Kalman Filtering with Adversarial Corruptions*

Sitan Chen
EECS, UC Berkeley
Berkeley, CA, USA
sitanc@berkeley.edu

Frederic Koehler
CS, Stanford University
Stanford, CA, USA
fkoehler@stanford.edu

Ankur Moitra
CSAIL & Math, MIT
Cambridge, MA, USA
moitra@mit.edu

Morris Yau
EECS, MIT
Cambridge, MA, USA
morrisy@mit.edu

ABSTRACT

Here we revisit the classic problem of linear quadratic estimation, i.e. estimating the trajectory of a linear dynamical system from noisy measurements. The celebrated Kalman filter gives an optimal estimator when the measurement noise is Gaussian, but is widely known to break down when one deviates from this assumption, e.g. when the noise is heavy-tailed. Many ad hoc heuristics have been employed in practice for dealing with outliers. In a pioneering work, Schick and Mitter gave provable guarantees when the measurement noise is a known infinitesimal perturbation of a Gaussian and raised the important question of whether one can get similar guarantees for large and unknown perturbations.

In this work we give a truly robust filter: we give the first strong provable guarantees for linear quadratic estimation when even a *constant* fraction of measurements have been adversarially corrupted. This framework can model heavy-tailed and even non-stationary noise processes. Our algorithm robustifies the Kalman filter in the sense that it competes with the optimal algorithm that knows the locations of the corruptions. Our work is in a challenging Bayesian setting where the number of measurements scales with the complexity of what we need to estimate. Moreover, in linear dynamical systems past information decays over time. We develop a suite of new techniques to robustly extract information across different time steps and over varying time scales.

CCS CONCEPTS

• **Mathematics of computing** → **Kalman filters and hidden Markov models**; Bayesian computation; • **Theory of computation** → Online learning theory.

KEYWORDS

Kalman filter, robust statistics, linear quadratic estimation, sum of squares, semidefinite programming, Bayesian statistics, time series

ACM Reference Format:
Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. 2022. Kalman Filtering with Adversarial Corruptions. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC '22), June 20–24, 2022, Rome, Italy.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3519935.3520050

*This work was supported in part by NSF Award 2103300, NSF CAREER Award CCF-1453261, NSF Large CCF-1565235, A. Moitra's ONR Young Investigator Award, E. Mossel's Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826, and a David and Lucile Packard Fellowship. Part of this work was completed while SC and FK were visiting the Simons Institute for the Theory of Computing.

STOC '22, June 20–24, 2022, Rome, Italy

ACM ISBN 978-1-4503-9264-8/22/06.
https://doi.org/10.1145/3519935.3520050

1 INTRODUCTION

In this work, we revisit the classic problem of linear quadratic estimation, i.e. estimating the trajectory of a linear dynamical system from noisy measurements. First we review the setup:

(1) There are known matrices $A \in \mathbb{R}^{d\times d}$ and $B \in \mathbb{R}^{m\times d}$, and an unknown initial state $x_0^* \in \mathbb{R}^d$ drawn from $\mathcal{N}(0, R^2\,\mathrm{Id})$.
(2) The trajectory $\{x_t^*\}_{0\le t<T}$ and the observations $\{y_t^*\}_{0\le t<T}$ are generated according to the following model:

$$x_i^* = Ax_{i-1}^* + w_i^*, \qquad y_i^* = Bx_i^* + v_i^*$$

where $\{w_i^*\}$ and $\{v_i^*\}$ are *process noise* and *observation noise* vectors in dimensions d and m, independently drawn from $\mathcal{N}(0, \sigma^2\,\mathrm{Id})$ and $\mathcal{N}(0, \tau^2\,\mathrm{Id})$ respectively.

The goal is to estimate the trajectory from the observations in either an offline or online sense, and to minimize the sum of squares of the error. The celebrated Kalman smoother and Kalman filter solve these two problems optimally. The main idea is that when the initialization and noise distributions are Gaussian, at any time the posterior distribution on the trajectory given the observations is a Gaussian process. It suffices to estimate the mean of the posterior, and this can be done by finding the least squares solution to a structured regression problem depending on A, B, and the observations. It turns out that there is an even more compact formulation in terms of equations that define the Kalman filter.

The Kalman filter [26, 27] is one of the crowning achievements in control theory. It has wide-ranging applications in robotics, navigation, signal processing and econometrics. It is also a key building block in algorithms for estimating the model parameters of a linear dynamical system, as well as in change-point detection and building optimal controllers. Famously, the Kalman filter provided navigation estimates that helped guide the landing of the Apollo 11 lunar module in the Sea of Tranquility. Rudolf Kalman was awarded the National Medal of Science in 2008.

But how brittle is the Kalman filter to assumptions of Gaussianity? This is by no means a new question. If we relax the distributional assumptions but instead restrict the disturbances $\{w_i^*\}$ and $\{v_i^*\}$ to have bounded norm, then the minimax optimal filter can be found by dynamic programming. The solution is called the H_∞ filter and has wide-ranging applications in its own right [25]. However in many settings the assumption that the disturbances are bounded in norm is not reasonable either. In such cases, it has often been repeated that the Kalman filter can fail catastrophically. This is an unfortunate state-of-affairs because it means even though we can find the optimal filter when the noise is nice and Gaussian, it can break down badly with even a single badly outlying observation.

There are many natural heuristics that have been employed for dealing with outliers. However we are aware of only one work that gives rigorous guarantees in the presence of outliers. In 1994, Schick and Mitter [47] studied a model where the observation noise is drawn from a distribution $(1-\eta)\mathcal{N}(0,\sigma^2\mathrm{Id}) + \eta\mathcal{H}$ where $\mathcal{H}$ is a sufficiently regular distribution, but is allowed to be heavy-tailed. This is essentially the Huber contamination model. They derived provable guarantees but under a number of strong assumptions: First, they assumed that the distribution $\mathcal{H}$ is known to the filter designer. Thus the filter can use information about $\mathcal{H}$ to correct for the exact ways in which the noise is non-Gaussian. Second, their guarantees are asymptotic in nature in the sense that they only hold in the limit as $\eta \to 0$. As Schick and Mitter [47] discussed, as $T \to \infty$ for fixed $\eta > 0$, the number of outliers (i.e. timesteps where the observation noise is sampled from $\mathcal{H}$) goes to infinity. However their analysis relies on the exponential asymptotic stability of the Kalman filter, whereby outside of a window around the current timestep, the influence of older observations is significantly attenuated. Thus as $\eta \to 0$ the number of outliers in any window can be bounded, even if the total number of outliers cannot. In fact, while their estimator is nonlinear, as is necessary for handling heavy-tailed noise, it is constructed through a bank of Kalman filters. Each filter ignores one of the observations, assuming that it is the lone outlier. The filters are then combined in a natural way.

In this work we seek a truly robust filter. We want to build a filter without knowledge of $\mathcal{H}$. Moreover we want the outliers to not merely be sampled from a heavy-tailed distribution, but allow for their values to be chosen, possibly adaptively, by an unbounded adversary. For instance, this captures situations where the process generating the outliers is non-stationary. Moreover we want to prove guarantees that hold for fixed noise rates, as opposed to guarantees in the $\eta \to 0$ limit. Finally, we will want our filter to be a robustification of the Kalman filter itself in the sense that when $\eta = 0$ we want to achieve the same exact guarantees as before. Thus our filter competes with the Kalman filter in a strong sense, but gracefully degrades in performance as we move away from the precise distributional assumptions underlying the classic theory.

