A statistical and computational theory for robust and sparse Kalman smoothing

Aleksandr Aravkin * **James V. Burke** ** **Gianluigi Pillonetto** ***

* *Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, Canada (e-mail: saravkin@eos.ubc.ca).*
** *Department of Mathematics, University of Washington, Seattle, USA (e-mail: burke@math.washington.edu)*
*** *Department of Information Engineering, University of Padova, Padova, Italy (e-mail: giapi@dei.unipd.it)*

Abstract: Kalman smoothers reconstruct the state of a dynamical system starting from noisy output samples. While the classical estimator relies on quadratic penalization of process deviations and measurement errors, extensions that exploit Piecewise Linear Quadratic (PLQ) penalties have been recently proposed in the literature. These new formulations include smoothers robust with respect to outliers in the data, and smoothers that keep better track of fast system dynamics, e.g. jumps in the state values. In addition to L_2, well known examples of PLQ penalties include the L_1, Huber and Vapnik losses. In this paper, we use a dual representation for PLQ penalties to build a statistical modeling framework and a computational theory for Kalman smoothing.
We develop a statistical framework by establishing conditions required to interpret PLQ penalties as negative logs of true probability densities. Then, we present a computational framework, based on interior-point methods, that solves the Kalman smoothing problem with PLQ penalties and maintains the linear complexity in the size of the time series, just as in the L_2 case. The framework presented extends the computational efficiency of the Mayne-Fraser and Rauch-Tung-Striebel algorithms to a much broader non-smooth setting, and includes many known robust and sparse smoothers as special cases.

Keywords: Piecewise linear quadratic penalties; nonsmooth optimization; L_1/Huber/Vapnik loss functions; interior point methods

1. INTRODUCTION

Consider the following discrete-time linear state-space model

$$\begin{aligned} x_1 &= x_0 + w_1 \\ x_k &= G_k x_{k-1} + w_k, \qquad k = 2, 3, \dots, N \\ z_k &= H_k x_k + v_k, \qquad k = 1, 2, \dots, N \end{aligned} \tag{1.1}$$

where $x_k \in \mathbb{R}^n$ is the state, x_0 is known, $z_k \in \mathbb{R}^m$ contains noisy output samples, G_k and H_k are known matrices. Further, $\{w_k\}$ and $\{v_k\}$ are mutually independent zero-mean random variables with covariances given by $\{Q_k\}$ and $\{R_k\}$, respectively.
The classical fixed-interval Kalman smoothing problem is to obtain the (unconditional) minimum variance linear estimator of the states $\{x_k\}_{k=1}^N$ as a function of $\{z_k\}_{k=1}^N$. It is well known that the structure of this estimator is related to the following optimization problem

$$\min_{\{x_k\}} \sum_{k=1}^{N} \|z_k - H_k x_k\|^2_{R_k^{-1}} + \|x_k - G_k x_{k-1}\|^2_{Q_k^{-1}} \tag{1.2}$$

where G_1 denotes the identity matrix and $\|a\|^2_\Sigma := a^\top \Sigma a$ for every column vector a. When data become available, the solution can be computed by the classical Kalman smoother with the number of operations linear in N. This procedure also provides the minimum variance estimate of the states when all the system noises are assumed to be Gaussian.
In many circumstances, linear estimators that rely on quadratic penalization of model deviation, such as (1.2), lead to unsatisfactory results. For instance, they are not robust with respect to the presence of outliers in the data [Huber, 1981, Aravkin et al., 2011a, Farahmand et al., 2011] and may have difficulties in reconstructing fast system dynamics, e.g. jumps in the state values [Ohlsson et al., 2011]. In addition, sparsity-promoting regularization is often used in order to extract from a large measurement or parameter vector a small subset having 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, such as feature selection, selective shrinkage, and compressed sensing [Hastie and Tibshirani, 1990, Efron et al., 2004, Donoho, 2006]. In many circumstances, when smoothing is considered, it can be interpreted as a sparse non Gaussian prior distribution on the noises entering the system. In these cases, the estimator (1.2) is often replaced by

$$\sum_{k=1}^{N} V\left(z_k - H_k x_k; R_k\right) + J\left(x_k - G_k x_{k-1}; Q_k\right) \tag{1.3}$$

where, for example, V can be the Huber or the Vapnik's ϵ-insensitive loss, used in support vector regression [Vapnik, 1998, Evgeniou et al., 2000], while J may be the ℓ_1-norm, as in the LASSO procedure [Tibshirani, 1996].

The interpretation of problems such as (1.3) in terms of Bayesian estimation 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 can be found e.g. in [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 in [Palmer et al., 2006] using convex-type and integral-type variational representations, with the latter being related to Gaussian scale mixtures. The fundamental novelty in this work is that, rather than taking this approach, we start with a particular class of losses, called PLQ penalties, well known from optimization literature [Rockafellar and Wets, 1998]. We establish conditions which allow these losses to be viewed as negative logs of true densities, ensuring that w_k and v_k in (1.1) come from true distributions. This in turn allows us to interpret the solution to the problem (1.3) as a MAP estimator when the loss functions V and J come from this subclass of PLQ penalties. We will show that this subclass includes the four key examples, the L_2, L_1, Huber, and Vapnik penalties.

The density characterization of PLQ penalties is achieved using a dual representation, which also underlies the development of algorithms for fitting models of the form (1.3). In particular, in the second part of the paper we derive the conditions, complimentary to those needed to set up the statistical framework, that allow the development of new and computationally efficient Kalman smoothers designed using non-smooth penalties on the process and measurement deviations. Amazingly, it turns out that the interior point method used in [Aravkin et al., 2011a] generalizes perfectly to the entire class of PLQ densities under a simple verifiable non-degeneracy condition. Hence, the solutions of all the PLQ Kalman smoothers can be computed with a number of operations that scales linearly in N, as in the quadratic case. Such theoretical foundation generalizes the results recently obtained in [Aravkin et al., 2011a,b, Farahmand et al., 2011, Ohlsson et al., 2011], framing them as particular cases of the framework presented here.

