SIAM J. CONTROL OPTIM.
Vol. 52, No. 5, pp. 2891–2916

ROBUST AND TREND-FOLLOWING STUDENT'S T KALMAN SMOOTHERS*

ALEKSANDR Y. ARAVKIN†, JAMES V. BURKE‡, AND GIANLUIGI PILLONETTO§

Abstract. We present a Kalman smoothing framework based on modeling errors using the heavy tailed Student's t distribution, along with algorithms, convergence theory, implementation, and several important applications. The computational effort per iteration grows linearly with the length of the time series, and all smoothers allow nonlinear process and measurement models. *Robust smoothers* form an important subclass of smoothers within this framework. These smoothers achieve reasonable performance when faced with noise misspecification or model error, for example, in situations where measurements are highly contaminated by outliers or include data unexplained by the forward model. Robust smoothers are developed by modeling *measurement* errors using the heavy tailed Student's t distribution and outperform the recently proposed ℓ_1-Laplace smoother in extreme situations with data containing 20% or more outliers. A second special application we consider in detail allows tracking sudden changes in the state. It is developed by modeling *process* noise using the Student's t distribution, and the resulting smoother can track sudden changes in the state. These features can be used separately or in tandem, and we present a general smoother algorithm and open source implementation, together with convergence analysis that covers a wide range of smoothers. A key ingredient of our approach is a technique to deal with the nonconvexity of the Student's t loss function. Numerical results for linear and nonlinear models illustrate the potential of the modeling framework proposed, showcasing new smoothers for robust and tracking applications, as well as for mixed problems that have both types of features.

Key words. trend-following smoothers, robust smoothers, Kalman smoothers, Student's t, convex-composite

AMS subject classifications. 60G35, 62F25, 65K10, 93E10, 93E11, 93E14

DOI. 10.1137/130918861

1. Introduction.

The Kalman filter is an efficient recursive algorithm for estimating the state of a dynamic system [22]. Traditional formulations are based on ℓ_2 penalties on model deviations and are optimal under assumptions of linear dynamics and Gaussian noise. Kalman filters are used in a wide array of applications, including navigation, medical technologies, and econometrics [13, 32, 36]. Many of these problems are nonlinear and may require smoothing over past data in both online and offline applications to significantly improve estimation performance [18].

This paper focuses on two important areas in Kalman smoothing: robustness with respect to outliers in measurement data, and improved tracking of quickly changing system dynamics. Robust filters and smoothers have been a topic of significant interest since the 1970s; see, e.g., [23, 27, 31]. Recent efforts have focused on building smoothers that are robust to outliers in the data [2, 3, 16], using convex loss functions such as ℓ_1, Huber, or Vapnik in place of the ℓ_2 penalty [20]. Here we use *robustness* in the statistical sense, where it means that the estimator achieves adequate perfor-

*Received by the editors April 29, 2013; accepted for publication (in revised form) June 30, 2014; published electronically September 23, 2014.
http://www.siam.org/journals/sicon/52-5/91886.html

†IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598 (saravkin@us.ibm.com).
‡Mathematics Department, University of Washington, Seattle, WA 98195 (burke@math.washington.edu).
§Control and Dynamic Systems, Department of Information Engineering, University of Padova, Padova, Italy (giapi@dei.unipd.it).

mance when faced with outliers or unexplained events; these may arise either as large measurement errors or due to misspecification of the dynamic model.

There have also been recent efforts to design smoothers able to better track fast system dynamics, e.g., jumps in the state values. A contribution can be found in [26], where the Laplace distribution, rather than the Gaussian, is used to model transition (process) noise. This introduces an ℓ_1 penalty on the state evolution in time, resulting in an estimator interpretable as a dynamic version of the well-known LASSO procedure [33].

For known dynamics, all of the smoothers mentioned above can be derived by modeling the process and the measurement noise using log-concave densities, taking the form

$$\mathbf{p}(\cdot) \propto \exp(-\rho(\cdot)), \quad \rho \text{ convex.} \tag{1.1}$$

Formulations exploiting (1.1) are nearly ubiquitous, in part because they correspond to convex optimization problems in the linear case. However, in order to model a regime with large outliers or sudden jumps in the state, we want to look beyond (1.1) and allow heavy tailed densities, i.e., distributions whose tails are not exponentially bounded. All such distributions necessarily have nonconvex loss functions [7, Theorem 2.1]. Nonetheless, models with heavy tailed densities have been very useful in applications related to glint noise [21], air turbulence [17], and asset returns [28] among others. Heavier tails are also a reasonable model for a contaminated normal distribution where "bad" measurements occur due to equipment malfunction, secondary noise sources, or other anomalies.

Several interesting candidates are possible; in this contribution we focus on the Student's t distribution for its computational properties in the context of the applications we consider. The Student's t distribution was successfully applied to a variety of robust inference applications in [24] and is closely related to redescending influence functions [19].

We emphasize that Student's t, as well as other heavy tailed distributions, are very useful *models* for errors. In particular, heavy tailed statistical models can take up the slack for significant deviations between observed and predicted data, allowing effective smoothing techniques in challenging contexts. It is essential to note that our concern is not with situations *where errors are distributed according to the Student's t distribution*; rather, we are interested in robust state estimation in a variety of difficult scenarios, including data contamination as well as sudden changes in the underlying trend that are not realizations from known or common distributions. This is reflected in the numerical experiments, where the efficacy of the approach is demonstrated using simulations where significant contamination is present, and where the contamination does *not* arise from the Student's t distribution.

In this work, we propose a broad smoothing framework that allows any component of the measurement residual errors or transition noise to be modeled using either Gaussians or heavy tailed distributions. We illustrate the framework for several applications, including robust and trend smoothing. An important simple example in this family of smoothers is the T-robust smoother, derived from a dynamic system with *measurement errors* modeled by the Student's t distribution. This is a further robustification of the estimator proposed by [2], which uses the Laplace density. The redescending influence function of the Student's t guarantees that outliers in the measurements have less of an effect on the smoothed estimate than any convex loss function. In practice, the T-robust smoother performs better than the smoother

of [2] for cases with a high proportion of outliers. A second important example is the T-trend smoother, derived starting from a dynamic system with *transition noise* modeled by the Student's t distribution. This allows T-trend to better track sudden changes in the state. One may consider using both aspects simultaneously; in addition, practitioners need the ability to distinguish between different measurements based on prior information of measurement fidelity, and between different states based on prior knowledge of trend stability.

In the context of Kalman filtering/smoothing, the idea of using Student's t distributions to model the system noise for robust and tracking applications was first proposed by [15]. However, our work differs from that approach in some important aspects. First, our analysis includes nonlinear measurement and process models. Second, we provide a novel approach to overcome the nonconvexity of the Student's t-loss function. Third, the approach we propose can be used to solve *any* smoothing problem that uses Student's t modeling for any process or measurement components.

The basic approach differs significantly from that of [15], who propose using the random information matrix (i.e., full Hessian) when possible, or its expectation (Fisher information) when the Hessian is indefinite. Instead, we propose a modified Gauss–Newton method which builds information about the curvature of the Student's t-log likelihood into the Hessian approximation and is guaranteed to be positive definite. As we show in section 5, the new approach is provably convergent and, unlike the approach in [15], uses information about the relative sizes of the residuals in computing descent directions, allowing us to control the effects of outliers on the Hessian approximation as the optimization proceeds (which is not true of methods using Fisher information).

The major computational tradeoff in using nonconvex penalties is that the loss function in the convex case is used directly [2], i.e., is not approximated, whereas in the nonconvex case, the loss function must be iteratively approximated with a local convex approximation. This requires a fundamental extension of the convergence analysis.

A conference proceeding previewing this paper appears in [6]. In the current work, we present a general smoothing framework that includes the two smoothers presented in [6] as special cases, together with a generalized convergence theory that covers the entire range of smoothers under discussion. In doing so, we correct a flaw in the statement of the main convergence result in [6]. We also provide an open-source implementation of the general algorithm [1], with a simple interface that enables the user to customize which measurement or process residual components to model using the Student's t penalty. Using this implementation, we present additional numerical experiments that show how robust and trend smoothing can be implemented simultaneously. Finally, we apply the smoothers to real data.

The paper is organized as follows. In section 2, we introduce the multivariate Student's t distribution, review its advantages for error modeling over log-concave distributions, and introduce the dynamic model class of interest for Kalman smoothing. In section 3, we describe a statistical modeling framework, where we can use Student's t to model any process or measurement residual components. We describe all objectives that can arise this way and provide a comprehensive method for obtaining approximate second order information for these objectives. In section 4, we provide details for three important special smoothers: T-robust (robust against large measurement noise), T-trend (able to follow sharp changes in the state), and the double-T smoother (incorporates both aspects). In section 5, we present the algo-

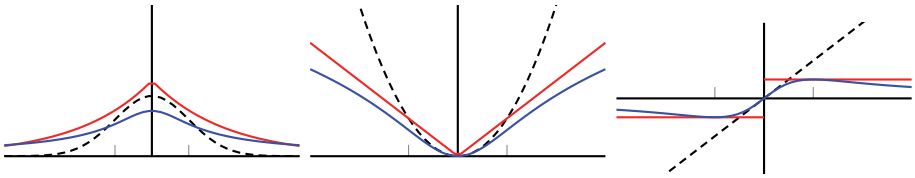

FIG. 1. *Gaussian, Laplace, and Student's t densities, corresponding negative log likelihoods, and influence functions.*

rithm and a convergence theory for the entire framework, which also extends the convergence theory developed in [2]. In section 6, we present numerical experiments that illustrate the behavior of all three special smoothers, including illustrations of linear and nonlinear models and results for real and simulated data. We end the paper with concluding remarks.

2. Error modeling with Student's t. For a vector $u \in \mathbb{R}^n$ and any positive definite matrix $M \in \mathbb{R}^{n\times n}$, let $\|u\|_M := \sqrt{u^{\mathrm{T}} M u}$. We use the following generalization of the Student's t distribution:

$$\mathbf{p}(v_k|\mu) = \frac{\Gamma(\frac{s+m}{2})}{\Gamma(\frac{s}{2})\det[\pi s R]^{1/2}} \left(1 + \frac{\|v_k - \mu\|^2_{R^{-1}}}{s}\right)^{\frac{-(s+m)}{2}}, \tag{2.1}$$

where μ is the mean, s is the degrees of freedom, m is the dimension of the vector v_k, and R is a positive definite matrix. A comparison of this distribution with the Gaussian and Laplacian distribution appears in Figure 1. Note that the Student's t distribution has much heavier tails than the others and that its influence function is redescending; see [25] for a discussion of influence functions. This means that as we pull a measurement further and further away, its "influence" decreases to 0, so it is eventually ignored by the model. Note also that the ℓ_1-Laplace is peaked at 0, while the Student's t distribution is not, and so a Student's t fit will not in general drive residuals to be exactly 0.

Before we proceed with the Kalman smoothing application, we review a result from [7], illustrating the fundamental modeling advantages of heavy tailed distributions.