1.1 Our Results

Model and objective. Let $0 \le \eta < 1/2$ be the *corruption fraction*. We will assume that for every timestep i, with probability η the observation y_i^* falls under the control of an all-powerful adversary. The adversary can replace all of the observations under his control with arbitrary values. Now let $\{y_t\}_{0\le t<T}$ denote the observation sequence that the learner ultimately receives. We stress that exactly which of these have been corrupted is unknown to the learner.

Note that the corrupted timesteps are *randomly chosen*, just like in the Huber contamination model. As a special case this captures the setting studied by Schick and Mitter. Moreover, because the adversary gets to coordinate his corruptions, our model also allows dependencies and captures situations where the observation noise is non-stationary over time. As we show in the full version of the paper, in the stronger corruption model where the adversary gets to choose which timesteps to corrupt, there are strong impossibility results. Thus our corruption model seems to be one of the strongest where we can still hope for meaningful guarantees.

Next we describe the objective. As discussed above, the Kalman filter can be thought of as computing the mean of the posterior distribution on the trajectory given the observations so far. When the noise is non-Gaussian, the posterior no longer need be a Gaussian process; it could be much more complex. *So how can we even define an optimization problem that generalizes that solved by Kalman filtering, if the posterior is non-Gaussian and, even worse, depends on $\mathcal{H}$ which is unknown to the filter and possibly changing over time?*

Our main idea is to compete with a strong oracle that knows which measurements are corrupted. Let $a_i^* \in \{0,1\}$ denote the indicator variable for whether round i is clean in the sense that its measurement error came from a Gaussian, rather than coming from $\mathcal{H}$ or being chosen by an adversary. When the a_i^*'s are known to the filter, the optimal estimator in a Bayesian sense is to estimate the mean of the posterior using only information from the uncorrupted observations. This leads us to the following objective:

$$L(\widehat{x}) = \frac{1}{T}\left(\sum_i (a_i^*\|B\widehat{x}_i - y_i\|^2/\tau^2 + \|\widehat{w}_i\|^2/\sigma^2) + \|\widehat{x}_0\|^2/R^2\right) \quad (1)$$

where the steps $\widehat{w}_i$ are defined in terms of the trajectory $\widehat{x}_i$, i.e. $\widehat{x}_i = A\widehat{x}_{i-1} + \widehat{w}_i$ for all $i > 0$. This is the *clean posterior negative log likelihood*. The best possible error we can achieve is given by $\mathrm{OPT} = \min_{\{\widehat{x}_i\}} L(\widehat{x})$ which is attained by the maximum a posteriori (MAP) estimator. This is the same as the posterior mean and can be explicitly computed from a Kalman smoother that knows the observations from clean rounds, i.e. y_i for the rounds i for which $a_i^* = 1$. For general $\{\widehat{x}_i\}$, we refer to $L(\widehat{x}) - \mathrm{OPT}$ as the *excess risk*.

We take a moment to explain the differences in our approach compared to the usual approaches in algorithmic robust statistics, for example in robust mean estimation [15, 23, 31]. Usually in a robust statistics problem, the objective of the estimator is to recover the *ground truth*, e.g. the true mean in the example of mean estimation, and the goal in the robust setting is to recover the ground truth perfectly as $\eta \to 0$ while optimizing the dependence on the corruption fraction η. This makes sense for mean estimation because when $\eta = 0$, the mean can always be estimated consistently by taking more samples. It does not make as much sense in the case of Kalman filtering: even without corruptions, we only ever get one observation per timestep, so we cannot hope to recover the ground truth trajectory arbitrarily well. Uncertainty about the true trajectory is unavoidable because the complexity of the trajectory grows with the number of observations we get to make, and also because information about the past is washed away over time. For this reason, we need to pick our measure of success carefully. From a Bayesian perspective, the clean posterior mean represents the *best possible* estimate we can make of the ground truth given the clean observations *given additional information about which rounds have been corrupted*. Thus it gives a natural way to quantify the distance of our estimator from optimality.

Main results. We show how to design an estimator which is robust to corruptions and competes with the optimum in the clean posterior log likelihood. When $\eta = 0$, our estimator gets asymptotically optimal posterior log likelihood, including the correct constant factor; its guarantee thus matches the posterior mean/Kalman smoother. In fact, with high probability our estimator will be exactly the same as the Kalman smoother (see Remark 4.2).

Theorem 1.1. *For $\eta \le 0.49$, and for a uniformly stable and completely observable dynamical system,*[1] *there is a polynomial-time algorithm which takes as input the corrupted observations $\{y_i\}$ and outputs trajectory $\{\widehat{x}_i\}$ with excess risk*

$$L(\widehat{x}) - \mathrm{OPT} \le C_d\,\eta\log(1/\eta)\left(m + d(\sigma^2/\tau^2)\log\log T\right) + o(1), \quad (2)$$

with high probability, where the steps $\widehat{w}_i$ are defined by $\widehat{x}_i = A\widehat{x}_{i-1} + \widehat{w}_i$, where C_d is a constant which is polynomially bounded in $\log d$ and the parameters of the system (see Section 3.1).

To compare, we remind the reader of the performance of the obvious baselines: the Kalman filter can have unbounded error if there is even a single corruption, and oblivious outlier removal makes error $\Theta(T)$ (see full version of paper for this and other lower bounds). Recall that σ^2 is the variance of the process noise and τ^2 is the variance of the observation noise. The dependence on the ratio σ^2/τ^2 is unavoidable, and the dependence on the dimensions d and m in (2) of Theorem 1.1 is also unavoidable.

Notably, we can obtain strong provable guarantees for any $\eta < 1/2$ (we wrote $\eta < 0.49$ above only to simplify the statement). Thus our estimator has an information-theoretically optimal breakdown point. Also, our result can handle the case where the eigenvalues of A are on or near the unit circle, e.g. $A = \mathrm{Id}$, a situation where the system is marginally stable but not strictly stable (see e.g. [49] for discussion of this terminology). This is an important distinction, because when the eigenvalues of A are all small, a relatively simple method based on truncating the Kalman filter can work (see full version of paper), but the performance of such a heuristic will degrade badly as the eigenvalues approach the unit circle, whereas the algorithm of Theorem 1.1 will still work (see Section 2 for more discussion)

The above result works in an offline setting. But what happens when our measurements come in an online fashion and we need to estimate the position at time i from only the observations up to that point? In this sense we want the filter to be *causal*. Fortunately, at a small loss in our overall guarantees, we are able to make our approach online too. We need to change the definition of the objective slightly to handle the fact that predictions are made online: for the online case, we look at the suboptimality in predicting the next state compared to the oracle Kalman filter.

Theorem 1.2. *There is a polynomial time and causal estimator $\hat{\hat{x}}_{i+1|i}$ satisfying the following guarantee with probability at least $1-\delta$. With $\hat{x}_{i+1|x}$ denoting the output of the oracle Kalman filter, the suboptimality is bounded by $\frac{r''}{1-\delta}(\eta + O(\sqrt{\eta\log(2/\delta)/T} + \log(2/\delta)/T))$ where $r'' = C_d\,\eta\left(m + d(\sigma^2/\tau^2)\log T\right) + o(1)$ and C_d is a constant polylogarithmic in d, m and depending on other system parameters.*

Our analysis combines our convex programming approach with standard properties of the Kalman filter, such as its exponential stability, building on an idea of Schick and Mitter [47].