The paper is organized as follows. In Section 2 we introduce the class of PLQ convex functions, and provide the conditions under which they can be interpreted as negative logs of corresponding densities. In Section 3 we present a new PLQ Kalman smoother theorem that generalizes the well known Mayne-Fraser two-filter and the Rauch-Tung-Striebel algorithm [Gelb, 1974] to nonsmooth formulations. This theorem is obtained by solving the Karush-Kuhn-Tucker (KKT) system for PLQ penalties using interior point methods, and exploiting the state space structure to obtain the solution in linear time. The necessary results and proofs supporting the main theorems appear in the Appendix. We end the paper with a few concluding remarks.

2. PIECEWISE LINEAR QUADRATIC PENALTIES AND DENSITIES

2.1 Preliminaries

We recall a few definitions from convex analysis.

- (Affine hull) Define the affine hull of any set S, denoted by aff S, as the smallest affine set that contains S.
- (Cone) For any set S, denote by cone S the set $\{ts | s \in S, t \in \mathbb{R}_+\}$.
- (Polar Cone) For any cone $K \subset \mathbb{R}^m$, the polar of K is defined to be
$$K^\circ := \{v | \langle v, w\rangle \leq 0 \; \forall \; w \in K\}.$$
- (Horizon cone). The (convex) Horizon cone C^∞ is the set of 'unbounded directions' for C, i.e. $d \in C^\infty$ if for any point $\bar{w} \in C$ we have $\{d | \bar{w} + \tau d \in \text{cl } C \; \forall \; \tau \geq 0\}$.

2.2 PLQ densities

We now introduce the PLQ penalties and densities that are the focus of this paper.

Definition 2.1. (piecewise linear-quadratic penalties) [Rockafellar and Wets, 1998]. For a nonempty polyhedral set $U \subset \mathbb{R}^m$ and a symmetric positive-semidefinite matrix $M \in \mathbb{R}^{m\times m}$ (possibly $M = 0$), the function $\theta_{U,M} : \mathbb{R}^m \to \overline{\mathbb{R}}$ defined by

$$\theta_{U,M}(w) := \sup_{u\in U} \left\{ \langle u, w\rangle - \frac{1}{2}\langle u, Mu\rangle \right\} \tag{2.1}$$

is proper, convex, and piecewise linear-quadratic. When $M = 0$, it is piecewise linear; $\theta_{U,0} = \sigma_U$, the support function of U. The effective domain of $\theta_{U,M}$, denoted by $\text{dom}(\theta_{U,M})$, is the set of $w \in \mathbb{R}^m$ where $\theta_{U,M}(w) < \infty$, and is given by $(U^\infty \cap \text{Null}(M))^\circ$. ■

In order to capture the full class of penalties of interest, we consider injective affine transformations into $\mathbb{R}^m$ of the form $b + By$. The requirements on B therefore are $m \geq n$ and $\text{Null}(B) = \{0\}$. The final technical requirement we impose is that $b \in \text{dom } \theta_{U,M}$.

Definition 2.2. (PLQ penalties with shifts and transforms) Using (2.1), define $\rho : \mathbb{R}^n \to \mathbb{R}$ as $\theta_{U,M}(b + By)$:

$$\rho_{U,M,b,B}(y) := \sup_{u\in U} \left\{ \langle u, b + By\rangle - \frac{1}{2}\langle u, Mu\rangle \right\} \tag{2.2}$$
■

The following result characterizes the effective domain of ρ (see Appendix for proof).

Theorem 2.3. Let ρ denote $\rho_{U,M,B,b}(y)$, and K denote $U^\infty \cap \text{Null}(M)$. 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 we have $(B^{\mathrm{T}} K)^\circ \subset \text{dom}(\rho)$ and $(B^{\mathrm{T}}(K \cap -K))^\perp = \text{aff}(\text{dom}(\rho))$. ■

Note that the functions ρ are still piecewise linear-quadratic. All of the important examples mentioned before can be represented in this way, as shown below.

Remark 2.4. (scalar examples). The L_2, ℓ_1, Huber, and Vapnik penalties are representable in the notation of Definition 2.2.

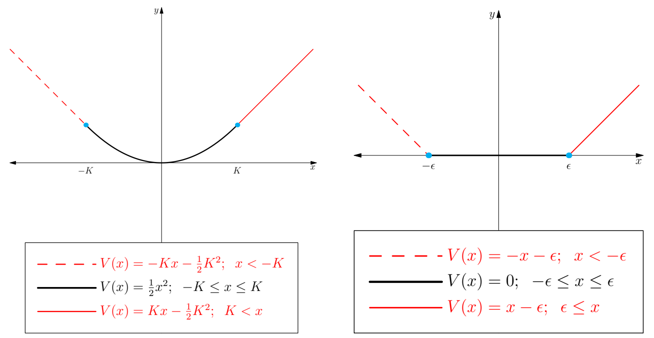

Fig. 1. Huber (left) and Vapnik (right) Penalties

(1) L_2: Take $U = \mathbf{R}$, $M = 1$, $b = 0$, and $B = 1$. We obtain $\rho(y) = \sup_{u \in \mathbf{R}} \left\langle uy - \frac{1}{2}u^2 \right\rangle$. The function inside the sup is maximized at $u = y$, whence $\rho(y) = \frac{1}{2}y^2$.
(2) ℓ_1: Take $U = [-1, 1]$, $M = 0$, $b = 0$, and $B = 1$. We obtain $\rho(x) = \sup_{u \in [-1,1]} \langle uy \rangle$. The function inside the sup is maximized by taking $u = \text{sign}(y)$, whence $\rho(x) = |y|$.
(3) Huber: Take $U = [-K, K]$, $M = 1$, $b = 0$, and $B = 1$. We obtain $\rho(y) = \sup_{u \in [-K,K]} \left\langle uy - \frac{1}{2}u^2 \right\rangle$. Take the derivative with respect to u and consider the following cases:
   (a) If $y < -K$, take $u = -K$ to obtain $-Ky - \frac{1}{2}K^2$.
   (b) If $-K \le y \le K$, take $u = y$ to obtain $\frac{1}{2}y^2$.
   (c) If $y > K$, take $u = K$ to obtain a contribution of $Ky - \frac{1}{2}K^2$.
   This is the Huber penalty with parameter K, shown in the left panel of Fig. 1.
(4) Vapnik: take $U = [0, 1] \times [0, 1]$, $M = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, $B = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$, and $b = \left[\begin{smallmatrix} -\epsilon \\ -\epsilon \end{smallmatrix}\right]$, for some $\epsilon > 0$. We obtain $\rho(y) = \sup_{u_1, u_2 \in [0,1]} \left\langle \begin{bmatrix} y - \epsilon \\ -y - \epsilon \end{bmatrix}, \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \right\rangle$. We can obtain an explicit representation by considering three cases:
   (a) If $|y| < \epsilon$, take $u_1 = u_2 = 0$. Then $\rho(y) = 0$.
   (b) If $y > \epsilon$, take $u_1 = 1$ and $u_2 = 0$. Then $\rho(y) = y - \epsilon$.
   (c) If $y < -\epsilon$, take $u_1 = 0$ and $u_2 = 1$. Then $\rho(y) = -y - \epsilon$.
   This is the Vapnik penalty with parameter ϵ, shown in the right panel of Fig. 1.

Note that the affine generalization (Definition 2.2) is already needed to express the Vapnik penalty. ■

In order to characterize PLQ penalties as negative logs of density functions, we need to ensure the integrability of said density functions. A function $\rho(x)$ is called *coercive* if $\lim_{\|x\| \to \infty} \rho(x) = \infty$, and coercivity turns out to be the key property to ensure integrability. The proof of this fact, and the characterization of coercivity for PLQ penalties using the notation of Def. 2.2, are the subject of the next two theorems (see Appendix for proofs).

Theorem 2.5. (PLQ Integrability). Suppose $\rho(y)$ is coercive, and let n_{aff} denote the dimension of $\text{aff}(\text{dom}\ \rho)$. Then the function $f(y) = \exp(-\rho(y))$ is integrable on $\text{aff}(\text{dom}\ \rho)$ with the n_{aff}-dimensional Lebesgue measure. ■

Theorem 2.6. (Coercivity of ρ). ρ is coercive if and only if $[B^{\mathrm{T}}\text{cone}(U)]^{\circ} = \{0\}$. ■

Theorem 2.6 can be used to show the coercivity of familiar penalties.

Corollary 2.7. The penalties L_2, L_1, Vapnik, and Huber are all coercive.

Proof: We show all of these penalties satisfy the hypothesis of Theorem 2.6.

(1) L_2: $U = \mathbf{R}$ and $B = 1$, so $[B^{\mathrm{T}}\text{cone}(U)]^{\circ} = \mathbf{R}^{\circ} = \{0\}$.
(2) ℓ_1: $U = [-1, 1]$, so $\text{cone}(U) = \mathbf{R}$, and $B = 1$, so proof reduces to that case 1.
(3) Huber: $U = [-K, K]$, so $\text{cone}(U) = \mathbf{R}$, and $B = 1$, so proof reduces to that of case 1.
(4) Vapnik: $U = [0, 1] \times [0, 1]$, so $\text{cone}(U) = \mathbf{R}^2_+$. $B = \left[\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right]$, so $B^{\mathrm{T}}\text{cone}(U) = \mathbf{R}$, and again we reduce to case 1. ■

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 the affine set $\text{aff}(\text{dom}\ \rho)$, characterized in Th. 2.3.

Definition 2.8. (Piecewise linear quadratic densities). Let ρ be any coercive piecewise linear quadratic function on $\mathbb{R}^n$ of the form $\rho_{U,M,B,b;}(y) = \theta_{U,M}(b + By)$. Define $\mathbf{p}(y)$ to be the following density on $\mathbb{R}^n$:

$$\mathbf{p}(y) = \begin{cases} c_1^{-1} \exp(-\rho(y)) & y \in \text{dom}\ \rho \\ 0 & \text{else,} \end{cases} \tag{2.3}$$

where

$$c_1 = \left(\int_{y \in \text{dom}\ \rho} \exp(-\rho(y)) dy \right),$$

and integral is with respect to the Lebesgue measure with dimension $\dim\Big(\text{aff}(\text{dom}\ \rho)\Big)$. ■

PLQ densities are true densities on the affine hull of the domain of ρ. The proof of Theorem 2.5 can be easily adapted to show that they have moments of all orders.

3. KALMAN SMOOTHING WITH PLQ PENALTIES

In this section, we consider the model (1.1), but in the general case where errors w_k and v_k can come from any of the densities introduced in the previous section. To this end, we first formulate the KS problem over the entire sequence $\{x_k\}$.

Given a sequence of column vectors $\{u_k\}$ and matrices $\{T_k\}$ we use the notation

$$\text{vec}(\{u_k\}) = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}, \quad \text{diag}(\{T_k\}) = \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_N \end{bmatrix}.$$

We make the following definitions.

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

We also introduce the matrices G and H:

$$G = \begin{bmatrix} \mathrm{I} & 0 & & \\ -G_2 & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -G_N & \mathrm{I} \end{bmatrix}, \quad H = \begin{bmatrix} H_1 & 0 & & \\ 0 & H_2 & \ddots & \\ & \ddots & \ddots & 0 \\ & & 0 & H_N \end{bmatrix}.$$

With this notation, model (1.1) can be written

$$\begin{aligned} \tilde{x}_0 &= Gx + w \\ z &= Hx + v , \end{aligned} \tag{3.1}$$

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 $\tilde{x}_0$ 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.

The general Kalman smoothing problem is described in the following proposition.

Proposition 3.1. Suppose that the noises w and v in the model (3.1) are PLQ densities with means 0, variances Q and R (see Def. 2.8). Then, for suitable U^w, M^w, b^w, B^w and U^v, M^v, b^v, B^v we have

$$\begin{aligned} \mathbf{p}(w) &\propto \exp(-\theta_{U^w,M^w}(b^w + B^w Q^{-1/2} w)) \\ \mathbf{p}(v) &\propto \exp(-\theta_{U^v,M^v}(b^v + B^v R^{-1/2} v)) \end{aligned} \tag{3.2}$$

while the MAP estimator of x in the model (3.1) is

$$\arg\min_{x \in \mathbb{R}^{nN}} \left\{ \begin{array}{l} \theta_{U^w,M^w}(b^w + B^w Q^{-1/2}(Gx - \tilde{x}_0)) \\ + \theta_{U^v,M^v}(b^v + B^v R^{-1/2}(Hx - z)) \end{array} \right\} \tag{3.3}$$

■

Note that since w_k and v_k are independent, problem (3.3) is decomposable into a sum of terms analogous to (1.2). This special structure is manifest in the block diagonal structure of H, Q, R, B^v, B^w, the bidiagonal structure of G, and the structure of sets U^w and U^v, and is key in proving the linear complexity result that will be derived in the next part of this section.

For our purposes, it is now important to recall that, when the sets U^w and U^v are polyhedral, (3.3) is an Extended Linear Quadratic program (ELQP), described in [Rockafellar and Wets, 1998, Example 11.43]. Rather than directly solving (3.3), we work with the Karush-Kuhn-Tucker (KKT) system. We present the system in the following lemma, and derive it in the Appendix.

Lemma 3.2. Suppose that the sets U^w and U^v are polyhedral, i.e. can be written

$$U^w = \{u | (A^w)^T u \le a^w\}, \quad U^v = \{u | (A^v)^T u \le a^v\} .$$

Then the necessary first-order conditions for optimality of (3.3) are given by

$$\begin{aligned} 0 &= (A^w)^{\mathrm{T}} u^w + s^w - a^w \; ; \quad 0 = (A^v)^{\mathrm{T}} u^v + s^v - a^v \\ 0 &= (s^w)^{\mathrm{T}} q^w \; ; \qquad\qquad\quad\;\; 0 = (s^v)^{\mathrm{T}} q^v \\ \\ 0 &= \tilde{b}^w + B^w Q^{-1/2} G \bar{d} - M^w \bar{u}^w - A^w q^w \\ 0 &= \tilde{b}^v - B^v R^{-1/2} H \bar{d} - M^v \bar{u}^v - A^v q^v \\ 0 &= G^{\mathrm{T}} Q^{-\mathrm{T}/2} (B^w)^{\mathrm{T}} \bar{u}^w - H^{\mathrm{T}} R^{-\mathrm{T}/2} (B^v)^{\mathrm{T}} \bar{u}^v \\ 0 &\le s^w, s^v, q^w, q^v. \end{aligned} \tag{3.4}$$

■

Our approach is to solve (3.4) directly using Interior Point (IP) methods. IP methods work by applying a damped Newton iteration to a relaxed version of (3.4), specifically relaxing the 'complementarity conditions':

$$\begin{aligned} (s^w)^{\mathrm{T}} q^w = 0 &\to Q^w S^w \mathbf{1} - \mu \mathbf{1} = 0 \\ (s^v)^{\mathrm{T}} q^v = 0 &\to Q^v S^v \mathbf{1} - \mu \mathbf{1} = 0 , \end{aligned}$$

where Q^w, S^w, Q^v, S^v are diagonal matrices with diagonals q^w, s^w, q^v, s^v respectively. The parameter μ is aggressively decreased to 0 as the IP iterations proceed. Typically, no more than 10 or 20 iterations of the relaxed system are required to obtain a solution of (3.4), and hence an optimal solution to (3.3). The following theorem is key and represents the main result of this section. It shows that the computational effort required (per IP iteration) is linear in the number of time steps whatever PLQ density enters the state space model.

Theorem 3.3. (PLQ Kalman Smoother Theorem) Suppose that all w_k and v_k in the Kalman smoothing model (1.1) come from PLQ densities that satisfy $\mathrm{Null}(M) \cap U^\infty = \{0\}$, i.e. their corresponding penalties are finite-valued. Then (3.3) can be solved using an IP method, with computational complexity $O(Nn^3 + Nm)$ time. ■

The proof is presented in the Appendix and shows that IP methods for solving (3.3) preserve the key block tridiagonal structure of the standard smoother. General smoothing estimates can therefore be computed in $O(Nn^3)$ time, as long as the number of IP iterations is fixed (as it usually is in practice, to 10 or 20).

It is important to observe that the motivating examples (see Remark 2.4) all satisfy the conditions of Theorem 3.3.

Corollary 3.4. The densities corresponding to L^1, L^2, Huber, and Vapnik penalties all satisfy the hypotheses of Theorem 3.3.

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 $U^\infty = \{0\}$.

4. CONCLUSIONS

We have presented a new theory for robust and sparse Kalman smoothing using nonsmooth PLQ penalties applied to process and measurement deviations. These smoothers can be designed within a statistical framework obtained by viewing PLQ penalties as negative logs of true probability densities, and we have presented necessary conditions that allow this interpretation. In this regard, the coercivity condition characterized in Th. 2.6 has been shown to play a central role. Notice that such a condition is also a nice example of how the statistical framework established in the first part of this paper gives an alternative viewpoint for an idea useful in machine learning. In fact, coercivity is also a fundamental prerequisite in sparse and robust estimation as it precludes directions for which the loss and the regularizer are insensitive to large parameter/state changes. Thus, the condition for a (PLQ) penalty to be a negative log of a true density also ensures that the problem is well posed and that the learning machine/smoother can control model complexity. In the second part of the paper, we have shown that solutions to PLQ Kalman smoothing formulations can be computed with a number of operations that is linear in the length of the time series, as in the quadratic case.