THEOREM 2.1. *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 t_0 with $\rho'(t_0) = \alpha_0 > 0$. Then for all $t_2 > t_1 \geq t_0$, the conditional tail distribution induced by $p(x)$ satisfies*

$$\Pr(|y| > t_2 \mid |y| > t_1) \leq \exp(-\alpha_0[t_2 - t_1]) \,. \tag{2.2}$$

When t_1 is large, the condition $|y| > t_1$ indicates that we are looking at an outlier. However, as shown by the theorem, *any* log-concave statistical model treats the outlier conservatively, dismissing the chance that $|y|$ could be significantly bigger than t_1. Contrast this behavior with that of the Student's t distribution. When $s = 1$, the Student's t distribution is simply the Cauchy distribution, with a density proportional to $1/(1 + y^2)$. Then we have that

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

Heavy tailed distributions thus provide a fundamental advantage in cases where outliers may be particularly large, or, in the second application we discuss, very sudden trend changes may be present.

We now turn to the Kalman smoothing framework. We use the following general model for the underlying dynamics: for $k = 1, \ldots, N$

$$\begin{aligned} x_k &= g_k(x_{k-1}) + w_k, \\ z_k &= h_k(x_k) + v_k \end{aligned} \tag{2.3}$$

with initial condition $g_1(x_0) = g_0 + w_1$, with g_0 a known constant, and where $g_k : \mathbb{R}^n \to \mathbb{R}^n$ are known smooth process functions, and $h_k : \mathbb{R}^n \to \mathbb{R}^m$ are known smooth measurement functions. Moreover, w_k and v_k are mutually independent, and with known covariance matrices $Q_k \in \mathbb{R}^{n\times n}$ and $R_k \in \mathbb{R}^{m\times m}$, respectively. Note that here we assume all the measurement vectors have consistent dimension m. The variable dimension case is a straightforward extension.

We now briefly explain how to use Student's t error modeling to design smoothers with two important characteristics. In order to obtain smoothers that are robust to heavily contaminated data, the vector $v_k \in \mathbb{R}^{m(k)}$ can be modeled zero-mean Student's t measurement noise (2.1) of known covariance $R_k \in \mathbb{R}^{m(k)\times m(k)}$ and degrees of freedom s. To design smoothers that can track sudden changes in the state, the process residuals w_k are modeled using Student's t noise. These features may be employed separately or in tandem, and we always assume that the vectors $\{w_k\} \cup \{v_k\}$ are all mutually independent.

In the next section, we design a smoother that finds the maximum a posteriori (MAP) estimates of $\{x_k\}$ for a general formulation, where Student's t or least squares modeling can be used for any or all process and measurement residuals. We then specialize it to recover the applications discussed above.

3. Generalized smoothing framework. Given a sequence of column vectors $\{u_k\}$ and matrices $\{T_k\}$ we use the notation

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

We also make the definitions

$$\begin{aligned} R &= \operatorname{diag}(\{R_k\}), \\ Q &= \operatorname{diag}(\{Q_k\}), \\ x &= \operatorname{vec}(\{x_k\}), \end{aligned} \qquad \begin{aligned} w(x) &= \operatorname{vec}(\{x_k - g_k(x_{k-1})\}), \\ v(x) &= \operatorname{vec}(\{z_k - h_k(x_k)\}), \end{aligned}$$

and, as in (2.3), write

$$w_k = x_k - g_k(x_{k-1}) \quad \text{and} \quad v_k = z_k - h_k(x_k)\ .$$

In the most general case, we suppose that any of the components w_k^i or v_k^i can be modeled using either Gaussian or Student's t distributions.

For the sake of modeling clarity, assume that *subcomponents* of measurement and process residuals are consistently modeled across time points k; this gives the user

the ability to select which *subvectors* of process and measurement residuals to model using Student's t, but not to assign different penalties to different time points.

Denote by $w_k^{\mathcal{N}}$ and $w_k^{\mathcal{S}}$ the subvectors of the process residuals w_k, and denote by $v_k^{\mathcal{N}}$ and $v_k^{\mathcal{S}}$ the subvectors of the measurement residuals v_k that are to be modeled using the Gaussian and Student's t distributions, respectively. Assume that all of these subvectors are mutually independent, and denote the corresponding covariance submatrices by $Q_k^{\mathcal{N}}$, $Q_k^{\mathcal{S}}$, $R_k^{\mathcal{N}}$, and $R_k^{\mathcal{S}}$. Maximizing the likelihood for this model is equivalent to minimizing the associated negative log likelihood

$$-\ln \mathbf{p}(\{v_k^{\mathcal{N}}\}, \{v_k^{\mathcal{S}}\}, \{w_k^{\mathcal{N}}\}, \{w_k^{\mathcal{S}}\}),$$

where $\mathbf{p}(\{v_k^{\mathcal{N}}\}, \{v_k^{\mathcal{S}}\}, \{w_k^{\mathcal{N}}\}, \{w_k^{\mathcal{S}}\})$ is the probability density function evaluated at the observed data, that is, it is the product of the individual Gaussian and Student t densities at the given data points. By dropping those terms associated with the normalizing constants, this negative log likelihood can be explicitly written as follows:

$$\sum_{k=1}^{N} s \ln\left[1 + \frac{\|v_k^{\mathcal{S}}\|^2_{(R_k^{\mathcal{S}})^{-1}}}{s}\right] + \|v_k^{\mathcal{N}}\|^2_{(R_k^{\mathcal{N}})^{-1}} + r \ln\left[1 + \frac{\|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}}{r}\right] + \|w_k^{\mathcal{N}}\|^2_{(Q_k^{\mathcal{N}})^{-1}}, \tag{3.1}$$

where s and r are degree of freedom parameters corresponding to $v_k^{\mathcal{S}}$ and $w_k^{\mathcal{S}}$.

A first-order accurate affine approximation to our model with respect to direction $d = \mathrm{vec}\{d_k\}$ near a fixed state sequence x is given by

$$\begin{aligned} \tilde{w}(x;d) &= \mathrm{vec}(\{x_k - g_k(x_{k-1}) - G_k d_k\}), \\ \tilde{v}(x;d) &= \mathrm{vec}(\{z_k - h_k(x_k) - H_k d_k\}), \end{aligned}$$

where

$$G_k = g_k^{(1)}(x_{k-1}) \quad \text{and} \quad H_k = h_k^{(1)}(x_k)\ .$$

Set $Q_{N+1} = I_n$ and $g_{N+1}(x_N) = 0$ (where I_n is the $n \times n$ identity matrix) so that the formulas are also valid for $k = N+1$.

We minimize the nonlinear nonconvex objective in (3.1) by iteratively solving quadratic programming (QP) subproblems of the form

$$\min \quad \tfrac{1}{2} d^{\mathrm{T}} C d + a^{\mathrm{T}} d \quad \text{w.r.t. } d \in \mathbb{R}^{nN}, \tag{3.2}$$

where a is the gradient of objective (4.1) with respect to x and C has the form

$$C = \begin{bmatrix} C_1 + \Phi_1 & A_2^{\mathrm{T}} & 0 & \\ A_2 & C_2 + \Phi_2 & A_3^{\mathrm{T}} & 0 \\ 0 & \ddots & \ddots & \ddots \\ & 0 & A_N & C_N + \Phi_N \end{bmatrix}. \tag{3.3}$$

Note that this matrix is symmetric block tridiagonal. This structure is essential to the computational results for a wide variety of Kalman filtering and smoothing algorithms; it was noted early on in [14, 38].

In order to fully describe C_k and A_k, first let $\mathcal{W}^{\mathcal{N}}$, $\mathcal{W}^{\mathcal{S}}$ denote the indices associated to all subvectors $w_k^{\mathcal{N}}$ and $w_k^{\mathcal{S}}$ within w_k. For example, if the Student's t density

is used for all measurement residuals, and the Gaussian penalty is used for all process residuals, then $\mathcal{W}^{\mathcal{N}} = \{1, \ldots, n\}$, $\mathcal{W}^{\mathcal{S}} = \emptyset$.

Now define $A_k, C_k, \Phi_k \in \mathbb{R}^{n \times n}$ as follows:

$$A_k(\mathcal{W}^{\mathcal{S}}, \mathcal{W}^{\mathcal{S}}) = -\frac{r(Q_k^{\mathcal{S}})^{-1} G_k^{\mathcal{S}}}{r + \|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}},$$

$$A_k(\mathcal{W}^{\mathcal{N}}, \mathcal{W}^{\mathcal{N}}) = -(Q_k^{\mathcal{N}})^{-1} G_k^{\mathcal{N}}, \tag{3.4}$$

$$C_k(\mathcal{W}^{\mathcal{N}}, \mathcal{W}^{\mathcal{N}}) = (G_{k+1}^{\mathcal{N}})^{\mathrm{T}} (Q_{k+1}^{\mathcal{N}})^{-1} G_{k+1}^{\mathcal{N}} + (Q_k^{\mathcal{N}})^{-1},$$

$$C_k(\mathcal{W}^{\mathcal{S}}, \mathcal{W}^{\mathcal{S}}) = \frac{r(G_{k+1}^{\mathcal{S}})^{\mathrm{T}} (Q_{k+1}^{\mathcal{S}})^{-1} G_{k+1}^{\mathcal{S}}}{r + \|w_{k+1}^{\mathcal{S}}\|^2_{(Q_{k+1}^{\mathcal{S}})^{-1}}} + \frac{r(Q_k^{\mathcal{S}})^{-1}}{r + \|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}}, \tag{3.5}$$

$$\Phi_k = \frac{s(H_k^{\mathcal{S}})^{\mathrm{T}} (R_k^{\mathcal{S}})^{-1} H_k^{\mathcal{S}}}{(s + \|v_k^{\mathcal{S}}\|^2_{(R_k^{\mathcal{S}})^{-1}})} + (H_k^{\mathcal{N}})^{\mathrm{T}} (R_k^{\mathcal{N}})^{-1} H_k^{\mathcal{N}}. \tag{3.6}$$

The entries of A_k and C_k not explicitly defined in (3.4) and (3.5) are set to 0.

The Hessian approximation terms Φ_k in (3.6) are motivated in section 5 and are crucial to both practical performance and theoretical convergence analysis. The solutions to the subproblem (3.2) have the form $d = -C^{-1}a$ and can be found in an efficient and numerically stable manner in $O(n^3 N)$ steps, since C is tridiagonal and positive definite (see [9]).

4. Special cases. We now show how the general framework of the previous section can be specialized to obtain three smoothers. The first two are T-robust and T-trend, which are presented in [6]. The third is a new smoother where *all* measurement and process residuals are modeled using Student's t.