1.2 Further Related Work

Robust Statistics and Sum-of-Squares. Our main algorithm for robust filtering is based on the Sum of Squares hierarchy [42], which has broad applications to both control theory (see e.g. [20, 44]) and to algorithmic robust statistics (see e.g. [4, 11, 14, 22, 29, 30]). It also builds upon a line of recent work in algorithmic robust statistics using both SoS and non-SoS methods (see e.g. [9, 12, 13, 15, 17, 31]). One of the techniques we use, introducing a positive semi-definiteness constraint utilizing matrix concentration bounds (see Technical Overview), is conceptually related to the main technical ingredient in [10] for robust linear regression. The recent work [6] also considers robust statistics in a Bayesian setting, namely community detection in the stochastic block model: Their method recovers the optimal detection threshold as the corruption level goes to zero, but not the optimal performance of the Bayes estimator (they only establish results for detection and not recovery).

Practical Approaches. Algorithms based on minimizing losses that are less sensitive to outliers, e.g. the Huber loss, have been widely applied in practical works on robust filtering (see e.g. [28, 34]) but lack strong theoretical guarantees. This is because, unlike in other contexts like robust/heavy-tailed mean estimation and regression [12, 23, 24, 35], the observations come from a non-stationary generative model and the length of the trajectory and number of observations are linked, so the existing proof techniques do not apply. There are many similar ad hoc methods which have been used in practice to handle outliers, especially by downweighting the Kalman filter updates when the innovation exceeds some threshold.

Other Notions of Robustness for Kalman Filtering. A popular variant of the standard Kalman filtering setup is to allow the variables y_i to be dropped independently with some probability p (analogous to our η); see e.g. [36, 37, 51]. This is like having an adversarial corruption model where the location of corruptions are known (and hence can be ignored); in this version of the problem, the Kalman filter remains optimal and the focus has been on understanding aspects of its behavior, e.g. in unstable systems.

We mention some other works in the direction of making the Kalman filter more robust. Some take the approach of assuming a particular parametric model for the noise that is non-Gaussian, such as a Student's t-distribution [45] or Lévy distribution [52], with the corresponding caveat that their results are limited to the model they assume. A major line of work in the control theory and filtering literature deals with robustness to uncertainty in the parameters of the system (e.g. A and B), which is quite different from the problem of handling outliers. For example [43], studies a version of this question in the context of the Wiener filter: the goal there is to choose the best linear filter given the uncertainty set, and so it cannot handle outlying observations where a nonlinear filter is required. Similarly, in recent works such as [50, 54] the authors study online methods for control theory problems where the goal is to compete with the best policy in a certain class (e.g. compete with the Kalman filter when the system is unknown, or compete with the best of the H_∞ controller and the H_2 controller without knowing the disturbance model), but in our setting the only filters which perform well are nonlinear and so competing with a class of linear filters like the Kalman filter is not sufficient.

[1] Uniform stability and complete observability are standard assumptions from the control theory literature, which we introduce in Section 2.1.

2 TECHNICAL OVERVIEW

We first review basic concepts from control theory like observability and stability that will play an important role in the later discussion. We then give an overview of the main challenges of handling adversarial corruptions when estimating a linear dynamical system and explain the techniques we develop to obtain our main results.

2.1 Control Theory Basics

Observability. Before we describe how to estimate a linear trajectory from corrupted observations, we first review how to do so in the absence of corruptions. Note that without additional assumptions on A and B, this may not be possible *a priori*. For instance, A might only act nontrivially on some subspace of $\mathbb{R}^d$, and the rows of B might simply be completely orthogonal to this subspace, in which case we can't hope to recover the trajectory.

In control theory, the standard way to ensure that the linear dynamical system at hand is not degenerate in this fashion is to assume it is *observable*, as originally defined by Kalman in [26]. Formally, given a parameter $s \in \mathbb{N}$, define the *observability matrix*

$$O_s \triangleq \sum_{i=0}^{s-1} (A^i)^\top B^\top B A^i.$$

To motivate this object, suppose momentarily that there were no observation or process noise, so that the trajectory is given by $x_i^* = Ax_{i-1}^*$ and the observations are given by $y_i^* = Bx_i^*$. Then our observations up to the s-th timestep are given by $y_i^* = BA^i x_0^*$ for $0 \le i < s$. Now note that $x_0^{*\top} O_s x_0^* = \sum_{i=0}^{s-1} \|BA^i x_0^*\|^2 = \sum_{i=0}^{s} \|y_i^*\|^2$. In particular, if O_s had nonzero kernel and x_0^* were an element of this kernel, then all of the observations up to time s would be zero, and we would get no information about $\{x_0^*, \dots, x_{s-1}^*\}$. Conversely, if O_s were full rank, then one can recover $\{x_0^*, \dots, x_{s-1}^*\}$ given $\{y_0^*, \dots, y_{s-1}^*\}$ by solving the appropriate linear system. In other words, non-degeneracy of the observability matrix O_s is a *necessary and sufficient condition* for being able to recover the trajectory up to time s from observations regardless of where the trajectory started.

More generally when there is observation and process noise, the natural quantitative analogue of non-degeneracy of O_s is an upper bound on its condition number (see e.g. [18, 38]):

Assumption 1 (Complete observability– informal, see Assumption 3)**.** *For some $s \in \mathbb{N}$, $O_s \triangleq \sum_{i=0}^{s-1} (A^i)^\top B^\top B A^i$ is well-conditioned.*

Stability. We will focus on models which satisfy the following weak stability assumption, often made in the control theory literature (including the work of Schick and Mitter [47]):

Assumption 2 (Uniform stability– informal, see Assumption 4)**.** *There is a constant $\rho \ge 1$ such that for any $i \in \mathbb{N}$, $\|A^i\| \le \rho$ (here $\|\cdot\|$ denotes the operator norm).*

Intuitively, uniform stability ensures that the system initialized at any point will not eventually blow up at some time in the future. In contrast, if A has an eigenvalue larger than one, the system is called *explosive* or *unstable* and the state will blow up at an exponential rate. Although Kalman filtering has also been studied in the case where A is unstable, we know from the work on *intermittent Kalman filtering* (see Section 1.2) that even the oracle Kalman filter, which knows the location of the corruptions, will diverge if the corruption level in the unstable case is above some critical value [36]. Since the setting we consider is strictly more difficult, there is no hope of closely tracking the trajectory in our setting.

2.2 Our Techniques

Corruptions Degrade Observability. The first complication that arises in our setting is that corruptions can degrade the observability of the system. To see this, again consider the setup where there is no process or observation noise, but now some of the observations have been corrupted. We're essentially given a linear system $\{y_i = BA^i x_0^*\}_{0<i<s}$ where some unknown subset $S_{\text{bad}} \subseteq \{0, \dots, s-1\}$ of equations have been altered adversarially. If we knew S_{bad}, then we could remove the corresponding equations and try solving for x_0^* with the rest. Then the matrix that we need to be non-degenerate is no longer O_s but rather

$$O_s' \triangleq \sum_{i \notin S_{\text{bad}}} (A^i)^\top B^\top B A^i$$

Of course, O_s being non-degenerate does not guarantee O_s' is as well. One might wonder then whether Assumption 1 must be significantly strengthened to ensure that the trajectory can be recovered from corrupted observations. This is indeed the case if the corruptions arrived at arbitrary timesteps. But if the corruptions arrive in a random fashion, we will demonstrate that no additional assumptions need to be made.

