51st IEEE Conference on Decision and Control
December 10-13, 2012. Maui, Hawaii, USA

Nonsmooth regression and state estimation using piecewise quadratic log-concave densities

Aleksandr Y. Aravkin, James V. Burke and Gianluigi Pillonetto

***Abstract*— We demonstrate that many robust, sparse and nonsmooth identification and Kalman smoothing problems can be studied using a unified statistical framework. This framework is built on a broad sub-class of log-concave densities, which we call PLQ densities, that include many popular models for regression and state estimation, e.g. ℓ_1, ℓ_2, Vapnik and Huber penalties. Using the dual representation for PLQ penalties, we review conditions that permit interpreting them as negative logs of true probability densities. This allows construction of non-smooth multivariate distributions with specified means and variances from simple scalar building blocks. The result is a flexible statistical modelling framework for a variety of identification and learning applications, comprising models whose solutions can be computed using interior point (IP) methods. For the special case of Kalman smoothing, the complexity of this method scales linearly with the number of time-points, exactly as in the quadratic (Gaussian) case.**

***Index Terms*— robust and sparse estimation; statistical modeling; nonsmooth optimization; Kalman smoothing; interior point methods**

I. INTRODUCTION

Consider the following classical Bayesian parametric regression problem [14], [21]. The unknown x is a random vector[1] with prior distribution specified by

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

where μ and the invertible matrix $G \in \mathbb{R}^{n\times n}$ are known, while the random vector w is zero-mean with covariance Q. We would like to define an estimator for x using the measurements vector z that corresponds to a linear transformation of x contaminated with additive noise v. In particular, we have

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

where $H \in \mathbb{R}^{\ell\times n}$ is a known matrix while v is zero-mean, with covariance R and independent of x. It is well known that, under Gaussian assumptions on w and v, the minimum variance estimator of x is given by

$$\underset{x}{\operatorname{argmin}} \quad (z-Hx)^{\mathrm{T}}R^{-1}(z-Hx)+(\mu-Gx)^{\mathrm{T}}Q^{-1}(\mu-Gx)\,. \tag{I.3}$$

Notably, (I.3) also includes estimation problems related to discrete-time state-space dynamic linear systems [1], [4]. To see this, it is sufficient to think of x as 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 solve (I.3) in $O(N)$ operations [10].

In many circumstances, the performance of the estimator (I.3) is not satisfactory. For instance, quadratic penalization on model deviation is not robust with respect to the presence of outliers in the data [13], [9], [2], [8] or may have difficulties in reconstructing fast system dynamics, e.g. jumps in the state values [16]. Furthermore, quadratic penalties do not induce sparse solutions while it is often desirable to extract a small subset from a large measurement or parameter vector having greatest impact on the predictive capability of the estimate. This sparsity principle is present in many recently developed machine learning techniques, such as variable selection, selective shrinkage, and compressed sensing [11], [6], [5]. For these reasons, in place of (I.3), the following more general estimator is often used:

$$\underset{x}{\operatorname{argmin}} \quad V\left(Hx-z;R\right)+W\left(Gx-\mu;Q\right)\,, \tag{I.4}$$

where the loss V may be the ℓ_2-norm, the Huber penalty [13], Vapnik's ε-insensitive loss (leading to support vector regression [25], [12]) or the hinge loss (defining support vector classifiers [7], [18], [22]). The regularizer W may be e.g. the ℓ_2-norm, the ℓ_1-norm (as in the LASSO [23]), or a weighted combination of the two, yielding the elastic net procedure [27].

The robust and sparse approaches mentioned above can often be given a Bayesian interpretation specifying non-Gaussian priors on w (or directly on x) and on the noise v. Indeed, the stochastic interpretation of (I.4) has been much studied recently [15], [24], [26]. A description of non-Gaussian model errors and priors defining a great variety of loss and penalty functions are also discussed in [17] using convex-type representations, and integral-type variational representations related to Gaussian scale mixtures.