The objective corresponding to T-robust is obtained from (3.1) by taking $w_k^{\mathcal{N}} = w_k$, $w_k^{\mathcal{S}} = 0$, $v_k^{\mathcal{N}} = 0$, $v_k^{\mathcal{S}} = v_k$:

$$\frac{1}{2} \sum_{k=1}^{N} s \ln \left[1 + \frac{\|v_k\|^2_{R_k^{-1}}}{s} \right] + \|w_k\|^2_{Q_k^{-1}}. \tag{4.1}$$

The terms A_k, C_k, Φ_k in (3.4)–(3.6) become

$$A_k = -Q_k^{-1} G_k,$$

$$C_k = Q_k^{-1} + G_{k+1}^{\mathrm{T}} Q_{k+1}^{-1} G_{k+1},$$

$$\Phi_k = \frac{s H_k^{\mathrm{T}} R_k^{-1} H_k}{(s + \|v_k\|^2_{R_k^{-1}})}. \tag{4.2}$$

The objective corresponding to T-trend is obtained from (3.1) by taking $w_k^{\mathcal{N}} = 0$, $w_k^{\mathcal{S}} = w_k$, $v_k^{\mathcal{N}} = v_k$, $v_k^{\mathcal{S}} = 0$:

$$\frac{1}{2} \sum_{k=1}^{N} r \ln \left[1 + \frac{\|w_k\|^2_{Q_k^{-1}}}{r} \right] + \|v_k\|^2_{R_k^{-1}}. \tag{4.3}$$

The terms A_k, C_k, Φ_k in (3.4)–(3.6) become

$$\begin{aligned} A_k &= -\frac{rQ_k^{-1}G_k}{r + \|w_k\|^2_{Q_k^{-1}}}, \\ C_k &= \frac{rQ_k^{-1}}{r + \|w_k\|^2_{Q_k^{-1}}} + \frac{rG_{k+1}^{\mathrm{T}}Q_{k+1}^{-1}G_{k+1}}{r + \|w_{k+1}\|^2_{Q_{k+1}^{-1}}}, \\ \Phi_k &= H_k^{\mathrm{T}}R_k^{-1}H_k. \end{aligned} \tag{4.4}$$

Finally, we can apply Student's t to all process and measurement residuals by taking $w_k^{\mathcal{N}} = 0$, $w_k^{\mathcal{S}} = w_k$, $v_k^{\mathcal{N}} = 0$, $v_k^{\mathcal{S}} = v_k$ to obtain

$$\frac{1}{2}\sum_{k=1}^{N} r_k \ln\left[1 + \frac{\|w_k\|^2_{Q_k^{-1}}}{r_k}\right] + s_k \ln\left[1 + \frac{\|v_k\|^2_{R_k^{-1}}}{s_k}\right]. \tag{4.5}$$

The terms A_k, C_k, Φ_k in (3.4)–(3.6) become

$$\begin{aligned} A_k &= -\frac{rQ_k^{-1}G_k}{r + \|w_k\|^2_{Q_k^{-1}}}, \\ C_k &= \frac{rQ_k^{-1}}{r + \|w_k\|^2_{Q_k^{-1}}} + \frac{r(G_{k+1})^{\mathrm{T}}Q_{k+1}^{-1}G_{k+1}}{r + \|w_{k+1}\|^2_{Q_{k+1}^{-1}}}, \\ \Phi_k &= \frac{sH_k^{\mathrm{T}}R_k^{-1}H_k}{(s + \|v_k\|^2_{R_k^{-1}})}. \end{aligned} \tag{4.6}$$

5. Algorithm and global convergence. When models g_k and h_k are linear, we can compare the algorithmic scheme proposed in the previous sections with the method in [15]. The latter uses the random information matrix (random Hessian) in place of the matrix C defined above and recommends using the expected (Fisher) information when the full Hessian is indefinite. When the densities for w_k and v_k are Gaussian, this is equivalent to using Newton's method when possible, and using Gauss–Newton when the Hessian is indefinite. In general, using the expected information is known as the method of Fisher's scoring. In the Student's t case, the scalar Fisher information matrix is computed in [24] to be

$$\frac{s+1}{s+3}\sigma^{-2}, \tag{5.1}$$

where σ^2 is the variance and s is the degrees of freedom. The authors of [15] proposed using (5.1) as the Hessian approximation when the full Hessian is indefinite. Implementing this approach would effectively replace the terms $\|w_k\|_2^2$ or $\|v_k\|_2^2$, present in the denominators of Φ_k and A_k (see (4.2) and (4.4)), with terms that depend only on s_k and r_k, the degrees of freedom. So while the random information (Hessian) matrix can become indefinite, the Fisher information is insensitive to outliers and fails to downweigh their contributions to the Hessian approximation.

To overcome these drawbacks and find a middle ground between the full Hessian and a very rough approximation, we propose a Gauss–Newton method that is able

to incorporate the relative size information of the residuals into the Hessian approximation. In the rest of this section we provide the details for the application of this method and a proof of convergence.

As in [2], the convergence theory is based upon the versatile convex-composite techniques developed in [10]. We begin by choosing the convex-composite structure for objective (3.1). We write it in the convex-composite form $K = \rho \circ F$, with smooth F and convex ρ:

$$\rho\begin{pmatrix} c \\ u \end{pmatrix} = |c| + \frac{1}{2}\|u\|^2_{B^{-1}}, \tag{5.2}$$

$$F(x) = \begin{pmatrix} f(x) \\ \begin{bmatrix} w^{\mathcal{N}}(x) \\ v^{\mathcal{N}}(x) \end{bmatrix} \end{pmatrix}, \tag{5.3}$$

$$f(x) = \frac{1}{2}\sum_{k=1}^{N} s \ln\left[1 + \frac{\|v_k^{\mathcal{S}}\|^2_{(R_k^{\mathcal{S}})^{-1}}}{s}\right] + \sum_{k=1}^{N} r \ln\left[1 + \frac{\|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}}{r}\right]. \tag{5.4}$$

Note that the range of f is $\mathbf{R}_+$, and ρ is coercive on its domain. The terms indexed with superscript S in (3.5) and (3.6) combine to form a positive definite approximation to the Hessian of f. To see this, consider the scalar function

$$\kappa(x) := \frac{1}{2}\ln(1 + x^2/r)\ .$$

The second derivative of this function in x is given by

$$\frac{(r + x^2) - 2x^2}{(r + x^2)^2} = \frac{r - x^2}{(r + x^2)^2} \tag{5.5}$$

and is only positive on $(-\sqrt{r}, \sqrt{r})$. There are two reasonable globally positive approximations to take. The first,

$$\frac{r}{(r + x^2)^2},$$

simply ignores the subtracted term $-x^2$. In practice, we found this approximation to be too aggressive. Instead, we drop the $2x^2$ from the left of (5.5) to obtain the approximation

$$\frac{(r + x^2)}{(r + x^2)^2} = \frac{1}{(r + x^2)}\ . \tag{5.6}$$

Similarly, the terms indexed by superscript S in (3.5) and (3.6) provide globally positive definite approximations to the Hessian of f, using the strategy in (5.6). This strategy offers a significant computational advantage—the Hessian approximation that is built up downweighs the contributions of outliers, helping the algorithm proceed faster to the solution. As we shall see, these terms are also essential for the general convergence theory.

Our approach exploits the objective structure by iteratively linearizing F about the iterates x^k and solving the *direction finding subproblem*

$$\min_{d \in \mathbb{R}^{nN}} \quad \rho(F(x^k) + F^{(1)}(x^k)d) + \tfrac{1}{2}d^T U(x^k)d, \tag{5.7}$$

where $U(x^k)$ is a symmetric positive semidefinite matrix that depends continuously on x^k. The matrix function $U(x)$ is intended to approximate the Hessian with respect to x of the convex-composite Lagrangian [11] at a point $(F(x), y) \in \mathrm{graph}(\partial\rho)$ where $\partial\rho$ is the convex subdifferential of ρ. At points x where ρ is differentiable at $F(x)$, $y = \nabla\rho(F(x))$ and $U(x) \approx \nabla^2_{xx} L(x, y) = \sum_{j=1}^{p} (\nabla\rho(F(x)))_j \nabla^2 F_j(x)$. For any smoother in the framework of section 3, problem (5.7) can be solved with a single block-tridiagonal solve of the system (3.2), yielding descent directions d for the objective $K(x)$.

We now develop a general convergence theory for convex-composite methods to establish the overall convergence to a stationary point of $K(x)$. This theory is in the spirit of [2] and [10] and allows the inclusion of the quadratic term $\frac{1}{2} d^T U(x^k) d$ in (5.7). This term was not necessary in [2] but is crucial here. Note that the theory does not rely at all on the technique used to solve the direction finding subproblem, and so the theory in this paper applies to the algorithm in [2] by taking $U = 0$.

Recall from [10] that the first-order necessary condition for optimality in the convex-composite problem involving $K(x)$ is

$$0 \in \partial K(x) = F^{(1)}(x) \partial\rho\left(F(x)\right),$$

where $\partial K(x)$ is the generalized subdifferential of K at x [30] and $\partial\rho\left(F(x)\right)$ is the convex subdifferential of ρ at $F(x)$ [29]. Elementary convex analysis gives us the equivalence

$$0 \in \partial K(x) \quad \Leftrightarrow \quad K(x) = \inf_d \rho\Big(F(x) + F^{(1)}(x) d \Big).$$

For the general smoothing class of interest, it is desirable to modify this objective by including curvature information, yielding the problem (5.7). We define the difference function

$$\Delta(x; d) = \rho\Big(F(x) + F^{(1)}(x) d \Big) + \frac{1}{2} d^{\mathrm{T}} U(x) d - K(x), \tag{5.8}$$

where $U(x)$ is positive semidefinite and varies continuously with x. Note that $\Delta(x; d)$ is a convex function of d that is bounded below; hence the optimal value

$$\Delta^*(x) = \inf_d \Delta(x;\ d) \tag{5.9}$$

is well defined regardless of the existence of a solution. If $\Delta^*(x) = 0$, then $0 \in \arg\min_d \Delta(x;\ d)$. Hence, by [10, Theorem 3.6], $\Delta^*(x) = 0$ if and only if $0 \in \partial K(x)$.

Given $\eta \in (0, 1)$, we define a set of search directions at x by

$$D(x, \eta) = \left\{\ d \ \middle|\ \Delta(x; d) \le \eta \Delta^*(x)\ \right\}. \tag{5.10}$$

Note that if there is a $d \in D(x, \eta)$ such that $\Delta(x;\ d) \ge -\eta\varepsilon$, then $\Delta^*(x) \ge -\varepsilon$. We use the following generalized Gauss–Newton algorithm to solve the problem.

ALGORITHM 5.1. *Generalized Gauss–Newton algorithm.*

The inputs to this algorithm are

- $x^0 \in \mathbb{R}^{Nn}$*: initial estimate of state sequence,*
- $\varepsilon \ge 0$*: overall termination criterion,*
- $\eta \in (0, 1)$*: search direction selection parameter,*
- $\beta \in (0, 1)$*: step size selection parameter,*
- $\gamma \in (0, 1)$*: line search step size factor.*

The steps are as follows:

1. *Set the iteration counter* $\nu = 0$.
2. *(Generalized Gauss–Newton step)*
   Find descent direction $d^\nu \in D(x^\nu, \eta)$ *in* (5.10) *by solving* (5.7).
   Set $\Delta_\nu = \Delta(x^\nu; d^\nu)$ *in* (5.8), *and Terminate if* $\Delta_\nu \geq -\varepsilon$.
3. *(Line search) Set*

$$\begin{aligned} t_\nu &= \max \gamma^i \\ &\quad s.t. \quad i \in \{0, 1, 2, \ldots\} \quad \textit{and} \\ &\quad s.t. \quad \rho\left(F(x^\nu + \gamma^i d^\nu)\right) \leq \rho\left(F(x^\nu)\right) + \beta\gamma^i \Delta_\nu. \end{aligned}$$

4. *(Iterate) Set* $x^{\nu+1} = x^\nu + t_\nu d^\nu$, *and return to Step* 2.

We now present a general global convergence theorem that covers any smoother in section 3. This theorem also generalizes [2, Theorem 5.1] to include positive semidefinite curvature terms in the Gauss–Newton framework.

THEOREM 5.1. *Define*

$$\Lambda := \{u | \rho(u) \leq K(x^0)\}, \tag{5.11}$$

and suppose that there exists a $\tau > 0$ *such that* $F^{(1)}$ *is bounded and uniformly continuous on the set*

$$S_0 := \overline{co}\left(F^{-1}(\Lambda)\right) + \tau\mathbb{B}, \tag{5.12}$$

where $\mathbb{B} := \{\, x \mid \|x\| \leq 1 \,\}$. *If* $\{x^\nu\}$ *is a sequence generated by the Gauss–Newton algorithm, Algorithm* 5.1, *with initial point* x^0 *and* $\varepsilon = 0$, *then one of the following must occur:*

(i) *The algorithm terminates finitely at a point* x^ν *with* $0 \in \partial K(x^\nu)$.
(ii) *The sequence* $\|d^\nu\|$ *diverges to* $+\infty$.
(iii) $\lim_{\nu\in I} \Delta_\nu = \lim_{\nu\in I} \Delta^*(x^\nu) = 0$ *for every subsequence* I *for which the set* $\{\, d^\nu \mid \nu \in I \,\}$ *is bounded.*

Moreover, if $\bar{x}$ *is any cluster point of a subsequence* $I \subset \mathbf{Z}_+$ *such that the subsequence* $\{d^\nu | \nu \in I\}$ *is bounded, then* $0 \in \partial K(\bar{x})$.

Remark 5.2. We note that this theorem also corrects a flaw in the statement of [2, Theorem 5.1]. In that theorem it was only stated that $F^{(1)}$ need be uniformly continuous on the set $\overline{\text{co}}(F^{-1}(\Lambda))$. Here we require that $F^{(1)}$ be both bounded and uniformly continuous on a slightly larger set. In particular, the proof given here fills a gap in the proof appearing in the technical report [5].

Proof. We will assume that none of (i), (ii), (iii) occur and establish a contradiction. Then there is a subsequence I such that

$$\sup_{\nu\in I} \|d_\nu\| < \infty \quad \text{and} \quad \sup_{\nu\in I} \Delta_\nu \leq \zeta < 0.$$

Since $K(x^\nu)$ is a decreasing sequence that is bounded below by 0, we know that the differences $K(x^{\nu+1}) - K(x^\nu) \to 0$. Therefore, by Step 3 of Algorithm 5.1, $\zeta t_\nu \Delta_\nu \to 0$, which implies that $t_{\nu\in I} \to 0$. Without loss of generality we may assume that $t_\nu \leq 1$ and $t_\nu \|d_\nu\| \leq \gamma\tau$ for all $\nu \in I$. Hence, for all $\nu \in I$,

$$\begin{aligned} \|F(x^\nu + t_\nu\gamma^{-1}d^\nu) - F(x^\nu)\| &\leq t_\nu\gamma^{-1} \int_0^1 \left\|F'(x^\nu + st_\nu\gamma^{-1}d^\nu)\right\| \|d^\nu\| ds \\ &\leq \tau M, \end{aligned}$$

where M is a bound on F' over S_0. Let K be a Lipschitz constant for ρ over the compact set $\Lambda + \tau M\mathbb{B}$. Again by Step 3 of Algorithm 5.1, for all $\nu \in I$,

$$\begin{aligned}\beta\gamma^{-1}t_\nu\Delta_\nu &\le \rho(F(x^\nu + t_\nu\gamma^{-1}d^\nu)) - \rho(F(x^\nu))\\ &\le t_\nu\gamma^{-1}\Delta_\nu + K\|F(x^\nu + t_\nu\gamma^{-1}d^\nu) - F(x^\nu) - t\nu\gamma^{-1}F^{(1)}(x^\nu)d^\nu\|\\ &= t_\nu\gamma^{-1}\Delta_\nu + t_\nu\gamma^{-1}K\left\|\int_0^1 \left(F^{(1)}(x^\nu + st_\nu\gamma^{-1}d_\nu) - F^{(1)}(x^\nu)\right)d^\nu ds\right\|\\ &\le t_\nu\gamma^{-1}\left(\Delta_\nu + K\omega(t_\nu\gamma^{-1}\|d_\nu\|)\|d_\nu\|\right),\end{aligned}$$

where ω is the modulus of continuity of F' on S_0. Rearranging, we obtain

$$0 \le (1-\beta)\Delta_\nu + K\omega(t_\nu\gamma^{-1}\|d_\nu\|)\|d_\nu\|.$$

Taking the limit for $\nu \in I$, we obtain the contradiction $0 \le (1-\beta)\zeta$. Hence, $\lim_{\nu\in I}\Delta_\nu = 0$, which implies that $\lim_{\nu\in I}\Delta^*(x^\nu) = 0$, since $\Delta_\nu \le \eta\Delta^*(x^\nu) \le 0$.

Finally, suppose that $\bar{x}$ is a cluster point of a sequence $I \subset \mathbf{Z}_+$ for which $\{d^\nu\}$ is bounded. Without loss of generality, there exists a $\bar{d}$ such that $(x^\nu, d^\nu)_{\nu\in I} \to (\bar{x}, \bar{d})$. For all $d \in \mathbb{R}^{Nn}$,

$$\begin{aligned}\Delta_\nu &= \rho\left(F(x^\nu) + F^{(1)}(x^\nu)d^\nu\right) + \tfrac{1}{2}\|d^\nu\|^2_{U^\nu} - \rho\left(F(x^\nu)\right)\\ &\le \eta\Delta^*(x^\nu)\\ &\le \eta\left(\rho\left(F(x^\nu) + F^{(1)}(x^\nu)d\right) + \tfrac{1}{2}\|d\|^2_{U^\nu} - \rho\left(F(x^\nu)\right)\right),\end{aligned}$$

where $U^\nu = U(x^\nu)$. Taking the limit over I gives

$$\begin{aligned}0 &= \rho\left(F(\bar{x}) + F^{(1)}(\bar{x})\bar{d}\right) + \tfrac{1}{2}\|\bar{d}\|^2_{\bar{U}} - \rho\left(F(\bar{x})\right)\\ &\le \eta\left(\rho\left(F(\bar{x}) + F^{(1)}(\bar{x})d\right) + \tfrac{1}{2}\|d\|^2_{\bar{U}} - \rho\left(F(\bar{x})\right)\right),\end{aligned}$$

where $\overline{U} = U(\bar{x})$. Since d was chosen arbitrarily, it must be the case that $\Delta^*(\bar{x}) = 0$, which implies that $0 \in \partial K$ by [10, Theorem 3.6]. ∎

We can guarantee convergence to a stationary point under additional assumptions. The details are given in the following corollary.

COROLLARY 5.2. *Suppose that* $F^{-1}(\Lambda) = \{\, x \mid F(x) \in \Lambda \,\}$ *is bounded, and there exists* $0 < \lambda_{\min}$ *such that*

$$\forall\, x \in F^{-1}(\Lambda), \quad 0 < \lambda_{\min}\|d\|^2 \le d^T U(x) d \quad \forall d \in \mathrm{Null}(F^{(1)}(x))\ . \tag{5.13}$$

If $\{x^\nu\}$ *is a sequence generated by Algorithm* 5.1 *with initial point* x^0 *and* $\varepsilon = 0$*, then* $\{x^\nu\}$ *and* $\{d^\nu\}$ *are bounded, and either the algorithm terminates finitely at a point* x^ν *with* $0 \in \partial K(x^\nu)$ *or* $\Delta_\nu \to 0$ *as* $\nu \to \infty$*, and every cluster point* $\bar{x}$ *of the sequence* $\{x^\nu\}$ *satisfies* $0 \in \partial K(\bar{x})$*.*

Proof. First note that $F^{-1}(\Lambda)$ is closed since F is continuous, and therefore $F^{-1}(\Lambda)$ is compact, since by assumption it is bounded. Hence S_0 (see (5.12)) is also compact. Therefore, $F^{(1)}$ is uniformly continuous and bounded on S_0, which implies that the hypotheses of Theorem 5.1 are satisfied, and so one of (i)–(iii) must hold. If (i) holds, we are done, and so we will assume that the sequence $\{x^\nu\}$ is infinite. Since $\{x^\nu\} \subset F^{-1}(\Lambda)$, this sequence is bounded. We now show that the sequence $\{d^\nu\}$ of search directions is also bounded.

Suppose that (5.13) holds. For any direction d^ν, note that d^ν satisfies

$$\rho\left(F(x^\nu)+F^{(1)}(x^\nu)d^\nu\right)+\frac{1}{2}\|d^\nu\|^2_{U^\nu}\le\rho\left(F(x^\nu)\right)\le\rho(F(x^0)). \tag{5.14}$$

Since $\rho\ge 0$, we have

$$\{F(x^\nu)\}\subset\Lambda \quad\text{and}\quad \{F(x^\nu)+F^{(1)}(x^\nu)d^\nu\}\subset\Lambda \tag{5.15}$$

and

$$\left\{\frac{1}{2}(d^\nu)^TU^\nu d^\nu\right\}\le\rho\left(F(x^0)\right)\quad\forall\nu. \tag{5.16}$$

Suppose that the $\{d^\nu\}$ is unbounded. Then, without loss of generality, there exist a subsequence I, a unit vector u, and a vector $\bar{x}\in F^{-1}(\Lambda)$ such that $\lim_{\nu\in I} d^\nu/\|d^\nu\|\to u$ and $\lim_{\nu\in I} x^\nu\to\bar{x}$. Since Λ is bounded, (5.15) implies that $F^{(1)}(\bar{x})u=0$, and so $u\in\mathrm{Nul}(F^{(1)}(\bar{x}))$, and therefore

$$0<\lambda_{\min}\le u^TU(\bar{x})u.$$

On the other hand, by (5.16), $\frac{1}{2}(\frac{d^\nu}{\|d^\nu\|})^TU^\nu(\frac{d^\nu}{\|d^\nu\|})\le\frac{\rho(F(x^0))}{\|d^\nu\|^2}$, and so in the limit we have the contradiction

$$0<\lambda_{\min}\le u^TU(\bar{x})u\le 0.$$

Hence d^ν are bounded. The result now follows from Theorem 5.1. ∎

We now show that all smoothers of section 3 satisfy the required assumptions of Theorem 5.1 and Corollary 5.2.

COROLLARY 5.3 (smoother satisfaction). *Suppose that the process and measurement functions g_k and h_k in (2.3) are twice continuously differentiable. Then for F given in (5.3), $F^{(1)}$ is bounded and uniformly continuous on S_0 in (5.12). Moreover, the hypotheses of Corollary 5.2 hold if for all x in $F^{-1}(\Lambda)$ and for all k there exists η such that*

$$0<\eta<\sigma_{\min}(G^{\mathcal{S}}(x)),\quad G^{\mathcal{S}}(x):=\begin{bmatrix} I & 0 & 0 & \\ -(G_2(x_1))^{\mathcal{S}} & I & 0 & 0\\ 0 & \ddots & \ddots & \ddots\\ & 0 & -(G_N(x_{N-1}))^{\mathcal{S}} & I\end{bmatrix}.$$

Proof. We first show that both Λ and $F^{-1}(\Lambda)$ are bounded. The first claim follows immediately by the coercivity of ρ in (5.2). To verify the second claim, we will show that for any sequence of x^ν with $\|x^\nu\|\to\infty$, we can find a subsequence J such that $\lim_{\nu\in J}\|w^\nu\|=\infty$, which implies the existence of subsequence I such that either $\lim_{\nu\in I}\|w^\nu\|=\infty$ or $\lim_{\nu\in I} f(x^\nu)=\infty$. In particular, there does not exist an unbounded sequence $\{x^\nu\}$ with $F(x^\nu)\subset\Lambda$, and therefore $F^{-1}(\Lambda)$ must be bounded.

If $\|x^\nu\|\to\infty$, we can find an index $k\subset[1,\dots,N]$ and subsequence J such that $\lim_{\nu\in J}\|x_k^\nu\|=\infty$. Either $\lim_{\nu\in J} w_k^\nu=\infty$ and we are done, or $\lim_{\nu\in J}\|g_k(x_{k-1}^\nu)\|=\infty$, and so $\lim_{\nu\in J}\|x_{k-1}^\nu\|=\infty$. Iterating this argument, we arrive at the limiting case $w_1^\nu=x_1^\nu-x_1^0$, and so if all $\|w_j^\nu\|$ are bounded for $j>1$, we can guarantee that $\lim_{\nu\in J}\|w_1^\nu\|=\infty$.

Since F is twice continuously differentiable by the hypotheses on g and h, the boundedness of $F^{-1}(\Lambda)$ establishes the boundedness and uniform continuity of $F^{(1)}$ on S_0 in (5.12) for any $\tau > 0$.

It remains to show that condition (5.13) is satisfied. Let $\mathcal{W}^{\mathcal{N}}$, $\mathcal{W}^{\mathcal{S}}$ denote the indices associated to all subvectors $w_k^{\mathcal{N}}$ and $w_k^{\mathcal{S}}$ within w_k. If $d \in \text{Null}(F^{(1)}(x))$, then necessarily $d_{\mathcal{W}^{\mathcal{N}}} = 0$. This is simply because $F^{(1)}$ is nonsingular on $\mathcal{W}^{\mathcal{N}}$, since it contains the submatrix

$$G^{\mathcal{N}}(x) := \begin{bmatrix} \mathrm{I} & 0 & & \\ -(G_2(x_1))^{\mathcal{N}} & \mathrm{I} & \ddots & \\ & \ddots & \ddots & 0 \\ & & -(G_N(x_{N-1}))^{\mathcal{N}} & \mathrm{I} \end{bmatrix},$$

which is the standard process matrix G projected to those coordinates where Gaussian modeling is applied. To finish the analysis, we present the full form of the matrix U restricted to $\mathcal{W}^{\mathcal{S}}$:

$$U = \begin{bmatrix} U_1 & A_2^{\mathrm{T}} & 0 & \\ A_2 & U_2 & A_3^{\mathrm{T}} & 0 \\ 0 & \ddots & \ddots & \ddots \\ & 0 & A_N & U_N \end{bmatrix} + \text{diag}(\{\Phi_k\}, \tag{5.17}$$

where

$$\begin{aligned} A_k &= -\frac{r(Q_k^{\mathcal{S}})^{-1} G_k^{\mathcal{S}}}{r + \|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}}, \\ U_k &= \frac{r(G_{k+1}^{\mathcal{S}})^{\mathrm{T}} (Q_{k+1}^{\mathcal{S}})^{-1} G_{k+1}^{\mathcal{S}}}{r + \|w_{k+1}^{\mathcal{S}}\|^2_{(Q_{k+1}^{\mathcal{S}})^{-1}}} + \frac{r(Q_k^{\mathcal{S}})^{-1}}{r + \|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}}, \\ \Phi_k &= \frac{s(H_k^{\mathcal{S}})^{\mathrm{T}} (R_k^{\mathcal{S}})^{-1} H_k^{\mathcal{S}}}{(s + \|v_k^{\mathcal{S}}\|^2_{(R_k^{\mathcal{S}})^{-1}})}. \end{aligned} \tag{5.18}$$

Note that we can write the first summand in (5.17) as $(G^{\mathcal{S}})^T \widetilde{Q}^{-1} G^{\mathcal{S}}$, where

$$\widetilde{Q}^{-1} := \text{diag}(\{\widetilde{Q}_k^{-1}\}), \quad \widetilde{Q}_k^{-1} = \frac{r(Q_k^{\mathcal{S}})^{-1}}{r + \|w_k^{\mathcal{S}}\|^2_{(Q_k^{\mathcal{S}})^{-1}}}.$$

Since $F^{-1}(\Lambda)$ is bounded, the denominators of $\widetilde{Q}_k^{-1}$ are bounded, and so eigenvalues of $\widetilde{Q}_k^{-1}$ are bounded from below, and the singular values of $G^{\mathcal{S}}$ are bounded from above.

We have

$$0 < \eta_{\min} \le \sigma_{\min}(G^{\mathcal{S}}) \le \sigma_{\max}(G^{\mathcal{S}}) \le \eta_{\max}$$

for all x, where the upper bound follows from [4, Theorem 2.2] together with compactness of $F^{-1}(\Lambda)$.

Then, by [4, Theorem 2.1], we have

$$\kappa((G^{\mathcal{S}})^T \widetilde{Q}^{-1} G^{\mathcal{S}}) \leq \frac{\lambda_{\max}(\widetilde{Q}^{-1})\eta_{\max}^2}{\lambda_{\min}(\widetilde{Q}^{-1})\eta_{\min}^2}$$

for all $x \in F^{-1}(\Lambda)$. This completes the proof. ∎

Remark 5.3. One can also consider conditions on the individual $g_k^{\mathcal{S}}$ that can produce a lower bound η on $G^{\mathcal{S}}$, as required by Corollary 5.3. One such condition is

$$0 < \eta \leq \left\{1 + \sigma_{\min}^2(g_{k+1}^{(1)}) - \sigma_{\max}(g_k^{(1)}) - \sigma_{\max}(g_{k+1}^{(1)})\right\}. \tag{5.19}$$

If this condition is satisfied, then by [4, Theorem 2.2], $\eta < \sigma_{\min}(G^{\mathcal{S}})$. This condition is sufficient, but not necessary in general.

6. Numerical experiments.

6.1. T-robust smoother: Function reconstruction using splines. In this section we compare the new T-robust smoother with the ℓ_2-Kalman smoother [9] and with the ℓ_1-Laplace robust smoother [2], both implemented in [1]. The *ground truth* for this simulated example is

$$x(t) = \begin{bmatrix} -\cos(t) & -\sin(t) \end{bmatrix}^{\mathrm{T}}.$$

The time between measurements is a constant Δt. We model the two components of the state as the first and second integrals of white noise, so that

$$g_k(x_{k-1}) = \begin{bmatrix} 1 & 0 \\ \Delta t & 1 \end{bmatrix} x_{k-1}, \qquad Q_k = \begin{bmatrix} \Delta t & \Delta t^2/2 \\ \Delta t^2/2 & \Delta t^3/3 \end{bmatrix}.$$

This stochastic model for function reconstruction underlies the Bayesian interpretation of cubic smoothing splines; see [35] for details.

The measurement model for the conditional mean of measurement z_k given state x_k is defined by

$$h_k(x_k) = \begin{bmatrix} 0 & 1 \end{bmatrix} x_k = x_{2,k}, \qquad R_k = \sigma^2,$$

where $x_{2,k}$ denotes the second component of x_k, $\sigma^2 = 0.25$ for all experiments, and the degrees of freedom parameter was set to 4 for the Student's t methods.

The measurements $\{z_k\}$ were generated as a sample from

$$z_k = x_2(t_k) + v_k, \quad t_k = 0.04\pi \times k,$$

where $k = 1, 2, \ldots, 100$. The measurement noise v_k was generated according to the following schemes.

1. *Nominal*: $v_k \sim \mathbf{N}(0, 0.25)$.
2. *Gaussian contamination*:

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

for $p \in \{0.1, 0.2, 0.5\}$ and $\phi \in \{1, 4, 10, 100\}$.

TABLE 1
Function reconstruction via spline: Median MSE over 1000 *runs and intervals containing* 95% *of MSE results.*

Outlier	p	ℓ_2 **MSE**	ℓ_1 **MSE**	**Student's t MSE**
Nominal	—	.04(.02, .1)	.04(.01, .1)	.04(.01, .09)
$\mathbf{N}(0,1)$	.1	.06(.02, .12)	.04(.02, .10)	.04(.02, .10)
$\mathbf{N}(0,4)$	.1	.09(.04, .29)	.05(.02, .12)	.04(.02, .11)
$\mathbf{N}(0,10)$	.1	.17(.05, .55)	.05(.02, .13)	.04(.02, .11)
$\mathbf{N}(0,100)$	.1	1.3(.30, 5.0)	.05(.02, .14)	.04(.02, .11)
$\mathbf{U}(-10,10)$	.1	.47(.12, 1.5)	.05(.02, .13)	.04(.02, .10)
$\mathbf{N}(0,10)$	.2	.32(.11, .95)	.06(.02, .19)	.05(.02, .16)
$\mathbf{N}(0,100)$	.2	2.9(.94, 8.5)	.07(.02, .22)	.05(.02, .14)
$\mathbf{U}(-10,10)$	.2	1.1(.36, 3.0)	.07(.03, .26)	.05(.02, .13)
$\mathbf{N}(0,10)$	.5	.74(.29, 1.9)	.13(.05, .49)	.10(.04, .45)
$\mathbf{N}(0,100)$	.5	7.7(2.9, 18)	.21(.06, 1.6)	.09(.03, .44)
$\mathbf{U}(-10,10)$	.5	2.6(1.0, 5.8)	.20(.06, 1.4)	.10(.03, .44)

3. *Uniform contamination*:

$$v_k \sim (1-p)\mathbf{N}(0,0.25) + p\mathbf{U}(-10,10)$$

for $p \in \{0.1, 0.2, 0.5\}$.

Each experiment was performed 1000 times. Table 1 presents the results for our simulated fitting showing the median mean squared error (MSE) value and a quantile interval containing 95% of the results. The MSE is defined by

$$\frac{1}{N}\sum_{k=1}^{N}[x_1(t_k) - \hat{x}_{1,k}]^2 + [x_2(t_k) - \hat{x}_{2,k}]^2, \tag{6.1}$$

where $\{\hat{x}_k\}$ is the corresponding estimating sequence.

From Table 1 one can see that T-robust and the ℓ_1-smoother perform as well as the (optimal) ℓ_2-smoother at nominal conditions and that both continue to perform at that same level for a variety of outlier generating scenarios. T-robust always performs at least as well as the ℓ_1-smoother, and it gains an advantage when either the probability of contamination is high or the contamination is uniform. This is likely due to the redescending influence function of the Student's t distribution; the smoother effectively throws out bad points rather than simply decreasing their impact to a certain threshold, as is the case for the ℓ_1-smoother. As an example, results coming from a single run for the case where 50% of measurements are contaminated with the uniform distribution on $[-10, 10]$ are displayed in Figure 2. Notice that T-robust has an advantage over the ℓ_1-smoother.

6.2. T-robust smoother: Van Der Pol oscillator. In this section, we present results for the Van Der Pol (VDP) oscillator, described in detail in [2]. The VDP oscillator is a coupled nonlinear ODE defined by

$$\begin{aligned}\dot{x}_1(t) &= x_2(t),\\ \dot{x}_2(t) &= \mu[1 - x_1(t)^2]x_2(t) - x_1(t).\end{aligned}$$

The process model here is the Euler approximation for $X(t_k)$ given $X(t_{k-1})$:

$$g_k(x_{k-1}) = \begin{pmatrix} x_{1,k-1} + x_{2,k-1}\Delta t \\ x_{2,k-1} + \{\mu[1 - x_{1,k}^2]x_{2,k} - x_{1,k}\}\Delta t \end{pmatrix}.$$

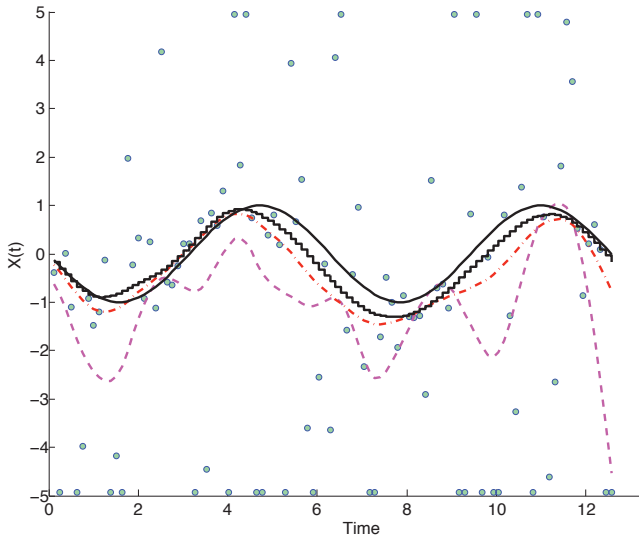

FIG. 2. *Function reconstruction via spline: Performance of ℓ_2 Kalman smoother (dash), ℓ_1-Laplace robust smoother (dash-dot), and T-robust (staircase solid) on contaminated normal model with 50% outliers distributed uniformly on $[-10, 10]$. True state x(t) is drawn as solid line. Measurements appear as "o" symbols, and all measurements visible off of the true state are outliers in this case. Values outside $[-5, 5]$ are plotted on the axis limits.*

For this simulation, the *ground truth* is obtained from a stochastic Euler approximation of the VDP oscillator. To be specific, with $\mu = 2$, $N = 164$, and $\Delta t = 16/N$, the ground truth state vector x_k at time $t_k = k\Delta t$ is given by $x_0 = (0, -0.5)^{\mathrm{T}}$, and for $k = 1, \ldots, N$, $x_k = g_k(x_{k-1}) + w_k$, where $\{w_k\}$ is a realization of independent Gaussian noise with variance 0.01.

In [2], the ℓ_1-Laplace smoother was shown to have a performance superior to that of the ℓ_2-smoother, both implemented in [1]. We compared the performance of the nonlinear T-robust and nonlinear ℓ_1-Laplace smoothers and found that T-robust gains an advantage in the extreme cases of 70% outliers. Figure 3 illustrates results coming from a single representative run. For 40% or fewer outliers, it is hard to differentiate the performance of the two smoothers for this nonlinear example.

6.3. T-robust smoother: Underwater tracking application. This application is described in detail in [2], so we just give a brief overview here. In [2] we used the application to test the ℓ_1-Laplace smoother. Here we use it for a qualitative comparison between the T-robust smoother, the ℓ_1-Laplace smoother, and the ℓ_2 smoother with outlier removal.

In this experiment, a tracking target was hung on a steel cable approximately 200 meters below a ship. The pilot was attempting to keep the ship in place (hold station) at specific coordinates, but the ship was pitching and rolling due to wave action. The measurements for the smoother were sound travel times between the tracking target and four bottom mounted transponders at known locations, and pressure readings from a gauge that was placed on the target. Tracking data was independently verified using a GPS antenna mounted on a ship, and the GPS system provided submeter accuracy in position.

Pressure measurements in absolute bars were converted to depth in meters by the

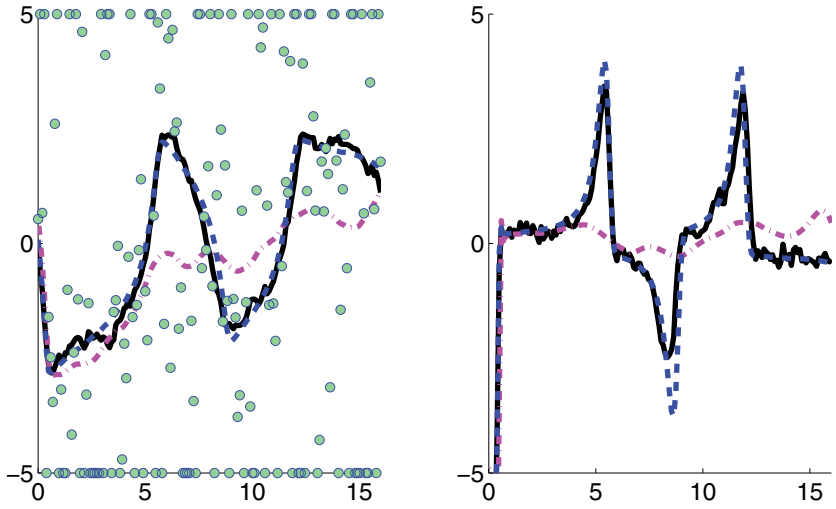

FIG. 3. *VDP oscillator: Smoother fits for X-component (left) and Y-component (right), with 70% outliers $N(0, 100)$. The black solid line is truth, the magenta dash-dot line is the ℓ_1 smoother result, and the blue dashed line is T-robust. Measurements on the X-component are shown as dots, with outliers outside the range $[-5, 5]$ plotted on top and bottom axes.*

formula

$$\text{depth} = 9.9184(\text{pressure} - 1).$$

We use N to denote the total number of time points at which we have tracking data. For $k = 1, \ldots, N$, the state vector at time t_k is defined by $x_k = (e_k, n_k, d_k, \dot{e}_k, \dot{n}_k, \dot{d}_k)^{\mathrm{T}}$, where (e_k, n_k, d_k) is the (east, north, depth) location of the object (in meters from the origin), and $(\dot{e}_k, \dot{n}_k, \dot{d}_k)$ is the time derivative of this location.

The measurement vector at time t_k is denoted by z_k. The first four components of z_k are the range measurements to the corresponding bottom mounted transponders, and the last component is the depth corresponding to the pressure measurement. For $j = 1, \ldots, 4$, the model for the mean of the corresponding range measurements was

$$h_{j,k}(x_k) = \|(e_k, n_k, d_k) - b_j\|_2 - \Delta r_j.$$

These measurements were assumed independent with standard deviation 3 meters. These depth measurements were assumed to have standard deviation of 0.05 meters. We use Δt_k to denote $t_{k+1} - t_k$. The model for the mean of x_{k+1} given x_k was

$$\begin{aligned} &g_{k+1}(x_k) \\ &= (e_k + \dot{e}_k \Delta t_k,\ n_k + \dot{n}_k \Delta t_k,\ d_k + \dot{d}_k \Delta t_k,\ \dot{e}_k,\ \dot{n}_k,\ \dot{d}_k)^{\mathrm{T}}. \end{aligned}$$

The process noise corresponding to east, north, and depth components of the conditional distribution of x_{k+1} given x_k was assumed to be Gaussian, with mean zero and standard deviation $.01\Delta t_k$. The process noise corresponding to the derivative vector of east, north, and depth components of the conditional mean x_{k+1} given x_k was also assumed Gaussian with mean zero and standard deviation $.2\Delta t_k$.

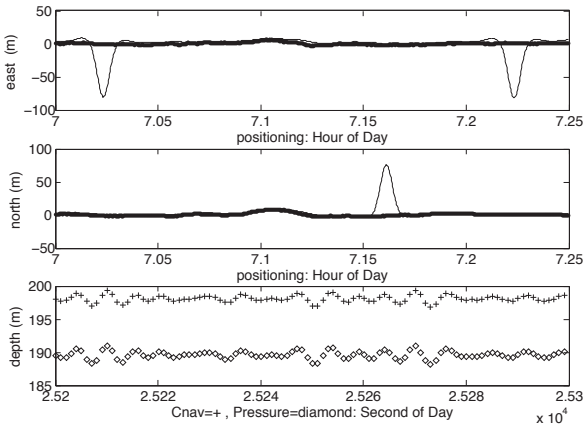

FIG. 4. *Track: Independent GPS verification (thick line and +); ℓ_2-smoother estimate (thin line). Note the large outliers in the data.*

ℓ_2-smoother results without outlier removal are shown in Figure 4. There are three large peaks (two in the east component and one in the north component of the state) that are due to measurement outliers, and require either an outlier removal strategy or robust smoothing.

Three fits are shown in Figure 5: ℓ_1-Laplace, T-robust, and ℓ_2-smoother with outlier removal. The darker curves appearing below the track are independent verifications using the GPS tracking near the top of the cable. A depth of 198 meters was added to the depth location of the GPS antenna so that the depth comparison can use the same axis for both the GPS data and the tracking results. Note that the time scale for the depth plots is different from (much finer than) the north, east, and down plots and demonstrates the accuracy of the GPS tracking as validated by the pressure sensor.

T-robust, like the ℓ_1-Laplace smoother, was able to use the whole data sequence, despite large outliers in the data. The fits look very similar, and it is clear that T-robust can also be used for smoothing in the presence of outliers. Note that the T-robust track (b) is smoother than the ℓ_1-Laplace track (a) but has more detail than the ℓ_2-smoother track with outlier removal (c). This is easiest to see by comparing the east coordinates in (a), (b), and (c) of Figure 5 between 7.2 and 7.25 hours.

The residual plots in Figure 5 are quite revealing. Outliers are defined as measurements corresponding to residuals with absolute value greater than three standard deviations from the mean. All outliers are shown as "o" characters, and those that fall outside the axis limits are plotted on the vertical axis limit lines. Note that the ℓ_2-smoother with outlier removal detects outliers after the first fit that are not outliers after the second fit. The peaks in Figure 4 are large enough to influence the entire fit, and so some points which are actually "good" measurements are removed by the 3-σ edit rule, resulting in "oversmoothing" of the outlier removal track and more detail in both of the robust smoothers in Figure 5.

The ℓ_1-Laplace smoother pushes more of the residuals to zero, particularly those corresponding to depth measurements, which are the most reliable and frequent. The T-robust smoother is somewhere in between; the residuals for the depth track are smaller in comparison to the residuals of the ℓ_2-smoother but are not set to zero as by

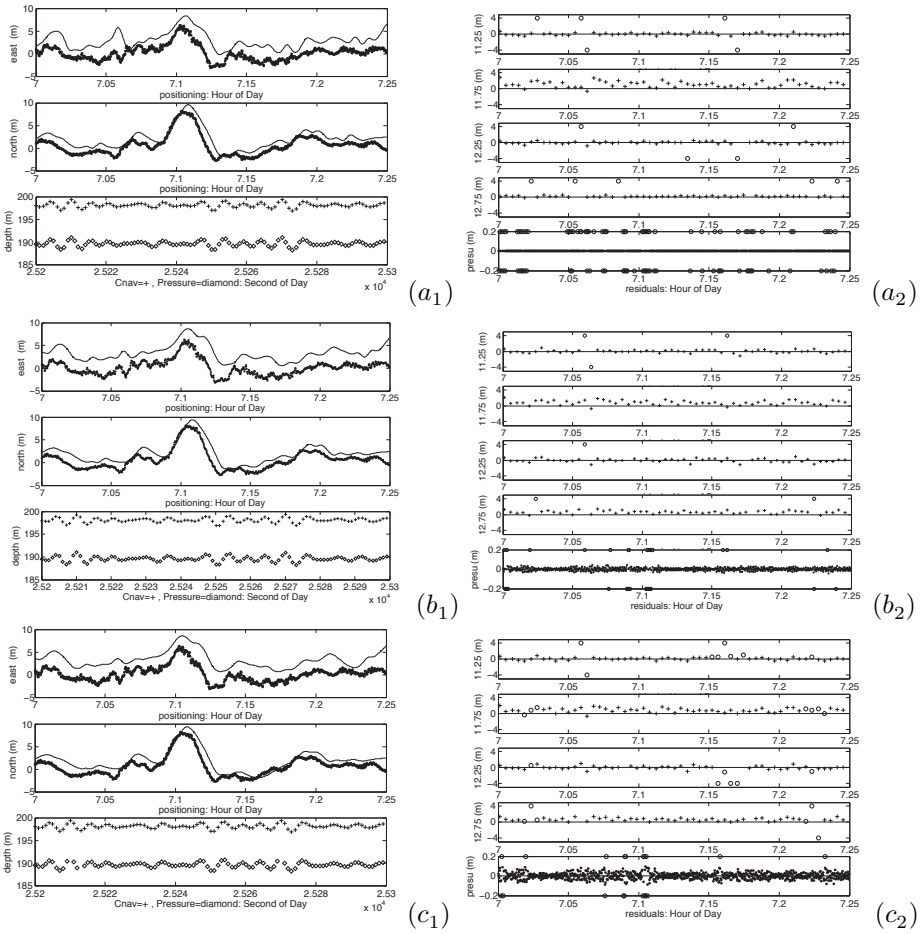

FIG. 5. *Track: Independent GPS verification (thick line and +) and residuals for* (a) ℓ_1*-Laplace smoother (thin line),* (b) *T-Robust smoother (thin line),* (c) ℓ_2*-smoother with outlier removal. All residuals lying outside of the given interval are graphed on the boundary of the interval.*

the ℓ_1-Laplace smoother. As discussed previously, these features are artifacts of the behavior of the distributions at zero, and the choice of smoother should be guided by particular applications.

6.4. T-trend smoother: Reconstruction of a sudden change in state. We present a proof of concept result for the T-trend smoother using two Monte Carlo studies of 200 runs. In the first study, the state vector, as well as the process and measurement models, are the same as in section 6.1. At any run, x_2 has to be reconstructed from 20 measurements corrupted by a white Gaussian noise of variance 0.05 and collected on $[0, 2\pi]$ using a uniform sampling grid. The top panel of Figure 6 reports the boxplot of the 200 root-MSE errors for the ℓ_2-, ℓ_1-, and T-trend smoothers, while the top right panel of Figure 6 displays the estimate obtained in a single run. It is apparent that the performance of the three estimators is very similar.

The second experiment is identical to the first except that we introduce a "jump" at the middle of the sinusoidal wave. The bottom panel of Figure 6 reveals the superior performance of the T-trend smoother under these perturbed conditions. The result

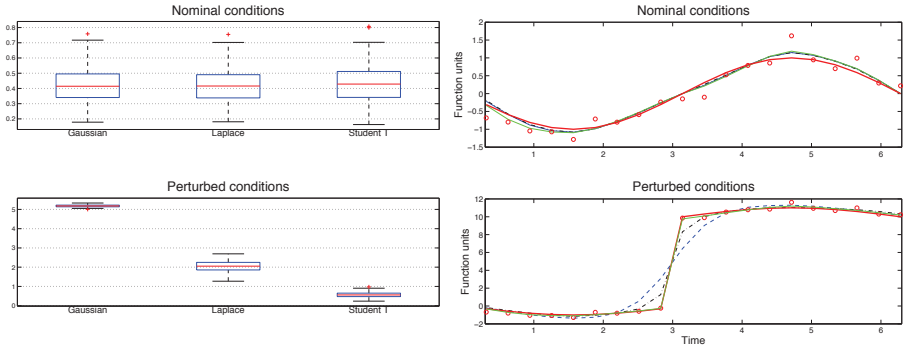

FIG. 6. *Reconstruction of a sudden change in state obtained by ℓ_2, ℓ_1, and T-trend smoothers. Left: Boxplot of reconstruction errors under nominal (top) and perturbed (bottom) conditions. Right: Reconstructions obtained using ℓ_2 (dashed), ℓ_1 (dashdot), and T-trend (thin line) smoother. The thick line is the true state.*

depicted in the bottom right panel of Figure 6 for a single run of the experiment is representative of the average performance of the estimators. The estimate achieved by the ℓ_2-smoother (dashed line) does not follow the jump well (the true state is the solid line). The ℓ_1-smoother (dashdot) does a better job than the ℓ_2-smoother, and the T-trend smoother outperforms the ℓ_1-smoother, following the jump very closely while still providing a good solution along the rest of the path.

6.5. Reconstruction of a sudden change in state in the presence of outliers. Until now, we have considered robust and trend applications separately in order to compare with previous robust smoothing formulations and to highlight the main features of the trend-filtering problem. A natural extension is to consider these features in tandem; in other words, can we smooth a track which has *both* outliers and a sudden change in state? In fact, smoothers of this nature (but exploiting convex formulations) have already been proposed [16].

The challenge to building such a strong smoother is that without prior knowledge, it is difficult to tell the difference between a bad measurement (an outlier) and a good measurement that may be consistent with a sudden change in the state. In many cases, the user will be aware that some sensors are reliable, while others are subject to contamination. This kind of prior information can now easily be incorporated using the generality and flexibility of section 3, so that the user may specify trustworthy sensors (by modeling corresponding residual *indices* with Gaussians) as well as stable state components (by modeling corresponding transition residual *indices* with Gaussians). Note that this is very different from specifying which of the individual *measurements* are reliable or which individual *transitions* follow the process model.

In this section, we consider a situation where we have a trustworthy sensor s_1 and an occasionally malfunctioning sensor s_2. Sensor s_2 gives frequent measurements, but some proportion of the time is subject to heavy contamination, while sensor s_1 gives measurements rarely, but they are trustworthy (i.e., only subject to small Gaussian noise). Using the flexible interface implemented in [1], we can model s_1 errors as Gaussian and s_2 errors as Student's t.

We use the setup in section 6.4 together with the Gaussian outlier contamination scheme described in section 6.1. Both measurements are direct, so the measurement

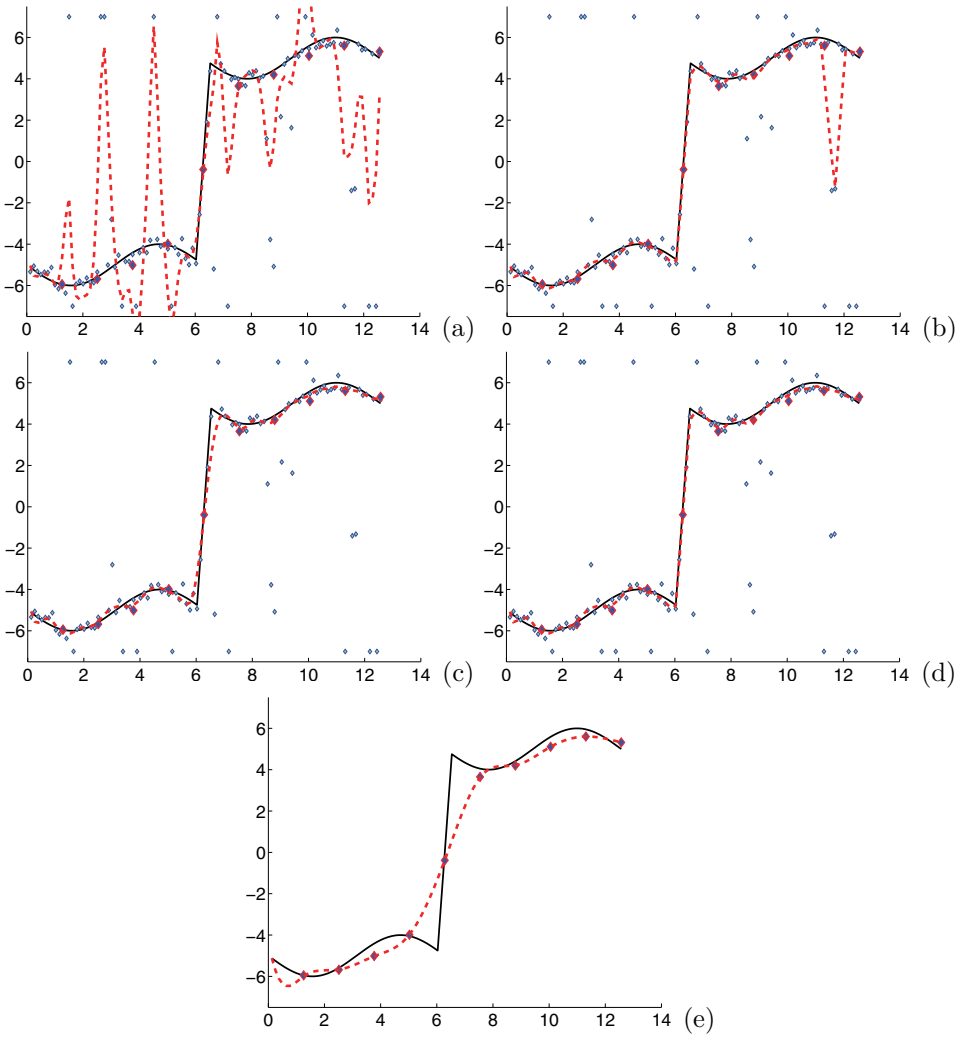

FIG. 7. *Tracking a sudden change in the presence of outliers. Measurements are plotted using diamonds, with s_1 measurements (rare, reliable) represented by large symbols, and s_2 measurements (frequent, contaminated) represented by small symbols. Outliers appear on the axes when they are out of range of the plot limits. Ground truth is shown using a sold black line, and smoother results are shown using a red dashed line.* (a) *Least squares smoother (Gaussian errors for process and measurements) is very vulnerable to outliers.* (b) *All-T smoother (Student's t errors for all components) follows sudden change in state quite well and ignores most outliers; however, it is fooled by a cluster of outliers, since it cannot distinguish them from sudden changes in state.* (c) *Differential T-robust (Student's t errors for s_2, Gaussian for all others) exploits differential properties of s_1 and s_2; however, because it uses the Gaussian model for transition errors, accuracy decreases around the jump.* (d) *Trend-following differential robust (Student's t errors for process components and s_2; Gaussian for s_1) ignores outliers and follows sudden change in state, using both information in the reliable measurements and appropriate process noise model to improve on* (b) *and* (c). (e) *Result obtained using only good (s_1) measurements with least squares smoothing to recover the track. The sparse (good) measurements alone do not give good recovery, which shows that smoothers in* (b)–(d) *must be extracting useful information from the noisy s_2 dataset.*

matrix in this case is

$$H_k x_k = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix} x_k \,, \qquad R_k = \begin{bmatrix} \sigma^2 & \\ & \sigma^2 \end{bmatrix} \,.$$

Since, in the ckbs interface, the user specifies R_k^{-1} rather than R_k, missing measurements are easily specified by setting the corresponding component of R_k^{-1} to 0.

For the contaminated sensor s_2, we consider $p = .2$ contamination level, and $\phi = 200$, a very large contaminating variance. We have s_2 measurements at every time step, but s_1 measurements only at every 10th time step. It is important to note that there is not enough information in s_2 measurements alone to recover the state; a smoother result using only s_2 measurements is shown in panel (e) of Figure 7. The challenge here is to supplement the rare reliable observations with information extracted from frequent but highly contaminated observations.

Measurements are plotted using diamonds, with s_2 measurements represented by small symbols, while s_1 measurements are represented by large symbols. Ground truth is shown using a sold black line, and smoother results are shown using a red dashed line. Results in panel (a) were obtained using the least squares smoother, which cannot handle outliers.

Results in panel (b) were obtained by the all-T smoother, which modeled all measurement and process residuals using Student's t. The all-T smoother does a great job, except for a problem that occurs around 12 seconds into the track. A clump of outliers fools the all-T smoother, which treats them as a true change in state. This result is pictorial proof of the claim made earlier—that one cannot perfectly distinguish between outliers and sudden changes; extra information is needed to make the call.

Results in panel (c) were obtained using a differential T-robust smoother, which used Student's t modeling for the contaminated s_2 measurement component and Gaussian for the process model as well as the s_1 measurement component. The resulting fit is quite good, but the classic model for transition errors results in a slight loss in accuracy around the jump. Note that this smoother exploits the additional information that s_1 measurements are good. Finally, results in panel (d) were obtained using Student's t modeling for all process residuals as well as for the s_2 measurements, and using a Gaussian model for the reliable s_1 measurement component. This smoother ignores the outliers and is able to follow the jump very well; it does not get fooled by the small outlier cluster.

The file used to generate the subplots in the figure is `noisy_jump_two_meas.m`, which can be accessed through the `example` subdirectory of [1].

7. Discussion and conclusions. We have presented a generalized Student's t smoothing framework, which allows modeling any process or measurement residuals using Student's t errors and includes T-robust, T-trend, and double-T smoothers as important special cases. All of the smoothers in the framework efficiently solve for the MAP estimates of the states in a state-space model with any selected set of residuals modeled using Student's t or Gaussian noise. We have shown that these features can be used independently and in tandem, work for linear and nonlinear process models, and can be used both for outlier-robust smoothing and for tracking sudden changes in the state.

Similar to contributions in other applications, e.g., sparse system identification [12, 34, 37], our results underscore the significant advantages of using *heavy tailed* distributions in statistical modeling. These distributions force the use of nonconvex

loss functions to solve for the associated MAP estimates [7, Theorem 2]. The resulting objective is nonconvex even when the system dynamics are linear, and an iterative smoother is required to solve it. The convergence analysis for these methods is still developed within the general framework of convex-composite optimization [10], although the details of the analysis differ.

Because the problems are nonconvex, iterative methods may converge to local rather than global minima. This problem can be mitigated by an appropriate initialization procedure; for example, in the presence of outliers, the ℓ_1-Laplace smoother can be used to obtain a starting point for the optimizer, in which case we can improve on the ℓ_1 solution when the data is highly contaminated with outliers. This approach was not taken in our numerical experiments, which used the same initial points. For all the linear experiments, the initial point was simply the null state sequence. For the VDP, the initial state x_0 was correctly specified in all experiments, and the remaining state estimates in the initial sequence were null.

The T-robust smoother compares favorably to the ℓ_1-Laplace smoother described in [2] and outperforms it in our experiments when the data is heavily contaminated by outliers. The T-trend smoother was designed for tracking signals that may exhibit sudden changes and therefore has many potential applications in areas such as navigation and financial trend tracking. It was demonstrated to follow a fast jump in the state better than a smoother with a convex penalty on model deviation. Finally, we demonstrated the power of a new method by tracking a fast change in the presence of outliers using the full flexibility of the presented framework, which allowed us to differentially model residuals for sensors which we knew to be reliable versus unreliable and to design a smoother that was robust to outliers yet able to track sudden changes.

We do not provide theoretical recovery guarantees for the proposed robust smoothers. Doing so would require making strong assumptions on the kinds of "unknown errors" one could observe, and we leave these developments to future work. Our empirical tests simulate outliers from contaminated Gaussian distributions that are quite unlike the model distributions used to model the smoother, and our empirical performance metric is the MSE between *ground truth signal* and the estimate.

Kalman filters and smoothers are known to be MSE optimal under the assumption that the dynamics are linear and the errors are Gaussian. Even though these mathematical assumptions rarely occur in practice, the Kalman filter has long been a valued tool in a number of applications. Our focus in this contribution is intended for cases where the linearity and Gaussian assumptions fail, and recovery through the classical Kalman filters and smoothers also fails. Nonetheless, the reader may be concerned that there may be a significant loss in the performance of robust smoothers in those cases where the dynamics are linear and the errors are indeed Gaussian. However, this does not appear to be the case in our experiments since robust smoothers have nearly identical performance to the standard RTS smoother under nominal conditions in Table 1.

An important question in the design and implementation of Student's t–based smoothers is how to estimate the degree of freedom parameter ν. In all of our experiments, we have treated this parameter as fixed and known. We note that there are established expectation maximization (EM)-based methods in the literature for estimating these parameters [15, 24], as well as other recently proposed methods [8], and we leave the implementation of these extensions in the Kalman smoothing framework to future work.

Acknowledgments. The authors would like to thank Bradley Bell and North Pacific Acoustic Laboratory (NPAL) investigators of the Applied Physics Laboratory, University of Washington, for the underwater tracking data used in this paper (NPAL is sponsored by the Office of Naval Research code 321OA). We are also grateful to Michael Gelbart for insightful discussions about the numerical experiments.

REFERENCES

[1] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, *CKBS: Matlab/Octave package for constrained and robust Kalman smoothing*, http://www.coin-or.org/CoinBazaar/ckbs/ckbs.xml, 2007–2013.

[2] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, *An ℓ_1-Laplace robust Kalman smoother*, IEEE Trans. Automat. Control, 56 (2011), pp. 2898–2911.

[3] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, *Learning using state space kernel machines*, in Proceedings of the IFAC World Congress 2011, Milan, Italy, 2011.

[4] A. Y. Aravkin, B. M. Bell, J. V. Burke, and G. Pillonetto, *New stability results and algorithms for block tridiagonal systems, with applications to Kalman smoothing*, http://arxiv.org/abs/1303.5237, 2013.

[5] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, *Robust and trend-following Kalman smoothers using Student's t*, preprint, http://arxiv.org/abs/1001.3907v3, 2011.

[6] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, *Robust and trend following Kalman smoothers using Student's t*, in Proceedings of SYSID, 2012.

[7] A. Y. Aravkin, M. P. Friedlander, F. Herrmann, and T. van Leeuwen, *Robust inversion, dimensionality reduction, and randomized sampling*, Math. Program., 134 (2012), pp. 101–125.

[8] A. Y. Aravkin and T. van Leeuwen, *Estimating nuisance parameters in inverse problems*, Inverse Problems, 28 (2012), 115016.

[9] B. M. Bell, J. V. Burke, and G. Pillonetto, *An inequality constrained nonlinear Kalman-Bucy smoother by interior point likelihood maximization*, Automatica J. IFAC, 45 (2009), pp. 25–33.

[10] J. V. Burke, *Descent methods for composite nondifferentiable optimization problems*, Math. Program., 33 (1985), pp. 260–279.

[11] J. V. Burke, *Second order necessary and sufficient conditions for convex composite NDO*, Math. Program., 38 (1987), pp. 287–302.

[12] A. Chiuso and G. Pillonetto, *Learning sparse dynamic linear systems using stable spline kernels and exponential hyperpriors*, in Advances in Neural Information Processing Systems (NIPS), available online from http://papers.nips.cc/book/advances-in-neural-information-processing-systems-23-2010, 2010.

[13] C. Chui and G. Chen, *Kalman Filtering*, Springer, New York, 2009.

[14] L. Fahrmeir and H. Kaufmann, *On Kalman filtering, posterior mode estimation, and Fisher scoring in dynamic exponential family regression*, Metrika, 38 (1991), pp. 37–60.

[15] L. Fahrmeir and R. Kunstler, *Penalized likelihood smoothing in robust state space models*, Metrika, 49 (1998), pp. 173–191.

[16] S. Farahmand, G. B. Giannakis, and D. Angelosante, *Doubly robust smoothing of dynamical processes via outlier sparsity constraints*, IEEE Trans. Signal Process., 59 (2011), pp. 4529–4543.

[17] J. Fernándes, J. L. Speyer, and M. Idan, *A stochastic controller for vector linear systems with additive Cauchy noise*, in Proceedings of the 52nd IEEE Conference on Decision and Control, Florence, Italy, 2013, pp. 1872–1879.

[18] A. Gelb, *Applied Optimal Estimation*, MIT Press, Cambridge, MA, 1974.

[19] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley Ser. Probab. Math. Statist., John Wiley and Sons, New York, 1986.

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

[21] G. A. Hewer, R. D. Martin, and J. Zeh, *Robust preprocessing for Kalman filtering of glint noise*, IEEE Trans. Aerospace Electron. Syst., 23 (1987), pp. 120–128.

[22] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. AMSE J. Basic Engrg., 82 (1960), pp. 35–45.

[23] S. A. Kassam and H. V. Poor, *Robust techniques for signal processing: A survey*, Proc. IEEE,

73 (1985), pp. 433–481.
[24] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, *Robust statistical modeling using the t distribution*, J. Amer. Statist. Assoc., 84 (1989), pp. 881–896.
[25] R. A. Maronna, D. Martin, and V. J. Yohai, *Robust Statistics*, Wiley Ser. Probab. Math. Statist., John Wiley and Sons, New York, 2006.
[26] H. Ohlsson, F. Gustafsson, L. Ljung, and S. Boyd, *Smoothed state estimates under abrupt changes using sum-of-norms regularization*, Automatica J. IFAC, 48 (2012), pp. 595–605.
[27] I. R. Petersen and A. V. Savkin, *Robust Kalman Filtering for Signals and Systems with Large Uncertainties*, Control Engineering, Birkhäuser, Basel, 1999.
[28] S. T. Rachev, ed., *Handbook of Heavy Tailed Distributions in Finance*, Elsevier Science, New York, 2003.
[29] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[30] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis, Vol.* 317, Springer, New York, 1998.
[31] I. C. Schick and S. K. Mitter, *Robust recursive estimation in the presence of heavy-tailed observation noise*, Ann. Statist., 22 (1994), pp. 1045–1080.
[32] J. C. Spall, *Estimation via Markov chain Monte Carlo*, IEEE Control Syst. Mag., 23 (2003), pp. 34–45.
[33] R. Tibshirani, *Regression shrinkage and selection via the LASSO*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[34] M. Tipping, *Sparse Bayesian learning and the relevance vector machine*, J. Mach. Learning Res., 1 (2001), pp. 211–244.
[35] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
[36] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*, 2nd ed., Springer, New York, 1999.
[37] D. P. Wipf and B. D. Rao, *An empirical Bayesian strategy for solving the simultaneous sparse approximation problem*, IEEE Trans. Signal Process., 55 (2007), pp. 3704–3716.
[38] S. J. Wright, *Solution of discrete-time optimal control problems on parallel computers*, Parallel Comput., 16 (1990), pp. 221–238.