Corruptions Subsample the Observability Matrix. To get a sense for how this could be possible, note that if S_{bad} comprises a random η fraction of the indices up to time s, then the expectation of O_s' is exactly $(1-\eta) \cdot O_s$. If we could argue that O_s' also concentrates around its expectation, then because O_s' is spectrally close to $(1-\eta) \cdot O_s$ and O_s is non-degenerate/well-conditioned by assumption, the corruptions don't actually impact the observability of the dynamical system in the presence of Huber contamination.

As the summands $(A^i)^\top B^\top B A^i$ are bounded in norm by Assumption 2, we can carry through this matrix concentration argument as long as s is sufficiently large. In Assumption 1 however, we make no assumptions about how large s is. Instead, we note that regardless of how large s in Assumption 1 is, by observing the system t steps at a time rather than s steps at a time for a large multiple t of s, we find that Assumption 1 also holds for t in place of s.

More precisely, if we consider the observability matrix $O_t \triangleq \sum_{i=0}^{t-1} (A^i)^\top B^\top B A^i$ for some moderately large t, then $O_t \succeq O_s$, and one can also easily check that $\|O_t\| \le (t\rho^2/s)\|O_s\|$. In other words, the condition number of O_t is at most $t\rho^2/s$ worse than that of O_s. So even if s from Assumption 1 is too small for O_s' to concentrate sufficiently around its expectation, it would appear that we can simply take t large enough that O_t' concentrates sufficiently around its expectation.

Key Challenge: Observable/Unobservable Subspaces. There is one essential wrinkle in the above argument: the fluctuations for matrix concentration for O_t' are of order $\sqrt{t}$, so at some point they may exceed the smallest singular value of O_t. In particular, it would appear that because of Huber contamination, we lose all control over how well we can estimate the component of the trajectory that lives in the span of the small singular vectors of O_t, and so no matter what value of t we choose, matrix concentration alone fails to show we can estimate the state successfully.

We now sketch a way to get around this problem. First note that in spite of the issue posed by the small singular vectors of O_t, the preceding discussion on matrix concentration does ensure that the projection of O'_t to the *large* singular vectors of O_t is sufficiently nondegenerate with high probability (see Lemma 4.9). One might therefore hope to be able to estimate the trajectory within this subspace. For this reason, we will refer to the span of the large singular vectors of O_t as the *observable subspace* and its orthogonal complement as the *unobservable subspace*; denote projectors to these subspaces by Π and $\Pi^\perp$ respectively. (The unobservable subspace should *not* be thought of as some kind of "invisible" subspace which doesn't affect the observations; based on the discussion above, it represents directions which are difficult to recover locally based on partially corrupted observations.)

So what do we do about the unobservable subspace? Here is the key idea: while we do lose control of the trajectory's component inside the unobservable subspace *within any fixed window of t steps*, we can consolidate information *across windows* to learn what goes on in the unobservable subspace. Before we can discuss how to implement this, we need to introduce our estimator.

An Inefficient Estimator. To motivate the design of our estimator, we will construct a system of constraints capturing salient features of the model. We begin by introducing vector-valued variables $\{x_0, \dots, x_{T-1}\}$ corresponding to our estimate for the trajectory, as well as variables $\{w_1, \dots, w_{T-1}\}$ and $\{v_0, \dots, v_{T-1}\}$ corresponding to our estimates for the process and observation noise. We also introduce Boolean variables $a_0, ..., a_{T-1} \in \{0, 1\}$, where a_i corresponds to our guess for whether the i-th observation was uncorrupted. Of course, the true values of the quantities that these variables represent are unknown to us, but there are a number of basic constraints they must satisfy. Firstly, because each observation is independently corrupted with probability η, we know by Chernoff that $\frac{1}{T}\sum_{i=0}^{T-1} a_i \geq (1-\eta) - o(1)$. Secondly, we know that the trajectory is given by a linear dynamical system, and in any uncorrupted timestep i, the observation y_i is a noisy linear measurement of the trajectory at that time, so $x_i = Ax_{i-1} + w_i$ and $a_i(y_i - Bx_{i-1} - v_i) = 0$. Additionally, we know the dynamics and observation noise is bounded with high probability, that is, $\|w_i\|^2 = O(\sigma^2 d), \quad \|v_i\| = O(\tau^2 m)$. Thus far the constraints have been fairly straightforward. We now describe a key constraint capturing the preceding discussion on matrix concentration. Recall that because the corruptions arrive in a random fashion, by matrix concentration the uncorrupted timesteps will "subsample" the observability matrix O_t in each window. In other words, in every window $\{\ell t, \dots, (\ell+1)t - 1\}$ of t timesteps, the following spectral lower bound holds with high probability

$$\textstyle\sum_{j=0}^{t-1}(1 - a_{\ell t+j})(A^j)^\top B^\top B A^j \preceq \eta \cdot O_t + O(\sqrt{t}) \cdot \mathrm{Id}$$

With these constraints in place, we consider the natural objective to optimize in light of (1). We want to minimize the "clean" negative log-likelihood achieved by $\{x_i\}$, where "clean" is defined with respect to the variables $\{a_i\}$ instead of the true indicators $\{a_i^*\}$:

$$\min \tfrac{1}{T}\left(\textstyle\sum_{i=1}^{T}(a_i\|Bx_i - y_i\|^2/\tau^2 + \|w_i\|^2/\sigma^2) + \|x_0\|^2/R^2\right)$$

Based on the discussion above, we might guess that $\|\Pi(x_i - x_i^*)\|^2$, the error from the observable subspace, can be bounded in the above program based on some kind of matrix concentration argument. It turns out this question itself is subtle, because we would need to rule out cancellations between the observable and unobservable subspace. However, even if we did argue that by itself, this is definitely not enough to make our ultimate objective (1), the clean negative posterior likelihood, small: for some of the clean observations (i with $a_i^* = 1$), they will be dependent on the information in the unobservable subspace, and the true state x_i^* can be very large in the unobservable subspace (see full version of paper for an illustrative example). The key difficulty to making the objective (1) small will be to argue that $\|x_i - x_i^*\|$, i.e. the error in the *whole* space, can be bounded.

Decomposing the Error. Roughly speaking, we will prove this by showing that the effect of errors in the past decays exponentially fast for our estimator. Intuitively, this argument will show that we do not make large errors in the unobservable subspace because the estimator will successfully propagate information from the past. Formally, we prove the following key inequality:

$$\|x_{i+t} - x_{i+t}^*\|^2 \leq \|x_i - x_i^*\|^2/2 + \text{noise}, \tag{3}$$

where here and throughout the rest of this overview, we use noise to denote a small quantity that is polynomially bounded in t and in the variance of the observation and process noise (see Lemma 4.10).