In contrast to the above approaches, as initiated in [3], in this paper we consider estimators containing penalty terms induced by a wide class of piecewise linear-quadratic (PLQ)

A.Y. Aravkin (saravkin@eos.ubc.ca) is with the Department of Earth and Ocean Sciences, University of British Columbia. Vancouver, BC, Canada. J.V. Burke (burke@math.washington.edu) is with the Department of Mathematics, University of Washington Seattle, WA, USA. G. Pillonetto (giapi@dei.unipd.it) is with the Department of Information Engineering, University of Padova, Padova, Italy.

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BPG, BP, Chevron, Conoco Phillips, Petrobras, PGS, Total SA, and WesternGeco. The research leading to these results has received funding also from the European Union Seventh Framework Programme [FP7/2007-2013] under grant agreement n° 257462 HYCON2 Network of excellence and n° 223866 FeedNetBack, by Progetto di Ateneo CPDA090135/09 funded by the University of Padova.

[1] All vectors are column vectors, unless otherwise specified

978-1-4673-2066-5/12/$31.00 ©2012 IEEE

functions starting from their dual representation [20]. This class includes, among others, ℓ_2, ℓ_1, hinge loss, Huber and Vapnik losses. We review the conditions which allow these losses to be viewed as negative logs of true probability densities. This ensures that the vectors w and v come from true distributions and allows us to interpret the solution to the problem (I.4) as a MAP estimator when the loss functions V and W come from this subclass of PLQ penalties. Then, we show that this viewpoint allows statistical modelling using non-smooth penalties, and in particular how multivariate densities with prescribed means and variances can be constructed using scalar PLQ penalties as building blocks.

In the second part of the paper, the Karush-Kuhn-Tucker (KKT) system for problem (I.4) as well as interior point (IP) methods to solve it are introduced. This allows a fundamentally smooth approach to many (non smooth) robust and sparse problems of interest to practitioners. Furthermore, we report a theorem showing that IP methods solve (I.4) when the noises v and w have PLQ densities, subject to sufficient additional hypotheses. The specialization of the result to the case of Kalman smoothing treated in [3] is also briefly reviewed.

The paper is organized as follows. In Section II we review the class of PLQ convex functions, and the sufficient conditions that allow us to interpret these functions as the negative logs of associated probability densities. In Section III we illustrate how to construct multivariate densities with prescribed means and variances using scalar building blocks. In Section IV we present the KKT system for PLQ penalties from [20], and report a theorem that guarantees convergence of IP methods under appropriate hypotheses. The Kalman smoothing dynamic case already described in [3] is also briefly reviewed. Conclusions are presented in Section V.

II. Piecewise Linear Quadratic Penalties and Densities

A. Preliminaries

We recall a few definitions from convex analysis, required to specify the domains of PLQ penalties. The reader is referred to [19], [20] 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 $\mathrm{cone}\, C$ the set $\{tr | r \in C, t \in \mathbb{R}_+\}$.
- (Domain) For $f(x) : \mathbb{R}^n \to \overline{\mathbb{R}} = \{\mathbb{R} \cup \infty\}$, $\mathrm{dom}(f) = \{x : f(x) < \infty\}$.
- (Polar cone) For any cone $K \subset \mathbb{R}^m$, the polar of K is defined to be
$$K^\circ := \{r | \langle r, d\rangle \le 0 \,\forall\, d \in K\}.$$
- (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, i.e. $d \in C^\infty$ if $C + d \subset C$.

B. PLQ Densities

Building a correspondence between penalties and densities allows us to establish a statistical and a computational framework for many estimation problems. Kalman smoothing provides a key example — the process and measurement covariance matrices are often known, and this information can be incorporated into the estimation problem through this correspondence. For PLQ penalties, the key to understanding their corresponding densities is their dual representation [20].

