, QS densities are defined over $\mathbb{R}^n$. The moments of these densities depend in a nontrivial way on the choice of parameters b, B, U, M. In practice, we would like to be able to construct these densities to have prescribed means and variances. We show how this can be done using scalar QS random variables as the building blocks. Suppose $y = \text{vec}(\{y_k\})$ is a vector of independent (but not necessarily identical) QS random variables with mean 0 and variance 1. Denote by b_k, B_k, U_k, M_k the specification for the densities of y_k. To obtain the density of y, we need only take

$$\begin{aligned} U &= U_1 \times U_2 \times \cdots \times U_N, \\ M &= \text{diag}(\{M_k\}), \\ B &= \text{diag}(\{B_k\}), \\ b &= \text{vec}(\{b_k\}) \, . \end{aligned}$$

For example, the standard Gaussian distribution is specified by $U = \mathbb{R}^n$, $M = I$, $b = 0$, $B = I$, while the standard ℓ_1-Laplace (Aravkin et al., 2011a) is specified by $U = [-1,1]^n$, $M = 0$, $b = 0$, $B = \sqrt{2}I$.

The random vector $\tilde{y} = Q^{1/2}(y + \mu)$ has mean μ and variance Q. If c is the normalizing constant for the density of y, then $c \det(Q)^{1/2}$ is the normalizing constant for the density of $\tilde{y}$.

Remark 13 *Note that only independence of the building blocks is required in the above result. This allows the flexibility to impose different QS densities on different errors in the model. Such flexibility may be useful for example when combining measurement data from different instruments, where some instruments may occasionally give bad data (with outliers), while others have errors that are modeled well by Gaussian distributions.*

We now show how to construct scalar building blocks with mean 0 and variance 1, that is, how to compute the key normalizing constants for any QS penalty. To this aim, suppose $\rho(y)$ is a scalar QS penalty that is symmetric about 0. We would like to construct a density $\mathbf{p}(y) = \exp\left[-\rho(c_2 y)\right]/c_1$ to be a true density with unit variance, that is,

$$\frac{1}{c_1}\int \exp\left[-\rho(c_2 y)\right] dy = 1 \quad \text{and} \quad \frac{1}{c_1}\int y^2 \exp\left[-\rho(c_2 y)\right] dy = 1, \tag{17}$$

where the integrals are over $\mathbb{R}$. Using u-substitution, these equations become

$$c_1 c_2 = \int \exp\left[-\rho(y)\right] dy \quad \text{and} \quad c_1 c_2^3 = \int y^2 \exp\left[-\rho(y)\right] dy.$$

Solving this system yields

$$\begin{aligned} c_2 &= \sqrt{\int y^2 \exp\left[-\rho(y)\right] dy \Big/ \int \exp\left[-\rho(y)\right] dy}, \\ c_1 &= \frac{1}{c_2}\int \exp\left[-\rho(y)\right] dy \, . \end{aligned}$$

These expressions can be used to obtain the normalizing constants for any particular ρ using simple integrals.

3.1 Huber Density

The scalar density corresponding to the Huber penalty is constructed as follows. Set

$$\mathbf{p_H}(y) = \frac{1}{c_1}\exp[-\rho_H(c_2 y)]\,,$$

where c_1 and c_2 are chosen as in (17). Specifically, we compute

$$\int \exp[-\rho_H(y)]\,dy = 2\exp\left[-\kappa^2/2\right]\frac{1}{\kappa} + \sqrt{2\pi}[2\Phi(\kappa)-1],$$
$$\int y^2\exp[-\rho_H(y)]\,dy = 4\exp\left[-\kappa^2/2\right]\frac{1+\kappa^2}{\kappa^3} + \sqrt{2\pi}[2\Phi(\kappa)-1]\,,$$

where Φ is the standard normal cumulative density function. The constants c_1 and c_2 can now be readily computed.

To obtain the multivariate Huber density with variance Q and mean μ, let $U = [-\kappa,\kappa]^n$, $M = I$, $B = I$ any full rank matrix, and $b = 0$. This gives the desired density:

$$\mathbf{p_H}(y) = \frac{1}{c_1^n \det(Q^{1/2})}\exp\left[-\sup_{u\in U}\left\{\left\langle c_2 Q^{-1/2}(y-\mu),u\right\rangle - \frac{1}{2}u^{\mathrm{T}}u\right\}\right].$$

3.2 Vapnik Density

The scalar density associated with the Vapnik penalty is constructed as follows. Set

$$\mathbf{p_V}(y) = \frac{1}{c_1}\exp\left[-\rho_V(c_2 y)\right]\,,$$

where the normalizing constants c_1 and c_2 can be obtained from

$$\int \exp[-\rho_V(y)]\,dy = 2(\varepsilon+1),$$
$$\int y^2\exp[-\rho_V(y)]\,dy = \frac{2}{3}\varepsilon^3 + 2(2-2\varepsilon+\varepsilon^2),$$

using the results in Section 3. Taking $U = [0,1]^{2n}$, the multivariate Vapnik distribution with mean μ and variance Q is

$$\mathbf{p_V}(y) = \frac{1}{c_1^n \det(Q^{1/2})}\exp\left[-\sup_{u\in U}\left\{\left\langle c_2 B Q^{-1/2}(y-\mu) - \varepsilon\mathbf{1}_{2n},u\right\rangle\right\}\right]\,,$$

where B is block diagonal with each block of the form $B = \left[\begin{smallmatrix}1\\-1\end{smallmatrix}\right]$, and $\mathbf{1}_{2n}$ is a column vector of 1's of length $2n$.

4. Optimization With PLQ Penalties

In the previous sections, QS penalties were characterized using their dual representation and interpreted as negative log likelihoods of true densities. As we have seen, the scope of such densities is extremely broad. Moreover, these densities can easily be constructed to possess specified moment

properties. In this section, we expand on their utility by showing that the resulting estimation problems (4) can be solved with high accuracy using standard techniques from numerical optimization for a large subclass of these penalties. We focus on PLQ penalties for the sake of simplicity in our presentation of an interior point approach to solving these estimation problems. However, the interior point approach applies in much more general settings (Nemirovskii and Nesterov, 1994). Nonetheless, the PLQ case is sufficient to cover all of the examples given in Remark 4 while giving the flavor of how to proceed in the more general cases.

We exploit the dual representation for the class of PLQ penalties (Rockafellar and Wets, 1998) to explicitly construct the Karush-Kuhn-Tucker (KKT) conditions for a wide variety of model problems of the form (4). Working with these systems opens the door to using a wide variety of numerical methods for convex quadratic programming to solve (4).

Let $\rho(U_v, M_v, b_v, B_v; y)$ and $\rho(U_w, M_w, b_w, B_w; y)$ be two PLQ penalties and define

$$V(v;R) := \rho(U_v, M_v, b_v, B_v; R^{-1/2}v) \tag{18}$$

and

$$W(w;Q) := \rho(U_w, M_w, b_w, B_w; Q^{-1/2}w). \tag{19}$$

Then (4) becomes

$$\min_{y\in\mathbb{R}^n} \rho(U, M, b, B; y), \tag{20}$$

where

$$U := U_v \times U_w,\ M := \begin{bmatrix} M_v & 0 \\ 0 & M_w \end{bmatrix},\ b := \begin{pmatrix} b_v - B_v R^{-1/2} z \\ b_w - B_w Q^{-1/2}\mu \end{pmatrix},$$

and

$$B := \begin{bmatrix} B_v R^{-1/2} H \\ B_w Q^{-1/2} G \end{bmatrix}.$$

Moreover, the hypotheses in (1), (2), (4), and (5) imply that the matrix B in (20) is injective. Indeed, $By = 0$ if and only if $B_w Q^{-1/2} G y = 0$, but, since G is nonsingular and B_w is injective, this implies that $y = 0$. That is, $\mathrm{nul}(B) = \{0\}$. Consequently, the objective in (20) takes the form of a PLQ penalty function (5). In particular, if (18) and (19) arise from PLQ densities (definition 12), then the solution to problem (20) is the MAP estimator in the statistical model (1)-(2).

To simplify the notational burden, in the remainder of this section we work with (20) directly and assume that the defining objects in (20) have the dimensions specified in (5);

$$U \in \mathbb{R}^m,\ M \in \mathbb{R}^{m\times m},\ b \in \mathbb{R}^m, \text{ and } B \in \mathbb{R}^{m\times n}.$$

The Lagrangian (Rockafellar and Wets, 1998)[Example 11.47] for problem (20) is given by

$$L(y,u) = b^{\mathrm{T}}u - \frac{1}{2}u^{\mathrm{T}}Mu + u^{\mathrm{T}}By\,.$$

By assumption U is polyhedral, and so can be specified to take the form

$$U = \{u : A^{\mathrm{T}}u \le a\}\,, \tag{21}$$

where $A \in \mathbb{R}^{m\times\ell}$. Using this representation for U, the optimality conditions for (20) (Rockafellar, 1970; Rockafellar and Wets, 1998) are

$$\begin{aligned} 0 &= B^{\mathrm{T}}u, \\ 0 &= b + By - Mu - Aq, \\ 0 &= A^{\mathrm{T}}u + s - a, \\ 0 &= q_i s_i \ , \ i = 1,\ldots,\ell \ , \ q,s \geq 0 \ , \end{aligned} \tag{22}$$

where the non-negative slack variable s is defined by the third equation in (22). The non-negativity of s implies that $u \in U$. The equations $0 = q_i s_i \ , \ i = 1,\ldots,\ell$ in (22) are known as the complementarity conditions. By convexity, solving the problem (20) is equivalent to satisfying (22). There is a vast optimization literature on working directly with the KKT system. In particular, interior point (IP) methods (Kojima et al., 1991; Nemirovskii and Nesterov, 1994; Wright, 1997) can be employed. In the Kalman filtering/smoothing application, IP methods have been used to solve the KKT system (22) in a numerically stable and efficient manner, see, for example, Aravkin et al. (2011b). Remarkably, the IP approach studied by Aravkin et al. (2011b) generalizes to the entire PLQ class. For Kalman filtering and smoothing, the computational efficiency is also preserved (see Section 6). Here, we show the general development for the entire PLQ class using standard techniques from the IP literature, see, for example, Kojima et al. (1991).

Let U, M, b, B, and A be as defined in (5) and (21), and let $\tau \in (0, +\infty]$. We define the τ *slice of the strict feasibility region for* (22) to be the set

$$\mathcal{F}_+(\tau) = \left\{ (s,q,u,y) \;\middle|\; \begin{array}{c} 0 < s,\ 0 < q,\ s^{\mathrm{T}}q \leq \tau, \text{ and} \\ (s,q,u,y) \text{ satisfy the affine equations in (22)} \end{array} \right\},$$

and the *central path for* (22) to be the set

$$\mathcal{C} := \left\{ (s,q,u,y) \;\middle|\; \begin{array}{c} 0 < s,\ 0 < q,\ \gamma = q_i s_i \ \ i = 1,\ldots,\ell, \text{ and} \\ (s,q,u,y) \text{ satisfy the affine equations in (22)} \end{array} \right\}.$$

For simplicity, we define $\mathcal{F}_+ := \mathcal{F}_+(+\infty)$. The basic strategy of a primal-dual IP method is to follow the central path to a solution of (22) as $\gamma \downarrow 0$ by applying a predictor-corrector damped Newton method to the function mapping $\mathbb{R}^\ell \times \mathbb{R}^\ell \times \mathbb{R}^m \times \mathbb{R}^n$ to itself given by

$$F_\gamma(s,q,u,y) = \begin{bmatrix} s + A^{\mathrm{T}}u - a \\ D(q)D(s)\mathbf{1} - \gamma\mathbf{1} \\ By - Mu - Aq + b \\ B^{\mathrm{T}}u \end{bmatrix},$$

where $D(q)$ and $D(s)$ are diagonal matrices with vectors q, s on the diagonal.

Theorem 14 *Let U, M, b, B, and A be as defined in* (5) *and* (21). *Given $\tau > 0$, let $\mathcal{F}_+$, $\mathcal{F}_+(\tau)$, and $\mathcal{C}$ be as defined above. If*

$$\mathcal{F}_+ \neq \emptyset \quad \textit{and} \quad \mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}, \tag{23}$$

then the following statements hold.

(i) $F_\gamma^{(1)}(s,q,u,y)$ is invertible for all $(s,q,u,y) \in \mathcal{F}_+$.

(ii) *Define* $\widehat{\mathcal{F}}_+ = \{(s,q) \,|\, \exists (u,y) \in \mathbb{R}^m \times \mathbb{R}^n \text{ s.t. } (s,q,u,y) \in \mathcal{F}_+ \}$. *Then for each* $(s,q) \in \widehat{\mathcal{F}}_+$ *there exists a unique* $(u,y) \in \mathbb{R}^m \times \mathbb{R}^n$ *such that* $(s,q,u,y) \in \mathcal{F}_+$.

(iii) *The set* $\mathcal{F}_+(\tau)$ *is bounded for every* $\tau > 0$.

(iv) *For every* $g \in \mathbb{R}^{\ell}_{++}$, *there is a unique* $(s,q,u,y) \in \mathcal{F}_+$ *such that* $g = (s_1 q_1, s_2 q_2, \ldots, s_\ell q_\ell)^{\mathrm{T}}$.

(v) *For every* $\gamma > 0$, *there is a unique solution* $[s(\gamma), q(\gamma), u(\gamma), y(\gamma)]$ *to the equation* $F_\gamma(s,q,u,y) = 0$. *Moreover, these points form a differentiable trajectory in* $\mathbb{R}^{\nu} \times \mathbb{R}^{\nu} \times \mathbb{R}^m \times \mathbb{R}^n$. *In particular, we may write*

$$\mathcal{C} = \{ [s(\gamma), q(\gamma), u(\gamma), y(\gamma)] \,|\, \gamma > 0 \} \ .$$

(vi) *The set of cluster points of the central path as* $\gamma \downarrow 0$ *is non-empty, and every such cluster point is a solution to* (22).

Please see the Appendix for proof. Theorem 14 shows that if the conditions (23) hold, then IP techniques can be applied to solve the problem (20). In all of the applications we consider, the condition $\mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}$ is easily verified. For example, in the setting of (20) with

$$U_v = \{u \,|\, A_v u \leq a_v\} \quad \text{and} \quad U_w = \{u \,|\, A_w u \leq b_w\} \tag{24}$$

this condition reduces to

$$\mathrm{null}(M_v) \cap \mathrm{null}(A_v^{\mathrm{T}}) = \{0\} \quad \text{and} \quad \mathrm{null}(M_w) \cap \mathrm{null}(A_w^{\mathrm{T}}) = \{0\}. \tag{25}$$

Corollary 15 *The densities corresponding to* ℓ_1, ℓ_2, *Huber, and Vapnik penalties all satisfy hypothesis* (25).

Proof We verify that $\mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = 0$ for each of the four penalties. In the ℓ_2 case, M has full rank. For the ℓ_1, Huber, and Vapnik penalties, the respective sets U are bounded, so $U^{\infty} = \{0\}$. ■

On the other hand, the condition $\mathcal{F}_+ \neq \emptyset$ is typically more difficult to verify. We show how this is done for two sample cases from class (4), where the non-emptiness of $\mathcal{F}_+$ is established by constructing an element of this set. Such constructed points are useful for initializing the interior point algorithm.

4.1 ℓ_1-ℓ_2

Suppose $V(v;R) = \left\|R^{-1/2} v\right\|_1$ and $W(w;Q) = \frac{1}{2}\left\|Q^{-1/2} w\right\|_2^2$. In this case

$$U_v = [-\mathbf{1}_m, \mathbf{1}_m],\ M_v = 0_{m\times m},\ b_v = 0_m,\ B_v = I_{m\times m},$$
$$U_w = \mathbb{R}^n,\ M_w = I_{n\times n},\ b_w = 0_n,\ B_w = I_{n\times n},$$

and $R \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are symmetric positive definite covariance matrices. Following the notation of (20) we have

$$U = [-\mathbf{1}, \mathbf{1}] \times \mathbb{R}^n,\ M = \begin{bmatrix} 0_{m\times m} & 0 \\ 0 & I_{n\times n} \end{bmatrix},\ b = \begin{pmatrix} -R^{-1/2} z \\ -Q^{-1/2}\mu \end{pmatrix},\ B = \begin{bmatrix} R^{-1/2} H \\ Q^{-1/2} G \end{bmatrix}.$$

The specification of U in (21) is given by

$$A^{\mathrm{T}} = \begin{bmatrix} I_{m\times m} & 0_{n\times n} \\ -I_{m\times m} & 0_{n\times n} \end{bmatrix} \text{ and } a = \begin{pmatrix} \mathbf{1} \\ -\mathbf{1} \end{pmatrix} .$$

Clearly, the condition $\mathrm{null}(M)\cap\mathrm{null}(A^{\mathrm{T}}) = \{0\}$ in (23) is satisfied. Hence, for Theorem 14 to apply, we need only check that $\mathcal{F}_+ \neq \emptyset$. This is easily established by noting that $(s,q,u,y) \in \mathcal{F}_+$, where

$$u = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ y = G^{-1}\mu,\ s = \begin{pmatrix} \mathbf{1} \\ \mathbf{1} \end{pmatrix},\ q = \begin{pmatrix} \mathbf{1} + [R^{-1/2}(Hy - z)]_+ \\ \mathbf{1} - [R^{-1/2}(Hy - z)]_- \end{pmatrix},$$

where, for $g \in \mathbb{R}^\ell$, g_+ is defined componentwise by $g_{+(i)} = \max\{g_i, 0\}$ and $g_{-(i)} = \min\{g_i, 0\}$.

4.2 Vapnik-Huber

Suppose that $V(v;R)$ and $W(w;Q)$ are as in (18) and (19), respectively, with V a Vapnik penalty and W a Huber penalty:

$$U_v = [0, \mathbf{1}_m] \times [0, \mathbf{1}_m],\ M_v = 0_{2m\times 2m},\ b_v = -\begin{pmatrix} \varepsilon\mathbf{1}_m \\ \varepsilon\mathbf{1}_m \end{pmatrix},\ B_v = \begin{bmatrix} I_{m\times m} \\ -I_{m\times m} \end{bmatrix},$$
$$U_w = [-\kappa\mathbf{1}_n, \kappa\mathbf{1}_n],\ M_w = I_{n\times n},\ b_w = 0_n,\ B_w = I_{n\times n} ,$$

and $R \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ are symmetric positive definite covariance matrices. Following the notation of (20) we have

$$U = ([0, \mathbf{1}_m] \times [0, \mathbf{1}_m]) \times [-\kappa\mathbf{1}_n, \kappa\mathbf{1}_n],\ M = \begin{bmatrix} 0_{2m\times 2m} & 0 \\ 0 & I_{n\times n} \end{bmatrix},$$
$$b = -\begin{pmatrix} \varepsilon\mathbf{1}_m + R^{-1/2}z \\ \varepsilon\mathbf{1}_m - R^{-1/2}z \\ Q^{-1/2}\mu \end{pmatrix},\ B = \begin{bmatrix} R^{-1/2}H \\ -R^{-1/2}H \\ Q^{-1/2}G \end{bmatrix} .$$

The specification of U in (21) is given by

$$A^{\mathrm{T}} = \begin{bmatrix} I_{m\times m} & 0 & 0 \\ -I_{m\times m} & 0 & 0 \\ 0 & I_{m\times m} & 0 \\ 0 & -I_{m\times m} & 0 \\ 0 & 0 & I_{n\times n} \\ 0 & 0 & -I_{n\times n} \end{bmatrix} \text{ and } a = \begin{pmatrix} \mathbf{1}_m \\ 0_m \\ \mathbf{1}_m \\ 0_m \\ \kappa\mathbf{1}_n \\ \kappa\mathbf{1}_n \end{pmatrix} .$$

Since $\mathrm{null}(A^{\mathrm{T}}) = \{0\}$, the condition $\mathrm{null}(M)\cap\mathrm{null}(A^{\mathrm{T}}) = \{0\}$ in (23) is satisfied. Hence, for Theorem 14 to apply, we need only check that $\mathcal{F}_+ \neq \emptyset$. We establish this by constructing an element (s,q,u,y) of $\mathcal{F}_+$. For this, let

$$u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix},\ s = \begin{pmatrix} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{pmatrix},\ q = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \\ q_6 \end{pmatrix},$$

and set

$$y = 0_n,\ u_1 = u_2 = \tfrac{1}{2}\mathbf{1}_\ell,\ u_3 = 0_n,\ s_1 = s_2 = s_3 = s_4 = \tfrac{1}{2}\mathbf{1}_\ell,\ s_5 = s_6 = \kappa\mathbf{1}_n,$$

and

$$\begin{aligned} q_1 &= \mathbf{1}_m - (\varepsilon\mathbf{1}_m + R^{-1/2}z)_-,\ q_2 = \mathbf{1}_m + (\varepsilon\mathbf{1}_m + R^{-1/2}z)_+, \\ q_3 &= \mathbf{1}_m - (\varepsilon\mathbf{1}_m - R^{-1/2}z)_-,\ q_4 = \mathbf{1}_m + (\varepsilon\mathbf{1}_m - R^{-1/2}z)_+, \\ q_5 &= \mathbf{1}_n - (Q^{-1/2}\mu)_-,\ q_6 = \mathbf{1}_n + (Q^{-1/2}\mu)_+ \,. \end{aligned}$$

Then $(s,q,u,y) \in \mathcal{F}_+$.

5. Simple Numerical Examples And Comparisons

Before we proceed to the main application of interest (Kalman smoothing), we present a few simple and interesting problems in the PLQ class. An IP solver that handles the problems discussed in this section is available through `github.com/saravkin/`, along with example files and ADMM implementations. A comprehensive comparison with other methods is not in our scope, but we do compare the IP framework with the Alternating Direction Method of Multipliers (ADMM), see Boyd et al. (2011) for a tutorial reference. We hope that the examples and the code will help readers to develop intuition about these two methods.

We focus on ADMM in particular because these methods enjoy widespread use in machine learning and other applications, due to their versatility and ability to scale to large problems. The fundamental difference between ADMM and IP is that ADMM methods have at best linear convergence, so they cannot reach high accuracy in reasonable time, see [Section 3.2.2] of Boyd et al. (2011). In contrast, IP methods have a superlinear convergence rate. In fact, some variants have 2-step quadratic convergence, see Ye and Anstreicher (1993) and Wright (1997).

In addition to accuracy concerns, IP methods may be preferable to ADMM when

- objective contains complex non-smooth terms, for example, $\|Ax - b\|_1$.
- linear operators within the objective formulations are ill-conditioned.

For formulations with well-conditioned linear operators and simple nonsmooth pieces (such as Lasso), ADMM can easily outperform IP. In these cases ADMM methods can attain moderate accuracy (and good solutions) very quickly, by exploiting partial smoothness and/or simplicity of regularizing functionals. For problems lacking these features, such as general formulations built from (nonsmooth) PLQ penalties and possibly ill-conditioned linear operators, IP can dominate ADMM, reaching the true solution while ADMM struggles.

We present a few simple examples below, either developing the ADMM approach for each, or discussing the difficulties (when applicable). We explain advantages and disadvantages of using IP, and present numerical results. A simple IP solver that handles all of the examples, together with ADMM code used for the comparisons, is available through `github.com/saravkin/`. The Lasso example was taken directly from `http://www.stanford.edu/~boyd/papers/admm/`, and we implemented the other ADMM examples using this code as a template.

5.1 Lasso Problem

Consider the Lasso problem

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1 ,$$

where $A\in\mathbb{R}^{n\times m}$. Assume that $m<n$. In order to develop an ADMM approach, we split the variables and introduce a constraint:

$$\min_{x,z} \frac{1}{2}\|Ax-b\|_2^2+\lambda\|z\|_1 \quad \text{s.t.} \quad x=z . \tag{26}$$

The augmented Lagrangian for (26) is given by

$$\mathcal{L}(x,z,y)=\frac{1}{2}\|Ax-b\|_2^2+\lambda\|z\|_1+\eta y^T(z-x)+\frac{\rho}{2}\|z-x\|_2^2 ,$$

where η is the augmented Lagrangian parameter. The ADMM method now comprises the following iterative updates:

$$\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_x \frac{1}{2}\|Ax-b\|_2^2+\frac{\eta}{2}\|x+y^k-z^k\|_2^2,\\
z^{k+1} &= \operatorname*{argmin}_z \lambda\|z\|_1+\frac{\eta}{2}\|z-x^{k+1}+y^k\|_2^2,\\
y^{k+1} &= y^k+(z^{k+1}-x^{k+1}) .
\end{aligned}$$

Turning our attention to the x-update, note that the gradient is given by

$$A^T(Ax-b)+\eta(x+y^k-z^k)=(A^TA+I)x-A^Tb+\eta(y^k-z^k) .$$

At every iteration, the update requires solving the same positive definite $m\times m$ symmetric system. Forming A^TA+I is $O(nm^2)$ time, and obtaining a Cholesky factorization is $O(m^3)$, but once this is done, every x-update can be obtained in $O(m^2)$ time by doing two back-solves.

The z-update has a closed form solution given by soft thresholding:

$$z^{k+1}=S(x^{k+1}-y^{k+1},\lambda/\eta) ,$$

which is an $O(n)$ operation. The multiplier update is also $O(n)$. Therefore, the complexity per iteration is $O(m^2+n)$, making ADMM a great method for this problem.

In contrast, *each iteration* of IP is dominated by the complexity of forming a dense $m\times m$ system A^TD^kA, where D^k is a diagonal matrix that depends on the iteration. So while both methods require an investment of $O(nm^2)$ to form and $O(m^3)$ to factorize the system, ADMM requires this only at the outset, while IP has to repeat the computation for every iteration. A simple test shows ADMM can find a good answer, with a significant speed advantage already evident for moderate (1000×5000) well-conditioned systems (see Table 1).

5.2 Linear Support Vector Machines

The support vector machine problem can be formulated as the PLQ (see Ferris and Munson, 2003, Section 2.1)

$$\min_{w,\gamma} \frac{1}{2}\|w\|^2+\lambda\rho_+(1-D(Aw-\gamma\mathbf{1})) , \tag{27}$$

Problem: Size	ADMM Iters	ADMM Inner	IP Iters	t_{ADMM} (s)	t_{IP} (s)	ObjDiff
Lasso: 1500×5000	15	—	18	2.0	58.3	0.0025
SVM: 32561×123						
$\text{cond}(A) = 7.7 \times 10^{10}$	653	—	77	41.2	23.9	0.17
H-Lasso: 1000×2000						
ADMM/ADMM						
$\text{cond}(A) = 5.8$	26	100	20	14.1	10.5	0.00006
$\text{cond}(A) = 1330$	27	100	24	40.0	13.0	0.0018
ADMM/L-BFGS						
$\text{cond}(A) = 5.8$	18	—	20	2.8	10.3	1.02
$\text{cond}(A) = 1330$	22	—	24	21.2	13.1	1.24
L1 Lasso: 500×2000						
ADMM/ADMM						
$\text{cond}(A) = 2.2$	104	100	29	57.4	5.9	0.06
$\text{cond}(A) = 1416$	112	100	29	81.4	5.6	0.21

Table 1: For each problem, we list iterations for IP, outer ADMM, cap for inner ADMM iterations (if applicable), total time for both algorithms and objective difference f(xADMM) - f(xIP). This difference is always positive, since in all experiments IP found a lower objective value, and its magnitude is an accuracy heuristic for ADMM, where lower difference means higher accuracy.

where ρ_+ is the hinge loss function, $w^T x = \gamma$ is the hyperplane being sought, $D \in \mathrm{R}^{m \times m}$ is a diagonal matrix with $\{\pm 1\}$ on the diagonals (in accordance to the classification of the training data), and $A \in \mathbb{R}^{m \times k}$ is the observation matrix, where each row gives the features corresponding to observation $i \in \{1, \ldots, m\}$. The ADMM details are similar to the Lasso example, so we omit them here. The interested reader can study the details in the file `linear_svm` available through `github/saravkin`.

The SVM example turned out to be very interesting. We downloaded the 9th Adult example from the SVM library at `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`. The training set has 32561 examples, each with 123 features. When we formed the operator A for problem (27), we found it was very poorly conditioned, with condition number 7.7×10^{10}. It should not surprise the reader that after running for 653 iterations, ADMM is still appreciably far away—its objective value is higher, and in fact the relative norm distance to the (unique) true solution is 10%.

It is interesting to note that in this application, high optimization accuracy does not mean better classification accuracy on the test set—indeed, the (suboptimal) ADMM solution achieves a lower classification error on the test set (18%, vs. 18.75% error for IP). Nonetheless, this is not an advantage of one method over another—one can also stop the IP method early. The point here is that from the optimization perspective, SVM illustrates the advantages of Newton methods over methods with a linear rate.

5.3 Robust Lasso

For the examples in this section, we take $\rho(\cdot)$ to be a robust convex loss, either the 1-norm or the Huber function, and consider the robust Lasso problem

$$\min_x \rho(Ax-b)+\lambda\|x\|_1 \ .$$

First, we develop an ADMM approach that works for both losses, exploiting the simple nature of the regularizer. Then, we develop a second ADMM approach when $\rho(x)$ is the Huber function by exploiting partial smoothness of the objective.

Setting $z = Ax-b$, we obtain the augmented Lagrangian

$$\mathcal{L}(x,z,y) = \rho(z)+\lambda\|x\|_1+\eta y^T(z-Ax+b)+\frac{\rho}{2}\|-z+Ax-b\|_2^2 \ .$$

The ADMM updates for this formulation are

$$\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_x \lambda\|x\|_1+\frac{\eta}{2}\|Ax-y^k-z^k\|_2^2,\\
z^{k+1} &= \operatorname*{argmin}_z \rho(z)+\frac{\eta}{2}\|z+y^k-Ax^{k+1}+b\|_2^2,\\
y^{k+1} &= y^k+(z^{k+1}-Ax^{k+1}+b) \ .
\end{aligned}$$

The z-update can be solved using thresholding, or modified thresholding, in $O(m)$ time when $\rho(\cdot)$ is the Huber loss or 1-norm. Unfortunately, the x-update now requires solving a LASSO problem. This can be done with ADMM (see previous section), but the nested ADMM structure does not perform as well as IP methods, even for well conditioned problems.

When $\rho(\cdot)$ is smooth, such as in the case of the Huber loss, the partial smoothness of the objective can be exploited by setting $x = z$, obtaining

$$\mathcal{L}(x,z,y) = \rho(Ax-b)+\lambda\|z\|_1+\eta y^T(zx)+\frac{\rho}{2}\|x-z\|_2^2 \ .$$

The ADMM updates are:

$$\begin{aligned}
x^{k+1} &= \operatorname*{argmin}_x \rho(Ax-b)+\frac{\eta}{2}\|x-z^k+y^k\|_2^2,\\
z^{k+1} &= \operatorname*{argmin}_z \lambda\|z\|_1+\frac{\eta}{2}\|z+(x^{k+1}+y^k)\|_2^2,\\
y^{k+1} &= y^k+(z^{k+1}-x^{k+1}) \ .
\end{aligned}$$

The problem required for the x-update is smooth, and can be solved by a fast quasi-Newton method, such as L-BFGS. L-BFGS is implemented using only matrix-vector products, and for well-conditioned problems, the ADMM/LBFGS approach has a speed advantage over IP methods. For ill-conditioned problems, L-BFGS has to work harder to achieve high accuracy, and inexact solves may destabilize the overall ADMM approach. IP methods are more consistent (see Table 1).

Just as in the Lasso problem, the IP implementation is dominated by the formation of A^TD^kA at every iteration with complexity $O(mn^2)$. However, a simple change of penalty makes the problem much harder for ADMM, especially when the operator A is ill-conditioned.

We hope that the toy problems, results, and code that we developed in order to write this section have given the reader a better intuition for IP methods. In the next section, we apply IP methods to Kalman smoothing, and show that these methods can be specifically designed to exploit the time series structure and preserve computational efficiency of classical Kalman smoothers.

6. Kalman Smoothing With PLQ Penalties

Consider now a dynamic scenario, where the system state x_k evolves according to the following stochastic discrete-time linear model

$$\begin{aligned} x_1 &= x_0 + w_1, \\ x_k &= G_k x_{k-1} + w_k, \qquad k = 2,3,\ldots,N \\ z_k &= H_k x_k + v_k, \qquad k = 1,2,\ldots,N \end{aligned} \tag{28}$$

where x_0 is known, z_k is the m-dimensional subvector of z containing the noisy output samples collected at instant k, G_k and H_k are known matrices. Further, we consider the general case where $\{w_k\}$ and $\{v_k\}$ are mutually independent zero-mean random variables which can come from any of the densities introduced in the previous section, with positive definite covariance matrices denoted by $\{Q_k\}$ and $\{R_k\}$, respectively.

In order to formulate the Kalman smoothing problem over the entire sequence $\{x_k\}$, define

$$\begin{aligned} x &= \text{vec}\{x_1,\cdots,x_N\}\,, & w &= \text{vec}\{w_1,\cdots,w_N\}\,, \\ v &= \text{vec}\{v_1,\cdots,v_N\}\,, & Q &= \text{diag}\{Q_1,\cdots,Q_N\}\,, \\ R &= \text{diag}\{R_1,\cdots,R_N\}\,, & H &= \text{diag}\{H_1,\cdots,H_N\}, \end{aligned}$$

and

$$G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix}.$$

Then model (28) can be written in the form of (1)-(2), that is,

$$\begin{aligned} \mu &= Gx + w, \\ z &= Hx + v\,, \end{aligned} \tag{29}$$

where $x \in \mathbb{R}^{nN}$ is the entire state sequence of interest, w is corresponding process noise, z is the vector of all measurements, v is the measurement noise, and μ is a vector of size nN with the first n-block equal to x_0, the initial state estimate, and the other blocks set to 0. This is precisely the problem (1)-(2) that began our study. The problem (3) becomes the classical Kalman smoothing problem with quadratic penalties. In this case, the objective function can be written

$$\|Gx-\mu\|^2_{Q^{-1}} + \|Hx-z\|^2_{R^{-1}},$$

and the minimizer can be found by taking the gradient and setting it to zero:

$$(G^T Q^{-1} G + H^T R^{-1} H)x = r\,.$$

One can view this as a single step of Newton's method, which converges to the solution because the objective is quadratic. Note also that once the linear system above is formed, it takes only $O(n^3 N)$ operations to solve due to special block tridiagonal structure (for a generic system, it would take $O(n^3 N^3)$ time). In this section, we will show that IP methods can preserve this complexity for much more general penalties on the measurement and process residuals. We first make a brief remark related to the statistical interpretation of PLQ penalties.

Remark 16 *Suppose we decide to move to an outlier robust formulation, where the 1-norm or Huber penalties are used, but the measurement variance is known to be R. Using the statistical interpretation developed in section 3, the statistically correct objective function for the smoother is*

$$\frac{1}{2}\|Gx-\mu\|^2_{Q^{-1}} + \sqrt{2}\|R^{-1}(Hx-z)\|_1 \ .$$

Analogously, the statistically correct objective when measurement error is the Huber penalty with parameter κ *is*

$$\frac{1}{2}\|Gx-\mu\|^2_{Q^{-1}} + c_2\rho(R^{-1/2}(Hx-z)) \ ,$$

where

$$c_2 = \frac{4\exp\left[-\kappa^2/2\right]\frac{1+\kappa^2}{\kappa^3} + \sqrt{2\pi}[2\Phi(\kappa)-1]}{2\exp\left[-\kappa^2/2\right]\frac{1}{\kappa} + \sqrt{2\pi}[2\Phi(\kappa)-1]}.$$

The normalization constant comes from the results in Section 3.1, and ensures that the weighting between process and measurement terms is still consistent with the situation regardless of which shapes are used for the process and measurement penalties. This is one application of the statistical interpretation.

Next, we show that when the penalties used on the process residual $Gx-w$ and measurement residual $Hx-z$ arise from general PLQ densities, the general Kalman smoothing problem takes the form (20), studied in the previous section. The details are given in the following remark.

Remark 17 *Suppose that the noises w and v in the model* (29) *are PLQ densities with means* 0*, variances Q and R (see Def. 12). Then, for suitable* U_w, M_w, b_w, B_w *and* U_v, M_v, b_v, B_v *and corresponding* ρ_w *and* ρ_v *we have*

$$\begin{aligned}\mathbf{p}(w) &\propto \exp\left[-\rho\left(U_w, M_w, b_w, B_w; Q^{-1/2}w\right)\right], \\ \mathbf{p}(v) &\propto \exp\left[-\rho(U_v, M_v, b_v, B_v; R^{-1/2}v)\right]\end{aligned}$$

while the MAP estimator of x in the model (29) *is*

$$\underset{x\in\mathbb{R}^{nN}}{\operatorname{argmin}} \left\{ \begin{array}{l} \rho\left[U_w, M_w, b_w, B_w; Q^{-1/2}(Gx-\mu)\right] \\ +\rho\left[U_v, M_v, b_v, B_v; R^{-1/2}(Hx-z)\right] \end{array} \right\}. \tag{30}$$

If U_w and U_v are given as in (24), then the system (22) decomposes as

$$\begin{array}{rcll} 0 &=& A_w^{\mathrm{T}}u_w + s_w - a_w \ ; & 0 = A_v^{\mathrm{T}}u_v + s_v - a_v, \\ 0 &=& s_w^{\mathrm{T}}q_w \ ; & 0 = s_v^{\mathrm{T}}q_v, \\ 0 &=& \tilde{b}_w + B_wQ^{-1/2}Gd - M_wu_w - A_wq_w, & \\ 0 &=& \tilde{b}_v - B_vR^{-1/2}Hd - M_vu_v - A_vq_v, & \\ 0 &=& G^{\mathrm{T}}Q^{-\mathrm{T}/2}B_w^{\mathrm{T}}u_w - H^{\mathrm{T}}R^{-\mathrm{T}/2}B_v^{\mathrm{T}}u_v, & \\ 0 &\leq& s_w, s_v, q_w, q_v. & \end{array} \tag{31}$$

See the Appendix and (Aravkin, 2010) for details on deriving the KKT system. By further exploiting the decomposition shown in (28), we obtain the following theorem.

Theorem 18 (PLQ Kalman smoother theorem) *Suppose that all w_k and v_k in the Kalman smoothing model (28) come from PLQ densities that satisfy*

$$\text{null}(M_k^w) \cap \text{null}((A_k^w)^{\mathrm{T}}) = \{0\}\ , \text{null}(M_k^v) \cap \text{null}((A_k^v)^{\mathrm{T}}) = \{0\}\ ,\ \forall k\ ,$$

that is, their corresponding penalties are finite-valued. Suppose further that the corresponding set $\mathcal{F}_+$ from Theorem 14 is nonempty. Then (30) *can be solved using an IP method, with computational complexity $O[N(n^3+m^3+l)]$, where l is the largest column dimension of the matrices $\{A_k^{\mathrm{v}}\}$ and $\{A_k^w\}$.*

Note that the first part of this theorem, the solvability of the problem using IP methods, already follows from Theorem 14. The main contribution of the result in the dynamical system context is the computational complexity. The proof is presented in the Appendix and shows that IP methods for solving (30) preserve the key block tridiagonal structure of the standard smoother. If the number of IP iterations is fixed ($10-20$ are typically used in practice), general smoothing estimates can thus be computed in $O[N(n^3+m^3+l)]$ time. Notice also that the number of required operations scales linearly with l, which represents the complexity of the PLQ density encoding.

7. Numerical Example

In this section, we illustrate the modeling capabilities and computational efficiency of PLQ Kalman smoothers on simulated and real data.

7.1 Simulated Data

We use a simulated example to test the computational scheme described in the previous section. We consider the following function

$$f(t) = \exp[\sin(8t)]$$

taken from Dinuzzo et al. (2007). Our aim is to reconstruct f starting from 2000 noisy samples collected uniformly over the unit interval. The measurement noise v_k was generated using a mixture of two Gaussian densities, with $p = 0.1$ denoting the fraction from each Gaussian; that is,

$$v_k \sim (1-p)\mathbf{N}(0, 0.25) + p\mathbf{N}(0, 25),$$

Data are displayed as dots in Figure 2. Note that the purpose of the second component of the Gaussian mixture is to simulate outliers in the output data and that all the measurements exceeding vertical axis limits are plotted on upper and lower axis limits (4 and -2) to improve readability.

The initial condition $f(0) = 1$ is assumed to be known, while the difference of the unknown function from the initial condition (i.e., $f(\cdot) - 1$) is modeled as a Gaussian process given by an integrated Wiener process. This model captures the Bayesian interpretation of cubic smoothing splines (Wahba, 1990), and admits a two-dimensional state space representation where the first component of $x(t)$, which models $f(\cdot) - 1$, corresponds to the integral of the second state component, modelled as Brownian motion. To be more specific, letting $\Delta t = 1/2000$, the sampled version of the state space model (Jazwinski, 1970; Oksendal, 2005) is defined by

$$G_k = \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix}, \qquad k = 2, 3, \ldots, 2000$$
$$H_k = \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad k = 1, 2, \ldots, 2000$$

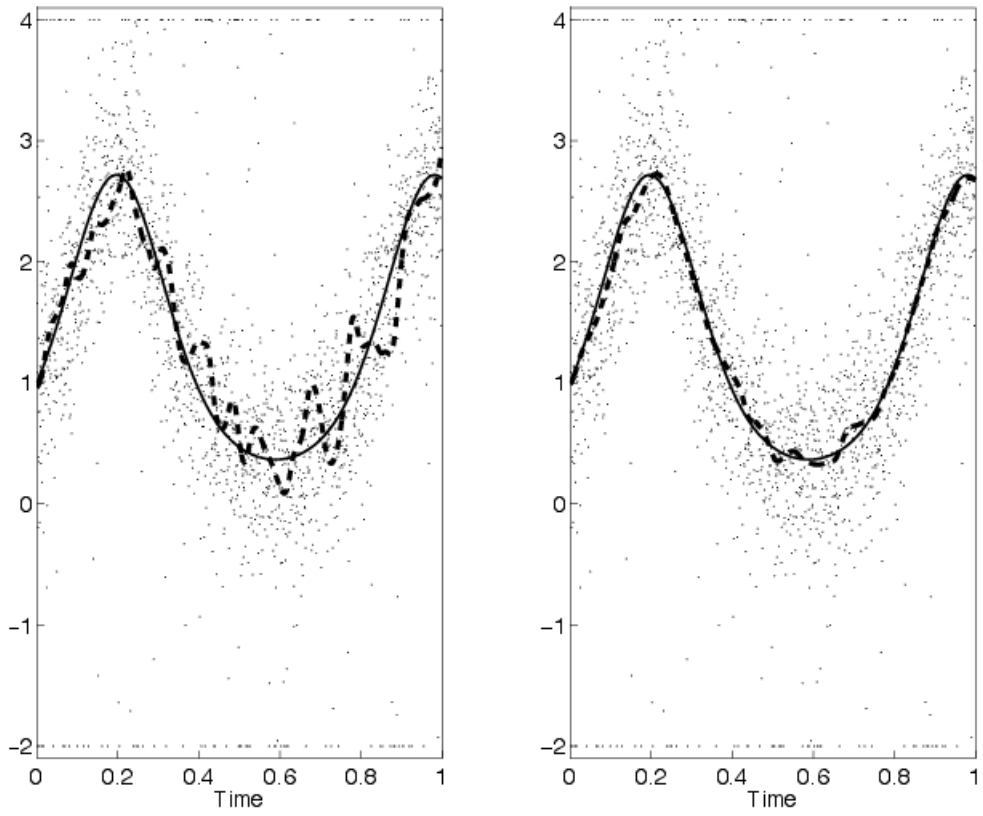

Figure 2: Simulation: measurements $(\cdot)$ with outliers plotted on axis limits (4 and -2), true function (continuous line), smoothed estimate using either the quadratic loss (dashed line, left panel) or the Vapnik's ε-insensitive loss (dashed line, right panel)

with the autocovariance of w_k given by

$$Q_k = \lambda^2 \begin{bmatrix} \Delta t & \frac{\Delta t^2}{2} \\ \frac{\Delta t^2}{2} & \frac{\Delta t^3}{3} \end{bmatrix}, \qquad k = 1, 2, \ldots, 2000\,,$$

where λ^2 is an unknown scale factor to be estimated from the data.

We compare the performance of two Kalman smoothers. The first (classical) estimator uses a quadratic loss function to describe the negative log of the measurement noise density and contains only λ^2 as unknown parameter. The second estimator is a Vapnik smoother relying on the ε-insensitive loss, and so depends on two unknown parameters: λ^2 and ε. In both cases, the unknown parameters are estimated by means of a cross validation strategy where the 2000 measurements are randomly split into a training and a validation set of 1300 and 700 data points, respectively. The Vapnik smoother was implemented by exploiting the efficient computational strategy described in the previous section; Aravkin et al. (2011b) provide specific implementation details. Efficiency is

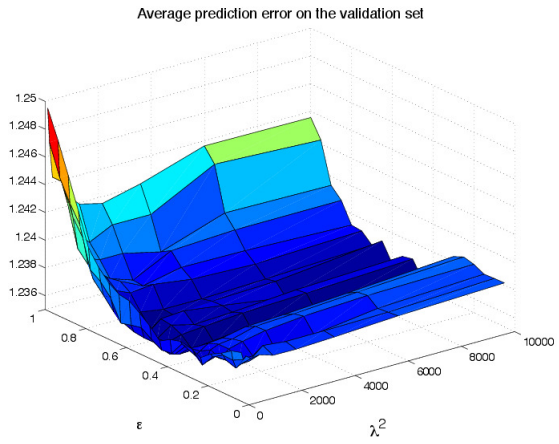

Figure 3: Estimation of the smoothing filter parameters using the Vapnik loss. Average prediction error on the validation data set as a function of the variance process λ^2 and ε.

particularly important here, because of the need for cross-validation. In this way, for each value of λ^2 and ε contained in a 10×20 grid on $[0.01, 10000] \times [0, 1]$, with λ^2 logarithmically spaced, the function estimate was rapidly obtained by the new smoother applied to the training set. Then, the relative average prediction error on the validation set was computed, see Figure 3. The parameters leading to the best prediction were $\lambda^2 = 2.15 \times 10^3$ and $\varepsilon = 0.45$, which give a sparse solution defined by fewer than 400 support vectors. The value of λ^2 for the classical Kalman smoother was then estimated following the same strategy described above. In contrast to the Vapnik penalty, the quadratic loss does not induce any sparsity, so that, in this case, the number of support vectors equals the size of the training set.