The proof of (3) starts by decomposing $x_{i+t} - x_{i+t}^*$ into

$$\left[(x_{i+t} - x_{i+t}^*) - A^t(x_i - x_i^*)\right] + A^t\Pi(x_i - x_i^*) + A^t\Pi_{V_\perp}(x_i - x_i^*) \tag{4}$$

The first term is the amount of new noise introduced between steps i and $i+t$: it will be small because we are only taking t steps, the process noise $\{w_i^*\}$ is small, and the corresponding program variables $\{w_i\}$ are also constrained to be small.

The remaining two terms in (4) account for the error propagated from the past: the second term $A^t\Pi(x_i - x_i^*)$ represents the error propagated from the observable error in the past, and the third term $A^t\Pi^\perp(x_i - x_i^*)$ represents the error propagated from the unobservable error in the past. We now show how both of those terms can be bounded, starting with the last of these terms (unobservable error).

Unobservable Error from the Past. We identify a simple but critical fact about observable linear dynamical systems: any vector in the unobservable subspace decreases in norm when it evolves forward by t timesteps. More formally, we show that for any $x \in \mathbb{R}^d$,

$$\|A^t\Pi^\perp x\|^2 \leq c \cdot \|\Pi^\perp x\|^2, \tag{5}$$

for small $c < 1$ provided that t is large enough relative to s (see Lemma 4.11).

In particular, by applying this to the vector given by the unobservable error in the past, we conclude that the error $\|A^t\Pi(x_i - x_i^*)\|^2$ from propagating the past unobservable error $\|\Pi(x_i - x_i^*)\|^2$ can be upper bounded by a small fraction of $\|\Pi(x_i - x_i^*)\|^2$.

Remark 2.1. We caution the reader that (5) does *not* mean the unobservable component of any vector x decays after x evolves over t timesteps, which would correspond to a bound of the form

$$\|\Pi_{V_\perp}A^t x\|^2 \leq C \cdot \|\Pi_{V_\perp} x\|^2. \tag{6}$$

for some $C < 1$. Of course, if (6) were true, it would make life much easier: by taking x to be any iterate in the trajectory, we would conclude that over time, the trajectory barely lives in the

unobservable subspace at all! Unfortunately, this is not the case, as it is easy to construct linear dynamical systems with a significant portion of the state in the unobservable subspace.

Observable Error from the Past. We now discuss how to handle the error $\|A^t\Pi(x_i - x_i^*)\|^2$ propagated from the observable error in the past. It is tempting to try the same approach as for the unobservable error here, namely to argue that $\|A^t\Pi(x_i - x_i^*)\|^2$ is a small fraction of the observable error in the past. But this is too much to hope for: together with (5) this would imply that $\|A^t\| < 1$, whereas we only assume that $\|A^t\| \le \rho$ for some $\rho \ge 1$ (as we show in the full version of the paper, it is quite easy to handle linear dynamical systems for which the former holds).

In the absence of an analogue of (5), our key insight is that we can instead relate $\|A^t\Pi(x_i - x_i^*)\|^2$ to the *unobservable* error in the past! Informally, any large errors in estimating the observable part of the state x_i^* must be explained by a large amount of interference from the unobservable part of the state. Essentially, in Lemma 4.13 we show that for some small constant $0 < c' < 1$,

$$\|\Pi(x_i - x_i^*)\|^2 \le c' \cdot \|\Pi^\perp(x_i - x_i^*)\|^2 + \text{noise}. \tag{7}$$

The proof is rather involved but can be distilled into two main ingredients. Firstly, as alluded to earlier, the fact that the random corruptions subsample the observability matrix lets us relate the error in estimating the state to the error in fitting the observations. We can then bound the latter using the following (see Lemma 4.12):

$$a_i^* a_i \|BA^i\Pi(x_0 - x_0^*)\|^2 \le 4a_i^* a_i \|BA^i\Pi^\perp(x_0 - x_0^*)\|^2 + \text{noise}. \tag{8}$$

Inequality (8) formalizes the following idea: because our estimator must match the observation at time i, any errors coming from the past are constrained to cancel between the observable and unobservable parts of the space, so if one is large the other is too.

Unfortunately, combining the two ingredients above naively would establish (7) with too large of a constant on the right hand side. It turns out however that we can greatly improve the constant by appealing to the aforementioned decay property of the unobservable subspace from (5). The details here are a bit subtle, and we refer to the proof of Lemma 4.10 for the formal argument.

Contraction of Error over Time. It is now easy to deduce (3) and conclude that the error incurred by the filter decays exponentially. Recalling that $\|(x_{i+t} - x_{i+t}^*) - A^t(x_i - x_i^*)\|^2$ will be some small noise term, we have by triangle inequality that

$$\|x_{i+t} - x_{i+t}^*\|^2 \le \text{noise} + 3\|A^t\Pi(x_i - x_i^*)\|^2 + 3\|A^t\Pi^\perp(x_i - x_i^*)\|^2$$

We can then use (5) and (7) to bound this by noise + $\|x_i - x_i^*\|^2/2$, proving (3) and establishing that over the course of t timesteps, the error of our estimate essentially decays geometrically. We remark that this doesn't give us any control over our error on the iterates *within* these timesteps, and in particular it remains to be shown how to use all of the tools we've introduced to compete with the Bayes-optimal predictor. It turns out however that by virtue of the program constraints ensuring that our estimate of the trajectory follows the linear dynamics and tries to fit the clean observations while minimizing the objective defined above, the excess risk introduced *within* any window of t is polynomially bounded by t and the variance of the noise (see Lemma 4.7). In this overview, we do not delve into the details of this as this particular argument is reminiscent of existing analyses in robust statistics.

Confidence Bands and Achieving $\log\log T$ *Excess Risk.* There is an important catch in the above discussion which we now address: We assumed that every noise vector u_i^* and w_i^* is bounded, and that subsampling holds in every window of size t. In order to ensure that these events hold across the entire trajectory of length T, we would need to take t to be logarithmic in T. This would translate into incurring a logarithmic overhead in the excess risk bound. *We are able to avoid paying the full price for this union bound.*

First, it turns out that our analysis only requires that the observation and process noise are bounded *on average* over the trajectory in an appropriate sense. The more serious issue is: How can we avoid assuming that subsampling holds in each window? This is a key component to being able to integrate information across different time windows. As a starting point, we observe that our initial estimator actually meets a stronger guarantee: It actually outputs a trajectory which is *pointwise* close to the true trajectory by a distance that scales polynomially in variance of the noise and polylogarithmically in T (see Corollary 4.14). In other words, it allows us to form a *confidence band* around the estimator of radius, and this logarithmic scaling is essentially optimal for any guarantee of this form.

We show how to exploit this confidence band. In particular, we engineer a second system of constraints which incorporates the output of our initial estimator and refines it to achieve our final $\log\log T$ excess risk bound in Theorem 1.1. The main idea is because the noise that accumulates over a window scales polynomially in the window size, we can consider windows over *shorter timescales* scaling doubly logarithmically rather than logarithmically in T. The goal is to achieve higher accuracy on shorter windows whenever subsampling holds. This might seem counterproductive: By shortening the windows, we are only making it more likely that subsampling will fail in any given window! Indeed, if we are now taking windows of doubly logarithmic length, then in roughly a $1/\text{polylog}(T)$ fraction of the windows, the random corruptions will fail to properly subsample the observability matrix. But this is where the confidence band comes in: Over these bad windows, we already know how to estimate those iterates pointwise to error $\text{polylog}(T)$, so the total contribution of the bad windows to the excess risk scales as $(1/\text{polylog}(T)) \cdot \text{polylog}(T) = O(1)$!