Definition 2.1 (extended PLQ penalties [20]): Define $\rho(U,M,b,B;\cdot) : \mathbb{R}^n \to \overline{\mathbb{R}}$ as

$$\rho(U,M,b,B;y) = \sup_{u \in U} \left\{ \langle u, b + By\rangle - \frac{1}{2}\langle u, Mu\rangle \right\}, \quad \text{(II.1)}$$

where $U \subset \mathbb{R}^m$ is a nonempty polyhedral set, $M \in \mathbb{R}^{m\times m}$ is a symmetric positive semidefinite matrix, and $b + By$ is an injective affine transformation, with $B \in \mathbb{R}^{m\times n}$, so, in particular, $m \le n$ and $\mathrm{null}(B) = \{0\}$.

■

The following result is taken from [3] and characterizes the effective domain of ρ.

Theorem 2.2 (effective domain of ρ [3]): Let ρ denote $\rho(U,M,B,b;y)$, and K denote $U^\infty \cap \mathrm{null}(M)$. Suppose $U \subset \mathbb{R}^m$ is a polyhedral set, $y \in \mathbb{R}^n$, $b \in K^\circ$, $M \in \mathbb{R}^{m\times m}$ is positive semidefinite, and $B \in \mathbb{R}^{m\times n}$ is injective. Then $(B^{\mathrm{T}}K)^\circ \subset \mathrm{dom}(\rho)$ and $[B^{\mathrm{T}}(K \cap -K)]^\perp = \mathrm{aff}[\mathrm{dom}(\rho)]$.

■

All the notable examples previously cited can be represented using an extended PLQ penalty ρ, as shown below.

Remark 2.3 (scalar examples): ℓ_2, ℓ_1, elastic net, Huber, hinge, and Vapnik penalties are all representable using the notation of Definition 2.1.

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 first left panel of Fig. 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 $u = \mathrm{sign}(y)$, hence $\rho(y) = |y|$, see second panel of Fig. 1.

3) Elastic net, $\ell_2 + \lambda\ell_1$. This is a weighted sum of the previous two examples, and so must be in the class. 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.

4) Huber: Take $U = [-\kappa, \kappa]$, $M = 1$, $b = 0$, and $B = 1$. We obtain $\rho(y) = \sup_{u \in U}\left\{uy - u^2/2\right\}$. We have the following cases:
   a) If $y < -\kappa$, take $u = -\kappa$ to obtain $-\kappa y - \frac{1}{2}\kappa^2$.

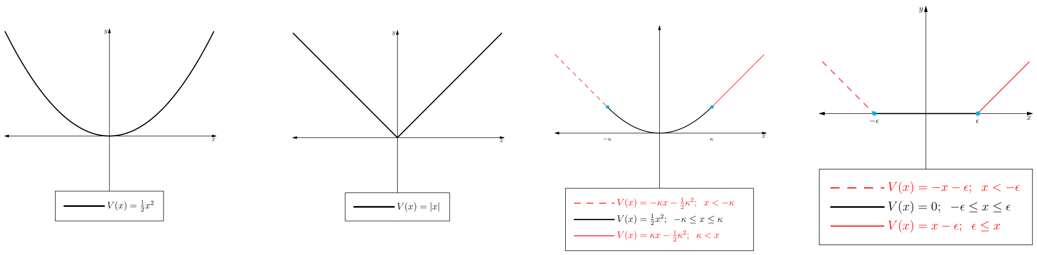

Fig. 1. Scalar ℓ_2 (first panel), ℓ_1 (second panel), Huber (third panel) and Vapnik (fourth panel) Penalties

b) If $-\kappa \le y \le \kappa$, take $u = y$ to obtain $\frac{1}{2}y^2$.
c) If $y > \kappa$, take $u = \kappa$ to obtain $\kappa y - \frac{1}{2}\kappa^2$.

This is the Huber penalty, see third panel of Fig. 1.

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

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

In fact, 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 in the fourth panel of Fig. 1.

■

From the above examples, note that the Vapnik penalty and the elastic net are obtained by summing together simpler PLQ penalties. These constructions are examples of a general pattern, as seen in the following remark.

Remark 2.4: Let $\rho_1(y)$ and $\rho_2(y)$ be two PLQ 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 PLQ penalty, with

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

■

To interpret PLQ penalties as negative logs of probability density functions, the integrability of said density functions has to be guaranteed. As discussed in [3], coercivity[2] is the key property to ensure integrability.