The left and right panels of Figure 2 display the function estimate obtained using the quadratic and the Vapnik losses, respectively. It is clear that the estimate obtained using the quadratic penalty is heavily affected by the outliers. In contrast, as expected, the estimate coming from the Vapnik based smoother performs well over the entire time period, and is virtually unaffected by the presence of large outliers.

7.2 Real Industrial Data

Let us now consider real industrial data coming from Syncrude Canada Ltd, also analyzed by Liu et al. (2004). Oil production data is typically a multivariate time series capturing variables such as flow rate, pressure, particle velocity, and other observables. Because the data is proprietary, the exact nature of the variables is not known. The data used by Liu et al. (2004) comprises two anonymized time series variables, called V14 and V36, that have been selected from the process data. Each time series consists of 936 measurements, collected at times $[1, 2, \ldots, 936]$ (see the top panels of Figure 4). Due to the nature of production data, we hypothesize that the temporal profile of the variables is smooth and that the observations contain outliers, as suggested by the fact that some observations differ markedly from their neighbors, especially in the case of V14.
Our aim is to compare the prediction performance of two smoothers that rely on ℓ_2 and ℓ_1 measure-

ment loss functions. For this purpose, we consider 100 Monte Carlo runs. During each run, data are randomly divided into three disjoint sets: training and a validation data sets, both of size 350, and a test set of size 236. We use the same state space model adopted in the previous subsection, with $\Delta t = 1$, and use a non-informative prior to model the initial condition of the system. The regularization parameter γ (equal to the inverse of λ^2 assuming that the noise variance is 1) is chosen using standard cross validation techniques. For each value of γ, logarithmically spaced between 0.1 and 1000 (30 point grid), the smoothers are trained on the training set, and the γ chosen corresponds to the smoother that achieves the best prediction on the validation set. After estimating γ, the variable's profile is reconstructed for the entire time series (at all times $[1,2,\ldots,936]$), using the measurements contained in the union of the training and the validation data sets. Then, the prediction capability of the smoothers is evaluated by computing the 236 relative percentage errors (ratio of residual and observation times 100) in the reconstruction of the test set.

In Figure 4 we display the boxplots of the overall 23600 relative errors stored after the 100 runs for V14 (bottom left panel) and V36 (bottom right panel). One can see that the ℓ_1-Kalman smoother outperforms the classical one, especially in case of V14. This is not surprising, since in this case prediction is more difficult due to the larger numbers of outliers in the time series. In particular, for V14, the average percentage errors are 1.4% and 2.4% while, for V36, they are 1% and 1.2% using ℓ_1 and ℓ_2, respectively.

8. Conclusions

We have presented a new theory for robust and sparse estimation using nonsmooth QS penalties. The QS class captures a variety of existing penalties, including all PLQ penalties, and we have shown how to construct natural generalizations based on general norm and cone geometries, and explored the structure of these functions using Euclidean projections.

Many penalties in the QS class can be interpreted as negative logs of true probability densities. Coercivity (characterized in Theorem 10) is the key necessary condition for this interpretation, as well as a fundamental prerequisite in sparse and robust estimation, since it precludes directions for which the sum of the loss and the regularizer are insensitive to large parameter changes. Thus, coercivity also ensures that the problem is well posed in the machine learning context, that is, the learning machine has enough control over model complexity.

It is straightforward to design new formulations in the QS framework. Starting with the requisite penalty *shape*, one can use results of Section 3 to obtain a standardized density, as well as the data structures required for the optimization problem in Section 4. The statistical interpretation for these methods allows specification of mean and variance parameters for the corresponding model.

In the second part of the paper, we presented a computational approach to solving estimation problems (4) using IP methods. We derived additional conditions that guarantee the successful implementation of IP methods to compute the estimator (4) when x and v come from PLQ densities, and characterized the convergence of IP methods for this class. The key condition for successful execution of IP iterations is for PLQ penalties to be finite valued, which implies non-degeneracy of the corresponding statistical distribution (the support cannot be contained in a lower-dimensional subspace). The statistical interpretation is thus strongly linked to the computational procedure.

We then applied both the statistical framework and the computational approach to the class of state estimation problems in discrete-time dynamic systems, extending the classical formulations to allow dynamics and measurement noise to come from any PLQ densities. We showed that clas-

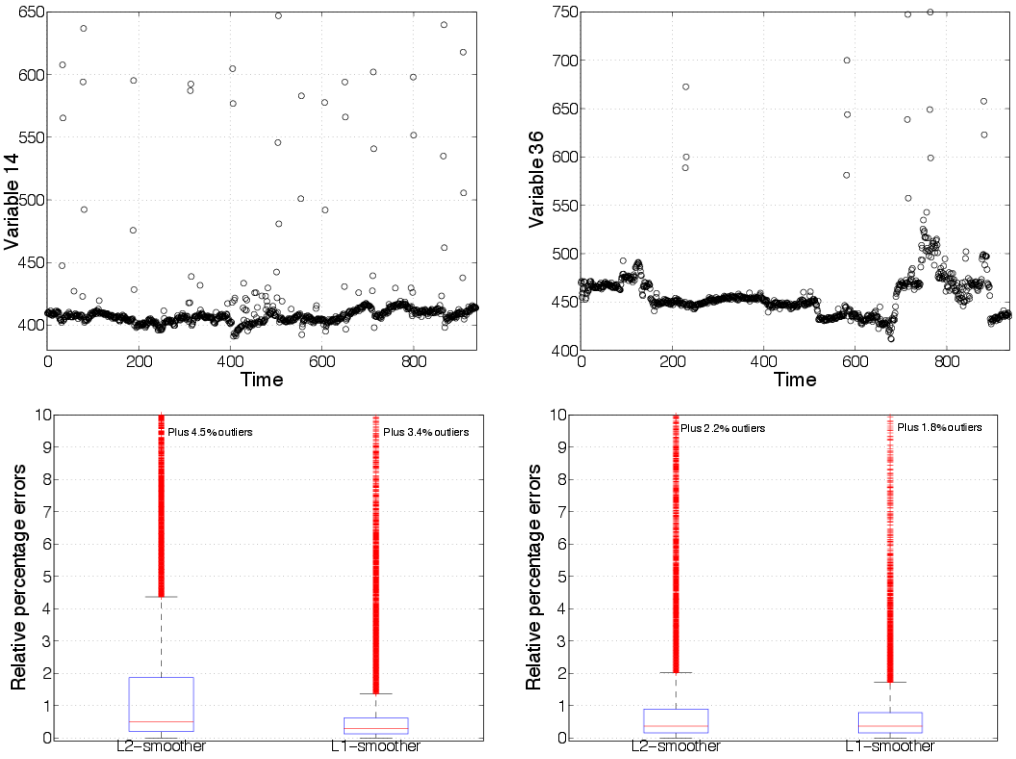

Figure 4: *Left panels*: data set for variable 14 (top) and relative percentage errors in the reconstruction of the test set obtained by Kalman smoothers based on the ℓ_2 and the ℓ_1 loss (bottom). *Right panels*: data set for variable 36 (top) and relative percentage errors in the reconstruction of the test set obtained by Kalman smoothers based on the ℓ_2 and the ℓ_1 loss (bottom).

sical computational efficiency results are preserved when the general IP approach is used for state estimation; specifically, the computational cost of PLQ Kalman smoothing scales linearly with the length of the time series, as in the quadratic case.

While we only considered convex formulations in this paper, the presented approach makes it possible to solve a much broader class of non-convex problems. In particular, if the functions Hx and Gx in (4) are replaced by nonlinear functions $g(x)$ and $h(x)$, the methods in this paper can be used to compute descent directions for the non-convex problem. For an example of this approach, see Aravkin et al. (2011a), which considers non-convex Kalman smoothing problems with nonlinear process and measurement models and solves by using the standard methodology of convex composite optimization (Burke, 1985). At each outer iteration the process and measurement models are linearized around the current iterate, and the descent direction is found by solving a particular subproblem of type (4) using IP methods.

In many contexts, it would be useful to estimate the parameters that define QS penalties; for example the κ in the Huber penalty or the ε in the Vapnik penalty. In the numerical examples presented in this paper, we have relied on cross-validation to accomplish this task. An alternative

method could be to compute the MAP points returned by our estimator for different filter parameters to gain information about the joint posterior of states and parameters. This strategy could help in designing a good proposal density for posterior simulation using, for example, particle smoothing filters (Ristic et al., 2004). We leave a detailed study of this approach to the QS modeling framework for future work.

Appendix A. Proofs

In this appendix, we present proofs for the new results given in the main body of the paper.

A.1 Proof of Theorem 3

Let $\rho(y) = \rho(U,M,I,0;y)$ so that $\rho(U,M,B,b;y) = \rho(b+By)$. Then $\mathrm{dom}\,(\rho(U,M,B,b;\cdot)) = B^{-1}(\mathrm{dom}\,(\rho)-b)$, hence the theorem follows if it can be shown that $\mathrm{bar}(U)+\mathrm{Ran}\,(M) \subset \mathrm{dom}\,(\rho) \subset [U^\infty \cap \mathrm{null}(M)]^\circ$ with equality when $\mathrm{bar}(U)+\mathrm{Ran}\,(M)$ is closed. Observe that if there exists $w \in U^\infty \cap \mathrm{null}(M)$ such that $\langle y, w\rangle > 0$, then trivially $\rho(y) = +\infty$ so $y \notin \mathrm{dom}\,(\rho)$. Consequently, $\mathrm{dom}\,(\rho) \subset [U^\infty \cap \mathrm{null}(M)]^\circ$. Next let $y \in \mathrm{bar}(U)+\mathrm{Ran}\,(M)$, then there is a $v \in \mathrm{bar}(U)$ and w such that $y = v+Mw$. Hence

$$\begin{aligned}
\sup_{u\in U}[\langle u, y\rangle - \tfrac{1}{2}\langle u, Mu\rangle] &= \sup_{u\in U}[\langle u, v+Mw\rangle - \tfrac{1}{2}\langle u, Mu\rangle] \\
&= \sup_{u\in U}[\langle u, v\rangle + \tfrac{1}{2}w^TMw - \tfrac{1}{2}(w-u)^TM(w-u)] \\
&\leq \delta^*(v\,|\,U) + \tfrac{1}{2}w^TMw \;<\; \infty\,.
\end{aligned}$$

Hence $\mathrm{bar}(U)+\mathrm{Ran}\,(M) \subset \mathrm{dom}\,(\rho)$.

If the set $\mathrm{bar}(U)+\mathrm{Ran}\,(M)$ is closed, then so is the set $\mathrm{bar}(U)$. Therefore, by (Rockafellar, 1970, Corollary 14.2.1), $(U^\infty)^\circ = \mathrm{bar}(U)$, and, by (Rockafellar, 1970, Corollary 16.4.2), $[U^\infty \cap \mathrm{null}(M)]^\circ = \mathrm{bar}(U)+\mathrm{Ran}\,(M)$, which proves the result.

In the polyhedral case, $\mathrm{bar}(U)$ is a polyhedral convex set, and the sum of such sets is also a polyhedral convex set (Rockafellar, 1970, Corollary 19.3.2).

A.2 Proof of Theorem 7

To see the first equation in (8) write $\rho(y) = \sup_u\left[\langle y, u\rangle - \left(\frac{1}{2}\|L^Tu\|_2^2 + \delta(u\,|\,U)\right)\right]$, and then apply the calculus of convex conjugate functions (Rockafellar, 1970, Section 16) to find that

$$\left(\tfrac{1}{2}\|L^T\cdot\|_2^2 + \delta(\cdot\,|\,U)\right)^*(y) = \inf_{s\in\mathbb{R}^k}\left[\tfrac{1}{2}\|s\|_2^2 + \delta^*(y-Ls\,|\,U)\right]\,.$$

The second equivalence in (8) follows from (Rockafellar, 1970, Theorem 14.5).

For the remainder, we assume that M is positive definite. In this case it is easily shown that $(MU)^\circ = M^{-1}U^\circ$. Hence, by Theorem 14.5 of Rockafellar (1970), $\gamma(\cdot\,|\,MU) = \delta^*\left(\cdot\,\middle|\,M^{-1}U^\circ\right)$. We use these facts freely throughout the proof.

The formula (9) follows by observing that

$$\tfrac{1}{2}\|s\|_2^2 + \delta^*(y-Ls\,|\,U) = \tfrac{1}{2}\|L^{-T}s\|_M^2 + \delta^*\left(M^{-1}y - L^{-T}s\,|\,MU\right)$$

and then making the substitution $v = L^{-T}s$. To see (10), note that the optimality conditions for (9) are $Ms \in \partial\delta^*\left(M^{-1}y - s \,|\, MU\right)$, or equivalently, $M^{-1}y - s \in N\left(Ms \,|\, MU\right)$, that is, $s \in U$ and

$$\left\langle M^{-1}y - s, u - s\right\rangle_M = \left\langle M^{-1}y - s, M(u-s)\right\rangle \leq 0 \;\forall\, u \in U,$$

which, by (7), tells us that $s = P_M\left(M^{-1}y \,|\, U\right)$. Plugging this into (9) gives (10).

Using the substitution $v = Ls$, the argument showing (11) and (12) differs only slightly from that for (9) and (10) and so is omitted.

The formula (13) follows by completing the square in the M-norm in the definition (5):

$$\begin{aligned}
\langle y, u\rangle - \tfrac{1}{2}\langle u, Mu\rangle &= \left\langle M^{-1}y, u\right\rangle_M - \tfrac{1}{2}\langle u, u\rangle_M \\
&= \tfrac{1}{2}y^T M^{-1}y - \tfrac{1}{2}[\left\langle M^{-1}y, M^{-1}y\right\rangle_M - 2\left\langle M^{-1}y, u\right\rangle_M + \langle u, u\rangle_M] \\
&= \tfrac{1}{2}y^T M^{-1}y - \tfrac{1}{2}\|M^{-1}y - u\|_M^2 \,.
\end{aligned}$$

The result as well as (14) now follow from Theorem 6. Both (15) and (16) follow similary by completing the square in the M^{-1}-norm.

A.3 Proof of Theorem 9

First we will show that if ρ is convex coercive, then for any $\bar{x} \in \operatorname{argmin} f \neq \emptyset$, there exist constants R and $K > 0$ such that

$$\rho(x) \geq \rho(\bar{x}) + K\|x - \bar{x}\| \quad \forall\, x \notin R\mathbb{B}\,. \tag{32}$$

Without loss of generality, we can assume that $0 = \rho(0) = \inf\rho$. Otherwise, replace $\rho(x)$ by $\hat{\rho}(x) = \rho(x + \bar{x}) - \rho(\bar{x})$, where $\bar{x}$ is any global minimizer of ρ.

Let $\alpha > 0$. Since ρ is coercive, there exists R such that $\operatorname{lev}_\rho(\alpha) \subset R\mathbb{B}$. We will show that $\frac{\alpha}{R}\|x\| \leq \rho(x)$ for all $x \notin R\mathbb{B}$.

Indeed, for all $x \neq 0$, we have $\rho(\frac{R}{\|x\|}x) \geq \alpha$. Therefore, if $x \notin R\mathbb{B}$, then $0 < \frac{R}{\|x\|} < 1$, and we have

$$\frac{\alpha}{R}\|x\| \leq \frac{\|x\|}{R}\rho\left(\frac{R}{\|x\|}x\right) \leq \frac{\|x\|}{R}\frac{R}{\|x\|}\rho(x) = \rho(x).$$

Then by (32),

$$\begin{aligned}
\int \exp(-\rho(x))dx &= \int_{\bar{x}+R\mathbb{B}} \exp(-\rho(x))dx + \int_{\|x-\bar{x}\|>R} \exp(-\rho(x))dx \\
&\leq C_1 + C_2 \int_{\|x-\bar{x}\|>R} \exp(-K\|x-\bar{x}\|)dx < \infty\,.
\end{aligned}$$

A.4 Proof of Theorem 10

First observe that $B^{-1}[\operatorname{cone}(U)]^\circ = [B^{\mathrm{T}}\operatorname{cone}(U)]^\circ$ by Corollary 16.3.2 of Rockafellar (1970).

Suppose that $\hat{y} \in B^{-1}[\operatorname{cone}(U)]^\circ$, and $\hat{y} \neq 0$. Then $B\hat{y} \in \operatorname{cone}(U)$, and $B\hat{y} \neq 0$ since B is injective, and we have

$$\begin{aligned}
\rho(t\hat{y}) &= \sup_{u\in U}\langle b + tB\hat{y}, u\rangle - \tfrac{1}{2}u^{\mathrm{T}}Mu \\
&= \sup_{u\in U}\langle b, u\rangle - \tfrac{1}{2}u^{\mathrm{T}}Mu + t\langle B\hat{y}, u\rangle \\
&\leq \sup_{u\in U}\langle b, u\rangle - \tfrac{1}{2}u^{\mathrm{T}}Mu \\
&\leq \rho(U, M, 0, I; b),
\end{aligned}$$

so $\rho(t\hat{y})$ stays bounded even as $t \to \infty$, and so ρ cannot be coercive.

Conversely, suppose that ρ is not coercive. Then we can find a sequence $\{y_k\}$ with $\|y_k\| > k$ and a constant P so that $\rho(y_k) \leq P$ for all $k > 0$. Without loss of generality, we may assume that $\frac{y_k}{\|y_k\|} \to \bar{y}$.

Then by definition of ρ, we have for all $u \in U$

$$\begin{array}{l} \langle b + By_k, u\rangle - \frac{1}{2}u^{\mathrm{T}}Mu \leq P, \\ \langle b + By_k, u\rangle \leq P + \frac{1}{2}u^{\mathrm{T}}Mu, \\ \langle \frac{b+By_k}{\|y_k\|}, u\rangle \leq \frac{P}{\|y_k\|} + \frac{1}{2\|y_k\|}u^{\mathrm{T}}Mu. \end{array}$$

Note that $\bar{y} \neq 0$, so $B\bar{y} \neq 0$. When we take the limit as $k \to \infty$, we get $\langle B\bar{y}, u\rangle \leq 0$. From this inequality we see that $B\bar{y} \in [\mathrm{cone}(U)]^\circ$, and so $\bar{y} \in B^{-1}[\mathrm{cone}(U)]^\circ$.

A.5 Proof of Theorem 14

Proof (i) Using standard elementary row operations, reduce the matrix

$$F_\gamma^{(1)} := \begin{bmatrix} I & 0 & A^{\mathrm{T}} & 0 \\ D(q) & D(s) & 0 & 0 \\ 0 & -A & -M & B \\ 0 & 0 & B^{\mathrm{T}} & 0 \end{bmatrix}$$

to

$$\begin{bmatrix} I & 0 & A^{\mathrm{T}} & 0 \\ 0 & D(s) & -D(q)A^{\mathrm{T}} & 0 \\ 0 & 0 & -T & B \\ 0 & 0 & B^{\mathrm{T}} & 0 \end{bmatrix},$$

where $T = M + AD(q)D(s)^{-1}A^{\mathrm{T}}$. The matrix T is invertible since $\mathrm{null}(M) \cap \mathrm{null}(C^{\mathrm{T}}) = \{0\}$. Hence, we can further reduce this matrix to the block upper triangular form

$$\begin{bmatrix} I & 0 & A^{\mathrm{T}} & 0 \\ 0 & D(s) & -D(q)C^{\mathrm{T}} & 0 \\ 0 & 0 & -T & B \\ 0 & 0 & 0 & -B^{\mathrm{T}}T^{-1}B \end{bmatrix}.$$

Since B is injective, the matrix $B^{\mathrm{T}}T^{-1}B$ is also invertible. Hence this final block upper triangular is invertible proving Part (i).

(ii) Let $(s,q) \in \widehat{\mathcal{F}}_+$ and choose (u_i, y_i) so that $(s,q,u_i,y_i) \in \mathcal{F}_+$ for $i = 1,2$. Set $u := u_1 - u_2$ and $y := y_1 - y_2$. Then, by definition,

$$0 = A^{\mathrm{T}}u,\ 0 = By - Mu, \text{ and } 0 = B^{\mathrm{T}}u\ . \tag{33}$$

Multiplying the second of these equations on the left by u and using the third as well as the positive semi-definiteness of M, we find that $Mu = 0$. Hence, $u \in \mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}$, and so $By = 0$. But then $y = 0$ as B is injective.

(iii) Let $(\hat{s},\hat{q},\hat{u},\hat{y}) \in \mathcal{F}_+$ and $(s,q,u,y) \in \mathcal{F}_+(\tau)$. Then, by (22),

$$\begin{aligned}(s-\hat{s})^{\mathrm{T}}(q-\hat{q}) &= [(a-A^{\mathrm{T}}u)-(a-A^{\mathrm{T}}\hat{u})]^{\mathrm{T}}(q-\hat{q}) \\ &= (\hat{u}-u)^{\mathrm{T}}(Aq-A\hat{q}) \\ &= (\hat{u}-u)^{\mathrm{T}}[(b+By-Mu)-(b+B\hat{b}-M\hat{u})] \\ &= (\hat{u}-u)^{\mathrm{T}}M(\hat{u}-u) \\ &\geq 0.\end{aligned}$$

Hence,

$$\tau+\hat{s}^{\mathrm{T}}\hat{q} \geq s^{\mathrm{T}}y+\hat{s}^{\mathrm{T}}\hat{q} \geq s^{\mathrm{T}}\hat{y}+y^{\mathrm{T}}\hat{s} \geq \xi\,\|(s,q)\|_1\,,$$

where $\xi = \min\{\hat{s}_i,\ \hat{q}_i \mid i=1,\ldots,\ell\} > 0$. Therefore, the set

$$\widehat{\mathcal{F}}_+(\tau) = \{(s,q) \mid (s,q,u,y) \in \mathcal{F}_+(\tau)\}$$

is bounded. Now suppose the set $\mathcal{F}_+(\tau)$ is not bounded. Then there exists a sequence $\{(s_\nu,q_\nu,u_\nu,y_\nu)\} \subset \mathcal{F}_+(\tau)$ such that $\|(s_\nu,q_\nu,u_\nu,y_\nu)\| \uparrow +\infty$. Since $\widehat{\mathcal{F}}_+(\tau)$ is bounded, we can assume that $\|(u_\nu,y_\nu)\| \uparrow +\infty$ while $\|(s_\nu,q_\nu)\|$ remains bounded. With no loss in generality, we may assume that there exits $(u,y) \neq (0,0)$ such that $(u_\nu,y_\nu)/\|(u_\nu,y_\nu)\| \to (u,y)$. By dividing (22) by $\|(u_\nu,y_\nu)\|$ and taking the limit, we find that (33) holds. But then, as in (33), $(u,y) = (0,0)$. This contradiction yields the result.

(iv) We first show existence. This follows from a standard continuation argument. Let $(\hat{s},\hat{q},\hat{u},\hat{y}) \in \mathcal{F}_+$ and $v \in \mathbb{R}^\ell_{++}$. Define

$$F(s,q,u,y,t) = \begin{bmatrix} s+A^{\mathrm{T}}u-a \\ D(q)D(s)\mathbf{1}-[(1-t)\hat{v}+tv] \\ By-Mu-Aq \\ B^{\mathrm{T}}u+b \end{bmatrix},$$

where $\hat{g} := (\hat{s}_1\hat{y}_1,\ldots,\hat{s}_\ell\hat{y}_\ell)^{\mathrm{T}}$. Note that

$$F(\hat{s},\hat{q},\hat{u},\hat{y},0) = 0 \text{ and, by Part (i), } \nabla_{(s,q,u,y)}F(\hat{s},\hat{q},\hat{u},\hat{y},0)^{-1} \text{ exists.}$$

The Implicit Function Theorem implies that there is a $\tilde{t} > 0$ and a differentiable mapping $t \mapsto (s(t),q(t),u(t),y(t))$ on $[0,\tilde{t})$ such that

$$F[s(t),q(t),u(t),y(t),t] = 0 \text{ on } [0,\tilde{t}).$$

Let $\bar{t} > 0$ be the largest such $\tilde{t}$ on $[0,1]$. Since

$$\{[s(t),q(t),u(t),y(t)] \mid t \in [0,\bar{t})\} \subset \mathcal{F}_+(\bar{\tau}),$$

where $\bar{\tau} = \max\{\mathbf{1}^{\mathrm{T}}\hat{g},\mathbf{1}^{\mathrm{T}}g\}$, Part (iii) implies that there is a sequence $t_i \to \bar{t}$ and a point $(\bar{s},\bar{q},\bar{u},\bar{y})$ such that $[s(t_i),q(t_i),u(t_i),y(t_i)] \to (\bar{s},\bar{q},\bar{u},\bar{y})$. By continuity $F(\bar{s},\bar{q},\bar{u},\bar{y},\bar{t}) = 0$. If $\bar{t} = 1$, we are done; otherwise, apply the Implicit Function Theorem again at $(\bar{s},\bar{q},\bar{u},\bar{y},\bar{t})$ to obtain a contradiction to the maximality of $\bar{t}$.

We now show uniqueness. By Part (ii), we need only establish the uniqueness of (s,q). Let $(s^\nu,q^\nu) \in \widehat{\mathcal{F}}_+$ be such that $g = (s_{j(1)}q_{j(1)},s_{j(2)}q_{j(2)},\ldots,s_{j(\ell)}q_{j(\ell)})^{\mathrm{T}}$, where $s_{j(i)}$ denotes the ith element of s_j, and $j = 1,2$. As in Part (iii), we have $(s_1-s_2)^{\mathrm{T}}(q_1-q_2) = (u_1-u_2)^{\mathrm{T}}M((u_1-u_2) \geq 0$,

and, for each $i = 1,\ldots,\ell$, $s_{1(i)}q_{1(i)} = s_{2(i)}q_{2(i)} = g_i > 0$. If $(s_1,q_1) \neq (s_2,q_2)$, then, for some $i \in \{1,\ldots,\ell\}$, $(s_{1(i)} - s_{2(i)})(q_{1(i)} - q_{2(i)}) \geq 0$ and either $s_{1(i)} \neq s_{2(i)}$ or $q_{1(i)} \neq q_{2(i)}$. If $s_{1(i)} > s_{2(i)}$, then $q_{1(i)} \geq q_{2(i)} > 0$ so that $g_i = s_{1(i)}q_{1(i)} > s_{2(i)}q_{2(i)} = g_i$, a contradiction. So with out loss in generality (by exchanging (s_1,q_1) with (s_2,q_2) if necessary), we must have $q_{1(i)} > q_{2(i)}$. But then $s_{1(i)} \geq s_{2(i)} > 0$, so that again $g_i = s_{1(i)}q_{1(i)} > s_{2(i)}q_{2(i)} = g_i$, and again a contradiction. Therefore, (s,q) is unique.
(v) Apply Part (iv) to get a point on the central path and then use the continuation argument to trace out the central path. The differentiability follows from the implicit function theorem.
(vi) Part (iii) allows us to apply a standard compactness argument to get the existence of cluster points and the continuity of $F_\gamma(s,q,u,y)$ in all of its arguments including γ implies that all of these cluster points solve (22). ■

A.6 Details for Remark 17

The Lagrangian for (30) for feasible (x,u_w,u_v) is

$$L(x,u_w,u_v) = \left\langle \begin{bmatrix} \tilde{b}_w \\ \tilde{b}_v \end{bmatrix}, \begin{bmatrix} u_w \\ u_v \end{bmatrix} \right\rangle - \frac{1}{2}\begin{bmatrix} u_w \\ u_v \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} M_w & 0 \\ 0 & M_v \end{bmatrix} \begin{bmatrix} u_w \\ u_v \end{bmatrix} - \left\langle \begin{bmatrix} u_w \\ u_v \end{bmatrix}, \begin{bmatrix} -B_wQ^{-1/2}G \\ B_vR^{-1/2}H \end{bmatrix} x \right\rangle$$

where $\tilde{b}_w = b_w - B_wQ^{-1/2}\tilde{x}_0$ and $\tilde{b}_v = b_v - B_vR^{-1/2}z$. The associated optimality conditions for feasible (x,u_w,u_v) are given by

$$\begin{array}{l} G^{\mathrm{T}}Q^{-\mathrm{T}/2}B_w^{\mathrm{T}}\bar{u}_w - H^{\mathrm{T}}R^{-\mathrm{T}/2}B_v^{\mathrm{T}}\bar{u}_v = 0, \\ \tilde{b}_w - M_w\bar{u}_w + B_wQ^{-1/2}G\bar{x} \in N_{U_w}(\bar{u}_w), \\ \tilde{b}_v - M_v\bar{u}_v - B_vR^{-1/2}H\bar{x} \in N_{U_v}(\bar{u}_v)\ , \end{array} \tag{34}$$

where $N_C(r)$ denotes the normal cone to the set C at the point r (Rockafellar, 1970).

Since U_w and U_v are polyhedral, we can derive explicit representations of the normal cones $N_{U_w}(\bar{u}_w)$ and $N_{U_v}(\bar{u}_v)$. For a polyhedral set $U \subset \mathbb{R}^m$ and any point $\bar{u} \in U$, the normal cone $N_U(\bar{u})$ is polyhedral. Indeed, relative to any representation

$$U = \{u | A^{\mathrm{T}}u \leq a\}$$

and the active index set $I(\bar{u}) := \{i | \langle A_i, \bar{u}\rangle = a_i\}$, where A_i denotes the ith column of A, we have

$$N_U(\bar{u}) = \left\{ q_1A_1 + \cdots + q_mA_m \;\middle|\; q_i \geq 0 \text{ for } i \in I(\bar{u}), \quad q_i = 0 \text{ for } i \notin I(\bar{u}) \right\}. \tag{35}$$

Using (35), Then we may rewrite the optimality conditions (34) more explicitly as

$$\begin{array}{l} G^{\mathrm{T}}Q^{-\mathrm{T}/2}B_w^{\mathrm{T}}\bar{u}_w - H^{\mathrm{T}}R^{-\mathrm{T}/2}B_v^{\mathrm{T}}\bar{u}_v = 0, \\ \tilde{b}_w - M_w\bar{u}_w + B_wQ^{-1/2}G\bar{d} = A_wq_w, \\ \tilde{b}_v - M_v\bar{u}_v - B_vR^{-1/2}H\bar{d} = A_vq_v, \\ \{q_v \geq 0 | q_{v(i)} = 0 \text{ for } i \notin I(\bar{u}_v)\}, \\ \{q^w \geq 0 | q_{w(i)} = 0 \text{ for } i \notin I(\bar{u}_w)\}\ . \end{array}$$

where $q_{v(i)}$ and $q_{w(i)}$ denote the ith elements of q_v and q_w. Define slack variables $s_w \geq 0$ and $s_v \geq 0$ as follows:

$$\begin{aligned} s_w &= a_w - A_w^{\mathrm{T}} u_w, \\ s_v &= a_v - A_v^{\mathrm{T}} u_v. \end{aligned}$$

Note that we know the entries of $q_{w(i)}$ and $q_{v(i)}$ are zero if and only if the corresponding slack variables $s_{v(i)}$ and $s_{w(i)}$ are nonzero, respectively. Then we have $q_w^{\mathrm{T}} s_w = q_v^{\mathrm{T}} s_v = 0$. These equations are known as the complementarity conditions. Together, all of these equations give system (31).

A.7 Proof of Theorem 18

IP methods apply a damped Newton iteration to find the solution of the relaxed KKT system $F_\gamma = 0$, where

$$F_\gamma \begin{pmatrix} s_w \\ s_v \\ q_w \\ q_v \\ u_w \\ u_v \\ x \end{pmatrix} = \begin{bmatrix} A_w^{\mathrm{T}} u_w + s_w - a_w \\ A_v^{\mathrm{T}} u_v + s_v - a_v \\ D(q_w) D(s_w) \mathbf{1} - \gamma \mathbf{1} \\ D(q_v) D(s_v) \mathbf{1} - \gamma \mathbf{1} \\ \tilde{b}_w + B_w Q^{-1/2} G d - M_w u_w - A_w q_w \\ \tilde{b}_v - B_v R^{-1/2} H d - M_v u_v - A_v q_v \\ G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} u_w - H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} \bar{u}_v \end{bmatrix}.$$

This entails solving the system

$$F_\gamma^{(1)} \begin{pmatrix} s_w \\ s_v \\ q_w \\ q_v \\ u_w \\ u_v \\ x \end{pmatrix} \begin{bmatrix} \Delta s_w \\ \Delta s_v \\ \Delta q_w \\ \Delta q_v \\ \Delta u_w \\ \Delta u_v \\ \Delta x \end{bmatrix} = -F_\gamma \begin{pmatrix} s_w \\ s_v \\ q_w \\ q_v \\ u_w \\ u_v \\ x \end{pmatrix}, \tag{36}$$

where the derivative matrix $F_\gamma^{(1)}$ is given by

$$\begin{bmatrix} I & 0 & 0 & 0 & (A_w)^{\mathrm{T}} & 0 & 0 \\ 0 & I & 0 & 0 & 0 & (A_v)^{\mathrm{T}} & 0 \\ D(q_w) & 0 & D(s_w) & 0 & 0 & 0 & 0 \\ 0 & D(q_v) & 0 & D(s_v) & 0 & 0 & 0 \\ 0 & 0 & -A_w & 0 & -M_w & 0 & B_w Q^{-1/2} G \\ 0 & 0 & 0 & -A_v & 0 & -M_v & -B_v R^{-1/2} H \\ 0 & 0 & 0 & 0 & G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} & -H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} & 0 \end{bmatrix}. \tag{37}$$

We now show the row operations necessary to reduce the matrix $F_\gamma^{(1)}$ in (37) to upper block triangular form. After each operation, we show only the row that was modified.

$$\begin{array}{l}
\text{row}_3 \leftarrow \text{row}_3 - D(q_w)\,\text{row}_1 \\
\begin{bmatrix} 0 & 0 & D(s_w) & 0 & -D(q_w)A_w^{\mathrm{T}} & 0 & 0 \end{bmatrix} \\
\text{row}_4 \leftarrow \text{row}_4 - D(q_v)\,\text{row}_2 \\
\begin{bmatrix} 0 & 0 & 0 & D(s_v) & 0 & -D(q_v)A_v^{\mathrm{T}} & 0 \end{bmatrix} \\
\text{row}_5 \leftarrow \text{row}_5 + A_w D(s_w)^{-1}\,\text{row}_3 \\
\begin{bmatrix} 0 & 0 & 0 & 0 & -T_w & 0 & B_w Q^{-1/2} G \end{bmatrix} \\
\text{row}_6 \leftarrow \text{row}_6 + A_v D(s_v)^{-1}\,\text{row}_4 \\
\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -T_v & -B_v R^{-1/2} H \end{bmatrix} .
\end{array}$$

In the above expressions,

$$\begin{aligned}
T_w &:= M_w + A_w D(s_w)^{-1} D(q_w) A_w^{\mathrm{T}}, \\
T_v &:= M_v + A_v D(s_v)^{-1} D(q_v) A_v^{\mathrm{T}} ,
\end{aligned} \tag{38}$$

where $D(s_w)^{-1}D(q_w)$ and $D(s_v)^{-1}D(q_v)$ are always full-rank diagonal matrices, since the vectors s_w, q_w, s_v, q_v. Matrices T_w and T_v are invertible as long as the PLQ densities for w and v satisfy (25).

Remark 19 (block diagonal structure of T in i.d. case) *Suppose that y is a random vector, $y = \text{vec}(\{y_k\})$, where each y_i is itself a random vector in $\mathbb{R}^{m(i)}$, from some PLQ density $\mathbf{p}(y_i) \propto \exp[-c_2\mathsf{p}(U_i, M_i, 0, I; \cdot)]$, and all y_i are independent. Let $U_i = \{u : A_i^{\mathrm{T}} u \le a_i\}$. Then the matrix T_p is given by $T_\mathsf{p} = M + ADA^{\mathrm{T}}$ where $M = \text{diag}[M_1, \cdots, M_N]$, $A = \text{diag}[A_1, \cdots, A_N]$, $D = \text{diag}[D_1, \cdots, D_N]$, and $\{D_i\}$ are diagonal with positive entries. Moreover, T_p is block diagonal, with ith diagonal block given by $M_i + A_i D_i A_i^{\mathrm{T}}$.*

From Remark 19, the matrices T_w and T_v in (38) are block diagonal provided that $\{w_k\}$ and $\{v_k\}$ are independent vectors from any PLQ densities.

We now finish the reduction of $F_\gamma^{(1)}$ to upper block triangular form:

$$\text{row}_7 \leftarrow \text{row}_7 + \left(G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} T_w^{-1}\right)\text{row}_5 - \left(H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} T_v^{-1}\right)\text{row}_6$$

$$\begin{bmatrix}
I & 0 & 0 & 0 & (A_w)^{\mathrm{T}} & 0 & 0 \\
0 & I & 0 & 0 & 0 & (A_v)^{\mathrm{T}} & 0 \\
0 & 0 & S_w & 0 & -Q_w(A_w)^{\mathrm{T}} & 0 & 0 \\
0 & 0 & 0 & S_v & 0 & -Q_v(A_v)^{\mathrm{T}} & 0 \\
0 & 0 & 0 & 0 & -T_w & 0 & B_w Q^{-1/2} G \\
0 & 0 & 0 & 0 & 0 & -T_v & -B_v R^{-1/2} H \\
0 & 0 & 0 & 0 & 0 & 0 & \Omega
\end{bmatrix}$$

where

$$\Omega = \Omega_G + \Omega_H = G^{\mathrm{T}} Q^{-\mathrm{T}/2} B_w^{\mathrm{T}} T_w^{-1} B_w Q^{-1/2} G + H^{\mathrm{T}} R^{-\mathrm{T}/2} B_v^{\mathrm{T}} T_v^{-1} B_v R^{-1/2} H.$$

Note that Ω is symmetric positive definite. Note also that Ω is block tridiagonal, since

1. Ω_H is block diagonal.

2. $Q^{-\mathrm{T}/2}B_w^{\mathrm{T}}T_w^{-1}B_wQ^{-1/2}$ is block diagonal, and G is block bidiagonal, hence Ω_G is block tridiagonal.

Solving system (36) requires inverting the block diagonal matrices T_v and T_w at each iteration of the damped Newton's method, as well as solving an equation of the form $\Omega\Delta x = \rho$. The matrices T_v and T_w are block diagonal, with sizes Nn and Nm, assuming m measurements at each time point. Given that they are invertible (see (25)), these inversions take $O(Nn^3)$ and $O(Nm^3)$ time. Since Ω is block tridiagonal, symmetric, and positive definite, $\Omega\Delta x = \rho$ can be solved in $O(Nn^3)$ time using the block tridiagonal algorithm in Bell (2000). The remaining four back solves required to solve (36) can each be done in $O(Nl)$ time, where we assume that $A_{v(k)} \in \mathbb{R}^{n\times l}$ and $A_{w(k)} \in \mathbb{R}^{m\times l}$ at each time point k.

References

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

A. Y. Aravkin, J. V. Burke, and M. P. Friedlander. Variational properties of value functions. *To Appear in Siam Journal of Optimization*, 2013.

A.Y. Aravkin. *Robust Methods with Applications to Kalman Smoothing and Bundle Adjustment*. PhD thesis, University of Washington, Seattle, WA, June 2010.

A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. An ℓ_1-laplace robust kalman smoother. *Automatic Control, IEEE Transactions on*, 56(12):2898–2911, dec. 2011a. ISSN 0018-9286. doi: 10.1109/TAC.2011.2141430.

A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. Learning using state space kernel machines. In *Proc. IFAC World Congress 2011*, Milan, Italy, 2011b.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

B.M. Bell. The marginal likelihood for parameters in a discrete Gauss-Markov process. *IEEE Transactions on Signal Processing*, 48(3):626–636, August 2000.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1): 1–122, January 2011. ISSN 1935-8237. doi: 10.1561/2200000016. URL `http://dx.doi.org/10.1561/2200000016`.

R. Brockett. *Finite Dimensional Linear Systems*. John Wiley and Sons, Inc., 1970.

J. V. Burke. An exact penalization viewpoint of constrained optimization. Technical report, Argonne National Laboratory, ANL/MCS-TM-95, 1987.

J.V. Burke. Descent methods for composite nondifferentiable optimization problems. *Mathematical Programming*, 33:260–279, 1985.

W. Chu, S. S. Keerthi, and O. C. Jin. A unified loss function in bayesian framework for support vector regression. In *In Proceeding of the 18th International Conference on Machine Learning*, pages 51–58, 2001.

F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39:1–49, 2001.

F. Dinuzzo. Analysis of fixed-point and coordinate descent algorithms for regularized kernel methods. *IEEE Transactions on Neural Networks*, 22(10):1576 –1587, 2011.

F. Dinuzzo, M. Neve, G. De Nicolao, and U. P. Gianazza. On the representer theorem and equivalent degrees of freedom of SVR. *Journal of Machine Learning Research*, 8:2467–2495, 2007.

D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

B. Efron, T. Hastie, L. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–150, 2000.

S. Farahmand, G.B. Giannakis, and D. Angelosante. Doubly robust smoothing of dynamical processes via outlier sparsity constraints. *IEEE Transactions on Signal Processing*, 59:4529–4543, 2011.

M.C. Ferris and T.S. Munson. Interior-point methods for massive support vector machines. *SIAM Journal on Optimization*, 13(3):783 – 804, 2003.

S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *J. Mach. Learn. Res.*, 2:243 –264, 2001.

J. Gao. Robust l1 principal component analysis and its Bayesian variational inference. *Neural Computation*, 20(2):555–572, February 2008.

A. Gelb. *Applied Optimal Estimation*. The M.I.T. Press, Cambridge, MA, 1974.

O. Güler and R. Hauser. Self-scaled barrier functions on symmetric cones and their classification. *Foundations of Computational Mathematics*, 2:121–143, 2002.

T. J. Hastie and R. J. Tibshirani. Generalized additive models. In *Monographs on Statistics and Applied Probability*, volume 43. Chapman and Hall, London, UK, 1990.

T. J. Hastie, R. J. Tibshirani, and J. Friedman. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. Springer, Canada, 2001.

P.J. Huber. *Robust Statistics*. Wiley, 1981.

A. Jazwinski. *Stochastic Processes and Filtering Theory*. Dover Publications, Inc, 1970.

T. Joachims, editor. *Making Large-Scale Support Vector Machine Learning Practical*. MIT Press, Cambridge, MA, USA, 1998.

S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):606 – 617, 2007.

M. Kojima, N. Megiddo, T. Noma, and A. Yoshise. *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*, volume 538 of *Lecture Notes in Computer Science*. Springer Verlag, Berlin, Germany, 1991.

C.J. Lin. On the convergence of the decomposition method for support vector machines. *IEEE Transactions on Neural Networks*, 12(12):1288 –1298, 2001.

H. Liu, S. Shah, and W. Jiang. On-line outlier detection and data cleaning. *Computers and Chemical Engineering*, 28:1635–1647, 2004.

S. Lucidi, L. Palagi, A. Risi, and M. Sciandrone. A convergent decomposition algorithm for support vector machines. *Comput. Optim. Appl.*, 38(2):217 –234, 2007.

D.J.C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

D.J.C. Mackay. Bayesian non-linear modelling for the prediction competition. *ASHRAE Trans.*, 100(2):3704–3716, 1994.

A. Nemirovskii and Y. Nesterov. *Interior-Point Polynomial Algorithms in Convex Programming*, volume 13 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, USA, 1994.

H. Ohlsson, F. Gustafsson, L. Ljung, and S. Boyd. Smoothed state estimates under abrupt changes using sum-of-norms regularization. *Automatica*, 48:595–605, 2012.

B. Oksendal. *Stochastic Differential Equations*. Springer, sixth edition, 2005.

J.A. Palmer, D.P. Wipf, K. Kreutz-Delgado, and B.D. Rao. Variational em algorithms for non-gaussian latent variable models. In *Proc. of NIPS*, 2006.

G. Pillonetto and B.M. Bell. Bayes and empirical Bayes semi-blind deconvolution using eigenfunctions of a prior covariance. *Automatica*, 43(10):1698–1712, 2007.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, 1998.

M. Pontil and A. Verri. Properties of support vector machines. *Neural Computation*, 10:955–974, 1998.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

B. Ristic, S. Arulampalam, and N. Gordon. *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House Publishers, 2004.

R.T. Rockafellar. *Convex Analysis*. Priceton Landmarks in Mathematics. Princeton University Press, 1970.

R.T. Rockafellar and R.J.B. Wets. *Variational Analysis*, volume 317. Springer, 1998.

S. Roweis and Z. Ghahramani. A unifying review of linear gaussian models. *Neural Computation*, 11:305–345, 1999.

S. Saitoh. *Theory of Reproducing Kernels and Its Applications*. Longman, 1988.

H. H. Schaefer. *Topological Vector Spaces*. Springe-Verlag, 1970.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (Adaptive Computation and Machine Learning). The MIT Press, 2001.

B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12:1207–1245, 2000.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. *Neural Networks and Computational Learning Theory*, 81:416–426, 2001.

A. J. Smola and B. Schölkopf. Bayesian kernel methods. In S. Mendelson and A. J. Smola, editors, *Machine Learning, Proceedings of the Summer School, Australian National University*, pages 65–117, Berlin, Germany, 2003. Springer-Verlag.

R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B.*, 58:267–288, 1996.

M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

P. Tseng and S. Yun. A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. *Comput. Optim. Appl.*, 47(2):1 –28, 2008.

V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, 1998.

G. Wahba. *Spline Models For Observational Data*. SIAM, Philadelphia, 1990.

G. Wahba. Support vector machines, reproducing kernel Hilbert spaces and randomized GACV. Technical Report 984, Department of Statistics, University of Wisconsin, 1998.

D.P. Wipf, B.D. Rao, and S. Nagarajan. Latent variable bayesian models for promoting sparsity. *IEEE Transactions on Information Theory*, 57:6236–6255, 2011.

S.J. Wright. *Primal-Dual Interior-Point Methods*. Siam, Englewood Cliffs, N.J., USA, 1997.

Y. Ye and K. Anstreicher. On quadratic and $o(\sqrt{nL})$ convergence of a predictor-corrector method for LCP. *Mathematical Programming*, 62(1-3):537–551, 1993.

E. H. Zarantonello. *Projections on Convex Sets in Hilbert Space and Spectral Theory*. Academic Press, 1971.

K. Zhang and J.T. Kwok. Clustered Nystrom method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576 –1587, 2010.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.