The key complication is that the algorithm designer doesn't actually know which windows subsampling failed in. Instead, we will set up a system of constraints similar to the one for our earlier estimator but with additional Boolean variables, one per window, corresponding to our guess for whether the random corruptions in that window successfully subsampled the observability matrix. We show how to integrate information *across the windows on which we correctly guessed that subsampling succeeded* to achieve our final guarantee. As the argument here is rather involved, we defer the details to the full version of the paper.

Efficient Algorithm via Sum-of-Squares. While the estimators we have described appear to be inefficient as they require solving certain systems of polynomial constraints, our proofs that the solutions to these systems satisfy the guarantees of Theorem 1.1 are simple in the sense that they can be implemented in the degree four sum-of-squares proof system [42, 48]. So instead of solving these polynomial systems which would *a priori* incur an exponential

runtime, it suffices to output a *pseudo-distribution* over solutions and round it to an integral solution in a straightforward way. These tools have become a mainstay in algorithm design in robust statistics more broadly [3–5, 16, 22, 30, 33]. We will explain the basics of the sum-of-squares proof system in Section 3.2.

Two-stage filter. Thus far we have been discussing the offline problem, where the estimator at some time instant is allowed to depend on the entire observation sequence $y_1, \dots, y_T$ – in other words, we designed a robust Kalman *smoother*. Our techniques can be used to solve the online filtering problem, where the prediction for x_t is only allowed to depend on $y_1, \dots, y_{t-1}$. It turns out that the transformation to online guarantees is simple: We use the offline smoother developed above, in particular the confidence band it outputs, as an outlier removal algorithm and combine it with the standard Kalman filter run on the observations $\tilde{y}_1, \dots, \tilde{y}_T$ which are not deleted by outlier removal. After the outlier removal, only small corruptions remain, so following an idea of Schick and Mitter [46, 47], we can then use stability estimates for the Kalman filter to establish accuracy guarantees on its prediction for the next state.

3 PRELIMINARIES

Given matrix M, let $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote its bottom and top singular values, and let $\kappa(M) \triangleq \frac{\lambda_{\min}(M)}{\lambda_{\max}(M)}$. We will sometimes also denote $\lambda_{\max}(M)$ by $\|M\|$.

3.1 Generative Model

For known matrices $A \in \mathbb{R}^{d\times d}$, $B \in \mathbb{R}^{m\times d}$, the underlying trajectory $\{x_i^*\}$ and uncorrupted observations $\{y_i^*\}$ are given by $x_0^* \sim \mathcal{N}(0, R^2 \cdot \mathrm{Id}_d)$ and

$$x_i^* = Ax_{i-1}^* + w_i^* \text{ for all } i > 0$$
$$y_i^* = Bx_i^* + v_i^* \text{ for all } i \geq 0$$

where the *dynamics noise* w_i^* is i.i.d. sampled from $\mathcal{N}(0, \sigma^2 \cdot \mathrm{Id}_d)$ and and the *observation noise* v_i^* is i.i.d. sampled from $\mathcal{N}(0, \tau^2 \cdot \mathrm{Id}_m)$, i.e. both types of noise in the system are isotropic Gaussian up to scaling. The assumption that the noise is isotropic simplifies notation greatly, and is largely without loss of generality in the following sense: if the noise covariance matrices are full rank, a change of basis will make the noise isotropic. Our results can also be extended to handle the case where the noise covariance is rank degenerate.

We now describe how the corrupted observations are formed. After the trajectory $\{x_i^*\}$ and observations $\{y_i^*\}$ have generated, an independent $\mathrm{Ber}(1-\eta)$ coin is flipped for every $0 \leq i < T$; let $a_i^* \in \{0, 1\}$ denote the outcome at time i. For all i for which $a_i^* = 1$, define $y_i = y_i^*$. For all i for which $a_i^* = 0$, a computationally unbounded adversary is allowed to set y_i arbitrarily. We can assume this adversary has full knowledge of the system, e.g. the full trajectory $\{x_0^*, \dots, x_{T_1}^*\}$, the full sequence of true observations $\{y_0^*, \dots, y_{T-1}^*\}$, etc.

The baseline estimation error which we will try to achieve approximately is

$$\mathrm{OPT} = \min_{\{\widehat{x}_i\}} \frac{1}{T}\left(\sum_{i=0}^{T-1}(a_i^*\|B\widehat{x}_i - y_i\|^2/\tau^2 + \|\widehat{w}_i\|^2/\sigma^2) + \|\widehat{x}_0\|^2/R^2\right) \tag{9}$$

where the minimum ranges over all trajectories $\widehat{x}_i$ with steps $\widehat{w}_i$ satisfying $\widehat{x}_i = A\widehat{x}_{i-1} + \widehat{w}_i$ for all $i > 0$. Note that the objective function here corresponds to the negative log-density of the (Gaussian) posterior of the trajectory given the clean observations, up to additive constants and multiplicative factors. In particular the minimum in (9) (i.e. the posterior MAP as well as the posterior mean, since the posterior is Gaussian) is achieved by iterates $\widehat{x}_i$ and steps $\widehat{w}_i$ given by running the offline Kalman filter (a.k.a. *Kalman smoother*) on the part of the trajectory indexed by i's for which $a_i^* = 1$. Since the algorithm does not know which times are uncorrupted (have $a_i^* = 1$), we cannot hope to exactly match the performance of OPT. However, we use it as a benchmark and bound the amount of excess error our algorithms make compared to this oracle.

We make the following assumptions which are standard in the filtering/control literature (see e.g. [1, 2, 47]):

Assumption 3 (Complete Observability)**.** *There exist constants $\alpha, \kappa > 0$ and $s \in \mathbb{N}$ for which*

$$O_s \triangleq \sum_{i=0}^{s-1}(A^i)^\top B^\top BA^i$$

satisfies $\lambda_{\min}(O_s) \geq \kappa s$ and $\lambda_{\max}(O_s) \leq \alpha s$.

Assumption 4 (Uniform stability)**.** *There is a constant $\rho \geq 1$ for which $\|A^j\| \leq \rho$ for all $j \in \mathbb{N}$.*

The following elementary manipulations will be useful:

Fact 3.1. *For $t \in \mathbb{N}$ divisible by s, $O_t = \sum_{r=0}^{t/s-1}(A^{rs})^\top O_s A^{rs}$. In particular,*

$$\|O_t\| \leq t \cdot \alpha \cdot \rho^2. \tag{10}$$

Fact 3.2. $x_t^* - A^t x_0^* = \sum_{j=1}^t A^{t-j} w_j^*$

3.2 Sum-of-Squares Basics

Here we give a quick review of the sum-of-squares algorithm; for a more thorough treatment, we refer the reader to [8, 21, 41].

Pseudoexpectations. The sum-of-squares SDP hierarchy is a series of increasingly tight SDP relaxations for solving polynomial systems $\mathcal{P} \triangleq \{p_i(x) \geq 0\}_{i=1}^N$. Although it is in general NP-hard to solve polynomial systems, the level-ℓ SoS SDP attempts to approximately solve $\mathcal{P}$ with increasing accuracy as ℓ increases by adding more constraints to the SDP. This improvement in approximation naturally comes at the expense of increasing runtime and space.