Theorem 2.5 (Coercivity [3]): A PLQ function ρ is *coercive* if and only if $[B^{\mathrm{T}}\mathrm{cone}(U)]^\circ = \{0\}$.

Theorem 2.6 (PLQ integrability [3]): Suppose $\rho(y)$ is coercive. Then the function $\exp[-\rho(y)]$ is integrable on $\mathrm{aff}[\mathrm{dom}(\rho)]$ with respect to the $\dim(\mathrm{aff}[\mathrm{dom}(\rho)])$-dimensional Lebesgue measure.

■

We can use Theorem 2.5 to show the coercivity of familiar penalties.

Corollary 2.7: The penalties ℓ_2, ℓ_1, elastic net, Vapnik, and Huber are all coercive.

Proof: We show all of these penalties satisfy the hypothesis of Theorem 2.5.

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 $B^{\mathrm{T}}\mathrm{cone}(U) = \mathbb{R}$.

■

The coercivity of the above examples can also be proved using their primal representations. However, our main objective is to establish a modeling framework where multidimensional penalties can be constructed from simple building blocks and then solved by a uniform approach, exploiting the dual representations alone.

We are now in a position to define a family of distributions on $\mathbb{R}^n$ interpreting PLQ penalties ρ as negative logs of corresponding densities. For this purpose, recall that the support of the distributions is always contained in the affine set $\mathrm{aff}(\mathrm{dom}\,\rho)$, characterized in Th. 2.2.

Definition 2.8: (PLQ densities). Let $\rho(U,M,B,b;y)$ be any coercive extended PLQ penalty on $\mathbb{R}^n$. Define $\mathbf{p}(y)$ to be the following density on $\mathbb{R}^n$:

$$\mathbf{p}(y) = \begin{cases} c^{-1}\exp[-\rho(y)] & y \in \mathrm{dom}\,\rho \\ 0 & \text{else,} \end{cases} \tag{II.2}$$

where

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

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

■

Thus, PLQ densities are true densities on the affine hull of the domain of ρ. In addition, Theorem 2.6 can be easily extended to show that they have moments of all orders.

[2] The function $\rho(y)$ is said to be *coercive* if $\lim_{\|y\|\to\infty}\rho(y) = \infty$

III. Constructing PLQ Densities

Given a sequence of column vectors $\{r_k\} = \{r_1, \ldots, r_N\}$ and matrices $\{\Sigma_k\} = \{\Sigma_1, \ldots, \Sigma_N\}$, let

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

In Definition 2.8, the PLQ 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, our aim is to construct densities having prescribed means and variances. We now illustrate how to do this using scalar PLQ random variables as the building blocks. Suppose $y = \mathrm{vec}(\{y_k\})$ is a vector of independent (but not necessarily identical) PLQ random variables having 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 just need to take

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

For example, the Gaussian distribution is specified by $U = \mathbb{R}^n$, $M = I$, $b = 0$, $B = I$, while the standard ℓ_1-Laplace (see [2]) 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. Letting c be the normalizing constant for the density of y, $c \det(Q)^{1/2}$ is the normalizing constant for the density of $\tilde{y}$.

To construct scalar building blocks with mean 0 and variance 1, we must be able to compute normalizing constants for any PLQ penalty. To this aim, if $\rho(y)$ is a scalar PLQ penalty symmetric about 0, we would like $\mathbf{p}(y) = \exp[-\rho(c_2 y)]/c_1$ to be a true density with unit variance, that is,

$$\frac{1}{c_1}\int_{\mathbb{R}} \exp[-\rho(c_2 y)]\,dy = 1, \quad \frac{1}{c_1}\int_{\mathbb{R}} y^2 \exp[-\rho(c_2 y)]\,dy = 1. \tag{III.1}$$

After u-substitution, these equations become

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

Solving this system yields

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

These expressions can be used to obtain the normalizing constants for any particular ρ using simple integrals. The Vapnik case is reported below.

A. Vapnik Density

The scalar Vapnik density is constructed as follows. Set

$$\mathbf{p_V}(y) = \frac{1}{c_1}\exp[-\rho_V(c_2 y)] \, , \tag{III.2}$$

where the normalizing constants c_1 and c_2 can be obtained from

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

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

$$\mathbf{p_V}(y) = \frac{\exp\left[-\sup_{u\in U}\left\{\left\langle c_2 B Q^{-1/2}(y - \mu) - \varepsilon \mathbf{1}_{2n}, u\right\rangle\right\}\right]}{c_1^n \det(Q^{1/2})} \tag{III.3}$$

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 $2n$.

IV. Optimization with PLQ Penalties

A. General Case

We now return to the estimation problem (I.4) where the functions V and W are to be taken from the class of PLQ penalties. In the previous sections, we showed how to construct PLQ densities with given moments to inform the optimization problem (I.4). We now show that the resulting problem can be solved with high accuracy *for the entire class* using standard techniques from numerical optimization. We exploit the dual representation for the class of PLQ penalties [20] to explicitly construct the Karush-Kuhn-Tucker (KKT) conditions for a wide variety of model problems of the form (I.4). Working with these systems opens the door to using a wide variety of numerical methods for convex quadratic programming to solve (I.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

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

and

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

Then (I.4) becomes

$$\min_{y\in\mathbb{R}^n} \rho(U, M, b, B; y), \tag{IV.3}$$

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 (I.1), (I.2), (I.4), and (II.1) imply that the matrix B in (IV.3) 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, $\mathrm{null}(B) = \{0\}$. Consequently, the objective in (IV.3) takes the form of a PLQ penalty function (II.1). In particular, if (IV.1) and (IV.2) arise from PLQ densities (definition 2.8), then the solution to problem (IV.3) is the MAP estimator in the statistical model (I.1)-(I.2).

To simplify the notational burden, in the remainder of this section we work with (IV.3) directly and assume that the defining objects in (IV.3) have the dimensions specified in (II.1);

$$U \in \mathbb{R}^m,\ M \in \mathbb{R}^{m\times m},\ b \in \mathbb{R}^m,\ \text{and}\ B \in \mathbb{R}^{m\times n}. \qquad \text{(IV.4)}$$

The Lagrangian [20][Example 11.47] for problem (IV.3) 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

$$U = \{u : A^{\mathrm{T}}u \le a\}\,, \qquad \text{(IV.5)}$$

where $A \in \mathbb{R}^{m\times \ell}$. Using this reprsentation for U, the optimality conditions for (IV.3) [19], [20] are

$$\begin{aligned} 0 &= B^{\mathrm{T}}u, & 0 &= b + By - Mu - Aq \\ 0 &= A^{\mathrm{T}}u + s - a, & 0 &= q_i s_i,\ q,s \ge 0\,, \end{aligned} \qquad \text{(IV.6)}$$

where $i = 1,\dots,\ell$ and the non-negative slack variable s is defined by the third equation in (IV.6). The non-negativity of s implies that $u \in U$. The equations $0 = q_i s_i\ \ i = 1,\dots,\ell$ in (IV.6) are known as the complementarity conditions. By convexity, solving the problem (IV.3) is equivalent to satisfying (IV.6). There is a vast optimization literature on working directly with the KKT system. Now, we show the general development for the entire PLQ class.

Let $U, M, b, B,$ and A be as defined in (II.1) and (IV.5), and let $\tau \in (0,+\infty]$. We define $\mathscr{F}_+(\tau)$, the τ *slice of the strict feasibility region for* (IV.6), to be the set

$$\left\{ (s,q,u,y) \;\middle|\; \begin{array}{c} 0 < s,\ 0 < q,\ s^{\mathrm{T}}q \le \tau,\ \text{and} \\ s,q,u,y \text{ satisfy the affine eq. in (IV.6)} \end{array} \right\},$$

and the $\mathscr{C}$, the *central path for* (IV.6), to be the set

$$\left\{ (s,q,u,y) \;\middle|\; \begin{array}{c} 0 < s,\ 0 < q,\ \gamma = q_i s_i\ \ i = 1,\dots,\ell,\ \text{and} \\ s,q,u,y \text{ satisfy the affine eq. in (IV.6)} \end{array} \right\}.$$

For simplicity, we define $\mathscr{F}_+ := \mathscr{F}_+(+\infty)$. The basic strategy of a primal-dual IP method is to follow the central path to a solution of (IV.6) 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

$$F_\gamma(s,q,u,y) = \begin{bmatrix} s + A^{\mathrm{T}}u - a \\ D(q)D(s)\mathbf{1} - \gamma\mathbf{1} \\ By - Mu - Aq + b \\ B^{\mathrm{T}}u \end{bmatrix}, \qquad \text{(IV.7)}$$

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

Theorem 4.1: Let $U, M, b, B,$ and A be as defined in (II.1) and (IV.5). Given $\tau > 0$, let $\mathscr{F}_+$, $\mathscr{F}_+(\tau)$, and $\mathscr{C}$ be as defined above. If

$$\mathscr{F}_+ \ne \emptyset \quad \text{and} \quad \mathrm{null}(M) \cap \mathrm{null}(A^{\mathrm{T}}) = \{0\}, \qquad \text{(IV.8)}$$

then the following statements hold.

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

(ii) Define

$$\widehat{\mathscr{F}_+} = \{(s,q) \mid \exists\,(u,y) \in \mathbb{R}^m \times \mathbb{R}^n \text{ s.t. } (s,q,u,y) \in \mathscr{F}_+\}$$

Then for each $(s,q) \in \widehat{\mathscr{F}_+}$ there exists a unique $(u,y) \in \mathbb{R}^m \times \mathbb{R}^n$ such that $(s,q,u,y) \in \mathscr{F}_+$.

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

(iv) For every $g \in \mathbb{R}^\ell_{++}$, there is a unique $(s,q,u,y) \in \mathscr{F}_+$ such that $g = (s_1q_1, s_2q_2, \dots, s_\ell q_\ell)^{\mathrm{T}}$.

(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

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

(vi) The set of cluster points of the central path as $\gamma \downarrow 0$ is non-empty, and every such cluster point solves (IV.6).

■

The proof is omitted due to space constraints. Theorem 4.1 shows that if the conditions (IV.8) hold, then IP techniques can be applied to solve the problem (IV.3). 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 (IV.3) with

$$U_v = \{u \mid A_v u \le a_v\} \quad \text{and} \quad U_w = \{u \mid A_w u \le b_w\} \qquad \text{(IV.9)}$$

this condition reduces to

$$\mathrm{null}(M_v) \cap \mathrm{null}(A_v^{\mathrm{T}}) = \{0\} \quad \text{and} \quad \mathrm{null}(M_w) \cap \mathrm{null}(A_w^{\mathrm{T}}) = \{0\}. \qquad \text{(IV.10)}$$

Corollary 4.2: The densities corresponding to ℓ_1, ℓ_2, Huber, and Vapnik penalties all satisfy hypothesis (IV.10).

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 $\mathscr{F}_+ \ne \emptyset$ is typically more difficult to verify, but it can be proved to be satisfied for all the popular losses, e.g. in the Vapnik and the Huber case. Details will be reported in future work

B. Kalman Smoothing with PLQ Penalties

The PLQ Kalman smoothing algorithm described in [3] can now be seen a special case of the theory described in the previous subsection. This is briefly reviewed below.

Consider a dynamic scenario, where the system state x_k evolves according to $x_1 = x_0 + w_1$, with x_0 known, and the following stochastic discrete-time linear model

$$\begin{aligned} x_k &= G_k x_{k-1} + w_k, & k &= 2,3,\dots,N \\ z_k &= H_k x_k + v_k, & k &= 1,2,\dots,N \end{aligned} \qquad \text{(IV.11)}$$

where 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, $\{w_k\}$ and $\{v_k\}$ are mutually independent zero-mean random variables. They can come from any of the densities introduced in the previous section and have positive definite covariance matrices denoted by $\{Q_k\}$ and $\{R_k\}$, respectively. In particular, one can write

$$\begin{aligned}\mathbf{p}(w_k) &\propto \exp\left[-\rho\left(U_k^w, M_k^w, b_k^w, B_k^w; Q_k^{-1/2} w_k\right)\right]\\ \mathbf{p}(v_k) &\propto \exp\left[-\rho(U_k^v, M_k^v, b_k^v, B_k^v; R_k^{-1/2} v_k)\right]\end{aligned} \tag{IV.12}$$

The following key result is then obtained (see [3]).

Theorem 4.3: (PLQ Kalman Smoother Theorem) Suppose that all w_k and v_k in the Kalman smoothing model (IV.11) come from PLQ densities that satisfy

$$\begin{aligned}&\mathrm{null}(M_k^w) \cap \mathrm{null}((A_k^w)^{\mathrm{T}}) = \{0\}\ ,\\ &\mathrm{null}(M_k^v) \cap \mathrm{null}((A_k^v)^{\mathrm{T}}) = \{0\}\ , \forall k\ .\end{aligned} \tag{IV.13}$$

i.e. their corresponding penalties are finite-valued. Suppose further that the corresponding set $\mathscr{F}_+$ from Theorem 4.1 is nonempty. Then, the MAP estimates of the states $\{x_k\}$ 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 $\{A_k^v\}$ and $\{A_k^w\}$ that define $\{U_k^w\}$ and $\{U_k^w\}$, respectively.

■

The main contribution of the result in the dynamical system context is the computational complexity: using IP methods the architecture of the MAP estimator preserves 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.

V. Conclusions

In this paper, we have complemented the theory initiated in [3] for robust and sparse estimation using nonsmooth PLQ penalties. Using the dual representation of PLQ functions, we first reviewed conditions allowing their interpretation as negative logs of true probability densities, thus establishing a statistical modelling framework. In the second part of the paper, we presented a broad computational approach to solving estimation problems (I.4) using interior point methods. Conditions that guarantee the successful implementation of such techniques, for solving (I.4) when x and v come from PLQ densities, have been derived. A theorem characterizing the convergence of IP methods for this class has been also stated. The key condition required for the successful execution of IP iterations was a requirement on 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). Thus, the statistical interpretation is strongly linked to the computational procedure. The specialization of this result for estimating states of discrete-time dynamic systems, subject to noises modeled by PLQ densities, has been also reviewed. In this context, our key result is that PLQ Kalman smoothing can always be performed with a number of operations linear in the length of the time series, as in the quadratic case.

The computational framework presented here allows the broad application of IP methods to a wide class of regression and Kalman smoothing problems of interest to practitioners. The powerful algorithmic scheme designed here, together with the statistical framework underlying it, underscores the practical utility and flexibility of this approach, making it suitable for many applications in the years ahead.

References

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

[2] A.Y. Aravkin, B.M. Bell, J.V. Burke, and G. Pillonetto. An ℓ_1-Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, 2011.

[3] A.Y. Aravkin, J.V. Burke, and G. Pillonetto. A statistical and computational theory for robust and sparse kalman smoothing. In *Proc. of the IFAC Symposium on System Identification (SysId 2012), Brussels, 2012*, 2012.

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

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

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

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

[8] 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.

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

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

[11] 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.

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

[13] P.J. Huber. *Robust Statistics*. Wiley, 1981.

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

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

[16] H. Ohlsson, F. Gustafsson, L. Ljung, and S. Boyd. State smoothing by sum-of-norms regularization. *Automatica (to appear)*, 2011.

[17] 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.

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

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

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

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

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

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

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

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

[26] D.P. Wipf, B.D. Rao, and S. Nagarajan. Latent variable bayesian models for promoting sparsity. *IEEE Transactions on Information Theory (to appear)*, 2011.

[27] 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.