A sufficient condition for the successful execution of IP iterations is that the PLQ penalties used should 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.

The computational framework presented allows a broad application of interior point methods to a wide class of smoothing problems of interest to practitioners. The powerful algorithmic scheme designed here, together with the breadth and significance of the new statistical framework presented, underscores the practical utility and flexibility of this approach. We believe that this perspective on model development and Kalman smoothing will be useful in a number of applications in the years ahead.

REFERENCES

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

A. 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.

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.

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.

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

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.

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

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

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

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.

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.

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

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

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

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

APPENDIX

Preliminaries

Definition 4.1. (Horizon cone, specialized to the convex setting by [Rockafellar and Wets, 1998, Theorem 3.6]). The Horizon cone C^∞ for a convex set C is convex, and for any point $\bar{w} \in C$ consists of the vectors $\{d|\bar{w} + \tau d \in \text{cl}\ C\ \forall\ \tau \geq 0\}$.

Definition 4.2. (Lineality). Define the lineality of convex cone K, denoted $\text{lin}(K)$, to be $K \cap -K$. Since K is a convex cone, $\text{lin}(K)$ is the largest subspace contained in K.

Lemma 4.3. (Characterization of lineality, [Rockafellar, 1970, Theorem 14.6]). Let K be any closed set containing the origin. Then $\text{lin}(K) = (K^\circ)^\perp$.

Definition 4.4. (Affine hull). Define the affine hull of any set S, denoted by aff S, as the smallest affine set that contains S.

Corollary 4.5. (Characterization of aff K°) Taking the perp of the characterization in Lemma 4.3, the affine hull of the polar of a closed convex cone K is given by aff $K^\circ = \text{lin}(K)^\perp$.

Proof of Theorem 2.3

Lemma 4.6. (Polars, linear transformations, and shifts) Let $K \subset \mathbb{R}^n$ be a closed convex cone, $b \in \mathbb{R}^n$, and $B \in \mathbb{R}^{n\times k}$. Then we have $(B^{\mathrm{T}}K)^\circ \subset B^{-1}(K^\circ - b)$ if $b \in K^\circ$.

Proof: Recall that a convex cone is closed under addition. Then for any $b \in K^\circ$, we have $K^\circ + b \subset K^\circ$, and hence $K^\circ \subset K^\circ - b$. By [Rockafellar, 1970, Corollary 16.3.2] we get

$$(B^{\mathrm{T}}K)^\circ = B^{-1}K^\circ \subset B^{-1}(K^\circ - b)\ .$$

■

Corollary 4.7. Let K be a closed convex cone, and $B \in \mathbb{R}^{n\times k}$. If $b \in K^\circ$, then $\left(B^{\mathrm{T}}(\text{lin}(K))\right)^\perp \subset \text{aff}(B^{-1}(K^\circ - b))$.

Proof: By Lemma 4.6, $\text{aff}(B^{-1}(K^\circ - b)) \supset \text{aff}(B^{\mathrm{T}}K)^\circ = \left(\text{lin}(B^{\mathrm{T}}K)\right)^\perp$ where the last equality is by Corollary 4.5. Since B^{T} is a linear transformation, we have $\text{lin}(B^{\mathrm{T}}K) = B^{\mathrm{T}}\text{lin}(K)$.

■

Lemma 4.8. Let $K \subset \mathbb{R}^n$ be a closed convex cone, $b \in \text{aff}(K)^\circ$, and $B \in \mathbb{R}^{n\times k}$. Then $\text{aff}(B^{-1}(K^\circ - b)) \subset B^{-1}(\text{lin}(K))^\perp \subset (B^{-1}\text{aff}(K^\circ - b))$.

Proof: If $w \in \text{aff}\left(B^{-1}(K^\circ - b)\right)$, for some finite N we can find sets $\{\lambda_i\} \subset \mathbb{R}$ and $\{w_i\} \subset B^{-1}(K^\circ - b)$ such that $\sum_{i=1}^N \lambda_i = 1$ and $\sum_{i=1}^N \lambda_i w_i = w$. For each w_i, we have $Bw_i \in K^\circ - b$, so $b + Bw_i \in K^\circ$. Then

$$b + Bw = \sum_{i=1}^N \lambda_i(b + Bw_i) \in \text{aff}(K^\circ) = \text{lin}(K)^\perp.$$

Since $b \in \text{lin}(K)^\perp$ by assumption, we have $Bw \in \text{lin}(K)^\perp$, and so $w \in B^{-1}(\text{lin}(K)^\perp)$.

Next, starting with $w \in B^{-1}(\text{lin}(K)^\perp)$ we have $Bw \in \text{lin}(K)^\perp$ and so $b + Bw \in \text{lin}(K)^\perp$ since $\text{lin}(K)^\perp$ is a subspace and $b \in \text{lin}(K)^\perp$. Then for some finite $\tilde{N}$ we can find sets $\{\lambda_i\} \subset \mathbb{R}$ and $\{v_i\} \subset K^\circ$ such that $\sum_{i=1}^{\tilde{N}} \lambda_i = 1$ and $\sum_{i=1}^{\tilde{N}} \lambda_i v_i = b+Bw$. Subtracting b from both sides, we have $\sum_{i=1}^{\tilde{N}} \lambda_i(v_i - b) = Bw$, so in particular $Bw \in \text{aff}(K^\circ - b)$. Then $w \in B^{-1}\text{aff}(K^\circ - b)$.

■

Theorem 4.9. Let $K \subset \mathbb{R}^n$ be a closed convex cone, $b \in \mathbb{R}^n$, and $B \in \mathbb{R}^{n\times k}$. If $b \in K^\circ$, then $(B^\mathrm{T}\text{lin}(K))^\perp = \text{aff}\left(B^{-1}(K^\circ - b)\right) = B^{-1}(\text{lin}(K)^\perp)$.

Proof: From Corollary 4.7 and Lemma 4.8, we immediately have

$$(B^\mathrm{T}\text{lin}(K))^\perp \subset \text{aff}\left(B^{-1}(K^\circ - b)\right) \subset B^{-1}(\text{lin}(K)^\perp).$$

Note that for any subspace S, $S^\perp = S^\circ$. Then by [Rockafellar, 1970, Corollary 16.3.2], $(B^\mathrm{T}\text{lin}(K))^\perp = B^{-1}(\text{lin}(K)^\perp)$.

■

The proof of Theorem 2.3 now follows from Lemma 4.6 and Theorem 4.9.

Proof of Theorem 2.5

Using the characterization of a piecewise quadratic function from [Rockafellar and Wets, 1998, Definition 10.20], the effective domain of $\rho(y)$ can be represented as the union of finitely many polyhedral sets U_i, relative to each of which $\rho(y)$ is given by an expression of the form $\frac{1}{2}\langle y, A_i y\rangle + \langle a_i, y\rangle + \alpha_i$ for some scalar $\alpha_i \in \mathbb{R}$, vector $a_i \in \mathbb{R}^n$ and symmetric positive semidefinite matrix $A_i \in \mathbb{R}^{n\times n}$. Since $\rho(y)$ is coercive, we claim that on each unbounded U_i there must be some constants N_i and $\beta_i > 0$ so that for $\|y\| \geq N_i$ we have $\rho(y) \geq \beta_i\|y\|$. Otherwise, we can find an index set J such that $\rho(y_j) \leq \beta_j\|y_j\|$, where $\beta_j \downarrow 0$ and $\|y_j\| \uparrow \infty$. Without loss of generality, suppose $\frac{y_j}{\|y_j\|}$ converges to $\bar{y} \in U_i^\infty$, by [Rockafellar, 1970, Theorem 8.2]. By assumption, $\frac{\rho(y_j)}{\|y_j\|} \downarrow 0$, and we have

$$\frac{\rho(y_j)}{\|y_j\|} = \|y_j\| \left\langle \frac{y_j}{\|y_j\|}, A_i \frac{y_j}{\|y_j\|} \right\rangle + \left\langle a_i, \frac{y_j}{\|y_j\|} \right\rangle + \frac{\alpha_i}{\|y_j\|}.$$

Taking the limit of both sides over J we see that $\|y_j\| \left\langle \frac{y_j}{\|y_j\|}, A_i \frac{y_j}{\|y_j\|} \right\rangle$ must converge to a finite value. But this is only possible if $\langle \bar{y}, A_i \bar{y}\rangle = 0$, so in particular we must have $\bar{y} \in \text{Null}(A_i)$. Note also that $\langle a_i, \bar{y}\rangle \leq 0$, by taking the limit over J of

$$\frac{\rho(y_j)}{\|y_j\|} \geq \left\langle a_i, \frac{y_j}{\|y_j\|} \right\rangle + \frac{\alpha}{\|y_i\|},$$

so for any $x_0 \in U_i$ and $\lambda > 0$ we have $x_0 + \lambda\bar{y} \in U_i$ since $\bar{y} \in U_i^\infty$ and

$$\rho(x_0 + \lambda\bar{y}) \leq \rho(x_0) + \alpha_i,$$

so in particular ρ stays bounded as $\lambda \uparrow \infty$ and cannot be coercive.

The integrability of $f(y)$ is now clear. First note that $f(y)$ is bounded below by 0. Recall that the effective domain of ρ can be represented as the union of finitely many polyhedral sets U_i, and for each unbounded such U_i we have shown $f(y) \leq \exp[-\beta_i\|y\|]$ off of some bounded subset of U_i. An elementary application of the bounded convergence theorem shows that f must be integrable.

Proof of Theorem 2.6

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

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

$$\begin{aligned}
\rho(\tau\hat{y}) &= \sup_{u\in U}\langle b + \tau B\hat{y}, u\rangle - \frac{1}{2}u^\mathrm{T}Mu \\
&= \sup_{u\in U}\langle b, u\rangle - \frac{1}{2}u^\mathrm{T}Mu + \tau\langle B\hat{y}, u\rangle \\
&\leq \sup_{u\in U}\langle b, u\rangle - \frac{1}{2}u^\mathrm{T}Mu \\
&\leq \theta_{U,M}(b),
\end{aligned}$$

so $\rho(\tau\hat{y})$ stays bounded even as $\tau \to \infty$, 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 K so that $\rho(y_k) \leq K$ 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 $u \in U$

$$\begin{aligned}
&\langle b + By_k, u\rangle - \frac{1}{2}u^\mathrm{T}Mu \leq K \\
&\langle b + By_k, u\rangle \leq K + \frac{1}{2}u^\mathrm{T}Mu \\
&\langle \frac{b + By_k}{\|y_k\|}, u\rangle \leq \frac{K}{\|y_k\|} + \frac{1}{2\|y_k\|}u^\mathrm{T}Mu
\end{aligned}$$

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 $B\bar{y} \in (\text{cone } U)^\circ$, and so $\bar{y} \in B^{-1}((\text{cone } U)^\circ)$.

Proof of Lemma 3.2

The Lagrangian for (3.3) for feasible (x, u^w, u^v) is

$$\begin{aligned}
L(x, u^w, u^v) = &\left\langle \begin{bmatrix}\tilde{b}^w \\ \tilde{b}^v\end{bmatrix}, \begin{bmatrix}u^w \\ u^v\end{bmatrix} \right\rangle - \frac{1}{2}\begin{bmatrix}u^w \\ u^v\end{bmatrix}^\mathrm{T} \begin{bmatrix}M^w & 0 \\ 0 & M^v\end{bmatrix} \begin{bmatrix}u^w \\ u^v\end{bmatrix} \\
&- \left\langle \begin{bmatrix}u^w \\ u^v\end{bmatrix}, \begin{bmatrix}-B^w Q^{-1/2}G \\ B^v R^{-1/2}H\end{bmatrix} x \right\rangle
\end{aligned} \tag{4.1}$$

where $\tilde{b}^w = b^w - B^w Q^{-1/2}\tilde{x}_0$ and $\tilde{b}^v = b^v - B^v R^{-1/2}z$. The associated optimality conditions for feasible (x, u^w, u^v) are given by

$$\begin{aligned}
&G^\mathrm{T}Q^{-\mathrm{T}/2}(B^w)^\mathrm{T}\bar{u}^w - H^\mathrm{T}R^{-\mathrm{T}/2}(B^v)^\mathrm{T}\bar{u}^v = 0 \\
&\tilde{b}^w - M^w\bar{u}^w + B^w Q^{-1/2}G\bar{x} \in N_{U^w}(\bar{u}^w) \\
&\tilde{b}^v - M^v\bar{u}^v - B^v R^{-1/2}H\bar{x} \in N_{U^v}(\bar{u}^v)\ ,
\end{aligned} \tag{4.2}$$

where $N_C(x)$ denotes the normal cone to the set C at the point x (see Rockafellar [1970] for details).

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

$$N_U(\bar{u}) = \left\{ \begin{array}{r} q_1 A_1 + \cdots + q_m A_m \mid q_i \geq 0 \text{ for } i \in I(\bar{u}) \\ q_i = 0 \text{ for} i \notin I(\bar{u}) \end{array} \right\}. \tag{4.3}$$

Using (4.3), Then we may rewrite the optimality conditions (4.2) more explicitly as

$$\begin{array}{l} G^{\mathrm{T}} Q^{-\mathrm{T}/2} (B^w)^{\mathrm{T}} \bar{u}^w - H^{\mathrm{T}} R^{-\mathrm{T}/2} (B^v)^{\mathrm{T}} \bar{u}^v = 0 \\ \tilde{b}^w - M^w \bar{u}^w + B^w Q^{-1/2} G \bar{d} = A^w q^w \\ \tilde{b}^v - M^v \bar{u}^v - B^v R^{-1/2} H \bar{d} = A^v q^v \\ \{q^v \geq 0 | q_i^v = 0 \text{ for } i \notin I(\bar{u}^v)\} \\ \{q^w \geq 0 | q_i^w = 0 \text{ for } i \notin I(\bar{u}^w)\} \end{array} \tag{4.4}$$

Define slack variables $s^w \geq 0$ and $s^v \geq 0$ as follows:

$$\begin{aligned} s^w &= a^w - (A^w)^{\mathrm{T}} u^w \\ s^v &= a^v - (A^v)^{\mathrm{T}} u^v. \end{aligned} \tag{4.5}$$

Note that we know the entries of q_i^w and q_i^v are zero if and only if the corresponding slack variables s_i^v and s_i^w 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 (3.4).

4.1 Proof of Theorem 3.3

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

$$F_\mu \begin{pmatrix} s^w \\ s^v \\ q^w \\ q^v \\ u^w \\ u^v \\ x \end{pmatrix} = \begin{bmatrix} (A^w)^{\mathrm{T}} u^w + s^w - a^w \\ (A^v)^{\mathrm{T}} u^v + s^v - a^v \\ D(q^w) D(s^w) \mathbf{1} - \mu \mathbf{1} \\ D(q^v) D(s^v) \mathbf{1} - \mu \mathbf{1} \\ \tilde{b}^w + B^w Q^{-1/2} G d - M^w u^w - A^w q^w \\ \tilde{b}^v - B^v R^{-1/2} H d - M^v u^v - A^v q^v \\ G^{\mathrm{T}} Q^{-\mathrm{T}/2} (B^w)^{\mathrm{T}} u^w - H^{\mathrm{T}} R^{-\mathrm{T}/2} (B^v)^{\mathrm{T}} \bar{u}^v \end{bmatrix}.$$

This entails solving the system

$$F_\mu^{(1)} \begin{pmatrix} s^w \\ s^v \\ q^w \\ q^v \\ u^w \\ u^v \\ d \end{pmatrix} \begin{bmatrix} \Delta s^w \\ \Delta s^v \\ \Delta q^w \\ \Delta q^v \\ \Delta u^w \\ \Delta u^v \\ \Delta d \end{bmatrix} = -F_\mu \begin{pmatrix} s^w \\ s^v \\ q^w \\ q^v \\ u^w \\ u^v \\ d \end{pmatrix}, \tag{4.6}$$

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

$$\begin{bmatrix} I & 0 & 0 & 0 & (A^w)^{\mathrm{T}} & 0 & 0 \\ 0 & I & 0 & 0 & 0 & (A^v)^{\mathrm{T}} & 0 \\ Q^w & 0 & S^w & 0 & 0 & 0 & 0 \\ 0 & Q^v & 0 & S^v & 0 & 0 & 0 \\ 0 & 0 & -A^w & 0 & -M^w & 0 & B^w Q^{-1/2} G \\ 0 & 0 & 0 & -A^v & 0 & -M^v & -B^v R^{-1/2} H \\ 0 & 0 & 0 & 0 & G^{\mathrm{T}} Q^{-\mathrm{T}/2} (B^w)^{\mathrm{T}} & -H^{\mathrm{T}} R^{-\mathrm{T}/2} (B^v)^{\mathrm{T}} & 0 \end{bmatrix} \tag{4.7}$$

We now show the row operations necessary to reduce the matrix $F_\mu^{(1)}$ in (4.7) to upper block triangular form. After each operation, we show only the row that was modified.

$\text{row}_3 \leftarrow \text{row}_3 - D(q^w)\, \text{row}_1$

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

$\text{row}_4 \leftarrow \text{row}_4 - D(q^v)\, \text{row}_2$

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

$\text{row}_5 \leftarrow \text{row}_5 + A^w D(s^w)^{-1}\, \text{row}_3$

$\begin{bmatrix} 0 & 0 & 0 & 0 & -T^w & 0 & B^w Q^{-1/2} G \end{bmatrix}$

$\text{row}_6 \leftarrow \text{row}_6 + A^v D(s^v)^{-1}\, \text{row}_4$

$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & -T^v & -B^v R^{-1/2} H \end{bmatrix}$.

In the above expressions,

$$\begin{aligned} T^w &:= M^w + A^w (S^w)^{-1} Q^w (A^w)^{\mathrm{T}} \\ T^v &:= M^v + A^v (S^v)^{-1} Q^v (A^v)^{\mathrm{T}}, \end{aligned} \tag{4.8}$$

where $(S^w)^{-1} Q^w$ and $(S^v)^{-1} Q^v$ are always full-rank diagonal matrices, since the vectors s^w, q^w, s^v, q^v are always strictly positive in IP iterations. The invertibility of T^w and T^v is charachterized in the following lemma.

Lemma 4.10. (Invertibility of T) Let $\theta_{U,M}(\cdot)$ be any PLQ penalty on $\mathbb{R}^k$, with $U = \{u \big| A^{\mathrm{T}} u \leq a\}$. Let $T_\theta := M + ADA^{\mathrm{T}}$, where D is any diagonal $k \times k$ matrix with positive entries on the diagonal. Then T_θ is invertible if and only if $\mathrm{Null}(M) \cap U^\infty = \{0\}$, or $\mathrm{dom}(\theta_{U,M})$ is $\mathbb{R}^k$.

Proof: Note that

$$\begin{aligned} \mathrm{Null}(M + ADA^{\mathrm{T}}) &= \{w \big| w^{\mathrm{T}} M w + w^{\mathrm{T}} ADA^{\mathrm{T}} w = 0\} \\ &= \{w \big| w \in \mathrm{Null}(M)\ ,\ w \in \mathrm{Null}(A^{\mathrm{T}})\} \\ &= \mathrm{Null}(M) \cap \mathrm{Null}(A^{\mathrm{T}}). \end{aligned}$$

The first claim now follows from the fact that $U^\infty = \mathrm{Null}(A^{\mathrm{T}})$. Recall that the effective domain of θ is given by $(\mathrm{Null}(M) \cap U^\infty)^\circ$, and it is immediate from the definition of 'polar' that $0^\circ = \mathbb{R}^k$. ■

Remark 4.11. (Block diagonal structure of T in i.d. case) Suppose that $\mathbf{y}$ is a random vector, $\mathbf{y} = (\mathbf{y_1} \cdots \mathbf{y_n})$, where each $\mathbf{y_i}$ is itself a random vector in $\mathbb{R}^{m_i}$, from some PLQ density $\mathbf{p}(y_i) \propto \exp[-c_2 \theta_{U_i, M_i}((\cdot))]$, and all $\mathbf{y_i}$ are independent. Let $U_i = \{u : A_i^T u \leq a_i\}$. Then the matrix T_θ is given by $T_\theta = M + ADA^T$ where $M = \mathrm{diag}[M_1, \cdots, M_N]$, $A = \mathrm{diag}[A_1, \cdots, A_N]$, $D = \mathrm{diag}[D_1, \cdots, D_N]$, and $\{D_i\}$ are diagonal with positive entries. Moreover, T_θ is block diagonal, with ith diagonal block given by $M_i + A_i D_i A_i^{\mathrm{T}}$. ■

Corollary 4.12. (T matrices in the Kalman smoothing context) The matrices T^w and T^v in (4.8) are block diagonal provided that $\{w_k\}$ and $\{v_k\}$ are independent vectors from any PLQ densities. Moreover, if these densities all satisfy the characterization in Lemma 4.10, these matrices are also invertible. ■

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

$$\text{row}_7 \leftarrow \text{row}_7 + \Big(G^{\mathrm{T}} Q^{-\mathrm{T}/2} (B^w)^{\mathrm{T}} (T^w)^{-1}\Big) \text{row}_5 - \Big(H^{\mathrm{T}} R^{-\mathrm{T}/2} (B^v)^{\mathrm{T}} (T^v)^{-1}\Big) \text{row}_6$$

$$\begin{bmatrix} I & 0 & 0 & 0 & (A^w)^{\mathrm{T}} & 0 & 0 \\ 0 & I & 0 & 0 & 0 & (A^v)^{\mathrm{T}} & 0 \\ 0 & 0 & S^w & 0 & -Q^w (A^w)^{\mathrm{T}} & 0 & 0 \\ 0 & 0 & 0 & S^v & 0 & -Q^v (A^v)^{\mathrm{T}} & 0 \\ 0 & 0 & 0 & 0 & -T^w & 0 & B^w Q^{-1/2} G \\ 0 & 0 & 0 & 0 & 0 & -T^v & -B^v R^{-1/2} H \\ 0 & 0 & 0 & 0 & 0 & 0 & \Phi \end{bmatrix}$$

where

$$\Phi = \Phi_G + \Phi_H = G^{\mathrm{T}} Q^{-\mathrm{T}/2} (B^w)^{\mathrm{T}} (T^w)^{-1} B^w Q^{-1/2} G + H^{\mathrm{T}} R^{-\mathrm{T}/2} (B^v)^{\mathrm{T}} (T^v)^{-1} B^v R^{-1/2} H. \tag{4.9}$$

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

(1) Φ_H is block diagonal.
(2) $Q^{-\mathrm{T}/2}(B^w)^{\mathrm{T}}(T^w)^{-1}B^w Q^{-1/2}$ is block diagonal, and G is block bidiagonal, hence Φ_G is block tridiagonal.

Solving system (4.6) 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 $\Phi \Delta x = \varrho$. We have already seen that Φ is block tridiagonal, symmetric, and positive definite, so $\Phi \Delta x = \varrho$ can be solved in $O(Nn^3)$ time using the block tridiagonal algorithm in [Bell, 2000]. The remaining four back solves required to solve (4.6) can each be done in $O(nN)$ time.