In particular, one can think of the SoS SDP as outputting a "distribution" μ over solutions to $\mathcal{P}$. However, there are two important caveats. Firstly, one can only access the degree-ℓ moments of the "distribution" and secondly there may be no true distribution with the corresponding degree ℓ moments. Thus we refer to μ as a *pseudodistribution*.

Definition 1. *A degree ℓ pseudoexpectation $\widetilde{\mathbb{E}} : \mathbb{R}[x]_{\leq \ell} \to \mathbb{R}$ satisfying $\mathcal{P}$ is a linear functional over polynomials of degree at most ℓ satisfying*

(1) *(Normalization)* $\widetilde{\mathbb{E}}[1] = 1$,
(2) *(Constraints of $\mathcal{P}$)* $\widetilde{\mathbb{E}}\left[p(x)a^2(x)\right] \geq 0$ *for all* $p \in \mathcal{P}$ *and polynomials* a *with* $\deg(a^2 \cdot p) \leq \ell$,
(3) *(Non-negativity on square polynomials)* $\widetilde{\mathbb{E}}\left[q(x)^2\right] \geq 0$ *whenever* $\deg(q^2) \leq \ell$.

For any fixed $\ell \in \mathbb{N}$, given a polynomial system, one can efficiently compute a degree ℓ pseudo-expectation in polynomial time.

Fact 3.3. *([32, 39, 42, 48]). For any n, $\ell \in \mathbb{Z}^+$, let $\widetilde{\mathbb{E}}_\zeta$ be degree ℓ pseudoexpectation satisfying a polynomial system $\mathcal{P}$. Then the following set has a $n^{O(\ell)}$-time weak separation oracle (in the sense of [19]):*

$$\left\{\widetilde{\mathbb{E}}_\zeta\left[(1, x_1, x_2, ..., x_n)^{\otimes \ell}\right] : \text{ degree } \ell \ \widetilde{\mathbb{E}}_\zeta \text{ satisfying } \mathcal{P}\right\}$$

Using this separation oracle, the ellipsoid algorithm finds a degree ℓ pseudoexpectation in time $n^{O(\ell)}$. We call this algorithm the degree ℓ sum-of-squares algorithm.

To reason about the properties of pseudo-expectations, we turn to the dual object of sum-of-squares proofs.

Sum-of-Squares Proofs. For any nonnegative polynomial $p(x) : \mathbb{R}^d \to \mathbb{R}$, one could hope to prove its nonnegativity by writing $p(x)$ as a sum of squares of polynomials $p(x) = \sum_{i=1}^m q_i(x)^2$ for a collection of polynomials $\{q_i(x)\}_{i=1}^m$. Unfortunately, there exist nonnegative polynomials with no sum-of-squares proof even for $d = 2$. Nevertheless, there is a generous class of nonnegative polynomials that admit a proof of positivity via a proof in the form of a sum of squares. The key insight of the sum-of-squares algorithm is that these sum-of-squares proofs of nonnegativity can be found efficiently provided the degree of the proof is not too large.

Definition 2. *(Sum-of-Squares Proof) Let $\mathcal{A}$ be a collection of polynomial inequalities $\{p_i(x) \geq 0\}_{i=1}^m$. A sum-of-squares proof that a polynomial $q(x) \geq 0$ for any x satisfying the inequalities in $\mathcal{A}$ takes on the form*

$$\left(1 + \sum_{k \in [m']} b_k^2(x)\right) \cdot q(x) = \sum_{j \in [m'']} s_j^2(x) + \sum_{i \in [m]} a_i^2(x) \cdot p_i(x)$$

where $\{s_j(x)\}_{j\in[m'']}, \{a_i(x)\}_{i\in[m]}, \{b_k(x)\}_{i\in[m']}$ are real polynomials. If such an expression were true, then $q(x) \geq 0$ for any x satisfying $\mathcal{A}$. We call these identities sum-of-squares proofs, and the degree of the proof is the largest degree of the involved polynomials $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$. Naturally, one can capture polynomial equalities in $\mathcal{A}$ with pairs of inequalities. We denote a degree ℓ sum-of-squares proof of the positivity of $q(x)$ from $\mathcal{A}$ as $\mathcal{A} \left|\frac{x}{\ell}\right. \{q(x) \geq 0\}$ where the superscript over the turnstile denote the formal variable over which the proof is conducted. This is often unambiguous and we drop the superscript unless otherwise specified.

A number of basic inequalities like Cauchy-Schwarz and Hölder's admit sum-of-squares proofs (see e.g. Appendix A of [22]).

Sum-of-squares proofs can also be strung together and composed according to the following convenient rules.

Fact 3.4. *For polynomial systems $\mathcal{A}$ and $\mathcal{B}$, if $\mathcal{A} \left|\frac{x}{d}\right. \{p(x) \geq 0\}$ and $\mathcal{B} \left|\frac{x}{d'}\right. \{q(x) \geq 0\}$ then $\mathcal{A} \cup \mathcal{B} \left|\frac{x}{\max(d,d')}\right. \{p(x) + q(x) \geq 0\}$. Also $\mathcal{A} \cup \mathcal{B} \left|\frac{x}{dd'}\right. \{p(x)q(x) \geq 0\}$*

Sum of squares proofs yield a framework to reason about the properties of pseudo-expectations, that are returned by the SoS SDP hierarchy.

Fact 3.5. *(Informal Soundness) If $\mathcal{A} \left|\frac{x}{r}\right. \{q(x) \geq 0\}$ and $\widetilde{\mathbb{E}}[\cdot]$ is a degree-ℓ pseudoexpectation operator for the polynomial system defined by $\mathcal{A}$, then $\widetilde{\mathbb{E}}[q(x)] \geq 0$.*

The following standard fact which follows by "SoS Cauchy-Schwarz" (see e.g. Lemma A.5 of [7]) will allow us to convert from pseudodistributions over solutions to a polynomial systems to integral solutions.

Lemma 3.6. *For any vector w^* and degree-2 pseudoexpectation $\widetilde{\mathbb{E}}[\cdot]$ over vector-valued variable w, we have that*

$$\|\widetilde{\mathbb{E}}[w] - w^*\|^2 \leq \widetilde{\mathbb{E}}[\|w - w^*\|^2]. \tag{11}$$

Proof. By the dual definition of L_2 norm, the left-hand side of (11) can be written as

$$\sup_{v \in \mathbb{S}^{d-1}} \langle \Sigma v, \widetilde{\mathbb{E}}[w] - w^* \rangle^2.$$

For any $v \in \mathbb{S}^{d-1}$, $\left\langle \Sigma v, \widetilde{\mathbb{E}}[w] - w^* \right\rangle^2 = \left(\widetilde{\mathbb{E}}[\langle \Sigma v, w - w^* \rangle]\right)^2 \leq \widetilde{\mathbb{E}}[\langle \Sigma v, w - w^* \rangle^2] \leq \widetilde{\mathbb{E}}[\|w - w^*\|_\Sigma^2]$, where the first inequality follows by the pseudoexpectation version of SoS Cauchy-Schwarz (see e.g. Lemma A.5 of [7]). Therefore, taking the maximum over all $v \in S^{d-1}$ proves the inequality. □

Finally, we will need the following elementary but crucial inequality which admits a degree-2 sum-of-squares proof. Roughly speaking, it captures the fact that if the sum of two vectors is small in norm, then either vector must have norm upper bounded in terms of the norm of the other vector:

Fact 3.7. *Let v_1, v_2 be d-dimensional vector-valued indeterminates. There is a degree-2 sum-of-squares proof of the inequality $\|v_1\|^2 \leq 4\|v_2\|^2 + \frac{4}{3}\varepsilon$ from the constraint $\|v_1 + v_2\|^2 \leq \varepsilon$.*

Proof. By expanding out the hypothesis, we have

$$\|v_1\|^2 + 2\langle v_1, v_2 \rangle + \|v_2\|^2 \leq \varepsilon.$$

By Cauchy-Schwarz, we also have

$$-2\langle v_1, v_2 \rangle \leq \frac{1}{4}\|v_1\|^2 + 4\|v_2\|^2.$$

Adding these two inequalities together and rearranging gives the desired inequality. □

3.3 Concentration Inequalities

Lemma 3.8 (Matrix Hoeffding, see e.g. Theorem 1.3 in [53])**.** *For any $\delta > 0$, given symmetric random matrices $M_1, \ldots, M_T \in \mathbb{R}^{d\times d}$ satisfying $\|M_t\| \leq 1$ almost surely for all t, if $M \triangleq \sum_t M_t$, then*

$$\mathbb{P}[\|M - \mathbb{E}[M]\| \geq \sqrt{8T \log(d/\delta)}] \leq \delta.$$

Lemma 3.9 (see e.g. [55]). *If $v \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \in \mathbb{R}^{d\times d}$, then with probability at least $1-\delta$,*

$$\|v\| \le O\left(\left(\sqrt{d} + \sqrt{\log(1/\delta)}\right)\|\Sigma\|^{1/2}\right)$$

We use concentration for Gaussian polynomials, which is a consequence of Gaussian hypercontractivity.

Lemma 3.10 (see e.g. [40]). *For degree-d polynomial $p : \mathbb{R}^m \to \mathbb{R}$, if $x \sim \mathcal{N}(0, \mathrm{Id})$, then*

$$\mathbb{P}\left[|p(x) - \mathbb{E}[p]| > t \cdot \sqrt{\mathbb{V}[p]}\right] \le \exp(-\Omega(t^{2/d})).$$

4 POLY-LOGARITHMIC EXCESS RISK AND CONFIDENCE BAND RECOVERY

In this section we show how to achieve excess risk scaling poly-logarithmically in the number of iterations. While this is worse than the final bound we will show (in the full version of the paper), it will introduce many of the important steps in the final analysis and also yield a warm start for our estimate of the trajectory which we will subsequently refine in the full version of the paper to get our final bound. Crucially, we show this algorithm can output a *confidence band* which with high probability (over the entire data generating process) contains the true trajectory.

The main result of this section is the following:

Theorem 4.1. *For any $\eta \le 0.49$, there is a polynomial-time algorithm that, given the corrupted observations $\{\widetilde{y}_i\}$, with probability $1-\delta$ over the randomness of the input, outputs a trajectory $\{\widehat{x}_i\}$ and steps $\{\widehat{w}_i\}$ for which $\widehat{x}_i = A\widehat{x}_{i-1} + \widehat{w}_i$ for every $i \in [T]$, and for which,*

$$\frac{1}{T}\left(\sum_{i=0}^{T-1}\left(a_i^*\|B\widehat{x}_i - y_i\|^2/\tau^2 + \|\widehat{w}_i\|^2/\sigma^2\right) + \|\widehat{x}_0\|^2/R^2\right) - \mathrm{OPT}$$
$$\lesssim \tau^{-2}\eta \cdot \left[E_{\mathrm{noise}} + \rho^2\left(\alpha + \|B\|^2\sqrt{\log(dT/t\delta)/t}\right)\cdot\right]$$
$$\left(\frac{\rho^6 E_{\mathrm{noise}} t}{\kappa} + \frac{R^2}{T/t}(d + \log(1/\delta))\right),$$

where

$$E_{\mathrm{noise}} \triangleq \tau^2\,(m + \log(T/\delta)) + t\rho^2\|B\|^2\sigma^2\,(d + \log(T/\delta)) \quad (12)$$
$$t \triangleq s \vee \widetilde{\Theta}(\kappa^{-2}\rho^{12}\|B\|^4\log(dT/\delta))). \quad (13)$$

4.1 Sum-of-Squares Relaxation

We now formulate the sum-of-squares program we work with in this section. We begin by introducing an important parameter, the so-called *window size* t. Recall from Assumption 3 that we assume that the observability matrix O_s is well-conditioned. We will take t to be a sufficiently large multiple of s such that, roughly speaking, the contribution to the observability matrix O_t from the uncorrupted time steps is also well-conditioned. We defer the tuning of t to later in the proof of Theorem 4.1. For convenience, given $0 \le i < T$, let $\ell(i) \triangleq \lfloor i/t \rfloor$ denote the index of the window to which iterate i belongs.

At this point, we can define our sum-of-squares relaxation:

Program 1. Let $\{y_i\}$ be the observations we are given, and let window size $t \in \mathbb{N}$ be a parameter to be tuned later. The program variables are d-dimensional vector-valued variables $\{x_i\}$ (trajectory estimates) and $\{w_i\}$ (process noise estimates), m-dimensional vector-valued variables $\{v_i\}$ (observation noise estimates), and Boolean variables $\{a_i\}$ (indicators for uncorrupted time steps), and the constraints are that for all $0 \le i < T$, *Boolean indicators for uncorrupted steps*

(1) $a_i^2 = a_i$

Trajectory estimate follow linear dynamics and fit y_i's on uncorrupted steps

(2) $x_i = Ax_{i-1} + w_i$
(3) $a_i(y_i - Bx_i - v_i) = 0$

Only η fraction of timesteps corrupted

(4) $\sum_{i=0}^{T-1} a_i \ge (1 - 1.01\eta)T$

Process and observation noise bounded

(5) $\|v_i\|^2 \le O(\tau^2(m + \log(T/\delta))$
(6) $\|w_i\|^2 \le O(\sigma^2(d + \log(T/\delta))$

Random corruptions subsample observability matrix in each window

(7) For all $0 \le \ell < T/t$,

$$\sum_{j=0}^{t-1}(1 - a_{\ell t+j})(A^j)^\top B^\top BA^j \le \eta \cdot O_t + O\left(\rho^2\|B\|^2\sqrt{t\log(dT/t\delta)}\right)\cdot \mathrm{Id}$$

Initial state bounded

(8) $\|x_0\|^2 \le R^2(d + O(\log(1/\delta)))$

The program objective is to minimize

$$\min \frac{1}{T}\widetilde{\mathbb{E}}\left[\sum_{i=0}^{T-1}\left(a_i\|Bx_i - y_i\|^2/\tau^2 + \|w_i\|^2/\sigma^2\right) + \|x_0\|^2/R^2\right] \quad (14)$$

over degree-4 pseudoexpectations satisfying the above constraints.

Remark 4.2 (Uncorrupted Case: Equivalence to Kalman Smoother). Suppose that we know there are no corruptions: then we can set $\eta = 0$ in the above program and therefore eliminate the variables a_i (they are all equal to 1). