Automatica 86 (2017) 63–86

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Survey paper

Generalized Kalman smoothing: Modeling and algorithms☆

Aleksandr Aravkin [a], James V. Burke [b], Lennart Ljung [c], Aurelie Lozano [d], Gianluigi Pillonetto [e]

[a] Department of Applied Mathematics, University of Washington, USA
[b] Department of Mathematics, University of Washington, Seattle, USA
[c] Division of Automatic Control, Linköping University, Linköping, Sweden
[d] IBM T.J. Watson Research Center Yorktown Heights, NY, USA
[e] Department of Information Engineering, University of Padova, Padova Italy

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 2 June 2017
Accepted 18 July 2017

ABSTRACT

State-space smoothing has found many applications in science and engineering. Under linear and Gaussian assumptions, smoothed estimates can be obtained using efficient recursions, for example Rauch–Tung–Striebel and Mayne–Fraser algorithms. Such schemes are equivalent to linear algebraic techniques that minimize a convex quadratic objective function with structure induced by the dynamic model.

These classical formulations fall short in many important circumstances. For instance, smoothers obtained using quadratic penalties can fail when outliers are present in the data, and cannot track impulsive inputs and abrupt state changes. Motivated by these shortcomings, generalized Kalman smoothing formulations have been proposed in the last few years, replacing quadratic models with more suitable, often nonsmooth, convex functions. In contrast to classical models, these general estimators require use of iterated algorithms, and these have received increased attention from control, signal processing, machine learning, and optimization communities.

In this survey we show that the optimization viewpoint provides the control and signal processing community great freedom in the development of novel modeling and inference frameworks for dynamical systems. We discuss general statistical models for dynamic systems, making full use of nonsmooth convex penalties and constraints, and providing links to important models in signal processing and machine learning. We also survey optimization techniques for these formulations, paying close attention to dynamic problem structure. Modeling concepts and algorithms are illustrated with numerical examples.

1. Introduction

The linear state space model

$$x_{t+1} = A_t x_t + B_t u_t + v_t \tag{1a}$$
$$y_t = C_t x_t + e_t \tag{1b}$$

is the bread and butter for analysis and design in discrete time systems, control and signal processing (Kalman, 1960; Kalman & Bucy, 1961). Application areas are numerous, including navigation, tracking, healthcare and finance, to name a few.

For a system model, $y_t \in \mathbb{R}^m$ and $u_t \in \mathbb{R}^p$ are, respectively, the output and input evaluated at the time instant t. The dimensions m and p may depend on t, but we treat them as fixed to simplify the exposition. In signal models, the input u_t may be absent. The state vectors $x_t \in \mathbb{R}^n$ are the variables of interest; A_t encodes the process transition, to the extent that it is known to the modeler, C_t is the observation model, and B_t describes the effect of the input on the transition. The *process disturbance* v_t models stochastic deviations from the linear model A_t, while e_t model *measurement errors*. We consider the *state estimation problem*, where the goal is to infer the values of x_t from the input–output measurements. Given measurements

$$\mathscr{Z}_0^N := \{u_0, y_1, u_1, y_2, \ldots, y_N, u_N\},$$

we are interested in obtaining an estimate $\hat{x}_t^N$ of x_t. If $N > t$ this is called a *smoothing* problem, if $N = t$ it is a *filtering* problem, and if $N < t$ it is a *prediction* problem.

The dimensions n and N vary with application. For many navigation examples, n is small; i.e. the state may have fewer than 20 elements at each time point. N, the number of time steps, can be in the thousands (e.g. when a sensor on a plant takes data at a high frequency or for a long time). In contrast, weather prediction

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Editor John Baillieul.

E-mail addresses: saravkin@uw.edu (A. Aravkin), burke@math.washington.edu (J.V. Burke), ljung@isy.liu.se (L. Ljung), aclozano@us.ibm.com (A. Lozano), giapi@dei.unipd.it (G. Pillonetto).

http://dx.doi.org/10.1016/j.automatica.2017.08.011
0005-1098/

models track a large state, and can have $n \gg 10^6$. The choice of method depends on the application; the survey is geared toward a small to moderate n.

How well the state estimate fits the true state depends upon the choice of models for the stochastic term v_t, error term e_t, and possibly on the initial distribution of x_0. While u_t is hereby seen as a known deterministic sequence, the observations y_t and states x_t are stochastic processes. We can consider using several estimators $\hat{x}_t^N$ of the state sequence $\{x_t\}$ (all functions of $\mathscr{Z}_0^N$):

$$E(x_t|\mathscr{Z}_0^N) \quad \text{conditional mean} \tag{2a}$$

$$\max_{x_t} \mathbf{p}(x_t|\mathscr{Z}_0^N) \quad \text{maximum } \textit{a posteriori}\text{(MAP)} \tag{2b}$$

$$\min_{\hat{x}_t} E(\|x_t - \hat{x}_t\|^2) \quad \text{minimum expected mean square error (MSE)} \tag{2c}$$

$$\min_{\hat{x}_t \in \text{span}\left(\mathscr{Z}_0^N\right)} E(\|x_t - \hat{x}_t\|^2) \quad \text{minimum linear expected MSE.} \tag{2d}$$

When e_t, v_t and the initial state x_0 are jointly Gaussian, all the four estimators coincide. In the general setting, the estimators (2a) and (2c) are the same. Indeed, the conditional mean represents the minimum variance estimate. In the general (non-Gaussian) case, computing (2a) may be difficult, while the MAP (2b) estimator can be computed efficiently using optimization techniques for a range of disturbance and error distributions.

Most models assume known means and variances for v_t, e_t, and x_0. In the classic settings, these distributions are Gaussian:

$$\begin{array}{ll} e_t & \sim \mathscr{N}(0, R_t) \\ v_t & \sim \mathscr{N}(0, Q_t), \\ x_0 & \sim \mathscr{N}(\mu, \Pi) \end{array} \qquad \text{all variables are mutually independent.} \tag{3}$$

Under this assumption, all the y_t and x_t become jointly Gaussian stochastic processes, which implies that the conditional mean (2a) becomes a linear function of the data $\mathscr{Z}_0^N$. This is a general property of Gaussian variables. Many explicit expressions and recursions for this linear filter have been derived in the literature, some of which are discussed in this article. We also consider a far more general setting, where the distributions in (3) can be selected from a range of densities, and discuss applications and general inference techniques.

We now make explicit the connection between *conditional mean* (2a) and *maximum a posteriori* (2b) in the Gaussian case. By Bayes' theorem and the independence assumptions (3), the posterior of the state sequence $\{x_t\}_{t=0}^N$ given the measurement sequence $\{y_t\}_{t=1}^N$ is

$$\begin{aligned} \mathbf{p}\left(\{x_t\}|\{y_t\}\right) &= \frac{\mathbf{p}\left(\{y_t\}|\{x_t\}\right)\mathbf{p}\left(\{x_t\}\right)}{\mathbf{p}\left(\{y_t\}\right)} \\ &= \frac{\mathbf{p}\left(x_0\right)\prod_{t=1}^N \mathbf{p}\left(y_t|x_t\right)\prod_{t=0}^{N-1}\mathbf{p}\left(x_{t+1}|x_t\right)}{\mathbf{p}\left(\{y_t\}\right)} \\ &\propto \mathbf{p}\left(x_0\right)\prod_{t=1}^N \mathbf{p}_{e_t}\left(y_t - C_t x_t\right)\prod_{t=0}^{N-1}\mathbf{p}_{v_t}\left(x_{t+1} - A_t x_t - B_t u_t\right), \end{aligned} \tag{4}$$

where we use $\mathbf{p}_{e_t}$ and $\mathbf{p}_{v_t}$ to denote the densities corresponding to e_t and v_t. Under Gaussian assumptions (3), and ignoring the normalizing constant, the posterior is given by

$$\begin{aligned} e^{-\frac{1}{2}\left\|\Pi^{-1/2}(x_0-\mu)\right\|^2}\prod_{t=0}^{N-1} e^{-\frac{1}{2}\left\|Q_t^{-1/2}(x_{t+1}-A_t x_t - B_t u_t)\right\|^2} \\ \times \prod_{t=1}^{N} e^{-\frac{1}{2}\left\|R_t^{-1/2}(y_t - C_t x_t)\right\|^2}. \end{aligned} \tag{5}$$

Note that state increments and measurement residuals appear explicitly in (5). Maximizing (5) is equivalent to minimizing its negative log:

$$\begin{aligned} \min_{x_0,\ldots,x_N} \left\|\Pi^{-1/2}(x_0-\mu)\right\|^2 + \sum_{t=1}^N \left\|R_t^{-1/2}(y_t - C_t x_t)\right\|^2 \\ + \sum_{t=0}^{N-1}\left\|Q_t^{-1/2}(x_{t+1} - A_t x_t - B_t u_t)\right\|^2. \end{aligned} \tag{6}$$

More general cases of correlated noise and singular covariance matrices are discussed in Appendix. This result is also shown in e.g. Bell (1994) and Kailath, Sayed, and Hassibi (2000, Sec. 3.5, 10.6) using a least squares argument. The solution can be derived using various structure-exploiting linear recursions. For instance, the Rauch–Tung–Striebel (RTS) scheme derived in Rauch, Tung, and Striebel (1965) computes the state estimates by forward–backward recursions, see also Ansley and Kohn (1982) for a simple derivation through projections onto spaces spanned by suitable random variables. The Mayne–Fraser (MF) algorithm uses a two-filter formula to compute the smoothed estimate as a linear combination of forward and backward Kalman filtering estimates (Fraser & Potter, 1969; Mayne, 1966). The nature of this recursion was clarified in Badawi, Lindquist, and Pavon (1975) through the concept of maximum-variance forward filter. A third scheme based on reverse recursion appears in Mayne (1966) under the name of Algorithm A. The relationships between these schemes, and their derivations from different perspectives are studied in Aravkin, Bell, Burke, and Pillonetto (2013) and Ljung and Kailath (1976). See also Chapter 15 in Lindquist and Picci (2015) for insights on how various smoothing formulas derive from different choices of coordinates in the frame space. Computational details for RTS and MF are presented in Section 2.

The *maximum a posteriori* (MAP) viewpoint (6) easily generalizes to new settings. Assume, for example, that the noises e_t and v_t are non-Gaussian, but rather have continuous probability densities defined by functions $V_t(\cdot)$ and $J_t(\cdot)$ as follows

$$\mathbf{p}_{e_t}(e) \propto \exp\left(-V_t\left(R_t^{-1/2}e\right)\right), \mathbf{p}_{v_t}(v) \propto \exp\left(-J_t\left(Q_t^{-1/2}v\right)\right). \tag{7}$$

From (4), we obtain that the analogous MAP estimation problem for (6) replaces all least squares $\|R_t^{-1/2}(y_t - C_t x_t)\|^2$ and $\|Q_t^{-1/2}(x_{t+1} - A_t x_t - B_t u_t)\|^2$ with more general terms $V_t\left(R_t^{-1/2}(y_t - C_t x_t)\right)$ and $J_t\left(Q_t^{-1/2}(x_{t+1} - A_t x_t - B_t u_t)\right)$, leading to

$$\begin{aligned} \min_{x_0,\ldots,x_N} -\log \mathbf{p}(x_0) + \sum_{t=1}^N V_t\left(R_t^{-1/2}(y_t - C_t x_t)\right) \\ + \sum_{t=0}^{N-1} J_t\left(Q_t^{-1/2}(x_{t+1} - A_t x_t - B_t u_t)\right). \end{aligned} \tag{8}$$

The initial distribution for x_0 can be non-Gaussian, and is specified by $\mathbf{p}(x_0)$. An algorithm to solve (8) is then required. In this paper, we will discuss general modeling of error distributions $\mathbf{p}_{e_t}$ and $\mathbf{p}_{v_t}$ in (7), as well as tractable algorithms for the solutions of these formulations.

Classic *Kalman* filters, predictors and smoothers have been enormously successful, and the literature detailing their properties and applications is rich and pervasive. Even if Gaussian assumptions (3) are violated, but the v_t, e_t are still white with covariances Q_t and R_t, problem (6) gives the best *linear* estimate, i.e. among all linear functions of the data $\mathscr{Z}_0^N$, the Kalman smoother residual has the smallest variance. However, this does not ensure successful performance, giving strong motivation to consider extensions to the Gaussian framework! For instance, impulsive disturbances often occur in process models, including target tracking, where one has to deal with force disturbances describing maneuvers for the tracked object, fault detection/isolation, where impulses model

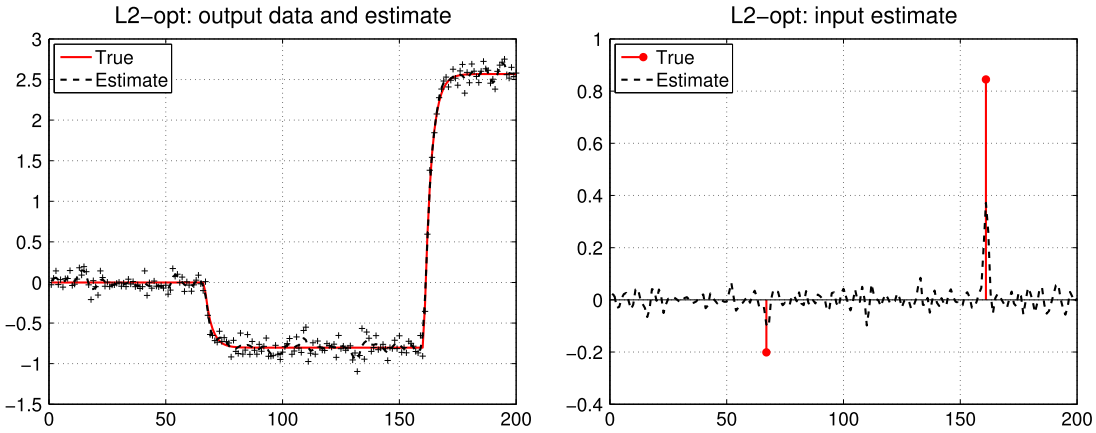


Fig. 1. DC motor and impulsive disturbances. *Left*: noiseless output (solid line), measurements (+) and output reconstruction by the optimal linear smoother L_2-opt (dashed line). *Right*: impulsive disturbance and reconstruction by L_2-opt (dashed line).

additive faults, and load disturbances. Unfortunately, smoothers that use the quadratic penalty on the state increments are not able to follow fast jumps in the state dynamics (Ohlsson, Gustafsson, Ljung, & Boyd, 2012). This problem is also relevant in the context of identification of switched linear regression models where the system states can be seen as time varying parameters which can be subject to abrupt changes (Niedzwiecki & Gackowski, 2013; Ohlsson & Ljung, 2013). In addition, constraints on the states arise naturally in many settings, and estimation can be improved by taking these constraints into account. Finally, estimates corresponding to quadratic losses applied to data misfit residuals are vulnerable to outliers, i.e. to unexpected deviations of the noise errors from Gaussian assumptions. In these cases, a Gaussian model for e gives poor estimates. Two examples are described below, the first focusing on impulsive disturbances, and second on measurement outliers.

1.1. DC motor example

A DC motor can be modeled as a dynamic system, where the input is applied torque while the output is the angle of the motor shaft, see also pp. 95–97 in Ljung (1999). The state comprises angular velocity and angle of the motor shaft, and with system parameters and discretization as in Section 8 of Ohlsson et al. (2012), we have the following discrete-time model:

$$\begin{aligned} x_{t+1} &= \begin{pmatrix} 0.7 & 0 \\ 0.084 & 1 \end{pmatrix} x_t + \begin{pmatrix} 11.81 \\ 0.62 \end{pmatrix} (u_t + d_t) \\ y_t &= \begin{pmatrix} 0 & 1 \end{pmatrix} x_t + e_t \end{aligned} \tag{9}$$

where d_t denotes a disturbance process while the measurements y_t are noisy samples of the angle of the motor shaft.

Impulsive inputs: In the DC motor system design, the disturbance torque acting on the motor shaft plays an important role and an accurate reconstruction of d_t can greatly improve model robustness with respect to load variations. Since the non observable input is often impulsive, we model the d_t as independent random variables such that

$$d_t = \begin{cases} 0 & \text{with probability } 1-\alpha \\ \mathcal{N}(0,1) & \text{with probability } \alpha. \end{cases}$$

According to (1), this corresponds to a zero-mean (non-Gaussian) noise v_t, with covariance $Q_t = \alpha \binom{11.81}{0.62}(11.81,\ 0.62)$. We consider the problem of reconstructing d_t from noisy output samples generated under the assumptions

$$x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ u_t = 0, \quad \alpha = 0.01, \quad e_t \sim \mathcal{N}(0, 0.1^2).$$

An instance of the problem is shown in Fig. 1. The left panel displays the noiseless output (solid line) and the measurements (+). The right panel displays the d_t (solid line) and their estimates (dashed line) obtained by the Kalman smoother[1] and given by $\hat{d}_t^N = (1/11.81\ \ 0)\left(\hat{x}_{t+1}^N - A_t \hat{x}_{t+1}^N\right)$.

This estimator, denoted L_2-opt, uses only information on the means and covariances of the noises. It solves problem (2d) and, hence, corresponds to the best linear estimator. However, it is apparent that the disturbance reconstruction is not satisfactory. The smoother estimates of the impulses are poor, and the largest peak, centered at $t = 161$, is highly underestimated.

Outliers corrupting output data: Consider now a situation where the disturbance d_t can be well modeled as a Gaussian process. So, there is no impulsive noise entering the system. In particular, we set $d_t \sim \mathcal{N}(0, 0.1^2)$, so that v_t is now Gaussian with covariance

$$Q_t = 0.1^2 \begin{pmatrix} 11.81 \\ 0.62 \end{pmatrix} \begin{pmatrix} 11.81 & 0.62 \end{pmatrix}.$$

The outputs y_t are instead contaminated by outliers, i.e. unexpected measurements noise model deviations. In particular, output data are corrupted by a mixture of two normals with a fraction of outliers contamination equal to $\alpha = 0.1$; i.e.,

$$e_t \sim (1-\alpha)\mathcal{N}(0, \sigma^2) + \alpha \mathcal{N}(0, (100\sigma)^2).$$

Thus, outliers occur with probability 0.1, and are generated from a distribution with standard deviation 100 times greater than that of the nominal. We consider the problem of reconstructing the angle of the motor shaft (the second state component which corresponds to the noiseless output) setting

$$x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},\ u_t = 0,\ \sigma^2 = 0.1^2.$$

An instance of the problem is shown in Fig. 2. The two panels display the noiseless output (solid line), the accurate measurements affected by the noise with nominal variance (denoted by +) and the outliers (denoted by ∘ with values outside the range ±6 displayed on the boundaries of the panel). The left panel displays the estimate (dashed line) obtained by the classical Kalman smoother, called L_2-nom, with the variance noise set to σ^2.

[1] Note that the covariance matrices Q_t are singular. In this case, the smoothed estimates have been computed using the RTS scheme (Rauch et al., 1965), as e.g. described in Section 2.C of Kitagawa and Gersch (1985), where invertibility of the transition covariance matrices are not required. This scheme provides the solution of the generalized Kalman smoothing objective (47), and is explained in Appendix.

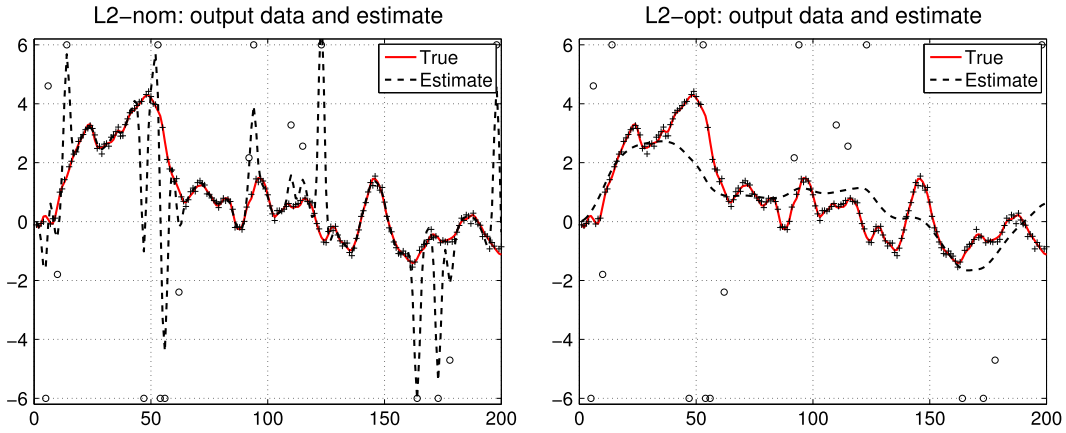


Fig. 2. DC motor with Gaussian disturbances and outliers in output measurements. Noiseless output (solid line), measurements (+) and outliers (○). *Left*: Kalman estimates (dashed line) with assumed nominal measurement error variance (0.01). *Right*: Kalman estimates (dashed line) from the optimal linear smoother which uses the correct measurement error variance (10.009).

Note that this estimator does not match any of the criteria (2a)–. In fact, this example represents a situation where the contamination is totally unexpected and the smoother is expected to work under nominal conditions. One can see that the reconstructed profile is very sensitive to outliers. The right panel shows the estimate (dashed line) returned by the optimal linear estimator L_2-opt (2d), obtained by setting the noise variance to $(1-\alpha)\sigma^2+\alpha(100\sigma)^2$.

In this case, the smoother is aware of the true variance of the signal; nonetheless, the reconstruction is still not satisfactory, since it cannot track the true output profile given the high measurement variance; the best linear estimate essentially averages the signal. Manipulating noise statistics is clearly not enough; to improve the estimator performance, we must change our model for the underlying distribution of the errors e_t.

1.2. Scope of the survey

In light of this discussion and examples, it is natural to turn to the optimization (MAP) interpretation (6) to design formulations and estimators that perform well in alternative and more general situations. The connection between numerical analysis and optimization and various kinds of smoothers has been growing stronger over the years (Aravkin, Bell et al., 2013; Bell & Cathey, 1993; Ljung & Kailath, 1976; Paige & Saunders, 1977). It is now clear that many popular algorithms in the engineering literature, including Rauch–Tung–Striebel (RTS) smoother and the Mayne–Fraser (MF) smoother, can be viewed as specific linear algebraic techniques to solve an optimization objective whose structure is closely tied to dynamic inference. Indeed, recently, Kalman smoothing has seen a remarkable renewal in terms of modern techniques and extended formulations based on emerging practical needs. This resurgence has been coupled with the development of new computational techniques and the intense progress in convex optimization in the last two decades has led to a vast literature on finding good state estimates in these more general cases. Many novel contributions to theory and algorithms related to Kalman smoothing, and to dynamic system inference in general, have come from statistics, engineering, and numerical analysis/optimization communities. However, while the statistical and engineering viewpoints are pervasive in the literature, the optimization viewpoint and its accompanying modeling and computational power is less familiar to the control community. Nonetheless, the optimization perspective has been the source of a wide range of astonishing recent advances across the board in signal processing, control, machine learning, and large-scale data analysis. In this survey, we will show how the optimization viewpoint allows the control and signal processing community great freedom in the development of novel modeling and inference frameworks for dynamical systems.

Recent approaches in dynamic systems inference replace quadratic terms, as in (6), with suitable convex functions, as in (8). In particular, new smoothing schemes deal with sparse dynamic models (Angelosante, Roumeliotis, & Giannakis, 2009), methods for tracking abrupt changes (Ohlsson et al., 2012), robust formulations (Aravkin, Bell, Burke, & Pillonetto, 2011; Farahmand, Giannakis, & Angelosante, 2011), inequality constraints on the state (Bell, Burke, & Pillonetto, 2009), and sum of norms (Ohlsson et al., 2012), many of which can be modeled using the general class called piecewise linear quadratic (PLQ) penalties (Aravkin, Burke, & Pillonetto, 2013b; Rockafellar & Wets, 1998). All of these approaches are based on an underlying body of theory and methodological tools developed in statistics, machine learning, kernel methods (Bottou, Chapelle, DeCoste, & Weston, 2007; Chan, Liao, & Tsui, 2011; Hofmann, Schölkopf, & Smola, 2008; Schölkopf & Smola, 2001), and convex optimization (Boyd & Vandenberghe, 2004). Advances in sparse tracking (Angelosante et al., 2009; Kim, Koh, Boyd, & Gorinevsky, 2009; Ohlsson et al., 2012) are based on LASSO (group LASSO) or elastic net techniques (Efron, Hastie, Johnstone, & Tibshirani, 2004; Tibshirani, 1996; Yuan & Lin, 2006; Zou & Hastie, 2005), which in turn use coordinate descent, see e.g. Bertsekas (1999), Dinuzzo (2011) and Friedman, Hastie, and Tibshirani (2010). Robust methods (Agamennoni, Nieto, & Nebot, 2011; Aravkin et al., 2011, 2013b; Chang, Hu, Chang, & Li, 2013; Farahmand et al., 2011) rely on Huber (Huber & Ronchetti, 2009) or Vapnik losses, leading to support vector regression (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997; Gunter & Zhu, 2007; Ho & Lin, 2012) for state space models, and take advantage of interior point optimization methods (Kojima, Megiddo, Noma, & Yoshise, 1991; Nemirovskii & Nesterov, 1994; Wright, 1997). Domain constraints are important for most applications, including camera tracking, fault diagnosis, chemical processes, vision-based systems, target tracking, biomedical systems, robotics, and navigation (Haykin, 2001; Simon, 2010). Modeling these constraints allows *a priori* information to be encoded into dynamic inference formulations, and the resulting optimization problems can also be solved using interior point methods (Bell et al., 2009).

Taking these developments into consideration, the aims of this survey are as follows. First, our goal is to firmly establish the connection between classical algorithms, including the RTS and MF smoothers, and the optimization perspective in the least squares case. This allows the community to view existing efficient algorithms as modular subroutines that can be exploited in new formulations. Second, we will survey modern regression

approaches from statistics and machine learning, based on new convex losses and penalties, highlighting their usefulness in the context of dynamic inference. These techniques are effective both in designing models for process disturbances v_t as well as robust statistical models for measurement errors e_t. Our final goal is two-fold: we want to survey algorithms for generalized smoothing formulations, but also to understand the theoretical underpinnings for the design and analysis of such algorithms. To this end, we include a self-contained tutorial of convex analysis, developing concepts of duality and optimality conditions from fundamental principles, and focused on the general Kalman smoothing context. With this foundation, we review optimization techniques to solve all general formulations of Kalman smoothers, including both first-order splitting methods, and second order (interior point) methods.

In many applications, process and measurement models may be nonlinear. These cases fall outside the scope of the current survey, since they require solving a nonconvex problem. In these cases, particle filters (Arulampalam, Maskell, Gordon, & Clapp, 2002) and unscented methods (Wan & Van Der Merwe, 2000) are very popular. An alternative is to exploit the composite structure of these problems, and apply a generalized Gauss–Newton method (Burke & Ferris, 1995). For detailed examples, see Aravkin et al. (2011) and Aravkin, Burke, and Pillonetto (2014).

Roadmap of the paper: In Section 2, we show the explicit connection between RTS and MF smoothers and the least squares formulation. This builds the foundation for efficient general methods that exploit underlying state space structure of dynamic inference. In Section 3, we present a general modeling framework where error distributions (3) can come from a large class of log-concave densities, and discuss important applications to impulsive disturbances and robust smoothing. We also show how to incorporate state-space constraints. In Section 4 basic techniques from optimization are presented, some of them possibly novel to the reader. Direct connections between the tutorial elements in this section and the smoothing problem are stressed. In Section 5, we present empirical results for the examples in the paper, showing the practical effect of the proposed methods. All examples are implemented using an open source software package IPsolve.[2] A few concluding remarks end the paper. Two appendices are provided. The first discusses smoothing under correlated noise and singular covariance matrices, and the second a brief tutorial on the tools from convex analysis that are useful to understand the algorithms presented in Section 4 and applied in Section 5.

2. Kalman smoothing, block tridiagonal systems of equations and classical schemes

To build an explicit correspondence between least squares problems and classical smoothing schemes, we first introduce data structures that explicitly embed the entire state sequence, measurement sequence, covariance matrices, and initial conditions into a simple form. Given a sequence of column vectors $\{v_k\}$ and matrices $\{T_k\}$ let

$$\operatorname{vec}(\{v_k\}) := \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}, \quad \operatorname{diag}(\{T_k\}) := \begin{bmatrix} T_1 & 0 & \cdots & 0 \\ 0 & T_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & T_N \end{bmatrix}.$$

[2] https://github.com/saravkin/IPsolve.

We make the following definitions:

$$\begin{aligned} R &:= \operatorname{diag}(\{R_1, R_2, \ldots, R_N\}) \in \mathbb{R}^{mN\times mN} \\ Q &:= \operatorname{diag}(\{\Pi, Q_0, Q_1, \ldots, Q_{N-1}\}) \in \mathbb{R}^{n(N+1)\times n(N+1)} \\ x &:= \operatorname{vec}(\{x_0, x_1, x_2, \ldots, x_N\}) \in \mathbb{R}^{n(N+1)\times 1} \\ y &:= \operatorname{vec}(\{y_1, y_2, \ldots, y_N\}) \in \mathbb{R}^{mN\times 1} \\ z &:= \operatorname{vec}(\{\mu, B_0u_0, \ldots, B_{N-1}u_{N-1}\}) \in \mathbb{R}^{n(N+1)\times 1} \end{aligned} \tag{10}$$

and

$$A := \begin{bmatrix} I & 0 & & \\ -A_0 & I & \ddots & \\ & \ddots & \ddots & 0 \\ & & -A_{N-1} & I \end{bmatrix}, \quad C := \begin{bmatrix} 0 & C_1 & 0 & \cdots & 0 \\ 0 & 0 & C_2 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & C_N \end{bmatrix}, \tag{11}$$

where $A \in \mathbb{R}^{n(N+1)\times n(N+1)}$ and $C \in \mathbb{R}^{mN\times n(N+1)}$. Using definitions (10) and (11) problem (6) can be efficiently stated as

$$\min_x \left\|R^{-1/2}(y - Cx)\right\|^2 + \left\|Q^{-1/2}(z - Ax)\right\|^2 . \tag{12}$$

The solution to (12) can be obtained by solving the linear system of equations

$$(C^\top R^{-1}C + A^\top Q^{-1}A)x = r \tag{13}$$

where

$$r := C^\top R^{-1}y + A^\top Q^{-1}z .$$

The linear operator in (13) is positive definite symmetric block tridiagonal (SBT). Direct computation gives

$$C^\top R^{-1}C + A^\top Q^{-1}A = \begin{bmatrix} F_0 & G_0^\top & 0 & \cdots & 0 \\ G_0 & F_1 & G_1^\top & \cdots & \vdots \\ \vdots & G_1 & \ddots & \ddots & \vdots \\ 0 & \cdots & \ddots & & G_{N-1}^T \\ 0 & \cdots & 0 & G_{N-1} & F_N \end{bmatrix},$$

a symmetric positive definite block tridiagonal system of equations in $\mathbb{R}^{n(N+1)\times n(N+1)}$, with $F_t \in \mathbb{R}^{n\times n}$ and $G_t \in \mathbb{R}^{n\times n}$ defined as follows:

$$\begin{aligned} F_0 &:= \Pi^{-1} + A_0^\top Q_0^{-1}A_0 \\ F_t &:= Q_{t-1}^{-1} + A_t^\top Q_t^{-1}A_t + C_t^\top R_t^{-1}C_t, \quad t = 1, \ldots, N \\ G_t &:= -Q_t^{-1}A_t, \quad t = 0, \ldots, N-1 \end{aligned}$$

using the convention $A_N^\top Q_N^{-1}A_N = 0$.

We now present two popular smoothing schemes, the RTS and MF. In our algebraic framework, both of them return the solution of the Kalman smoothing problem (12) by efficiently solving the block tridiagonal system (13), which can be rewritten as

$$\begin{pmatrix} F_0 & G_0^\top & 0 & \cdots & 0 \\ G_0 & F_1 & G_1^\top & \cdots & \vdots \\ \vdots & G_1 & \ddots & \ddots & \vdots \\ 0 & \cdots & \ddots & & G_{N-1}^T \\ 0 & \cdots & 0 & G_{N-1} & F_N \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \\ x_N \end{pmatrix} = \begin{pmatrix} r_0 \\ r_1 \\ \vdots \\ r_{N-1} \\ r_N \end{pmatrix}. \tag{14}$$

In particular, the RTS scheme coincides with the forward–backward algorithm as described in Bell (2000, algorithm 4), see also Bell and Cathey (1993). The MF scheme can be seen as a block tridiagonal solver which uses elements of both forward and backward algorithms so that it can exploit two filters running in parallel (Aravkin, Bell et al., 2013 Section 7). In Aravkin, Bell et al. (2013) one can find full numerical analysis of these and also other smoothing algorithms.

Algorithm 1 Rauch–Tung–Striebel (Forward Block Tridiagonal scheme)

The inputs to this algorithm are $\{G_t\}_{t=0}^{N-1}$, $\{F_t\}_{t=0}^{N}$, and $\{r_t\}_{t=0}^{N}$ where, for each t, $G_t \in \mathbb{R}^{n\times n}$, $F_t \in \mathbb{R}^{n\times n}$, and $r_t \in \mathbb{R}^m$. The output is the sequence $\{\hat{x}_t^N\}_{t=0}^{N}$ that solves equation (14), with each $\hat{x}_t^N \in \mathbb{R}^n$.

(1) Set $d_0^f = F_0$ and $s_0^f = r_0$.
For $t = 1$ to N :
- Set $d_t^f = F_t - G_{t-1}(d_{t-1}^f)^{-1}G_{t-1}^\top$.
- Set $s_t^f = r_t - G_{t-1}(d_{t-1}^f)^{-1}s_{t-1}^f$.

(2) Set $\hat{x}_N^N = (d_N^f)^{-1}s_N$.
For $t = N-1$ to 0 :
- Set $\hat{x}_t^N = (d_t^f)^{-1}(s_t^f - G_t^\top \hat{x}_{t+1}^N)$.

Algorithm 2 Mayne–Fraser (Two Filter Block Tridiagonal scheme)

The inputs to this algorithm are $\{G_t\}_{t=0}^{N-1}$, $\{F_t\}_{t=0}^{N}$, and $\{r_t\}_{t=0}^{N}$ where, for each t, $G_t \in \mathbb{R}^{n\times n}$, $F_t \in \mathbb{R}^{n\times n}$, and $r_t \in \mathbb{R}^m$. The output is the sequence $\{\hat{x}_t^N\}_{t=0}^{N}$ that solves equation (14), with each $\hat{x}_t^N \in \mathbb{R}^n$.

(1) Set $d_0^f = F_0$ and $s_0^f = r_0$.
For $t = 1$ to N :
- Set $d_t^f = F_t - G_{t-1}(d_{t-1}^f)^{-1}G_{t-1}^\top$.
- Set $s_t^f = r_t - G_{t-1}(d_{t-1}^f)^{-1}s_{t-1}$.

(2) Set $d_N^b = F_N$ and $s_N^b = r_N$.
For $t = N-1, \ldots, 0$,
- Set $d_t^b = F_t - G_t^\top(d_{t+1}^b)^{-1}G_t$.
- Set $s_t^b = r_t - G_t^\top(d_{t+1}^b)^{-1}s_{t+1}^b$.

(3) For $t = 0, 1, \ldots, N$
- Set $\hat{x}_t^N = (d_t^f + d_t^b - F_t)^{-1}(s_t^f + s_t^b - r_t)$.

3. General formulations: convex losses and penalties, and statistical properties of the resulting estimators

In the previous section, we showed that Gaussian assumptions on process disturbances v_t and measurement errors e_t lead to least squares formulations (6) or (12). One can then view classic smoothing algorithms as numerical subroutines for solving these least squares problems. In this section, we generalize the Kalman smoothing model to allow log-concave distributions for v_t and e_t in model (3). This allows more general *convex* disturbance and error measurement models, and the log-likelihood (MAP) problem (12) becomes a more general *convex* inference problem.

In particular, we consider the following general convex formulation:

$$\min_{x\in\mathscr{X}} V\left(R^{-1/2}(y - Cx)\right) + \gamma J\left(Q^{-1/2}(z - Ax)\right) \tag{15}$$

where $x \in \mathscr{X}$ specifies a feasible domain for the state, $V : \mathbb{R}^{mN} \to \mathbb{R}$ measures the discrepancy between observed and predicted data (due to noise and outliers), while $J : \mathbb{R}^{n(N+1)} \to \mathbb{R}$ measures the discrepancies between predicted and observed state transitions, due to the net effect of factors outside the process model; we can think of these discrepancies as 'process noise'. The structure of this problem is related to Tikhonov regularization and inverse problems (Bertero, 1989; Tikhonov & Arsenin, 1977; Vito, Rosasco, Caponnetto, De Giovannini, & Odone, 2005). In this context, γ is called the *regularization parameter* and has a link to the (typically unknown) scaling of the pdfs of e_t and v_t in (7). The choice of γ controls the tradeoff between bias and variance, and it has to be tuned from data. Popular tuning methods include cross-validation or generalized cross-validation (Golub, Heath, & Wahba, 1979; Hastie, Tibshirani, & Friedman, 2001; Rice, 1986).

Problem (15) is overly general. In practice we restrict V and J to be functions following the block structure of their arguments, i.e. sums of terms $V_t\left(R_t^{-1/2}(y_t - C_t x_t)\right)$ and $J_t\left(Q_t^{-1/2}(x_{t+1} - A_t x_t - B_t u_t)\right)$, leading to the objective already reported in (8). The terms $V_t : \mathbb{R}^m \to \mathbb{R}$ and $J_t : \mathbb{R}^n \to \mathbb{R}$ can then be linked to the MAP interpretation of the state estimate (7)–(8), so that V_t is a version of $-\log \mathbf{p}_{e_t}$ and J_t is a version of $-\log \mathbf{p}_{v_t}$. Possible choices for such terms are depicted in Figs. 4(a)–4(f) and 5.

Domain constraints $x \in \mathscr{X}$ provide a disciplined framework for incorporating prior information into the inference problem, which improves performance for a wide range of applications. To complement this, general J and V allow the modeler to incorporate information about *uncertainty*, both in the process and measurements. This freedom in designing (15) has numerous benefits. The modeler can choose J to reflect prior knowledge on the structure of the process noise; important examples include sparsity (see Fig. 1) and smoothness. In addition, she can robustify the formulation in the presence of outliers or non-gaussian errors (see Fig. 2), by selecting penalties V that perform well in spite of data contamination. To illustrate, we present specific choices for the functions V and J and explain how they can be used in a range of modeling scenarios; we also highlight the potential for constrained formulations.

3.1. General functions J for modeling process noise

As mentioned in the introduction, a widely used assumption for the process noise is that it is Gaussian. This yields the quadratic loss $\|Q^{-1/2}(z - Ax)\|^2$. However, in many applications prior knowledge on the process disturbance dictates alternative loss functions. A simple example is the DC motor in Section 1.1. We assumed that the process disturbance v_t is impulsive. One therefore expects that the disturbance v_t should be zero most of the time, while taking non-zero values at a few unknown time points. If each v_t is scalar, a natural way to regulate the number of non-zero components in vec $(\{v_t\})$ is to use the ℓ_0 norm for J in (15):

$$J(z - Ax; Q) = \|Q^{-1/2}(z - Ax)\|_0,$$

where $\|z\|_0$ counts the number of nonzero elements of z.

Sparsity promotion via ℓ_1 norm. The ℓ_0 norm, however, is non-convex, and solving optimization problems involving the ℓ_0 norm is NP-hard (combinatorial). Tractable approaches can be designed by replacing the ℓ_0 norm with a convex relaxation, the ℓ_1 norm, $\|x\|_1 = \sum|x_i|$. The ℓ_1 norm is nonsmooth and encourages sparsity, see Fig. 4(b). The use of the ℓ_1 norm in lieu of the ℓ_0 norm is now common practice, especially in compressed sensing (Candès & Tao, 2006; Donoho, 2006) and statistical learning, see e.g. Hastie et al. (2001). The reader can gain some intuition by considering the intersection of a general hyperplane with the ℓ_1 ball and ℓ_2 ball in Fig. 3. The intersection is likely to land on a corner, which means that adding a ℓ_1 norm constraint (or penalty) tends to select solutions with many zero elements.

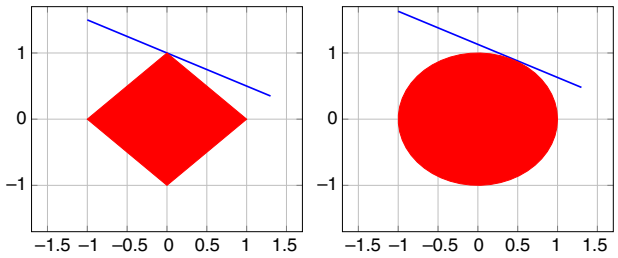


Fig. 3. When minimizing $\|Ax - b\|$ subject to a ℓ_1-norm constraint (left panel), the solution tends to land on a corner, where many coordinates are 0; in 2D the cartoon, the x-coordinate is zero. An ℓ_2-norm constraint (right panel) does not have this effect.

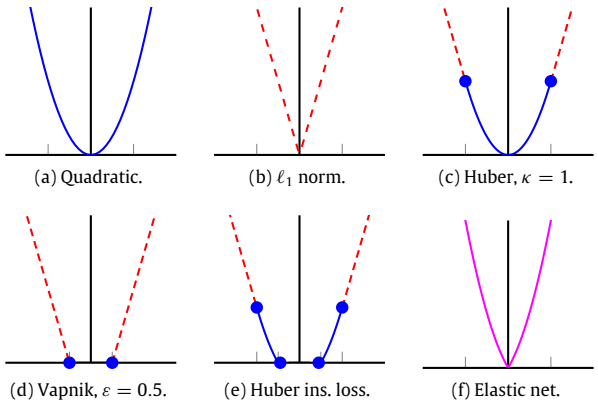


Fig. 4. Important penalties for errors and process models.

For the case of scalar-valued process disturbance v_t, we can set J to be the ℓ_1 norm and obtain the problem

$$\min_x \frac{1}{2}\|R^{-1/2}(y - Cx)\|^2 + \gamma\|Q^{-1/2}(z - Ax)\|_1, \tag{16}$$

where γ is a penalty parameter controlling the tradeoff between measurement fit and number of non-zero components in process disturbance—larger γ implies a larger number of zero process disturbance elements, at the cost of increasing the bias of the estimator.

Note that the vector norms in (16) translate to term-wise norms of the time components as in (8). Problem (16) is analogous to the LASSO problem (Tibshirani, 1996), originally proposed in the context of linear regression. Indeed, the LASSO problem minimizes the sum of squared residuals regularized by the ℓ_1 penalty on the regression coefficients. The LASSO estimates can be interpreted as the Bayes posterior mode under independent Laplace priors for the coefficients (Kyung, Gill, Ghosh, Casella, et al., 2010). In the context of regression, the LASSO has been shown to have strong statistical guarantees, including prediction error consistency (Van de Geer & Buhlmann, 2009), consistency of the parameter estimates in ℓ_2 or some other norm (Meinshausen & Yu, 2009; Van de Geer & Buhlmann, 2009), as well as variable selection consistency (Meinshausen & Bühlmann, 2006; Wainwright, 2009; Zhao & Yu, 2006). However, this connection is limited in the dynamic context: if we think of Kalman smoothing as linear regression, note from (16) that the measurement vector y is a *single* observation of the parameter (state sequence) x, so asymptotic consistency results are not relevant. More important is the general idea of using the ℓ_1 norm to promote sparsity of the right object, in this case, the residual $Q^{-1/2}(z - Ax)$, which corresponds to our model of impulsive disturbances.

Elastic net penalty. Suppose we need a penalty that is nonsmooth at the origin, but has quadratic growth in the tails. For example, taking J with these properties is useful in the context of our model for impulsive disturbances, if we believed them to be sparse, and also considered large disturbances unlikely. The *elastic net* shown in Fig. 4(f) has these properties—it is a weighted sum $\alpha\|\cdot\|_1 + (1 - \alpha)\|\cdot\|_2^2$. The elastic net penalty has been widely used for sparse regularization with correlated predictors (De Mol, De Vito, & Rosasco, 2009; Li, Lin, et al., 2010; Zou & Hastie, 2005). Using an elastic net constraint has a *grouping effect* (Zou & Hastie, 2005). Specifically, when minimizing $\frac{1}{2}\|Ax - b\|^2$ with an elastic net constraint, the distance between estimates $\hat{x}_i$ and $\hat{x}_j$ is proportional to $\sqrt{1 - \kappa_{ij}}$, where κ_{ij} is the correlation between the corresponding columns of A. In our context, in case of nearly perfectly correlated impulsive disturbances (either all present or all absent), the elastic net can discover the entire group, while the ℓ_1 norm alone usually picks a single member of the group.

Group sparsity. If the process disturbance is known to be grouped (e.g. a disturbance vector is always present or absent for each time point), $J(\cdot)$ can be set to the mixed $\ell_{2,1}$ norm, where the ℓ_2 norm is applied to each block of $Q_t^{-1/2}(z_t - A_t x_t)$, yielding the following Kalman smoothing formulation:

$$\min_x \|R^{-1/2}(y - Cx)\|^2 + \gamma \sum_{t=1}^{N} \left\| Q_t^{-1/2}(z_t - A_t x_t) \right\|_2, \tag{17}$$

where γ is again a penalty parameter controlling the tradeoff between measurement fit and number of non-zero components in process disturbance. Note that the objective is still of the type (8) with a penalty term that now corresponds to the sparsity inducing ℓ_1 norm applied to groups of process disturbances v_t, where the ℓ_2 norm used as the intra-group penalty. This group penalty has been widely used in statistical learning where it is referred to as the "group-LASSO" penalty. Its purpose is to select important factors, each represented by a group of derived variables, for joint model selection and estimation in regression. In the state estimation context, the estimator (17) was proposed in Ohlsson et al. (2012) and will be used later on in Section 5.2 to solve the impulsive inputs problem described in Section 1.1. The group $\ell_{2,1}$ penalty was originally proposed in the context of linear regression in Yuan and Lin (2006). The resulting estimates have a Bayesian interpretation: they can be obtained using a hierarchical model where the regression coefficients of each group are assigned a Gaussian prior whose variance is controlled with a gamma prior (Kyung et al., 2010). General $\ell_{q,1}$ regularized least squares formulations (with $q \geq 2$) were subsequently studied in Jacob, Obozinski, and Vert (2009), Tropp, Gilbert, and Strauss (2006), Yuan and Lin (2006), Zhao, Rocha, and Yu (2009) and shown to have strong statistical guarantees, including convergence rates in ℓ_2-norm (Baraniuk, Cevher, Duarte, & Hegde, 2008; Lounici, Pontil, Tsybakov, & van de Geer, 2009) as well as model selection consistency (Negahban & Wainwright, 2009; Obozinski, Wainwright, & Jordan, in press). Similarly to the LASSO, however, such results are not applicable in the dynamic context.

3.2. General functions V to model measurement errors

Gaussian assumption on measurement errors is not valid in many cases. Indeed, heavy tailed errors are frequently observed in applications such glint noise (Hewer, Martin, & Zeh, 1987), air turbulence (Fernández, Speyer, & Idan, 2013), and asset returns (Rachev, 2003) among others. The resulting state estimation problems can be addressed by adopting the penalties J introduced above. But, in addition, corrupted measurements might occur due to equipment malfunction, secondary sources of noise or other anomalies. The quadratic loss is not robust with respect to the

presence of outliers in the data (Aravkin et al., 2011; Farahmand et al., 2011; Gao, 2008; Huber & Ronchetti, 2009), as seen in Fig. 2, leading to undesirable behavior of resulting estimators. This calls for the design of new losses V.

One way to derive a robust approach is to assume that the noise comes from a probability density with tail probabilities larger (heavier) than those of the Gaussian, and consider the maximum a posteriori (MAP) problem derived from the corresponding negative log likelihood function. For instance the Laplace distribution $c\exp(-\|x\|_1)$ corresponds to the ℓ_1 loss function by this approach, see Fig. 4(b). The *tail probabilities* $P(|x| > t)$ of the standard Laplace distribution are greater than that of the Gaussian; so larger observations are more likely under this error model. Note, however, that the ℓ_1 loss also has a nonsmooth feature at the origin (which is exactly why we considered it as a choice for J in the previous section). In the current context, when applied to the measurement residual $Hx - z$, the approach will sparsify the residual, i.e. fit a portion of the data exactly. Exact fitting of some of the data may be reasonable in some contexts, but undesirable in many others, where we mainly care about guarding against outliers, and so only the *tail behavior* is of interest. In such settings, the Huber Loss (Huber & Ronchetti, 2009) (see Fig. 4(c)) is a more suitable model, as it combines the ℓ_2 loss for small errors with the absolute loss for larger errors. Huber (Huber & Ronchetti, 2009) showed that this loss is optimal over a particular class of errors

$$(1-\varepsilon)\mathcal{N} + \varepsilon\mathcal{M},$$

where $\mathcal{N}$ is Gaussian, and $\mathcal{M}$ is unknown; the level ε is then related to the Huber parameter κ. The Huber loss has a Bayesian interpretation, as a mixture between Gaussian and Laplacian loss functions.

Another important loss function is the Vapnik ε-insensitive loss (Drucker et al., 1997), sometimes known as the 'deadzone' penalty, see Fig. 4(d), defined as

$$V_\varepsilon(r) := \max\{0, |r| - \varepsilon\},$$

where r is the (scalar) residual. The ε-insensitive loss was originally considered in support vector regression (Drucker et al., 1997), where the 'deadzone' helps identify active support vectors, i.e. data elements that determine the solution. This penalty has a Bayesian interpretation, as a mixture of Gaussians that may have nonzero means (Pontil, Mukherjee, & Girosi, 2000). In particular, its use yields smoothers that are robust to minor fluctuations below a noise floor (as well as to large outliers). Note that the radius of the deadzone ε defines a noise floor beyond which one cannot resolve the signal. This penalty can also be 'huberized', yielding a penalty called 'smooth insensitive loss' (Chu, Keerthi, & Ong, 2001; Dekel, Shalev-Shwartz, & Singer, 2005; Lee, Hsieh, & Huang, 2005), see Fig. 4(e).

The process of choosing penalties based on behavior in the tail, near the origin, or at other specific regions of their subdomains makes it possible to customize the formulation of (15) to address a range of situations. We can then associate statistical densities to all the penalties in Fig. 4(a)–4(f), and use this perspective to incorporate prior knowledge about mean and variance of the residuals and process disturbances (Aravkin et al., 2013b Section 3). This allows one to incorporate variance information on process components; as e.g. available in the example of Fig. 2.

Asymmetric extensions. All of the PLQ losses in Fig. 4(a)–4(f) have asymmetric analogues. For example, the asymmetric 1-norm (Koenker & Bassett Jr, 1978) and asymmetric Huber (Aravkin, Lozano, Luss, & Kambadur, 2014) have been used for analysis of heterogeneous datasets, especially in high dimensional inference.

Beyond convex approaches. All of the penalty options for J and V presented so far are *convex*. Convex losses make it possible to provide strong guarantees—for example, if both J and V are convex in (15), then any stationary point is a global minimum. In addition, if J has compact level sets (i.e. there are no directions where it stays bounded), then at least one global minimizer exists. From a modeling perspective, however, it may be beneficial to choose a non-convex penalty in order to strengthen a particular feature. In the context of residuals, the need for non-convex loss is motivated by considering the *influence function*. This function measures the derivative of the loss with respect to the residual, quantifying the effect of the size of a residual on the loss. For nonconstant convex losses, linear growth is the limiting case, and this gives each residual constant influence. Ideally the influence function should redescend towards zero for large residuals, so that these are basically ignored. But redescending influence corresponds to sublinear growth, which excludes convex loss functions. We refer the reader to Hampel, Ronchetti, Rousseeuw, and Stahel (1986) for a review of influence-function approaches to robust statistics, including redescending influence functions. An illustration is presented in Fig. 5, contrasting the density, negative log-likelihood, and influence function of the heavy-tailed student's t penalty with those of gaussian (least squares) and laplace (ℓ_1) densities and penalties. More formally, consider any scalar density p arising from a symmetric convex coercive and differentiable penalty ρ via $p(x) = \exp(-\rho(x))$, and take any point x_0 with $\rho'(x_0) = \alpha_0 > 0$.

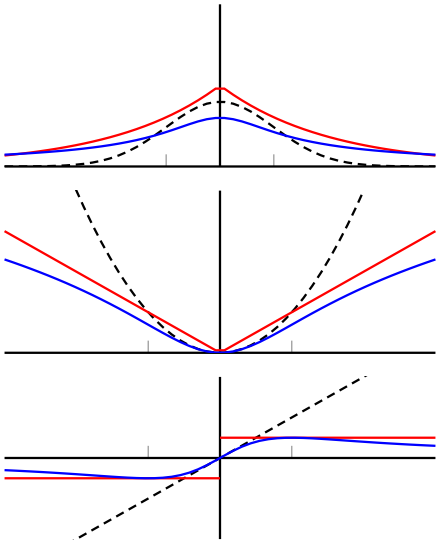


Fig. 5. Gaussian (black dashed), Laplace (red solid), and Student's t (blue solid) Densities, Corresponding Negative Log Likelihoods, and Influence Functions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Then, for all $x_2 > x_1 \geq x_0$ it is shown in Aravkin, Friedlander, Herrmann, and Van Leeuwen (2012) that the conditional tail distribution induced by $\mathbf{p}(x)$ satisfies

$$\Pr(|y| > x_2 \mid |y| > x_1) \leq \exp(-\alpha_0[x_2 - x_1]) \,. \tag{18}$$

When x_1 is large, the condition $|y| > x_1$ indicates that we are looking at an outlier. However, as shown by (18), *any* log-concave statistical model treats the outlier conservatively, dismissing the chance that $|y|$ could be significantly bigger than x_1. Contrast this behavior with that of the Student's t-distribution. With one degree of freedom, the Student's t-distribution is simply the Cauchy distribution, with a density proportional to $1/(1+y^2)$. Then we have that

$$\lim_{x\to\infty} \Pr(|y| > 2x \mid |y| > x) = \lim_{x\to\infty} \frac{\frac{\pi}{2} - \arctan(2x)}{\frac{\pi}{2} - \arctan(x)} = \frac{1}{2}.$$

See Aravkin, Burke, and Pillonetto (2014) for a more detailed discussion of non-convex robust approaches to Kalman smoothing using the Student's t distribution.

Non-convex functions J have also been frequently applied to modeling process noise. In particular, see Wipf and Nagarajan (2007), Wipf and Rao (2007) and Wipf, Rao, and Nagarajan (2011) for a link between penalized regression problems like LASSO and Bayesian methods. One classical approach is ARD (Mackay, 1994), which exploits hierarchical hyperpriors with 'hyperparameters' estimated via maximizing the marginal likelihood, following the Empirical Bayes paradigm (Maritz & Lwin, 1989). In addition, see Aravkin, Burke, Chiuso, and Pillonetto (2014) and Loh and Wainwright (2013) for statistical results in the nonconvex case. Although the nonconvex setting is essential in this context, it is important to point out that solution methodologies in the above examples are based on iterative convex approximations, which is our main focus.

3.3. Incorporating constraints

Constraints can be important for improving estimation. In state estimation problems, constraints arise naturally in a variety of ways. When estimating biological quantities such as concentration, or physical quantities such as height above ground level, we know these to be *non-negative*. Prior information can induce other constraints; for example, if maximum velocity or acceleration is known, this gives *bound constraints*. Some problems also offer up other interesting constraints: in the absence of maintenance, physical systems degrade (rather than improve), giving *monotonicity constraints* (Simon & Chia, 2002). Both unimodality and monotonicity can be formulated using linear inequality constraints (Aravkin, Burke, & Pillonetto, 2013a).

All of these examples motivate the constraint $x \in \mathscr{X}$ in (15). Since we focus only on the convex case, we require that $\mathscr{X}$ should be convex. In this paper, we focus on two types of convex sets:

(1) $\mathscr{X}$ is polyhedral, i.e. given by $\mathscr{X} = \{x : D^T x \leq d\}$.
(2) $\mathscr{X}$ has a simple projection operator $\text{proj}_{\mathscr{X}}$, where

$$\text{proj}_{\mathscr{X}}(y) := \arg\min_{x \in \mathscr{X}} \frac{1}{2}\|x - y\|_2^2.$$

The cases are not mutually exclusive, for example box constraints are polyhedral and easy to project onto. The set $\mathbb{B}_2 := \{x : \|x\|_2 \leq 1\}$ is not polyhedral, but has an easy projection operator:

$$\text{proj}_{\mathbb{B}_2}(y) = \begin{cases} y/\|y\|_2 & \text{if} \quad \|y\|_2 > 1 \\ y & \text{else.} \end{cases}$$

In general, we let $\mathbb{B}$ denote a closed unit ball for a given norm, and for the ℓ_p norms, this unit ball is denoted by $\mathbb{B}_p$. These approaches extend to the nonconvex setting. A class of nonconvex Kalman smoothing problems, where $\mathscr{X}$ is given by functional inequalities, is studied in Bell et al. (2009). We restrict ourselves to the convex case, however.

4. Efficient algorithms for generalized Kalman smoothing

In this section, we present an overview of smooth and nonsmooth methods for convex problems, and tailor them specifically to the Kalman smoothing case. The section is organized as follows. We begin with a few basic facts about convex sets and functions, and review gradient descent and Newton methods for smooth convex problems. Next, extensions to nonsmooth convex functions are discussed, beginning with a brief exposition of sub-gradient descent and its associated (slow) convergence rate. We conclude by showing how first- and second-order methods can be extended to develop efficient algorithms for the nonsmooth case using the proximity operator, splitting techniques, and interior point methods.

All of these methods are iterative, that is, we cannot obtain the solution after a single pass through the data, as in the classic Kalman filter and RTS smoother. However, the (block tridiagonal) structure of the dynamic problem plays a key role in all iterative methods.

4.1. Convex sets and functions

A subset $\mathscr{C}$ of $\mathbb{R}^n$ is said to be convex if it contains every line segment whose endpoints are in $\mathscr{C}$, i.e.,

$$(1-\lambda)x + \lambda y \in \mathscr{C} \quad \forall\, \lambda \in [0,1] \quad \text{whenever } x, y \in \mathscr{C}.$$

For example, the unit ball $\mathbb{B}$ for any norm is a convex set.

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is said to be convex if the secant line between any two points on the graph of f always lies above the graph of the function, i.e. $\forall \lambda \in [0,1]$:

$$f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y), \quad \forall\, x, y \in \mathbb{R}^n.$$

These ideas are related by the epigraph of f:

$$\text{epi}\,(f) := \{(x, \mu) \mid f(x) \leq \mu\} \subset \mathbb{R}^n \times \mathbb{R}.$$

A function $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex if and only if epi (g) is a convex set. A function f is called *closed* if epi (f) is a closed set, or equivalently, if f is *lower semicontinuous* (lsc).

Facts about convex sets can be translated into facts about convex functions. The reverse is also true with the aid of the convex indicator functions:

$$\delta_{\mathscr{C}}(x) := \begin{cases} 0 & \text{if} \quad x \in \mathscr{C} \\ \infty & \text{else.} \end{cases} \tag{19}$$

Examples of convex sets include subspaces and their translates (affine sets) as well as the lower level sets of convex functions:

$$\text{lev}_f(\tau) := \{x \mid f(x) \leq \tau\}.$$

Just as with closed sets, the intersection of an arbitrary collection of convex sets is also convex. For this reason we define the convex hull of a set $\mathscr{E}$ to be the intersection of all convex sets that contain it, denoted by conv $(\mathscr{E})$.

The convex sets of greatest interest to us are the convex polyhedra,

$$\mathscr{W} := \left\{x \;\middle|\; H^T x \leq h\right\} \quad \text{for some } H \in \mathbb{R}^{n\times m} \text{ and } h \in \mathbb{R}^m,$$

while the convex functions of greatest interest are the piecewise linear–quadratic (PLQ) penalties, shown in Fig. 4(a)–4(f). As discussed in Section 3, these penalties allow us to model impulsive disturbances in the process (see Fig. 4(b) and 4(f)), to develop robust distributions for measurements (see Fig. 4(c)) and implement support vector regression (SVR) in the context of dynamic systems (see Fig. 4(d)).

4.2. Smooth case: first- and second-order methods

Consider the problem

$$\min_x f(x),$$

together with an iterative procedure indexed by κ that is initialized at x^1. When f is a C^1-smooth function with β-Lipschitz continuous gradient, i.e. β-*smooth*:

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta\|x - y\|, \;\; \beta \geq 0, \tag{20}$$

f admits the upper bounding quadratic model

$$f(x) \leq m_\kappa(x) := f(x^\kappa) + \langle \nabla f(x^\kappa), x - x^\kappa\rangle + \frac{\beta}{2}\|x - x^\kappa\|^2. \tag{21}$$

If we minimize $m_\kappa(x)$ to obtain $x^{\kappa+1}$, this gives the steepest descent iteration

$$x^{\kappa+1} := x^\kappa - \frac{1}{\beta}\nabla f(x^\kappa).$$

The upper bound (21) shows we have strict descent:

$$\begin{aligned} f(x^{\kappa+1}) &\leq f(x^\kappa) - \langle \nabla f(x^\kappa), \beta^{-1}\nabla f(x^\kappa)\rangle + \frac{\beta}{2}\|\beta^{-1}\nabla f(x^\kappa)\|^2 \\ &= f(x^\kappa) - \frac{\|\nabla f(x^\kappa)\|^2}{2\beta}. \end{aligned}$$

If, in addition, f is convex, and a minimizer $\hat{x}$ exists, we obtain

$$\begin{aligned} f(x^\kappa) - \frac{\|\nabla f(x^\kappa)\|^2}{2\beta} &\leq \hat{f} + \langle \nabla f(x^\kappa), x^\kappa - \hat{x}\rangle - \frac{\|\nabla f(x^\kappa)\|^2}{2\beta} \\ &= \hat{f} + \frac{\beta}{2}\left(\|x^\kappa - \hat{x}\|^2 - \|x^{\kappa+1} - \hat{x}\|^2\right), \end{aligned}$$

where $\hat{f} = f(\hat{x})$ is the same at any minimizer by convexity.

Adding up, we get an $O\left(\frac{1}{\kappa}\right)$ convergence rate on function values:

$$f(x^\kappa) - \hat{f} \leq \frac{\beta\|x^1 - \hat{x}\|^2}{2\kappa}.$$

For the least squares Kalman smoothing problem (12), we also know that f is α-*strongly convex*, i.e. $f(x) - \frac{\alpha}{2}\|x\|^2$ is convex with $\alpha \geq 0$. Strong convexity can be used to obtain a much better rate for steepest descent:

$$f(x^\kappa) - \hat{f} \leq \frac{\beta}{2}(1 - \alpha/\beta)^\kappa \|x^1 - \hat{x}\|^2.$$

Note that $0 \leq \frac{\alpha}{\beta} \leq 1$, since the strong convexity constant α is the curvature of the quadratic lower bound, and so is necessarily smaller than the Lipschitz constant of the gradient β, which is the curvature of the quadratic upper bound.

When minimizing a strongly convex function, the minimizer $\hat{x}$ is unique, and we can also obtain a rate on the squared distance between x^κ and $\hat{x}$:

$$\|x^\kappa - \hat{x}\|^2 \leq (1 - \alpha/\beta)^\kappa \|x^1 - \hat{x}\|^2.$$

These rates can be further improved by considering *accelerated-gradient* methods (see e.g. Nesterov, 2004) which achieve the much faster rate $(1 - \sqrt{\alpha/\beta})^\kappa$.

Each iteration of steepest descent in the classic least squares formulation (12) of the Kalman smoothing problem gives a fractional reduction in both function value and distance to optimal solution. The gradient is computed using matrix–vector products, which require $O(Nn^2)$ arithmetic operations. Thus, either gradient descent or conjugate gradient (which has the same rate as accelerated gradient methods in the least squares case) is a reasonable option for large n.

The solution to (12) can also be obtained by solving a linear system of equations using $O(Nn^3)$ arithmetic operations, since (13) is block-tridiagonal positive definite. This complexity is tractable for moderate state-space dimension n. The approach is equivalent to a single iteration on the full quadratic model of the Newton's method, discussed below.

Consider the problem of minimizing a C^2-smooth function f. Finding a critical point x of f can be recast as the problem of solving the nonlinear equation $\nabla f(x) = 0$. For a smooth function $G : \mathbb{R}^n \to \mathbb{R}^n$, Newton's method is designed to locate solutions to the equation $G(x) = 0$. Given a current iterate x^κ, Newton's method linearizes G at x^κ and solves the equation $G(x^\kappa) + \nabla G(x^\kappa)(y - x^\kappa) = 0$ for y. Provided that $\nabla G(x^\kappa)$ is invertible, the Newton iterate is given by

$$x^{\kappa+1} := x^\kappa - [\nabla G(x^\kappa)]^{-1} G(x^\kappa). \tag{22}$$

When $G := \nabla f$, the Newton iterate (22) is the unique critical point of the best quadratic approximation of f at x^κ, namely

$$\begin{aligned} Q(x^\kappa; y) := f(x^\kappa) &+ \langle \nabla f(x^\kappa), y - x^\kappa\rangle \\ &+ \frac{1}{2}\langle \nabla^2 f(x^\kappa)(y - x^\kappa), y - x^\kappa\rangle, \end{aligned}$$

provided that the Hessian $\nabla^2 f(x^\kappa)$ is invertible.

If G is a C^1-smooth function with β-Lipschitz Jacobian ∇G that is locally invertible for all x near a point $\hat{x}$ with $G(\hat{x}) = 0$, then near $\hat{x}$ the Newton iterates (22) satisfy

$$\|x^{\kappa+1} - \hat{x}\| \leq \frac{\beta}{2}\|\nabla G(x^\kappa)^{-1}\| \cdot \|x^\kappa - \hat{x}\|^2.$$

Once we are *close enough* to a solution, Newton's method gives a *quadratic* rate of convergence. Consequently, locally the number of correct digits double for each iteration. Although the solution may not be obtained in one step (as in the quadratic case), only a few iterations are required to converge to machine precision.

In the remainder of the section, we generalize steepest descent and Newton's methods to nonsmooth problems of type (15). In Section 4.3, we describe the *sub-gradient* descent method, and show that it converges very slowly. In Section 4.4, we describe the proximity operator and proximal-gradient methods, which are applicable when working with separable nonsmooth terms in (15). In Section 4.5, we show how to solve more general nonsmooth problems (15) using splitting techniques, including ADMM and Chambolle–Pock iterations. Finally, in Section 4.6, we show how second-order interior point methods can be brought to bear on all problems of interest of type (15).

4.3. Nonsmooth case: subgradient descent

Given a convex function f, a vector v is a *subgradient* of f at a point x if

$$f(y) \geq f(x) + \langle v, y - x\rangle \quad \forall y. \tag{23}$$

The set of all subgradients at x is called the *subdifferential*, and is denoted by $\partial f(x)$. Subgradients generalize the notion of gradient; in particular, $\partial f(x) = \{v\} \iff v = \nabla f(x)$ (Rockafellar, 1974). A more comprehensive discussion of the subdifferential is presented in Appendix A.2.

Consider the absolute value function shown in Fig. 4(b). It is differentiable at all points except for $x = 0$, and so the subdifferential is precisely the gradient for all $x \neq 0$. The subgradients at $x = 0$ are the slopes of lines passing through the origin and lying below the graph of the absolute value function. Therefore, $\partial|\cdot|(0) = [-1, 1]$.

Consider the following simple algorithm for minimizing a Lipschitz continuous (but nonsmooth) convex f. Given an oracle that delivers some $v^\kappa \in \partial f(x^\kappa)$, set

$$x^{\kappa+1} := x^\kappa - \alpha_\kappa v^\kappa, \tag{24}$$

for a judiciously chosen stepsize α_κ. Suppose we are minimizing $|x|$ and start at $x = 0$, the global minimum. The oracle could return any value $v \in [-1, 1]$, and so we will move away from 0 when using (24)! In general, the function value need not decrease at each iteration, and we see that α_κ must decrease to 0 for any hope of convergence. On the other hand, if $\sum_\kappa \alpha_\kappa = R < \infty$, we can never reach $\hat{x}$ if $\|x^1 - \hat{x}\| > R$, where x^1 is the initial point and $\hat{x}$ the minimizer. Therefore, we also must have $\sum_\kappa \alpha_\kappa = \infty$.

Setting $l_\kappa := f(x^\kappa) + \langle v^\kappa, \hat{x} - x^\kappa\rangle$, by (23) we have $l_\kappa \leq f(\hat{x}) \leq f(x^\kappa)$ for $v \in \partial f(x^\kappa)$. The subgradient method closes the gap between l_κ and $f(x^\kappa)$. The Lipschitz continuity of f implies that

$\|v^\kappa\| \leq L$, and so, by (23),

$$\begin{aligned} 0 \leq \|x^{\kappa+1} - \hat{x}\|^2 &= \|x^\kappa - \hat{x}\|^2 + 2\alpha_\kappa \langle v^\kappa, \hat{x} - x^\kappa \rangle + \alpha_\kappa^2 \|v^\kappa\|^2 \\ &\leq \|x^1 - \hat{x}\|^2 + \sum_{i=1}^{\kappa} 2\alpha_i \langle v^i, \hat{x} - x^i \rangle + L^2 \sum_{i=1}^{\kappa} \alpha_i^2 \\ &= \|x^1 - \hat{x}\|^2 + \sum_{i=1}^{\kappa} 2\alpha_i (l_i - f(x^i)) + L^2 \sum_{i=1}^{\kappa} \alpha_i^2. \end{aligned}$$

Rewriting this inequality gives

$$\begin{aligned} 0 \leq \min_{i=1,\ldots,\kappa} (f(x^i) - l_i) &\leq \sum_{i=1}^{\kappa} \frac{\alpha_i}{\sum_{i=1}^{\kappa} \alpha_i} (f(x^i) - l_i) \\ &\leq \frac{\|x^1 - \hat{x}\|^2 + L^2 \sum_{i=1}^{\kappa} \alpha_i^2}{2 \sum_{i=1}^{\kappa} \alpha_i}. \end{aligned} \tag{25}$$

In particular, if $\{\alpha_\kappa\}$ are square summable but not summable, convergence of $\min_{i=1,\ldots,\kappa}\{f(x^i) - l_i\}$ to 0 is guaranteed. But there is a fundamental limitation of the subgradient method. In fact, suppose that we know $\|x^1 - \hat{x}\|$, and want to choose steps α_i to minimize the gap in t iterations. By minimizing the right hand side of (25), we find that the optimal step sizes (with respect to the error bound) are

$$\alpha_i = \frac{\|x^1 - \hat{x}\|}{L\sqrt{\kappa}}.$$

Plugging these back in, and defining $f^\kappa_{\text{best}} = \min_{i=1,\ldots,\kappa} f(x^i)$, we have

$$f^\kappa_{\text{best}} - \hat{f} \leq \frac{\|x^1 - \hat{x}\| L}{\sqrt{\kappa}}.$$

Consequently, the best provable subgradient descent method is extremely slow. This rate can be significantly improved by exploiting the structure of the nonsmoothness in f.

4.4. Proximal gradient methods and accelerations

For many convex functions, and in particular for a range of general smoothing formulations (15), we can design algorithms that are much faster than $O(1/\sqrt{\kappa})$. Suppose we want to minimize the sum

$$f(x) + g(x),$$

where f is convex and β-smooth (20), while g is any convex function. Using the bounding model (21) for f, we can get a global upper bound for the sum:

$$\begin{aligned} &f(x) + g(x) \leq m_\kappa(x) \\ &m_\kappa(x) := f(x^\kappa) + \langle \nabla f(x^\kappa), x - x^\kappa \rangle + \frac{\beta}{2}\|x - x^\kappa\|^2 + g(x). \end{aligned}$$

We immediately see that setting

$$x^{\kappa+1} := \arg\min_x \; m_\kappa(x) \tag{26}$$

ensures descent for $f + g$, since

$$f(x^{\kappa+1}) + g(x^{\kappa+1}) \leq m_\kappa(x^{\kappa+1}) \leq m_\kappa(x^\kappa) = f(x^\kappa) + g(x^\kappa).$$

One can check that $m_\kappa(x^{\kappa+1}) = m_\kappa(x^\kappa)$ if and only if x^κ is a global minimum of $f + g$. Rewriting (26) as

$$x^{\kappa+1} := \arg\min_x \beta^{-1} g(x) + \frac{1}{2}\left\| x - \left(x^\kappa - \frac{1}{\beta}\nabla f(x^\kappa)\right)\right\|^2,$$

and define the *proximity* operator for ηg (Bauschke & Combettes, 2011) by

$$\text{prox}_{\eta g}(y) := \arg\min_x \eta g(x) + \frac{1}{2}\|x - y\|^2, \tag{27}$$

where η is any positive scalar. We see that (26) is precisely the proximal gradient method:

$$x^{\kappa+1} := \text{prox}_{\beta^{-1} g}\left(x^\kappa - \frac{1}{\beta}\nabla f(x^\kappa)\right). \tag{28}$$

The proximal gradient iteration (28) converges with the same rate as gradient descent, in particular with rate $O(1/\kappa)$ for convex functions and $O((1 - \alpha/\beta)^\kappa)$ for α-strongly convex functions. These rates are in a completely different class than the $O\left(1/\sqrt{\kappa}\right)$ rate obtained by the subgradient method, since they exploit the additive structure of $f + g$. Proximal gradient algorithms can also be accelerated, achieving rates of $O\left(1/\kappa^2\right)$ and $O\left((1 - \sqrt{\alpha/\beta})^\kappa\right)$ respectively, using techniques from Nesterov (2004).

In order to implement (28), we must be able to efficiently compute the proximity operator for ηg. For many nonsmooth functions g, this operator can be computed in $O(n)$ or $O(n \log n)$ time. An important example is the *convex indicator* function (19). In this case, the proximity operator is the projection operator:

$$\begin{aligned} \text{prox}_{\eta\delta_{\mathscr{C}}(x)}(y) &= \delta_{\mathscr{C}}(x) + \min_x \frac{1}{2}\|x - y\|^2 \\ &= \min_{x \in \mathscr{C}} \frac{1}{2}\|x - y\|^2 = \text{proj}_{\mathscr{C}}(y). \end{aligned} \tag{29}$$

In particular, when minimizing f over a convex set $\mathscr{C}$, iteration (28) recovers the *projected gradient* method if we choose $g(x) = \delta_{\mathscr{C}}(x)$.

Many examples and identities useful for computing proximal operators are collected in Combettes and Pesquet (2011). One important example is the Moreau identity (see e.g. Rockafellar & Wets, 1998):

$$\text{prox}_f(y) + \text{prox}_{f^*}(y) = y\,. \tag{30}$$

Here, f^* denotes the *convex conjugate* of f:

$$f^*(\omega) := \sup_y (\langle y, \omega \rangle - f(y)), \tag{31}$$

whose properties are explained in Appendix A.2, in the context of convex duality. Identity (30) shows that the prox of f can be used to compute the prox of f^*, and vice versa.

The identity (30) is a direct consequence of Fenchel's inequality (50), derived in Appendix A.2.

Example (*Proximity Operator for the ℓ_1-Norm*)**.** Consider the example $g(x) = \|x\|_1$, often used in applications to induce sparsity of x. The proximity operator of this function can be computed by reducing to the 1-dimensional setting and considering cases. Here, we show how to compute it using (30):

$$\text{prox}_{\eta\|\cdot\|_1}(y) = y - \text{prox}_{(\eta\|\cdot\|_1)^*}(y).$$

The convex conjugate of the scaled 1-norm is given by

$$(\eta\|\cdot\|_1)^*(\omega) = \sup_x \langle x, \omega \rangle - \eta\|x\|_1 = \begin{cases} 0 & \text{if } \|\omega\|_\infty \leq \eta \\ \infty & \text{otherwise,} \end{cases} \tag{32}$$

which is precisely the indicator function of $\eta\mathbb{B}_\infty$, the scaled ∞-norm unit ball. As previously observed, the proximity operator for an indicator function is the projection. Consequently, the identity (30) simplifies to

$$\text{prox}_{\eta\|\cdot\|_1}(y) = y - \text{proj}_{\eta\mathbb{B}_\infty}(y)$$

whose ith element is given by

$$\text{prox}_{\eta\|\cdot\|_1}(y)_i = \begin{cases} y_i - y_i = 0 & \text{if } |y_i| \leq \eta \\ y_i - \eta\,\text{sign}(y_i) & \text{if } |y_i| > \eta \end{cases} \tag{33}$$

which corresponds to *soft-thresholding*. Computing the proximal operator for the 1-norm and projection onto the ∞-norm ball both require $O(n)$ operations. Projection onto the 1-norm ball $\mathbb{B}_1$ can be implemented using a sort, and so takes $O(n \log(n))$ operations, see e.g. Van den Berg and Friedlander (2008). ■

To illustrate the method in the context of Kalman smoothing, consider taking the general formulation (15) with V and J both smooth, $\gamma = 1$, and $x \in \tau\mathbb{B}$ any norm-ball for which we have a fast projection (common cases are 2-norm, 1-norm, or ∞-norm):

$$\min_{x\in\tau\mathbb{B}} f(x) := V(R^{-1/2}(y - Cx)) + J(Q^{-1/2}(z - Ax)).$$

Algorithm 3 Proximal Gradient for Kalman Smoothing, J and V Huber or quadratic

(1) Initialize $x^1 = 0$, $\kappa = 0$, compute $d^1 = \nabla f(x^1)$. Let $\beta = \|C^TR^{-1}C + A^TQ^{-1}A\|_2$.

(2) While $\|\text{prox}_g(x^\kappa - d^\kappa)\| > \varepsilon$

- Set $\kappa = \kappa + 1$.
- update $x^\kappa = \text{prox}_{\beta^{-1}g}(x^{\kappa-1} - \beta^{-1}d^{\kappa-1})$.
- Compute $d^\kappa = \nabla f(x^\kappa)$.

(3) Output x^κ.

Algorithm 4 FISTA for Kalman Smoothing, J and V Huber or quadratic

(1) Initialize $x^1 = 0$, $\omega = 0$, $\kappa = 0$, $s_1 = 1$, compute $d^1 = \nabla f(x^1)$. Let $\beta = \|C^TR^{-1}C + A^TQ^{-1}A\|_2$.

(2) While $\|\text{prox}_g(\omega^\kappa - d^\kappa)\| > \varepsilon$

- Set $\kappa = \kappa + 1$.
- update $x^\kappa = \text{prox}_{\beta^{-1}g}(\omega^{\kappa-1} - \alpha d^{\kappa-1})$.
- set $s_\kappa = \frac{1+\sqrt{1+4s_{\kappa-1}^2}}{2}$
- set $\omega^\kappa = x^\kappa + \frac{s_{\kappa-1}-1}{s_\kappa}(x_\kappa - x_{\kappa-1})$.
- Compute $g^\kappa = \nabla f(x^\kappa)$.

(3) Output x^κ.

The gradient for the system of equations is given by

$$\nabla f(x) = C^TR^{-1/2}\nabla V(R^{-1/2}(Cx - y)) + A^TQ^{-1/2}\nabla J(Q^{-1/2}(Ax - z)).$$

When V and J are quadratic or Huber penalties, the Lipschitz constant β of ∇f is bounded by the largest singular value of $C^TR^{-1}C + A^TQ^{-1}A$, which we can obtain using power iterations. This system is block tridiagonal, so matrix–vector multiplications are far more efficient than for general systems of equations. Specifically, for Kalman smoothing, the matrices C, Q, R are block diagonal, while A is block bidiagonal. As a result, products with A, A^T, C, $Q^{-1/2}$, $R^{-1/2}$ can all be computed using $O(Nn^2)$ arithmetic operations, rather than $O(N^2n^2)$ operations as for a general system of the same size. A simple proximal gradient method is given by Algorithm 3. Note that soft thresholding for Kalman smoothing has complexity $O(nN)$, while e.g. projecting onto the 1-norm ball has complexity $O(nN\log(nN))$. Therefore the $O(n^2N)$ cost of computing the gradient $\nabla f(x^\kappa)$ is dominant.

Algorithm 3 has at worst $O\left(\kappa^{-1}\right)$ rate of convergence. If J is taken to be a quadratic, f is strongly convex, in which case we achieve the much faster rate $O\left((1-\alpha/\beta)^\kappa\right)$.

Algorithm 4 illustrates the FISTA scheme (Beck & Teboulle, 2009) applied to Kalman smoothing. This acceleration uses two previous iterates rather than just one, and achieves a worst case rate of $O\left(\kappa^{-2}\right)$. This can be further improved to $O\left((1-\sqrt{\alpha/\beta})^\kappa\right)$ when J is a convex quadratic using techniques in Nesterov (2004), or periodic restarts of the step-size sequence s_κ.

4.5. Splitting methods

Not all smoothing formulations (15) are the sum of a smooth function and a separable nonsmooth function. In many cases, the composition of a nonsmooth penalty with a general linear operator can preclude the approach of the previous section; this is the case for the *robust* Kalman smoothing problem in Aravkin et al. (2011):

$$\min_x \|R^{-1/2}(y - Cx)\|_1 + \frac{1}{2}\|Q^{-1/2}(z - Ax)\|^2. \tag{34}$$

Replacing the quadratic penalty with the 1-norm allows the development of a robust smoother when a portion of (isolated) measurements are contaminated by outliers. The composition of the nonsmooth 1-norm with a general linear form makes it impractical to use the proximal gradient method since the evaluation of the prox operator

$$\text{prox}_{\eta\|y-C(\cdot)\|_1}(y) = \arg\min_x \frac{1}{2}\|y - x\|^2 + \eta\|y - Cx\|_1$$

requires an iterative solution scheme for general C. However, it is possible to design a primal–dual method using a range of strategies known as splitting methods.

Splitting methods can be generally viewed as fixed-point iterations for nonlinear operators derived from optimality conditions; see the discussion at the end of Appendix A.2.

A well-known splitting method, popularized by Boyd, Parikh, Chu, Peleato, and Eckstein (2011), is the Alternating Direction Method of Multipliers (ADMM), which is equivalent to Douglas–Rachford splitting on an appropriate dual problem (Lions & Mercier, 1979). The ADMM scheme is applicable to general problems of type

$$\min_{x,\omega} f(x) + g(\omega) \quad \text{s.t.} \quad K_1x + K_2\omega = c. \tag{35}$$

A fast way to derive the approach is to consider the Augmented Lagrangian (Rockafellar, 1974) dualizing the equality constraint in (35):

$$\mathscr{L}(x, \omega, u, \tau) := f(x) + g(\omega) + u^T(K_1x + K_2\omega - c) + \frac{\tau}{2}\|K_1x + K_2\omega - c\|^2,$$

where $\tau > 0$. The ADMM method proceeds by using alternating minimization of $\mathscr{L}$ in x and ω with appropriate dual updates (which is equivalent to the Douglas–Rachford method on the dual of (35)). The iterations are explained fully in Algorithm 5.

ADMM has convergence rate $O(1/\kappa)$, but can be accelerated under sufficient regularity conditions (see e.g. Davis & Yin, 2015). For the Laplace ℓ_1 smoother (34), the transformation to template (35) is given by

$$\min_{x,\omega}\left\{\|\omega\|_1 + \frac{1}{2}\|Q^{-1/2}(z - Ax)\|^2 \;\middle|\; \omega + R^{-1/2}Cx = R^{-1/2}y\right\}. \tag{36}$$

ADMM specialized to (36) is given by Algorithm 6.

We make two observations. First, note that the x-update requires solving a least squares problem, in particular inverting $A^TQ^{-1}A + C^TR^{-1}C$. Fortunately, in problem (36) this system of equations does not change between iterations, and can be factorized once in $O(n^3N)$ arithmetic operations and stored. Each iteration of the x-update can be obtained in $O(n^2N)$ arithmetic operations which has the same complexity as a matrix–vector product. Splitting schemes that avoid factorizations are described below. However, avoiding factorizations is not always the best strategy since the choice of splitting scheme can have a dramatic

Algorithm 5 ADMM algorithm for (35)

(1) Input x^1, $\omega^0 \neq \omega^1$, $\kappa = 0$. Input $\tau > 0$, ε.

(2) While $\|K_1x^\kappa + K_2\omega^\kappa - c\| > \varepsilon$ and $\|\tau K_1^T K_2(\omega^{\kappa+1} - \omega^\kappa)\| > \varepsilon$

- Set $\kappa := \kappa + 1$.
- update
$$x^{\kappa+1} := \arg\min_x \left\{ \begin{array}{l} f(x) + (u^\kappa)^T K_1 x \\ \quad + \dfrac{\tau}{2}\|K_1x + K_2\omega^\kappa - c\|^2 \end{array} \right\}$$
- update
$$\omega^{\kappa+1} := \arg\min_\omega \left\{ \begin{array}{l} g(\omega) + (u^\kappa)^T K_2 \omega \\ \quad + \dfrac{\tau}{2}\|K_1x^{\kappa+1} + K_2\omega - c\|^2 \end{array} \right\}$$
- update $u^{\kappa+1} := u^\kappa + \tau(K_1x^{\kappa+1} + K_2\omega^{\kappa+1} - c)$

(3) Output $(x^\kappa, \omega^\kappa)$.

Algorithm 6 ADMM algorithm for (36)

(1) Input x^1, $\omega^0 \neq \omega^1$, $\kappa = 0$. Input $\tau > 0$, ε.

(2) While $\|\omega^\kappa + R^{-1/2}Cx^\kappa - R^{-1/2}y\| > \varepsilon$ and $\|\tau C^T R^{-T/2}(\omega^{\kappa+1} - \omega^\kappa)\| > \varepsilon$

- Set $\kappa := \kappa + 1$.
- update
$$x^{\kappa+1} := \arg\min_x \frac{1}{2}\|Q^{-1/2}(z - Ax)\|^2 + x^T u^\kappa + \frac{\tau}{2}\|R^{-1/2}(Cx - y) + \omega^\kappa\|^2$$
- update
$$\omega^{\kappa+1} := \arg\min_\omega \|\omega\|_1 + \frac{\tau}{2}\left\|\omega + u^\kappa/\tau + R^{-1/2}(Cx^{\kappa+1} - y)\right\|^2$$
- update
$$u^{\kappa+1} := u^\kappa + \tau(R^{-1/2}Cx^{\kappa+1} + \omega^{\kappa+1} - R^{-1/2}y)$$

(3) Output x^κ.

effect on the performance. Performance differences between various splitting are explored in the numerical section. Second, the ω-update has a convenient closed form representation in terms of the proximity operator (27):

$$\omega^{\kappa+1} := \text{prox}_{\tau^{-1}\|\cdot\|_1}(u^\kappa/\tau + R^{-1/2}(Cx^{\kappa+1} - y)).$$

The overall complexity of each iteration of the ADMM ℓ_1-Kalman smoother is $O(n^2N)$, after the initial $O(n^3N)$ investment to factorize $A^TQ^{-1}A + C^TR^{-1}C$.

There are several types of splitting schemes, including Forward–Backward (Passty, 1979), Peaceman–Rachford (Lions & Mercier, 1979), and others. A survey of these algorithms is beyond the scope of this paper. See Bauschke and Combettes (2011) and Davis and Yin (2015) for a discussion of splitting methods and the relationships between them. See also Davis and Yin (2014), for a detailed analysis of convergence rates of several splitting schemes under regularity assumptions.

We are not aware of a detailed study or comparison of these techniques for general Kalman smoothing problems, and future work in this direction can have a significant impact in the community. To give an illustration of the numerical behavior and variety of splitting algorithms, we present the algorithm of Chambolle–Pock (CP) (Chambolle & Pock, 2011), for convex problems of type

$$\min_x f(Kx) + g(x), \tag{37}$$

where f and g are convex functions with computable proximity operators. The CP iteration is specified in Algorithm 7; the quantity L used in the algorithm is the largest singular value of K in (37).

Algorithm 7 Chambolle–Pock algorithm for (37)

(1) Input $x^0 \neq x^1$, $\omega^0 \neq \omega^1$, $\kappa = 0$. Input τ, σ s.t. $\tau\sigma L^2 < 1$. Input ε.

(2) While $(\|\omega^{\kappa+1} - \omega^\kappa\| + \|x^{\kappa+1} - x^\kappa\| > \varepsilon)$

- Set $\kappa = \kappa + 1$.
- update $\omega^{\kappa+1} = \text{prox}_{\sigma f^*}(\omega^\kappa + \sigma K(2x^\kappa - x^{\kappa-1}))$
- update $x^{\kappa+1} = \text{prox}_{\tau g}(x^\kappa - \tau K^T\omega^{\kappa+1})$

(3) Output x^κ.

Algorithm 7 requires only the proximal operators for f^* and g to be implementable. Like ADMM, it has a convergence rate of $O(1/\kappa)$, and can be accelerated to $O(1/\kappa^2)$ under specific regularity assumptions. When g is strongly convex, one such acceleration is presented in Chambolle and Pock (2011).

There are multiple ways to apply the CP scheme to a given Kalman smoothing formulation. Some schemes allow CP to solve large-scale smoothing problems (15) using only matrix–vector products, avoiding large-scale matrix solves entirely. However, this may not be the best approach, as we show in our numerical study in the following section. General splitting schemes such as Chambolle–Pock can achieve at best $O(1/\kappa^2)$ convergence rate for general nonsmooth Kalman formulations. Faster rates require much stronger assumptions, e.g. smoothness of the primal or dual problems (Chambolle & Pock, 2011). When these conditions are present, the methods can be remarkably efficient.

4.6. Formulations using Piecewise Linear Quadratic (PLQ) penalties

When the state size n is moderate, so that $O(n^3N)$ is an acceptable cost to pay, we can obtain fast methods for general Kalman smoothing systems. We recover *second-order behavior* and fast local convergence rates by developing interior point methods for the entire class (15). These methods can be developed for any piecewise linear quadratic V and J, and allow the inclusion of polyhedral constraints that link adjacent time points. This can be accomplished using $O(n^3N)$ arithmetic operations, the same complexity as solving the least squares Kalman smoother.

To see how to develop second-order interior point methods for these PLQ smoothers, we first define the general PLQ family and consider its conjugate representation and optimality conditions.

Definition 1 (*PLQ Functions and Penalties*)**.** A piecewise linear quadratic (PLQ) function is any function $\rho(h, H, b, B, M; \cdot) : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ admitting representation

$$\begin{aligned} \rho(h, H, b, B, M; x) &:= \sup_{v \in \mathscr{V}} \left\{ \langle v, b + Bx\rangle - \frac{1}{2}\langle v, Mv\rangle \right\} \\ &= \left(\frac{1}{2}\|\cdot\|_M^2 + \delta_{\mathscr{V}}(\cdot) \right)^* (b + Bx)\,, \end{aligned} \tag{38}$$

where $\mathscr{V}$ is the polyhedral set specified by $H \in \mathbb{R}^{k\times\ell}$ and $h \in \mathbb{R}^\ell$ as follows:

$$\mathscr{V} = \{v : H^T v \leq h\}\,,$$

$M \in \mathscr{S}_+^k$ the set of real symmetric positive semidefinite matrices, $b + Bx$ is an injective affine transformation in x, with $B \in \mathbb{R}^{k\times n}$, so, in particular, $n \leq k$ and $null(B) = \{0\}$. If $0 \in \mathscr{V}$, then the PLQ is necessarily non-negative and hence represents a *penalty*.

The last equation in (38) is seen immediately using (31). In what follows we reserve the symbol ρ for a PLQ penalty often writing $\rho(x)$ and suppressing the litany of parameters that precisely define the function. When detailed knowledge of these parameters is required, they will be specified.

Below we show how the six loss functions illustrated in Fig. 4(a)–4(f) can be represented as members of the PLQ class. In each case, the verification of the representation is straightforward. These dual (conjugate) representations facilitate the general optimization approach.

Examples of scalar PLQ

(1) **quadratic** (ℓ_2) penalty, Fig. 4(a):

$$\sup_{v\in\mathbb{R}} \left\{ vx - \frac{1}{2}v^2 \right\}$$

(2) **absolute value** (ℓ_1) penalty, Fig. 4(b):

$$\sup_{v\in[-1,1]} \{vx\}$$

(3) **Huber** penalty, Fig. 4(c):

$$\sup_{v\in[-\kappa,\kappa]} \left\{ vx - \frac{1}{2}v^2 \right\}$$

(4) **Vapnik** penalty, Fig. 4(d):

$$\sup_{v\in[0,1]^2} \left\{ \left\langle \begin{bmatrix} x-\varepsilon \\ -x-\varepsilon \end{bmatrix}, v \right\rangle \right\}$$

(5) **Huber insensitive** loss, Fig. 4(e):

$$\sup_{v\in[0,1]^2} \left\{ \left\langle \begin{bmatrix} x-\varepsilon \\ -x-\varepsilon \end{bmatrix}, v \right\rangle - \frac{1}{2}v^T v \right\}$$

(6) **Elastic net**, Fig. 4(f):

$$\sup_{v\in[0,1]\times\mathbb{R}} \left\{ \left\langle \begin{bmatrix} 1 \\ 1 \end{bmatrix} x, v \right\rangle - \frac{1}{2} v^T \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} v \right\}.$$

Note that the set $\mathscr{V}$ is shown explicitly, and in each case can be easily represented as $\mathscr{V} := \{v \ : \ D^T v \leq d\}$. In addition, H and M are very sparse in all examples. ■

Consider now optimizing a PLQ penalty subject to inequality constraints:

$$\begin{aligned} &\min_x \quad \rho(x) \\ &\text{s.t.} \quad D^T x \leq d. \end{aligned} \tag{39}$$

Using the techniques of convex duality theory developed in Appendix A.2, the Lagrangian for (39) is given by

$$\mathscr{L}(x, v, \omega) = \langle \omega, \ D^T x - d \rangle - \delta_{\mathbb{R}^{n_1}_+}(\omega) + \langle v, \ b + Bx \rangle - \frac{1}{2} v^T M v - \delta_{\mathbb{R}^{n_2}_-}\left(H^T v - h\right),$$

where n_1 and n_2 are dimensions of d and c.

See (54) in Appendix A.2.

The dual problem associated to this Lagrangian is

$$\begin{aligned} &\min_{(v,\omega)} \quad \langle d, \ \omega \rangle + \frac{1}{2} v^T M v - \langle b, \ v \rangle \\ &\text{s.t.} \quad B^T v + D\omega = 0, \quad H^T v \leq h, \quad 0 \leq \omega\,. \end{aligned} \tag{40}$$

When the primal and dual problems have finite optimal values, strong duality holds in the PLQ case (see Appendix A.2). The conditions characterizing the optimal primal–dual pair (Theorem 4) are then given by

$$\begin{aligned} &\omega, \ w \geq 0 \\ &D\omega + B^T v = 0 \\ &Mv + Hw = Bx + b \\ &H^T v \leq h, \quad D^T x \leq d \\ &\omega_j (D^T x - d)_j = 0, \ j = 1, \ldots, n_1 \\ &w_j (H^T v - h)_j = 0, \ j = 1, \ldots, n_2. \end{aligned} \tag{41}$$

The final two conditions in (41) are called *complementary slackness* conditions. If $(\overline{x}, \overline{v}, \overline{\omega}, \overline{w})$ satisfy all of the conditions in (41), then $\overline{x}$ solves the primal problem (39) and $(\overline{v}, \overline{\omega})$ solves the dual problem (40). The optimality criteria (41) are known as the Karush–Kuhn–Tucker (KKT) conditions for (39) and are used in the interior point method described in the next section.

4.7. Interior point (second-order) methods for PLQ functions

Interior point methods directly target the KKT equations (41). In essence, they apply a damped Newton's method to a relaxed KKT system of equations (Kojima et al., 1991; Nemirovskii & Nesterov, 1994; Wright, 1997), recovering second-order behavior (i.e. superlinear convergence rates) for nonsmooth problems.

To develop an interior point method for the previous section, we first introduce slack variables

$$s := d - D^T x \geq 0 \quad \text{and} \quad r := h - H^T v \geq 0\,.$$

Complementarity slackness conditions (41) can now be stated as

$$\Omega S = 0 \quad \text{and} \quad WR = 0,$$

where Ω, S, W, R are diagonal matrices with diagonals ω, s, w, r, respectively. Let $\mathbf{1}$ denote the vector of all ones of the appropriate dimension. Given $\mu > 0$, we apply damped Newton iterations to the *relaxed* KKT system of equations

$$F_\mu(x, v, s, r, \omega, w) := \begin{bmatrix} D\omega + B^T v \\ Mv + Hw - Bx - b \\ D^T x - d + s \\ H^T v - h + r \\ \Omega s - \mu \mathbf{1} \\ Wr - \mu \mathbf{1} \end{bmatrix} = 0,$$

where $\omega, s, w, r \geq 0$ is enforced by the line search.

Interior point methods apply damped Newton iterations to find a solution to $F_\mu = 0$ (with ω, s, w, r nonnegative) as μ is driven to 0, so that cluster points are necessarily KKT points of the original problem. Damped Newton iterations take the following form. Let $\xi := [x^T, v^T, s^T, r^T, \omega^T, w^T]^T$. Then the iterations are given by

$$\xi^{\kappa+1} := \xi^\kappa - \gamma (F^{(1)}_{\mu_\kappa})^{-1} F_{\mu_\kappa},$$

where $F^{(1)}_{\mu_\kappa}$ is the Jacobian of the relaxed KKT system F_μ. The γ is chosen to satisfy two conditions: (1) $\omega^{\kappa+1}, w^{\kappa+1}, s^{\kappa+1}, r^{\kappa+1}$ remain positive, and (2) $\|F_{\mu_\kappa}(\xi^{\kappa+1})\|$ decreases. The homotopy parameter μ_κ is decreased at each iteration in a manner that preserves a measure of centrality within the feasible region.

While interior point methods have a long history (see e.g. Nemirovskii & Nesterov, 1994; Wright, 1997), using them in this manner to solve any PLQ problem in a uniform way was proposed in Aravkin et al. (2013b), and we refer the reader to this reference for implementation details. In particular, the Kalman smoothing case is developed in Section 6. Each iteration of the resulting conjugate-PLQ interior point method can be implemented with a complexity of $O(N(n^3 + m^3))$, which scales linearly in with N, just as for the classic smoother. The local convergence rate for IP methods is superlinear or quadratic in many circumstances (Ye, 2011), which in practice means that few iterations are required.

5. Numerical experiments and illustrations

We now present a few numerical results to illustrate the formulations and algorithms discussed above. In Section 5.1, we consider a nonsmooth Kalman formulation and compare the subgradient method, Chambolle–Pock, and interior point methods. In Section 5.2, we show how nonsmooth formulations can be used to address the motivating examples in the introduction. Finally, in Section 5.3, we show how to construct general piecewise linear quadratic Kalman smoothers (with constraints) using the open-source package IPsolve.[3]

5.1. Algorithms and convergence rates

In this section, we consider a particular signal tracking problem, where the underlying smooth signal is a sine wave, and a portion of the measurements are outliers.

The synthetic ground truth function is given by $x(t) = \sin(-t)$. We reconstruct it from direct noisy samples taken at instants multiple of Δt.

We use $N = 100$ time steps for this example.

We track this smooth signal by modeling it as an integrated Brownian motion which is equivalent to using cubic smoothing splines (Wahba, 1990). The state space model (sampled at instants where data are collected) is given by Bell et al. (2009), Jazwinski (1970) and Oksendal (2005)

$$\begin{bmatrix} \dot{x}_{t+1} \\ x_{t+1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_t \\ x_t \end{bmatrix} + v_t$$

where the model covariance matrix of v_t is

$$Q_t = \begin{bmatrix} \Delta t & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t^3/3 \end{bmatrix}.$$

The goal is to reconstruct the signal function from direct noisy measurements y_t, given by

$$y_t = C_t x_t + e_t, \quad C_t = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

We solve the following constrained modification of (34):

$$\min_{x \in \mathscr{C}} \|R^{-1/2}(y - Cx)\|_1 + \frac{1}{2}\|Q^{-1/2}(z - Ax)\|^2, \tag{42}$$

where z is constructed as in (10). For the sine wave, $\mathscr{C}$ is a simple bounding box, forcing each component to be in $[-1, 1]$. Our goal is to compare three algorithms discussed in Section 4:

(1) Projected subgradient method. We use the step size $\alpha_\kappa := \frac{1}{\kappa}$, and apply projected subgradient:

$$x^{\kappa+1} := \text{proj}_{\mathscr{C}}\left(x^\kappa - \frac{1}{\kappa} v^\kappa\right),$$

where $v^\kappa \in \partial f(x^\kappa)$ is any element in the subgradient.
(2) Chambolle–Pock (two variants described below).
(3) Interior point formulation for (39).

Multiple splitting methods can be applied, including ADMM (customized to deal with two nonsmooth terms), or the three-term splitting algorithm of Davis and Yin (2015). We focus instead on a simple comparison of two variants of Chambolle–Pock with extremely different behaviors.

[3] https://github.com/UW-AMO/IPsolve.

To apply Chambolle–Pock, we first write the optimization problem (42) using the template

$$\min_x f(Kx - r) + g(x).$$

The Chambolle–Pock iterations (see Algorithm 7) are given by

$$\begin{aligned} \omega^{\kappa+1} &:= r + \text{prox}_{\sigma f^*}(\omega^\kappa + \sigma K(2x^\kappa - x^{\kappa-1}) - r) \\ x^{\kappa+1} &:= \text{proj}_{\tau g}(x^\kappa - \tau K^T \omega^{\kappa+1}), \end{aligned}$$

where τ and σ are stepsizes that must satisfy $\tau\sigma L < 1$, and L is the squared operator norm of K. Choices for K give rise to two different CP algorithms, denoted by CP-V1 and CP-V2 below.

CP-V1. One way to make the assignment is as follows:

$$f(\omega_1, \omega_2) = \|\omega_1\|_1 + \frac{1}{2}\|\omega\|_2^2, \quad g(x) = \delta_{\mathscr{C}}(x)$$

$$f^*(\eta_1, \eta_2) = \delta_{\mathbb{B}_\infty}(\eta_1) + \frac{1}{2}\|\eta_2\|^2.$$

$$K = \begin{bmatrix} R^{-1/2}C \\ Q^{-1/2}A \end{bmatrix}, \quad r = \begin{bmatrix} R^{-1/2}y \\ Q^{-1/2}z \end{bmatrix}.$$

The conjugate of $\|\cdot\|_1$ is computed in (32), and it is easy to see that the function $\frac{1}{2}\|\cdot\|^2$ is its own conjugate using definition (31).

To understand the ω-step, observe that

$$\text{prox}_{\sigma(f_1^*(x_1)+f_2^*(x_2))}\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) = \begin{bmatrix} \text{prox}_{\sigma f_1^*}(y_1) \\ \text{prox}_{\sigma f_2^*}(y_2) \end{bmatrix} = \begin{bmatrix} \text{proj}_{\mathbb{B}_\infty}(y_1) \\ \dfrac{1}{1+\sigma}\, y_2 \end{bmatrix}.$$

The proximity operator for the indicator function is derived in (29), and the proximity operator for $\frac{1}{2}\|\cdot\|^2$ can be easily obtained. The x-step requires a projection onto the set $\mathscr{C}$, which is the unit box for the sine example.

CP-V2. Here we treat $\frac{1}{2}\|Q^{-1/2}(Ax - z)\|^2$ as a unit, and assign in to g. As a result, the behavior of A plays no role in the convergence rate of the algorithm.

$$f(\omega_1, \omega_2) = \|\omega\|_1 + \delta_{\mathscr{C}}(\omega_2), \quad g(x) = \frac{1}{2}\|Q^{-1/2}(Ax - z)\|^2$$

$$f^*(\eta_1, \eta_2) = \delta_{\mathbb{B}_\infty}(\eta_1) + \|\eta_2\|_1.$$

$$K = \begin{bmatrix} R^{-1/2}C \\ I \end{bmatrix}, \quad r = \begin{bmatrix} R^{-1/2}y \\ 0 \end{bmatrix}.$$

The proximity operator for g is obtained by solving a linear system of equations:

$$\text{prox}_{\tau g}(y) = (\tau A^T Q^{-1} A + I)^{-1}(y + \tau A^T Q^{-1} z).$$

The matrix $\tau A^T Q^{-1} A + I$ is block tridiagonal positive definite, and its eigenvalues are bounded away from 0. Since it does not change between iterations, we compute its Cholesky factorization once and use it to implement the inversion at each iteration. This requires a single factorization using $O(n^3 N)$ arithmetic operations, followed by multiple $O(n^2 N)$ iterations (same cost as matrix–vector products with a block tridiagonal system).

The ω-step for CP-V2 is also different from the ω-step in CP-V1, but still very simple and efficient:

$$\begin{aligned} \text{prox}_{\sigma(f_1^*(x_1)+f_2^*(x_2))}\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) &= \begin{bmatrix} \text{prox}_{\sigma f_1^*}(y_1) \\ \text{prox}_{\sigma f_2^*}(y_2) \end{bmatrix} \\ &= \begin{bmatrix} \text{proj}_{\mathbb{B}_\infty}(y_1) \\ \text{prox}_{\sigma\|\cdot\|_1}(y_2) \end{bmatrix}. \end{aligned}$$

The proximity operator for $\sigma\|\cdot\|_1$ is derived in (33).

The results are shown in Fig. 6. The subgradient method is disastrously slow. Given a simple step size schedule, e.g. $\alpha_\kappa = \frac{1}{\kappa}$, it may waste tens of thousands of iterations before the objective starts to decrease. In the left panel of Fig. 6, it took over 10,000 iterations before any noticeable impact. Moreover, as the step

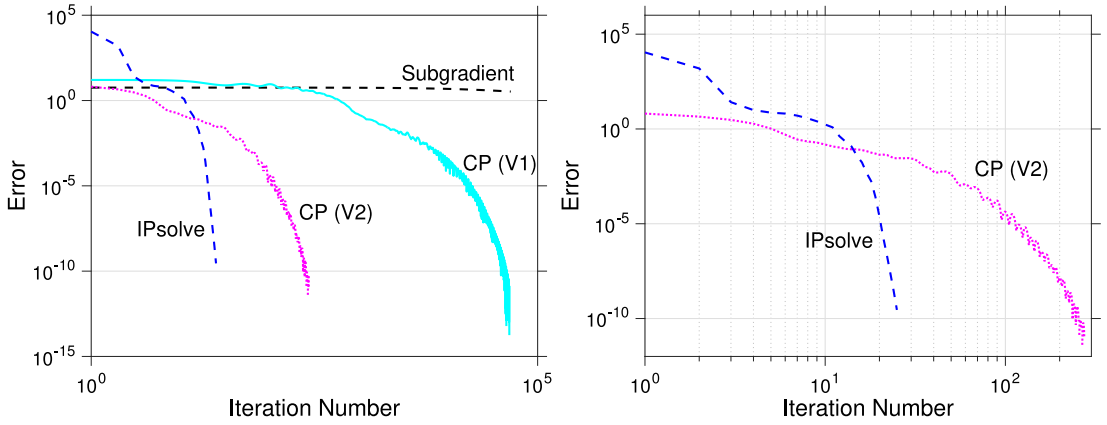


Fig. 6. Convergence rate comparisons. The y-axis shows $f(x^t) - f(x^*)$, while x-axis shows the iteration count. *Left*: Convergence rates for subgradient, CP-V1, CP-V2, and Interior Point methods, after 50,000 iterations. *Right*: Comparison for CP-V2 and IPsolve, after 300 iterations. Note that the methods have different complexities: subgradient and CP-V1 use only matrix–vector products; CP-V2 requires a single factorization and then back-substitution at each iteration, and IPsolve solves linear systems of equations at each iteration. The run times were: 4.11 s, 1.9 s 0.03 s and 0.05 s for subgradient, CP-V1, CP-V2, and IPsolve. The subgradient method terminated because it hit the maximum number of iterations (50,000); the remaining methods converged.

sizes become small, it can stagnate, and while in theory it should continue to slowly improve the objective, in practice it stalls on the example problem.

CP-V1 is able to make some progress, but the results are not impressive. Even though the algorithm requires only matrix–vector products, it is adversely impacted by the conditioning of the problem. In particular, the ODE term for the Kalman smoothing problem (i.e. the A) can be poorly conditioned, and in the CP-V1 scheme, it sits inside K. As a result, we see very slow convergence. Interestingly, the rate itself looks linear, but the constants are terrible, so it requires 50,000 iterations to fully solve the problem.

In contrast, CP-V2 performs extremely well. The algorithm treats the quadratic ODE term as a unit, and the ill-conditioning of A does not impact the convergence rate. The price we pay is having to solve a linear system of equations at each iteration. However, since the system does not change, we factorize it, with a one-time cost of $O(n^3N)$ operations, and then use back-substitution to implement prox_g at each iteration, at a cost of $O(n^2N)$ operations at each iteration. The resulting empirical convergence rate is also linear, but with a significant improvement in the constant: CP-V2 needs only 300 iterations to reach 10^{-10} accuracy (gap to the minimum objective value), see the right plot of Fig. 6.

Finally, IPsolve has a *super-linear* rate, and finishes in 27 iterations. It is not possible to pre-factorize any linear systems of equations, so the complexity is $O(n^3N)$ for each iteration. For moderate problem sizes (specifically, smaller n), this approach is fast and generalizes to any PLQ losses V and J and any constraints. For large problem sizes, CP-V2 will win; however, it is very specific to the current problem. In particular, if we change J in (15) from the quadratic to the 1-norm or Huber, we would need to develop a different splitting approach. The more general CP-V1 approach is far less effective.

The following sections focus on modeling and the resulting behavior of the estimates. Section 5.2 presents the results for the motivating examples in the introduction.

5.2. DC motor: robust solutions using ℓ_1 losses and penalties

We now solve the problems described in Section 1.1 using two different smoothing formulations based on the ℓ_1 norm.

Impulsive inputs: Let $E_1 = \begin{pmatrix} 1 & 0 \end{pmatrix}$, $E_2 = \begin{pmatrix} 0 & 1 \end{pmatrix}$. To reconstruct the disturbance torque d_t acting on the motor shaft, we use the LASSO-type estimator proposed in Ohlsson et al. (2012):

$$\min_{x_1,\ldots,x_N} \sum_{t=1}^{N} (y_t - E_2 x_t)^2 + \gamma \sum_{t=0}^{N-1} |d_t| \quad \text{subject to the dynamics (9).} \tag{43}$$

Since $u_t = 0$, this corresponds to the optimization problem

$$\min_{x_1,\ldots,x_N} \quad \sum_{t=1}^{N} (y_t - E_2 x_t)^2 + \frac{\gamma}{2}\left[\sum_{t=0}^{N-1} \frac{|E_1(x_{t+1} - A_t x_t)|}{11.81} + \frac{|E_2\,(x_{t+1} - A_t x_t)|}{0.625}\right]$$
$$\text{subject to} \quad \frac{E_1\,(x_{t+1} - A_t x_t)}{11.81} = \frac{E_2\,(x_{t+1} - A_t x_t)}{0.625}.$$

The regularization parameter γ is tuned using 5-fold cross validation on a grid consisting of 20 values, logarithmically spaced between 0.1 and 10. The resulting smoother is dubbed LASSO-CV.

The right panel of Fig. 7 shows the estimate of d_t obtained by LASSO-CV starting from the noisy outputs in the left panel. Note that we recover the impulsive disturbance, and that the LASSO smoother outperforms the optimal linear smoother L_2-opt, shown in Fig. 1. To further examine the improved performance of the LASSO smoother in this setting, we performed a Monte Carlo study of 200 runs, comparing the fit measure

$$100\left(1 - \frac{\|\hat{d} - d\|}{\|d\|}\right),$$

where $d = [d_1 \ldots d_{200}]$ is the true signal and $\hat{d}$ is the estimate returned by L_2-opt or by LASSO-CV. Fig. 8 shows Matlab boxplots of the 200 fits obtained by these estimators. The rectangle contains the inter-quartile range (25%–75% percentiles) of the fits, with median shown by the red line. The "whiskers" outside the rectangle display the upper and lower bounds of all the numbers, not counting what are deemed outliers, plotted separately as "+". The effectiveness of the LASSO smoother is clearly supported by this study.

Presence of outliers: To reconstruct the angular velocity, we use the following smoother based on the ℓ_1 loss:

$$\min_{x_1,\ldots,x_N} \sum_{t=1}^{N} \frac{|y_t - E_2 x_t|}{\sigma} + \frac{1}{0.1^2} \sum_{t=0}^{N-1} d_t^2 \quad \text{subject to the dynamics (9).} \tag{44}$$

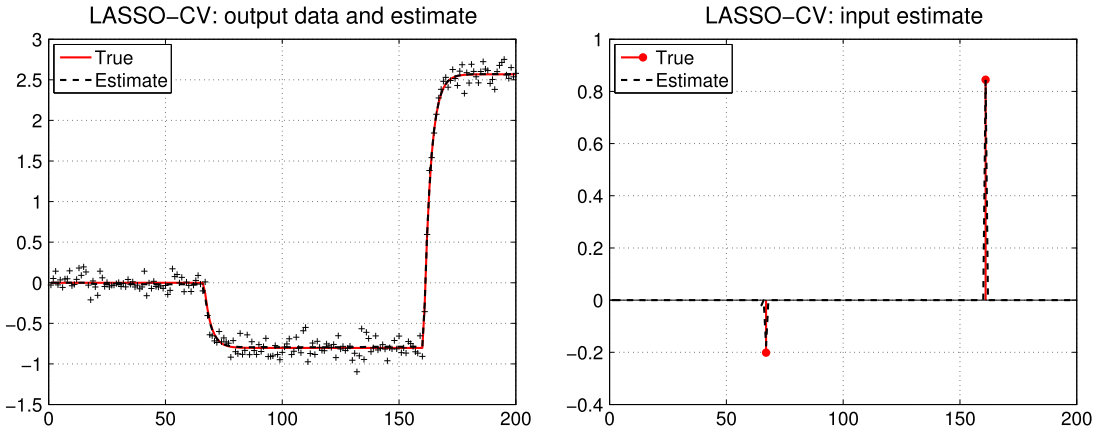


Fig. 7. DC motor and impulsive disturbances. *Left*: noiseless output (solid line), measurements (+) and output reconstruction by the LASSO smoother (dashed line). *Right*: impulsive disturbance and reconstruction by the LASSO smoother (dashed line).

Recall that $d_t \sim \mathcal{N}(0, 0.1^2)$, so now there is no impulsive input. The ℓ_1 loss used in (44) is shown in Fig. 4(b). It can also be viewed as a limiting case of Huber (Fig. 4(c)) and Vapnik (Fig. 4(d)) losses, respectively, when their breakpoints κ and ε are set to zero.

Over the state space domain, problem (44) is equivalent to

$$\min_{x_1,\ldots,x_N} \sum_{t=1}^{N} \frac{|y_t - E_2 x_t|}{\sigma} + \frac{1}{0.1^2}\left[\sum_{t=0}^{N-1} \frac{(E_1(x_{t+1} - A_t x_t))^2}{11.81} + \frac{(E_2(x_{t+1} - A_t x_t))^2}{0.625}\right]$$
$$\text{subject to } \frac{E_1\,(x_{t+1} - A_t x_t)}{11.81} = \frac{E_2\,(x_{t+1} - A_t x_t)}{0.625}.$$

Note that the ℓ_1 loss uses the nominal standard deviation $\sigma = 0.1$ as weight for the residuals, so that we call this estimator L_1-nom.

The left panel of Fig. 9 displays the estimate of the angle returned by L_1-nom. The profile is very close to truth, revealing the robustness of the smoother to the outliers. Here, we have also performed a Monte Carlo study of 200 runs, using the fit measure

$$100\left(1 - \frac{\|\hat{y} - y\|}{\|y\|}\right),$$

where $y = [y_1 \ldots y_{200}]$ is the true value while $\hat{y}$ are the estimates returned by L_2-nom, L_2-opt or L_1-nom. The boxplots in the right panel of Fig. 9 compare the fits of the three estimators, and illustrate the robustness of L_1-nom.

Finally, we repeated the same Monte Carlo study setting $\alpha = 0$, generating no outliers in the output measurements. Under these assumptions, L_2-nom and L_2-opt coincide and represent the best estimator among all the possible smoothers. Fig. 10 shows Matlab boxplots of the 200 fits obtained by L_2-nom and L_1-nom. Remarkably, the robust smoother has nearly identical performance to the optimal smoother, so there is little loss of performance under nominal conditions.

5.3. Modeling with PLQ using Ipsolve

In this section, we include several modeling examples that combine robust penalties with constraints. Each example is implemented using IPsolve. The solver and examples are available online[4] ; see in particular `515Examples/KalmanDemo.m`. In all examples, the ground truth function of interest is given by $x(t) = \exp(\sin(4t))$, and we reconstruct it from direct and noisy samples taken at instants multiple of Δt. The function $x(t)$ is smooth and periodic, but the exponential accelerates the transitions around the maximum and minimum values. The process and measurement models are the same as in Section 5.1. Four smoothers (15) are compared in this example using IPsolve. The L_2 smoother uses the quadratic penalty for both V and J, and no constraints. The cL_2 smoother uses least squares penalties with constraints including the information that $\exp(-1) \leq x(t) \leq \exp(1)\ \forall t$. The Huber smoother uses Huber penalties ($\kappa = 1$) for both V and J, without constraints, while cHuber uses Huber penalties ($\kappa = 1$) together with constraints. The results are shown in Fig. 11. 90% of the measurement errors are generated from a Gaussian with nominal standard deviation 0.05, while 10% of the data are large outliers generated using a Gaussian with standard deviation 10. The smoother is given the nominal standard deviation.

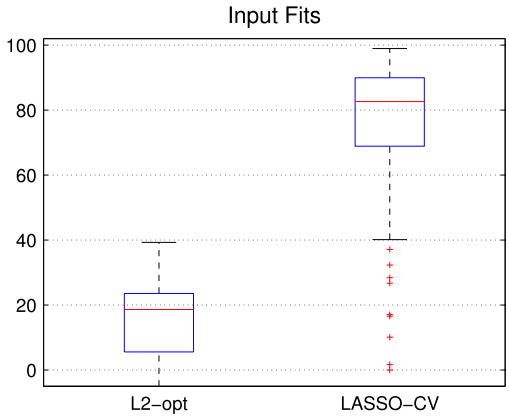


Fig. 8. DC motor and impulsive disturbances. Boxplot of the fits returned by optimal linear smoother (left) and by the LASSO smoother (right).

The least squares smoother L_2 without constraints does a very poor job. The Huber smoother obtains a much better fit. Interestingly, cL_2 is much better than L_2, indicating that domain constraints can help a lot, even when using quadratic penalties. Combining constraints and robustness in cHuber gives the best fit since the inclusion of constraints eliminates the constraint violations of Huber at instant 3 in the left plot of Fig. 11.

The run times for $N = 1000$, 10 000, and 100 000 are shown in Table 1. The run times scale linearly with N, as expected.

The calls to IPsolve are given below:

[4] https://github.com/saravkin/IPsolve.

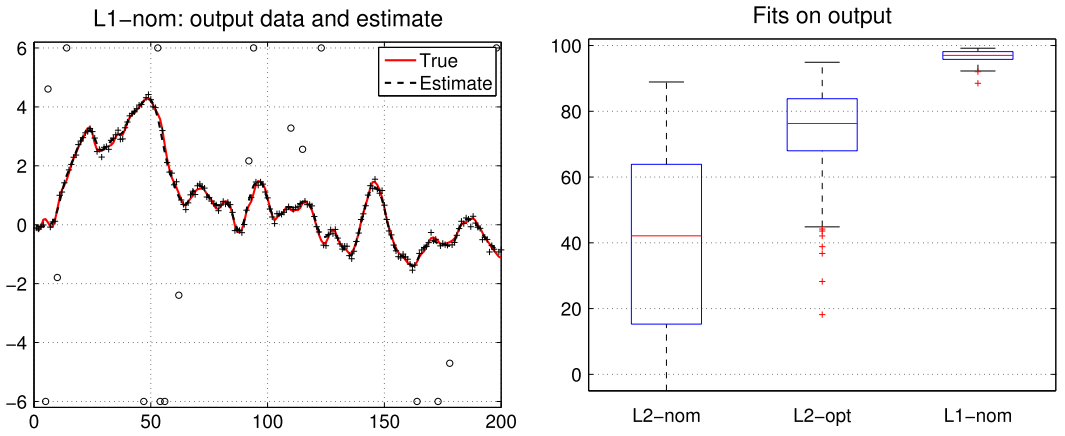


Fig. 9. DC motor and outliers in the output measurements. *Left*: noiseless output (solid line), measurements (+), outliers (∘) and output reconstruction by the robust smoother L_1-nom which uses the ℓ_1 loss and the nominal noise standard deviation $\sigma = 0.1$ as weight for the residuals (dashed). *Right*: boxplot of the output fits returned by the linear smoother L_2-nom which uses the nominal standard deviation $\sigma = 0.1$ as weight for the residuals, by the optimal linear smoother L_2-opt and by the robust smoother L_1-nom.

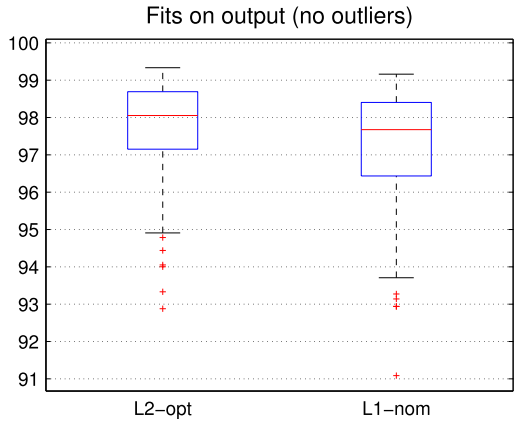


Fig. 10. DC motor and output reconstruction without outliers corrupting the measurements. Boxplot of the output fits returned by the optimal linear smoother L_2-opt and by the robust smoother L_1-nom.

(1) L_2:

```
params.K = Gmat; params.k = w;
L2 = run_example( Hmat, meas, 'l2', 'l2', ...
[], params );
```

(2) Huber:

```
params.K = Gmat; params.k = w;
Huber = run_example( Hmat, meas, 'huber', ...
'huber', [], params );
```

The only difference required to run the HH smoother is to replace the names of the PLQ penalties in the calling sequence.

(3) cL_2:

```
params.K = Gmat; params.k = w;
params.constraints = 1; conA = [0 1; 0 -1];
cona = [exp(1); -exp(-1)];
params.A = kron(speye(N), conA)';
params.a = kron(ones(N,1), cona);
cL2 = run_example( Hmat, meas, 'l2', 'l2',...
 [], params );
```

For constraints, we need to create the constraint matrix and also pass it in using the `params` structure.

Table 1
Run times (s) of IPsolve for Huber, least squares with box constraints, and Huber with box constraints.

N	L2	Huber	cL2	cHuber
1000	.001	0.125	.053	0.15
10 000	.008	0.7	0.2	1.3
100 000	.1	10.4	2.14	15.8

(4) cHuber:

```
params.K = Gmat; params.k = w;
params.constraints = 1; conA = [0 1; 0 -1];
cona = [exp(1); -exp(-1)];
params.A = kron(speye(N), conA)';
params.a = kron(ones(N,1), cona);
cHuber = run_example( Hmat, meas, 'huber',...
 'huber',[], params );
```

The constrained Huber call sequence requires only a name change for the PLQ penalties. A library of PLQ penalties is included in IPsolve; users can construct their own custom penalties once they understand the conjugate representation (38).

Above, one can see that the names of PLQ measurements are arguments to the file `run_example`, which builds the combined PLQ model object that it passes to the interior point method. The measurement matrix and observations vector are also passed directly to the solver. The process terms are passed through the auxiliary `params` structure. Full details for constructing the matrices are provided in the online demo `KalmanDemo` cited above.

6. Concluding remarks

Various aspects of the state estimation problem in the linear system (1) have been treated over many years in a very extensive literature. One reason for the richness of the literature is the need to handle a variety of realistic situations to characterize the signals v and e in (1). This has led to deviations from the classical situation with Gaussian signals where the estimation problem is a linear–quadratic optimization problem. This survey attempts to give a comprehensive and systematic treatment of the main issues in this large literature. The key has been to start with a general formulation (15) that contains the various situations as special cases of the functions V and J. An important feature is that (15) still is a convex optimization problem under mild and natural assumptions. This

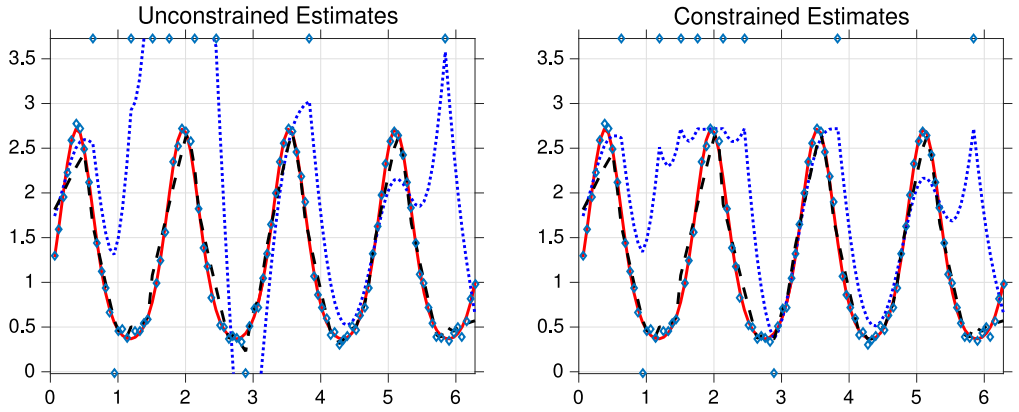


Fig. 11. Results of four smoothers. *Left*: Ground truth (solid red) and unconstrained results for L2 (dashed blue) and Huber (densely dashed black). *Right*: Ground truth (solid red) and constrained results for cL2 (dashed blue) and cHuber (densely dashed black). Constraints can be very helpful in dealing with contamination. Best results are obtained when we use both robust penalties and constraints on the domain. Here, $n = 2$ and $N = 100$. Run times for $N = 1000$, 10 000, and 100 000 are shown in Table 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

opens the huge area of convex optimization as a fruitful arena for state estimation. In a way, this alienates the topic from the original playground of Gaussian estimation techniques and linear algebraic solutions. The survey can therefore also be read as a tutorial on convex optimization techniques applied to state estimation.

Acknowledgments

The authors wish to thank Damek Davis for insightful conversations about operator splitting. This research was supported in part by the National Science Foundation grant no. DMS-1514559, by the Washington Research Foundation Data Science Professorship, by the WRF Data Science Professorship, by the MIUR FIRB project RBFR12M3AC-Learning meets time: a new computational approach to learning in dynamic systems, by the Progetto di Ateneo CPDA147754/14-New statistical learning approach for multi-agents adaptive estimation and coverage control as well as by the Linnaeus Center CADICS, funded by the Swedish Research Council 353-2012-5860, and the ERC advanced grant LEARN, no 287381, funded by the European Research Council.

Appendix

A.1. Optimization viewpoint on Kalman smoothing under correlated noise and singular covariances

In some applications, the noises $\{e_t, v_t\}_{t=1}^N$ are correlated. Assume that e_t and v_t are still jointly Gaussian, but with a cross-covariance denoted by S_t. For $t = 1, \ldots, N$, this implies that the last assumption in (3) can be replaced by

$$E(v_s e_t^\top) = \begin{cases} S_t & \text{if } t = s \\ 0 & \text{otherwise,} \end{cases}$$

while v_0 is assumed independent of $\{e_t, v_t\}_{t=1}^N$.

We now reformulate the objective (6) under this more general model. Define the process $\tilde{v}_0 = v_0$ and

$$\tilde{v}_t = v_t - E(v_t|e_t) = v_t - S_t R_t^{-1} e_t, \quad t \geq 1$$

which, by basic properties of Gaussian estimation, is independent of e_t and consists of white noise with covariance

$$\tilde{Q}_t = Q_t - S_t R_t^{-1} S_t^\top, \quad t \geq 1.$$

Since v_t is correlated only with e_t, we have that all the $\{\tilde{v}_t\}$ and $\{e_t\}$ form a set of mutually independent Gaussian noises. Also, since $e_t = y_t - C_t x_t$, model (1) can be reformulated as

$$x_{t+1} = \tilde{A}_t x_t + B_t u_t + S_t R_t^{-1} y_t + \tilde{v}_t \tag{45a}$$

$$y_t = C_t x_t + e_t \tag{45b}$$

where we define $\tilde{A}_0 x_0 + S_0 R_0^{-1} y_0 = A_0 x_0$ while

$$\tilde{A}_t = A_t - S_t R_t^{-1} C_t, \quad t \geq 1.$$

Note that (45) has the same form as the original system (1) except for the presence of an additional input given by the output injection $S_t R_t^{-1} y_t$.

Assuming also the initial condition x_0 independent of the noises, the joint density of $\{\tilde{v}_t\}$, $\{e_t\}$ and x_0 turns out to be

$$\mathbf{p}(x_0, \{e_t\}, \{\tilde{v}_t\}) = \mathbf{p}(x_0) \prod_{t=1}^{N} \mathbf{p}_{e_t}(e_t) \prod_{t=0}^{N-1} \mathbf{p}_{\tilde{v}_t}(\tilde{v}_t),$$

where we use $\mathbf{p}_{e_t}$ and $\mathbf{p}_{\tilde{v}_t}$ to denote the densities corresponding to e_t and $\tilde{v}_t$. Since $\{x_t\}_{t=0}^N$ and $\{y_t\}_{t=1}^N$ are a linear transformation of $\{v_t\}_{t=0}^N$, $\{e_t\}_{t=1}^N$ and x_0, the joint posterior of states and outputs is proportional to

$$\mathbf{p}(x_0) \prod_{t=1}^{N} \mathbf{p}_{e_t}(y_t - C_t x_t) \prod_{t=0}^{N-1} \mathbf{p}_{\tilde{v}_t}\left(x_{t+1} - \tilde{A}_t x_t - S_t R_t^{-1} y_t - B_t u_t\right).$$

Consequently, maximizing the posterior of the states given the output measurements is equivalent to solving

$$\begin{aligned} \min_{x_0, \ldots, x_N} \quad & \|\Pi^{-1/2}(x_0 - \mu)\|^2 + \sum_{t=1}^{N} \|R_t^{-1/2}(y_t - C_t x_t)\|^2 \\ & + \sum_{t=0}^{N-1} \|\tilde{Q}_t^{-1/2}(x_{t+1} - \tilde{A}_t x_t - S_t R_t^{-1} y_t - B_t u_t)\|^2. \end{aligned} \tag{46}$$

Next consider the case where some of the covariance matrices are singular. If some of the matrices Q_t or R_t are not invertible, problems (46) and (6) are not well-defined. In this case, one can proceed as follows. First, $\tilde{v}_t$, $\tilde{Q}_t$ and $\tilde{A}_t$ can be defined in the same way where R_t^{-1} is replaced by its pseudoinverse $R_t^\dagger$. The objective can then be reformulated by replacing $\tilde{Q}_t^{-1}$ and R_t^{-1} by $\tilde{Q}_t^\dagger$ and $R_t^\dagger$, respectively. Linear constraints can be added to prevent the state evolution in the null space of $\tilde{Q}_t$ and R_t. By letting I_Q and I_R be the sets with the time instants associated with singular $\tilde{Q}_t$ and R_t,

problem (46) can be rewritten as

$$\begin{aligned}\min_{x_0,\dots,x_N}\ & \|\Pi^{-1/2}(x_0-\mu)\|^2+\sum_{t=1}^{N}\|(R_t^\dagger)^{1/2}(y_t-C_tx_t)\|^2\\ &+\sum_{t=0}^{N-1}\|(\tilde{Q}_t^\dagger)^{1/2}(x_{t+1}-\tilde{A}_tx_t-S_tR_t^{-1}y_t-B_tu_t)\|^2 \qquad (47)\\ &\text{subject to } R_t^\perp\,(y_t-C_tx_t)=0 \text{ for } t\in I_R \text{ and}\\ &\tilde{Q}_t^\perp\left(x_{t+1}-A_tx_t-S_tR_t^\dagger y_t-B_tu_t\right)=0 \text{ for } t\in I_Q,\end{aligned}$$

where $R_t^\perp = I - R_tR_t^\dagger$ and $\tilde{Q}_t^\perp = I-\tilde{Q}_t\tilde{Q}_t^\dagger$ provide the projections onto the null-space of R_t and $\tilde{Q}_t$, respectively.

A.2. Convex analysis and optimization

Some of the background in convex analysis and optimization used in the previous sections is briefly reviewed in this section. In particular, the fundamentals used in the development and analysis of algorithms for (15) are reviewed.

Many members of the broader class of penalties (15) do not yield least squares objectives since they include nonsmooth penalties and constraints; however, they are convex. Convexity is a fundamental notion in optimization theory and practice and gives access to globally optimal solutions as well as extremely efficient and reliable numerical solution techniques that scale to high dimensions. The relationship between convex sets and functions was presented in Section 4.1.

Fundamental objects in convex analysis

We begin by developing a duality theory for the general objective (15). This is key for both algorithm design and sensitivity analysis. Duality is a consequence of the separation theory for convex sets.

Separation: We say that a hyperplane (i.e. an affine set of codimension 1) separates two sets if they lie on opposite sides of the hyperplane. To make this idea precise, we introduce the notion of *relative interior*. The affine hull of a set $\mathscr{E}$, denoted $\text{aff}(\mathscr{E})$, is the intersection of all affine sets that contain $\mathscr{E}$. Given $\mathscr{E}\subset\mathbb{R}^n$ the relative interior of $\mathscr{E}$ is

$$\text{ri}(\mathscr{E}) := \{x\in\mathscr{E} \mid \exists\varepsilon>0 \text{ s.t. } (x+\varepsilon\mathbb{B})\cap\text{aff}(\mathscr{E})\subset\mathscr{E}\}.$$

For example, $\text{ri}\{(2,x)\mid -1\le x\le 1\} = \{(2,x)\mid -1<x<1\}$.

Let $\text{cl}(\mathscr{E})$ denote the closure of set $\mathscr{E}$, and $\text{intr}(\mathscr{E})$ denote the interior. Then the boundary of $\mathscr{E}$ is given by $\text{bdry}(\mathscr{E}) := \text{cl}(\mathscr{E})\setminus\text{intr}(\mathscr{E})$, and the relative boundary $\text{rbdry}(\mathscr{C})$ is given by $\text{cl}(\mathscr{C})\setminus\text{ri}(\mathscr{C})$.

Theorem 2 (*Separation*). *Let $\mathscr{C}\subset\mathbb{R}^n$ be nonempty and convex, and suppose $\bar{y}\notin\text{ri}(\mathscr{C})$. Then there exist $z\neq 0$ such that*

$$\langle z,\bar{y}\rangle > \langle z,y\rangle \quad \forall\, y\in\text{ri}(\mathscr{C}).$$

Support Function: Apply Theorem 2 to a point $\bar{x}\in\text{rbdry}(\mathscr{C})$ to obtain a nonzero vector z for which

$$\langle z,\bar{x}\rangle = \sigma_{\mathscr{C}}(z) := \sup\{\langle z,x\rangle \mid x\in\mathscr{C}\} > \inf\{\langle z,x\rangle \mid x\in\mathscr{C}\}. \qquad (48)$$

The function $\sigma_{\mathscr{C}}$ is called the *support function* for $\mathscr{C}$, and the nonzero vector z is said to be a support vector to $\mathscr{C}$ at $\bar{x}$. When $\mathscr{C}$ is polyhedral, $\sigma_{\mathscr{C}}$ is an example of a PLQ function, with (48) a special case of (38) with $M=0$.

Example (*Dual Norms*). Given a norm $\|\cdot\|$ on $\mathbb{R}^n$ with unit ball $\mathbb{B}$, the dual norm is given by

$$\|z\|_\circ := \sup_{\|x\|\le 1}\langle z,x\rangle = \sigma_{\mathbb{B}}(z).$$

For example, the 2-norm is self dual, while the dual norm for $\|\cdot\|_1$ is $\|\cdot\|_\infty$.

This definition implies that $\|x\| = \sigma_{\mathbb{B}^\circ}(x)$, where

$$\mathbb{B}^\circ := \{z\mid \langle z,x\rangle\le 1\ \forall\, x\in\mathbb{B}\}.$$

The set $\mathbb{B}^\circ$ is the closed unit ball for the dual norm $\|\cdot\|_\circ$. This kind of relationship between the unit ball of a norm and that of its dual generalizes to *polars* of sets and cones.

Polars of sets and cones: For any set $\mathscr{C}$ in $\mathbb{R}^n$, the set

$$\mathscr{C}^\circ := \{z\mid \langle z,x\rangle\le 1\ \forall\, x\in\mathscr{C}\}$$

is called the *polar* of $\mathscr{C}$, and we have $(\mathscr{C}^\circ)^\circ = \text{cl}(\text{conv}(\mathscr{C}\cup\{0\}))$. Hence, if $\mathscr{C}$ is a closed convex set containing the origin, then $(\mathscr{C}^\circ)^\circ = \mathscr{C}$. If $\mathscr{K}\subset\mathbb{R}^n$ is a convex cone ($\mathscr{K}$ is a convex and $\lambda\mathscr{K}\subset\mathscr{K}$ for all $\lambda>0$), then, by rescaling,

$$\mathscr{K}^\circ = \{z\mid \langle z,x\rangle\le 0\ \forall\, x\in\mathscr{K}\} \text{ and } (\mathscr{K}^\circ)^\circ = \text{cl}(\mathscr{K}).$$

In particular, this implies that $\sigma_{\mathscr{K}} = \delta_{\mathscr{K}^\circ}$.

Subdifferential: For nonsmooth convex functions, the notion of derivative can be captured by examining support vectors to their epigraph. Define the domain of the function f to be the set $\text{dom}(f) := \{x\mid f(x)<\infty\}$. Using the fact that

$$\text{ri}(\text{epi}(f)) = \{(x,\mu)\mid x\in\text{ri}(\text{dom}(f)) \text{ and } f(x)<\mu\},$$

Theorem 2 tells us that, for every $\bar{x}\in\text{ri}(\text{dom}(f))$, there is a support vector to $\text{epi}(f)$ at $(\bar{x},f(\bar{x}))$ of the form $(z,-1)$, which separates the points in the epigraph from the points in a half space below the epigraph:

$$\langle(z,-1),(\bar{x},f(\bar{x}))\rangle \ge \langle(z,-1),(x,f(x))\rangle \quad \forall\, x\in\text{dom}(f),$$

or equivalently,

$$f(\bar{x})+\langle z,x-\bar{x}\rangle\le f(x) \quad \forall\, x\in\text{dom}(f). \qquad (49)$$

This is called the *subgradient inequality*. The vectors z satisfying (49) are said to be subgradients of f at $\bar{x}$, and the set of all such subgradients is called the *subdifferential* of f at $\bar{x}$, denoted $\partial f(\bar{x})$. This derivation shows that $\partial f(\bar{x})\neq\emptyset$ for all $\bar{x}\in\text{ri}(\text{dom}(f))$ when f is *proper*, i.e. $\text{dom}(f)$ is nonempty, with $f(x)>-\infty$. In addition, it can be shown that $\partial f(\bar{x})$ is a singleton if and only if f is differentiable at $\bar{x}$ with the gradient equal to the unique subgradient.

For example, the absolute value function on $\mathbb{R}$ is not differentiable at zero so there is no tangent line to its graph at zero; however, every line passing through the origin having slope between -1 and 1 defines a support vector to the epigraph at the origin. In this case, we can replace the notion of derivative by the set of slopes of hyperplanes at the origin. Each of these slopes is a subgradient, and the set of all these is the *subdifferential* of $|\cdot|$ at the origin.

Necessary and Sufficient Conditions for Optimality: An immediate consequence of the subgradient inequality is that

$$0\in\partial f(\bar{x}) \quad \text{if and only if} \quad \bar{x}\in\text{argmin} f.$$

That is, a first-order necessary and sufficient condition for optimality in convex optimization is that the zero vector is an element of the subdifferential. Returning to the absolute value function on $\mathbb{R}$, note that the zero slope hyperplane supports the epigraph at zero and zero is the global minimizer of $|\cdot|$.

Theorem 3 (*Convex Optimality*). *Let $f:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be a closed proper convex function. Then the following conditions are*

equivalent:

(i) $\bar{x}$ *is a global solution to the problem* $\min_x f$.
(ii) $\bar{x}$ *is a local solution to the problem* $\min_x f$.
(iii) $0 \in \partial f(\bar{x})$.

Convex conjugate: Again consider the support functions defined in (48). By construction, $z \in \partial f(x)$ if and only if

$$\langle (z, -1),\ (x, f(x)) \rangle = \sigma_{\mathrm{epi}(f)}\left((z, -1)\right) = \sup_y (\langle z,\ y\rangle - f(y)) = f^*(z),$$

or equivalently, $f(x) + f^*(z) = \langle z,\ x\rangle$. When f is a proper convex function, the conjugate function f^* (defined in (31)), is a closed, proper, convex function, since it is the pointwise supremum of the affine functions $z \to \langle z,\ y\rangle - f(y)$ over the index set $\mathrm{dom}\,(f)$. Consequently we have

$$\partial f(x) = \left\{ z \;\middle|\; f(x) + f^*(z) \le \langle z,\ x\rangle \right\}.$$

Due to the symmetry of this expression for the subdifferential, it can be shown that $(f^*)^* = f$ and $\partial f^* = (\partial f)^{-1}$, i.e.,

$$z \in \partial f(x) \iff x \in \partial f^*(z) \tag{50}$$

whenever f is a closed proper convex function. These relationships guide us to focus on the class of functions

$$\Gamma_n := \left\{ f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \mid f \text{ is closed proper and convex} \right\}.$$

For example, if $\mathscr{C} \subset \mathbb{R}^n$ is a nonempty closed convex set, then $\delta_{\mathscr{C}} \in \Gamma_n$, where $\delta_{\mathscr{C}}$ is defined in (19). It is easily seen that $\delta^*_{\mathscr{C}} = \sigma_{\mathscr{C}}$ and, for $x \in \mathscr{C}$,

$$\partial \delta_{\mathscr{C}}(x) = \{ z \mid \langle z,\ y - x\rangle \le 0\ \forall\, y \in \mathscr{C} \} =: N(x \mid \mathscr{C}),$$

where $N(x \mid \mathscr{C})$ is called the normal cone to $\mathscr{C}$ at x.

Calculus for PLQ: Just as in the smooth case, subdifferentials and conjugates become useful in practice by developing a calculus for their ready computation. Here we focus on calculus rules for PLQ functions ρ defined in (38) which are well established in Rockafellar and Wets (1998). In particular, if we set $q(v) := \frac{1}{2} v^T M v + \delta_{\mathscr{V}}(v)$, then, by Rockafellar and Wets (1998, Corollary 11.33), either $\rho \equiv \infty$ or

$$\rho^*(y) = \inf_{B^T v = y} \left[q(v) - \langle b,\ v\rangle \right] \ \text{ and } \ \partial \rho(z) = B^T \partial q^*(Bz + b), \tag{51}$$

which can be reformulated as

$$\partial \rho(z) = \left\{ B^T v \mid v \in \mathscr{V} \ \text{ and } \ Bz - Qv + b \in N(v \mid \mathscr{V}) \right\}.$$

In addition, we have from Aravkin et al. (2013b, Theorem 3) that

$$\mathrm{dom}\left(\rho^*\right) = B^T \mathscr{V} \ \text{ and}$$
$$\mathrm{dom}\,(\rho) = B^{-1}\left([\mathscr{V}^\infty \cap \mathrm{Nul}\,(M)]^\circ - b \right), \tag{52}$$

where $\mathscr{V}^\infty$ is the *horizon cone* of $\mathscr{V}$. As the name suggests, $\mathscr{V}^\infty$ is a closed cone, and, when $\mathscr{V}$ is nonempty convex, it is a nonempty closed convex cone satisfying $\mathscr{V}^\infty = \{ w \mid \mathscr{V} + w \subset \mathscr{V} \}$. In particular, $\mathscr{V}$ is bounded if and only if $\mathscr{V}^\infty = \{0\}$.

The reader can verify by inspection of Fig. 4(a)–4(f) that the domain of each scalar PLQ is $\mathbb{R}$. This is also immediate from (52). Four of the six penalties have bounded sets $\mathscr{V}$, so that $\mathscr{V}^\infty = \{0\}$, the polar is the range of B, and so the result follows immediately. The quadratic penalty has $\mathscr{V}^\infty = \mathbb{R}$, but $\mathrm{Nul}\,(M) = \{0\}$. We leave the elastic net as an exercise.

More importantly, (51) gives explicit expressions for derivatives and subgradients of PLQ functions in terms of v. Consider the Huber function, Fig. 4(c). From (51), we have

$$\partial \rho(z) = \{ v \mid v \in \kappa[-1, 1] \ \text{ and } \ z - v \in N(v \mid \kappa[-1, 1]) \}.$$

From this description, we immediately have $\partial \rho(z) = \nabla \rho(z) = z$ for $|z| < \kappa$, and $\kappa\ \mathrm{sgn}(z)$ for $|z| > \kappa$.

Convex duality

There are many approaches for convex duality theory (Rockafellar & Wets, 1998). For our purposes, we choose one based on the convex-composite Lagrangian (Burke, 1985).

Primal objective: Let $f \in \Gamma_m$, $g \in \Gamma_n$, and $K \in \mathbb{R}^{m \times n}$ and consider the primal convex optimization problem

$$\mathfrak{P} \qquad \min_x \mathrm{p}(x) := f(Kx) + g(x), \tag{53}$$

where we call $\mathrm{p}(x)$ the *primal* objective.

The structure of the problem (53) is the same as that used to develop the celebrated Fenchel–Rockafellar Duality Theorem (Rockafellar, 1970 Section 31) (Theorem 4). It is sufficiently general to allow an easy translation to several formulations of the problem (15) depending on how one wishes to construct an algorithmic framework. This variability in formulation is briefly alluded to in Section 4.5. In this section, we focus on general duality results for (53) leaving the discussion of specific reformulation of (15) to the discussion of algorithms.

We now construct the *dual* to the convex optimization problem $\mathfrak{P}$. In general, the dual is a concave optimization problem, but, as we show, it is often beneficial to represent it as a convex optimization problem.

Lagrangian: First, define the *Lagrangian* $\mathscr{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \to \mathbb{R} \cup \{-\infty\}$ for $\mathfrak{P}$ by setting

$$\mathscr{L}(x, w, v) := \langle w,\ Kx\rangle - f^*(w) + \langle v,\ x\rangle - g^*(v). \tag{54}$$

The definition of the conjugate immediately tells us that the primal objective is given by maximizing the Lagrangian over the dual variables:

$$f(Kx) + g(x) = \sup_{w,v} \mathscr{L}(x, w, v).$$

Dual objective: Conversely, the dual objective is obtained by minimizing the Lagrangian over the primal variables:

$$\mathrm{d}(w, v) := \inf_x \mathscr{L}(x, w, v) = \begin{cases} -f^*(w) - g^*(v), & K^T w + v = 0, \\ -\infty, & K^T w + v \neq 0. \end{cases}$$

The corresponding dual optimization problem is

$$\max_{w,v} \mathrm{d}(w, v) \quad = \quad \max_{K^T w + v = 0} -f^*(w) - g^*(v).$$

One can eliminate v from the dual problem and reverse sign to obtain a simplified version of the dual problem:

$$\mathfrak{D} \qquad \min_w \tilde{\mathrm{d}}(w) := f^*(w) + g^*(-K^T w). \tag{55}$$

Weak and strong duality: By definition, $\max \mathrm{d}(w, v) \le \min \mathrm{p}(x)$, or equivalently, $0 \le (\min \tilde{\mathrm{d}}(w)) + (\min \mathrm{p}(x))$. This inequality is called *weak duality*. If equality holds, we say the duality gap is zero. If solutions to both $\mathfrak{P}$ and $\mathfrak{D}$ exist with zero duality gap, then we say *strong duality* holds. In general, a zero duality gap and strong duality require additional hypotheses called *constraint qualifications*. Constraint qualifications for the problem $\mathfrak{P}$ are given as conditions (a) and (b) in the following theorem (examples of primal–dual problem pairs in sparsity promotion are given in Table 2).

Theorem 4 (*Fenchel–Rockafellar Duality Theorem Rockafellar, 1970, Corollary 31.2.1*)**.** *Let* $f \in \Gamma_m$, $g \in \Gamma_n$, *and* $K \in \mathbb{R}^{m \times n}$. *If either*

(a) *there exists* $x \in \mathrm{ri}\,(\mathrm{dom}\,(g))$ *with* $Kx \in \mathrm{ri}\,(\mathrm{dom}\,(f))$, *or*
(b) *there exists* $w \in \mathrm{ri}\,(\mathrm{dom}\,(f^*))$ *with* $-K^T w \in \mathrm{ri}\,(\mathrm{dom}\,(g^*))$,

Table 2
We show three common variants of sparsity promoting formulations, and compute the dual in each case using the relationships between (53) and (55). Strong duality holds for all three examples.

	$\overset{f}{g}$	$\overset{f^*}{g^*}$	$\mathfrak{P}$	$\mathfrak{D}$
Basis	$\delta_{\tau\mathbb{B}_2}(\cdot - s)$	$\tau\|\cdot\|_2 + \langle w, s\rangle$	$\min \|x\|_1$	$\min \tau\|w\|_2 + \langle w, s\rangle$
Pursuit	$\|\cdot\|_1$	$\delta_{\mathbb{B}_\infty}(\cdot)$	s.t. $\|Kx - s\|_2 \le \tau$	s.t. $\left\|K^T w\right\|_\infty \le 1$
LASSO	$\frac{1}{2}\|\cdot - s\|_2^2$	$\langle\cdot, s\rangle + \frac{1}{2}\|\cdot\|_2^2$	$\min \frac{1}{2}\|Kx - s\|_2^2$	$\min \frac{1}{2}\|w\|_2^2 + \kappa\left\|K^T w\right\|_\infty + \langle w, s\rangle$
	$\delta_{\kappa\mathbb{B}_1}(\cdot)$	$\kappa\|\cdot\|_\infty$	s.t. $\|x\|_1 \le \kappa$	
Lagrangian	$\frac{1}{2}\|\cdot - s\|_2^2$	$\langle\cdot, s\rangle + \frac{1}{2}\|\cdot\|_2^2$	$\min \frac{1}{2}\|Kx - s\|_2^2 + \lambda\|x\|_1$	$\min \frac{1}{2}\|w + s\|_2^2 - \frac{1}{2}\|s\|_2^2$
	$\lambda\|\cdot\|_1$	$\delta_{\lambda\mathbb{B}_\infty}(\cdot)$		s.t. $\left\|K^T w\right\|_\infty \le \lambda$

hold, then $\min \mathfrak{p} + \min \tilde{\mathfrak{d}} = 0$ *with finite optimal values. Under condition (a),* $\operatorname{argmin}\tilde{\mathfrak{d}}$ *is nonempty, while under (b),* $\operatorname{argmin}\mathfrak{p}$ *is nonempty. In particular, if both (a) and (b) hold, then strong duality between* $\mathfrak{P}$ *and* $\mathfrak{D}$ *holds in the sense that* $\min \mathfrak{p} + \min \tilde{\mathfrak{d}} = 0$ *with finite optimal values that are attained in both* $\mathfrak{P}$ *and* $\mathfrak{D}$*. In this case, optimal solutions are characterized by*

$$\left\{\begin{array}{r} \bar{x} \text{ solves } \mathfrak{P} \\ \bar{w} \text{solves } \mathfrak{D} \\ \min \mathfrak{p} + \min \tilde{\mathfrak{d}} = 0 \end{array}\right\} \iff \left\{\begin{array}{r} \bar{w} \in \partial f(K\bar{x}) \\ -K^T\bar{w} \in \partial g(\bar{x}) \end{array}\right\}$$

$$\iff \left\{\begin{array}{r} \bar{x} \in \partial g^*(-K^T\bar{w}) \\ K\bar{x} \in \partial f^*(\bar{w}) \end{array}\right\}$$

where (50) can be used to obtain the second set of conditions from the first. When f and g are piecewise linear–quadratic, finite optimal values for $\min \mathfrak{p}$ and $\min \mathfrak{d}$ imply strong duality holds (Rockafellar & Wets, 1998 Theorem 11.42).

Rewriting the characterization in a fully symmetric way, we obtain

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \underbrace{\left(\begin{bmatrix} 0 & K \\ -K & 0 \end{bmatrix} + \begin{bmatrix} \partial f^* & 0 \\ 0 & \partial g \end{bmatrix}\right)}_{\mathscr{A}} \underbrace{\begin{bmatrix} x \\ w \end{bmatrix}}_{z}.$$

Formally, $0 \in \mathscr{A}z$ whenever $z \in (I + \mathscr{A})z$. Primal–dual splitting methods, including ADMM (Algorithm 6) and Chambolle–Pock (Algorithm 7), can be derived as fixed-point iterations for the nonlinear operator $(I + \mathscr{A})$. For additional reading, see e.g. Bauschke and Combettes (2011).

References

Agamennoni, G., Nieto, J., & Nebot, E. (2011). An outlier-robust Kalman filter. In *2011 IEEE international conference on robotics and automation, ICRA* (pp. 1551–1558).

Angelosante, D., Roumeliotis, S. I., & Giannakis, G. B. (2009). Lasso-Kalman smoother for tracking sparse signals. In *2009 Conference record of the forty-third asilomar conference on signals, systems and computers* (pp. 181–185), http://doi.org/10.1109/ACSSC.2009.5470133, ISSN: 1058-6393.

Ansley, C. F., & Kohn, R. (1982). A geometrical derivation of the fixed interval smoothing algorithm. *Biometrika*, *69*, 486–487.

Aravkin, A., Bell, B. B., Burke, J. V., & Pillonetto, G. (2013). Kalman smoothing and block tridiagonal systems: new connections and numerical stability results, arXiv preprint arXiv:1303.5237.

Aravkin, A., Bell, B. M., Burke, J. V., & Pillonetto, G. (2011). An ℓ_1-Laplace robust Kalman smoother. *IEEE Transactions on Automatic Control*, *56*(12), 2898–2911.

Aravkin, A., Burke, J., Chiuso, A., & Pillonetto, G. (2014). Convex vs. nonconvex approaches for sparse estimation: GLASSO, multiple kernel learning, and HGLASSO. *Journal of Machine Learning Research (JMLR)*, *15*, 217–252.

Aravkin, A., Burke, J. V., & Pillonetto, G. (2013a). Generalized system identification with stable spline kernels, arXiv preprint arXiv:1309.7857.

Aravkin, A., Burke, J. V., & Pillonetto, G. (2013b). Sparse/robust estimation and Kalman smoothing with nonsmooth log-concave densities: Modeling, computation, and theory. *Journal of Machine Learning Research (JMLR)*, *14*, 2689–2728.

Aravkin, A., Burke, J. V., & Pillonetto, G. (2014). Robust and trend-following Student's t Kalman smoothers. *SIAM Journal on Control and Optimization*, *52*(5), 2891–2916.

Aravkin, A., Friedlander, M., Herrmann, F., & Van Leeuwen, T. (2012). Robust inversion, dimensionality reduction, and randomized sampling. *Mathematical Programming*, *134*(1), 101–125.

Aravkin, A., Lozano, A., Luss, R., & Kambadur, P. (2014). Orthogonal matching pursuit for sparse quantile regression. In *2014 IEEE international conference on data mining* (pp. 11–19). IEEE.

Arulampalam, M. S., Maskell, S., Gordon, N., & Clapp, T. (2002). A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, *50*(2), 174–188.

Badawi, F., Lindquist, A., & Pavon, M. (1975). A stochastic realization approach to the smoothing problem. *IEEE Transactions on Automatic Control*, *24*, 878–888.

Baraniuk, R.G., Cevher, V., Duarte, M.F., & Hegde, C. (2008). Model-based compressive sensing, Technical report, Rice University. Available at arxiv:0808.3572.

Bauschke, H. H., & Combettes, P. L. (2011). *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media.

Beck, A., & Teboulle, M. (2009). Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, *18*(11), 2419–2434.

Bell, B. (1994). The iterated Kalman smoother as a Gauss-Newton method. *SIAM Journal on Optimization*, *4*(3), 626–636.

Bell, B., & Cathey, F. (1993). The iterated Kalman filter update as a Gauss-Newton method. *IEEE Transactions on Automatic Control*, *38*(2), 294–297.

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

Bell, B. M., Burke, J. V., & Pillonetto, G. (2009). An inequality constrained nonlinear Kalman–Bucy smoother by interior point likelihood maximization. *Automatica*, *45*(1), 25–33.

Bertero, M. (1989). Linear inverse and ill-posed problems. *Advances in Electronics and Electron Physics*, *75*, 1–120.

Bertsekas, D. (1999). *Nonlinear programming*. (2nd ed.). Athena Scientific.

Bottou, L., Chapelle, O., DeCoste, D., & Weston, J. (Eds.). (2007). *Large scale kernel machines*. Cambridge, MA, USA: MIT Press.

Boyd, S., Parikh, N., Chu, E., Peleato, B., & Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1–122.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

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

Burke, J. V., & Ferris, M. C. (1995). A Gauss-Newton method for convex composite optimization. *Mathematical Programming*, *71*(2), 179–194.

Candès, E., & Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies. *IEEE Transactions on Information Theory*, *52*(12), 5406–5425.

Chambolle, A., & Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, *40*(1), 120–145.

Chan, S., Liao, B., & Tsui, K. (2011). Bayesian Kalman filtering, regularization and compressed sampling. In *2011 IEEE 54th international midwest symposium on circuits and systems, MWSCAS* (pp. 1–4).

Chang, L., Hu, B., Chang, G., & Li, A. (2013). Robust derivative-free Kalman filter based on Huber's M-estimation methodology. *Journal of Process Control*, *23*(10), 1555–1561.

Chu, W., Keerthi, S. S., & Ong, C. J. (2001). A unified loss function in Bayesian framework for support vector regression. *Epsilon*, *1*(1.5), 2.

Combettes, P. L., & Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering* (pp. 185–212). Springer.

Davis, D., & Yin, W. (2014). Faster convergence rates of relaxed Peaceman–Rachford and ADMM under regularity assumptions. arXiv preprint arXiv:1407.5210.

Davis, D., & Yin, W. (2015). A three-operator splitting scheme and its optimization applications. arXiv preprint arXiv:1504.01032.

De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, *25*(2), 201–230.

Dekel, O., Shalev-Shwartz, S., & Singer, Y. (2005). Smooth ε-insensitive regression by loss symmetrization. In *Journal of Machine Learning Research* (pp. 711–741).

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

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

Drucker, H., Burges, C., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems*.

Efron, B., Hastie, T., Johnstone, L., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*, 407–499.

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

Fernándes, J., Speyer, J. L., & Idan, M. (2013). A stochastic controller for vector linear systems with additive Cauchy noise. In *52nd IEEE conference on decision and control* Florence, Italy, (pp. 1872–1879).

Fraser, D. C., & Potter, J. E. (1969). The optimum linear smoother as a combination of two optimum linear filters. *IEEE Transactions on Automatic Control*, 387–390.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw*, *33*(1), 1–22.

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

Golub, G., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Gunter, L., & Zhu, J. (2007). Efficient computation and model selection for the support vector regression. *Neural Computation*, *19*, 1633–1655.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. John Wiley and Sons.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning. data mining, inference and prediction*. Canada: Springer.

Haykin, S. (2001). *Kalman filtering and neural networks*. Wiley Online Library.

Hewer, G. A., Martin, R. D., & Zeh, J. (1987). Robust preprocessing forKalman filtering of glint noise. *IEEE Transactions on Aerospace and Electronic Systems*, *AES-23*(1), 120–128.

Ho, C., & Lin, C. (2012). Large-scale linear support vector regression. *Journal of Machine Learning Research*, *13*, 3323–3348.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, *36*(3), 1171–1220.

Huber, P., & Ronchetti, E. (2009). *Robust statistics*. New York, NY, USA: John Wiley and Sons.

Jacob, L., Obozinski, G., & Vert, J. P. (2009). Group lasso with overlap and graph lasso. In *International conference on machine learning, ICML* (pp. 433–440).

Jazwinski, A. (1970). *Stochastic processes and filtering theory*. Dover Publications, Inc.

Kailath, T., Sayed, A., & Hassibi, B. (2000). *Linear estimation*. Prentice Hall.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the AMSE - Journal of Basic Engineering*, *82*(D), 35–45.

Kalman, R. E., & Bucy, R. S. (1961). New results in linear filtering and prediction theory. *Transactions of the AMSE Journal of Basic Engineering*, *83*, 95–108.

Kim, S., Koh, K., Boyd, S., & Gorinevsky, D. (2009). ℓ_1 trend filtering. *SIAM Reviews*, *51*(2), 339–360.

Kitagawa, G., & Gersch, W. (1985). A smoothness priors time-varying AR coefficient modeling of nonstationary covariance time series. *IEEE Transactions on Automatic Control*, *30*(1), 48–56.

Koenker, R., & Bassett Jr., G. (1978). Regression quantiles. *Journal of the Econometric Society*, 33–50.

Kojima, M., Megiddo, N., Noma, T., & Yoshise, A. (1991). *Lecture notes in computer science*: *vol. 538*. *A unified approach to interior point algorithms for linear complementarity problems*. Berlin, Germany: Springer Verlag.

Kyung, M., Gill, J., Ghosh, M., Casella, G., et al. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411.

Lee, Y.-J., Hsieh, W.-F., & Huang, C.-M. (2005). ε-SSVR: a smooth support vector machine for ε;-insensitive regression. *IEEE Trans. Knowl. Data Eng.*, *17*(5), 678–685.

Li, Q., Lin, N., et al. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170.

Lindquist, A., & Picci, G. (2015). *Linear stochastic systems: A geometric approach to modeling, estimation and identification*. Springer Berlin Heidelberg.

Lions, P.-L., & Mercier, B. (1979). Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, *16*(6), 964–979.

Ljung, L. (1999). *System identification - theory for the user*. (2nd ed.). Upper Saddle River, N.J.: Prentice-Hall.

Ljung, L., & Kailath, T. (1976). A unified approach to smoothing formulas. *Automatica*, *12*(2), 147–157.

Loh, P.-L., & Wainwright, M. J. (2013). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems* (pp. 476–484).

Lounici, K., Pontil, M., Tsybakov, A. B., & van de Geer, S. (2009). Taking advantage of sparsity in multi-task learning. In *Technical Report* arXiv:0903.1468, ETH Zurich.

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

Maritz, J. S., & Lwin, T. (1989). *Empirical Bayes method*. Chapman and Hall.

Mayne, D. Q. (1966). A solution of the smoothing problem for linear dynamic systems. *Automatica*, *4*, 73–92.

Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, *34*, 1436–1462.

Meinshausen, N., & Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, *37*(1), 246–270.

Negahban, S., & Wainwright, M. J. (2009). *Simultaneous support recovery in high-dimensional regression: Benefits and perils of $\ell_{1,\infty}$-regularization*. UC Berkeley: Department of Statistics.

Nemirovskii, A., & Nesterov, Y. (1994). *Studies in applied mathematics*: *vol. 13*. *Interior-point polynomial algorithms in convex programming*. Philadelphia, PA, USA: SIAM.

Nesterov, Y. (2004). *Applied optimization*: *vol. 87*. *Introductory lectures on convex optimization* (p. xviii+236). Boston, MA: Kluwer Academic Publishers, A basic course.

Niedzwiecki, M., & Gackowski, S. (2013). New approach to noncausal identification of nonstationary stochastic FIR systems subject to both smooth and abrupt parameter changes. *IEEE Transactions on Automatic Control*, *58*(7), 1847–1853.

Obozinski, G., Wainwright, M. J., & Jordan, M. I. (2010). Union support recovery in high-dimensional multivariate regression. *The Annals of Statistics* (in press).

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

Ohlsson, H., & Ljung, L. (2013). Identification of switched linear regression models using sum-of-norms regularization. *Automatica*, *49*(4), 1045–1050.

Oksendal, B. (2005). *Stochastic differential equations*. (6th ed.). Springer.

Paige, C. C., & Saunders, M. A. (1977). Least squares estimation of discrete linear dynamic systems using orthogonal transformations. *SIAM Journal on Numerical Analysis*, *14*(2), 180–193.

Passty, G. B. (1979). Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *Journal of Mathematical Analysis and Applications*, *72*(2), 383–390.

Pontil, M., Mukherjee, S., & Girosi, F. (2000). On the noise model of support vector machines regression. In *Algorithmic learning theory* (pp. 316–324). Springer.

Rachev, S. T. (Ed.). (2003). *Handbook of heavy tailed distributions in finance*. Elsevier Science.

Rauch, H. E., Tung, F., & Striebel, C. T. (1965). Maximum likelihood estimates of linear dynamic systems. *AIAA Journal*, *3*(8), 1145–1150.

Rice, J. (1986). Choice of smoothing parameter in deconvolution problems. *Contemporary Mathematics*, *59*, 137–151.

Rockafellar, R. (1970). *Priceton landmarks in mathematics*. *Convex analysis*. Princeton University Press.

Rockafellar, R. T. (1974). Augmented Lagrangian multiplier functions and duality in nonconvex programming. *SIAM Journal on Control*, *12*, 268–285.

Rockafellar, R. T., & Wets, R. J. B. (1998). *Variational analysis, Vol. 317*. Springer.

Schölkopf, B., & Smola, A. J. (2001). *(Adaptive computation and machine learning). Learning with Kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press.

Simon, D. (2010). Kalman filtering with state constraints: a survey of linear and nonlinear algorithms. *IET Control Theory & Applications*, *4*(8), 1303–1318.

Simon, D., & Chia, T. L. (2002). Kalman filtering with state equality constraints. *IEEE Transactions on Aerospace and Electronic Systems*, *38*(1), 128–136.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B.*, *58*(1), 267–288.

Tikhonov, A., & Arsenin, V. (1977). *Solutions of Ill-Posed problems*. Washington, D.C.: Winston/Wiley.

Tropp, J. A., Gilbert, A. C., & Strauss, M. J. (2006). Algorithms for simultaneous sparse approximation. *Signal Processing*, *86*, 572–602 Special issue on "Sparse approximations in signal and image processing".

Van de Geer, S., & Buhlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, *3*, 1360–1392.

Van den Berg, E., & Friedlander, M. P. (2008). Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, *31*(2), 890–912.

Vito, E. D., Rosasco, L., Caponnetto, A., De Giovannini, U., & Odone, F. (2005). Learning from examples as an Inverse problem. *Journal of Machine Learning Research (JMLR)*, *6*, 883–904.

Wahba, G. (1990). *Spline models for observational data*. Philadelphia: SIAM.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, *55*, 2183–2202.

Wan, E. A., & Van Der Merwe, R. (2000). The unscented Kalman filter for nonlinear estimation. In *Adaptive systems for signal processing, communications, and control symposium 2000. AS-SPCC. the IEEE 2000* (pp. 153–158). Ieee.

Wipf, D., & Nagarajan, S. (2007). A new view of automatic relevance determination. In *Proc. of NIPS*.

Wipf, D., & Rao, B. (2007). An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Transactions on Signal Processing*, *55*(7), 3704–3716.

Wipf, D. P., Rao, B. D., & Nagarajan, S. (2011). Latent variable Bayesian models for promoting sparsity. *IEEE Transactions on Information Theory*, *57*(9), 6236–6255.

Wright, S. J. (1997). *Primal-dual interior-point methods*. Englewood Cliffs, N.J., USA: Siam.

Ye, Y. (2011). *Interior point algorithms: theory and analysis, Vol. 44*. John Wiley & Sons.

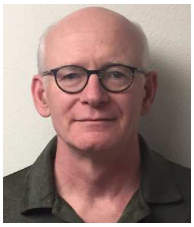
Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B.*, *68*, 49–67.

Zhao, P., Rocha, G., & Yu, B. (2009). Grouped and hierarchical model selection through composite absolute penalties. *The Annals of Statistics*, *37*(6A), 3468–3497.

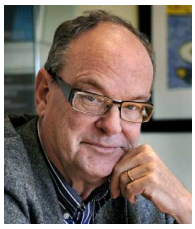
Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research (JMLR)*, *7*, 2541–2567.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *67*(2), 301–320.

Aleksandr Aravkin received B.S. degrees in Mathematics and Computer Science from the University of Washington in 2004. He then received an M.S. in Statistics and a Ph.D. in Mathematics from the University of Washington in 2010. He was a joint postdoctoral fellow in Earth and Ocean Sciences and Computer Science at the University of British Columbia from 2010–2012, and a research staff member at the IBM T.J. Watson Research Center from 2012–2015. During this time he also worked at Columbia as an Adjunct Professor in Computer Science and IEOR. In 2015, Dr. Aravkin joined the faculty at UW Applied Mathematics, where he works on theoretical and practical problems connected to data science, including convex and variational analysis, statistical modeling, and algorithm design.



James V. Burke received his Ph.D. in mathematics from the University of Illinois in 1983, and has been a member of the mathematics faculty at the University of Washington since 1985. He is published widely in convex and nonsmooth analysis and optimization with an emphasis on numerical methods. His recent research has focused on Kalman smoothers with non-Gaussian densities and state constraints, smoothing methods for convex and non convex problems, and non-symmetric eigenvalue optimization problems.



Lennart Ljung received his Ph.D. in Automatic Control from Lund Institute of Technology in 1974. Since 1976 he is Professor of the chair of Automatic Control In Linkoping, Sweden. He has held visiting positions at IPU (Moscow), Stanford, MIT, Berkeley and Newcastle University (NSW) and has written several books on System Identification and Estimation. He is an IEEE Fellow, an IFAC Fellow and an IFAC Advisor. He is a member of the Royal Swedish Academy of Sciences (KVA), a member of the Royal Swedish Academy of Engineering Sciences (IVA), an Honorary Member of the Hungarian Academy of Engineering, an Honorary Professor of the Chinese Academy of Mathematics and Systems Science, and a Foreign Member of the US National Academy of Engineering (NAE) as well as a member of the Academia Europaea. He has received honorary doctorates from the Baltic State Technical University in St Petersburg, from Uppsala University, Sweden, from the Technical University of Troyes, France, from the Catholic University of Leuven, Belgium and from Helsinki University of Technology, Finland. In 2003 he received the Hendrik W. Bode Lecture Prize from the IEEE Control Systems Society, and in 2007 the IEEE Control Systems Award. He received the Quazza Medal in 2002 and the Nichols Medal in 2017, both from IFAC.

Aurélie C. Lozano is a Research Staff Member at the IBM T.J. Watson Research Center. She received the M.S./Dipl.Ing. degree in Communication Systems from the Swiss Federal Institute of Technology Lausanne (EPFL) in 2001, and the M.A. and Ph.D. degrees in Electrical Engineering from Princeton University respectively in 2004 and 2007. She was an Adjunct Associate Professor in the Computer Science Department and the Industrial Engineering and Operations Research Department at Columbia University from 2014 to 2016. Her research interests include machine learning, statistics and optimization. Her current focus is on high dimensional data analysis and predictive modeling, with applications including biology, environmental sciences, business and infrastructure analytics, and social media analytics. She was a recipient of the best paper award at the conference on Uncertainty in Artificial Intelligence (UAI) 2013 and of the IBM Research Pat Goldberg Memorial best paper award, 2013.

Gianluigi Pillonetto was born on January 21, 1975 in Montebelluna (TV), Italy. He received the Doctoral degree in Computer Science Engineering cum laude from the University of Padova in 1998 and the Ph.D. degree in Bioengineering from the Polytechnic of Milan in 2002. In 2000 and 2002 he was visiting scholar and visiting scientist, respectively, at the Applied Physics Laboratory, University of Washington, Seattle. From 2002 to 2005 he was Research Associate at the Department of Information Engineering, University of Padova, becoming an Assistant Professor in 2005. He is currently an Associate Professor of Control and Dynamic Systems at the Department of Information Engineering, University of Padova. His research interests are in the field of system identification, estimation and machine learning. He currently serves as Associate Editor for Automatica and IEEE Transactions on Automatic Control. In 2003 he received the Paolo Durst award for the best Italian Ph.D. thesis in Bioengineering, and he was the 2017 recipient of the Automatica Prize.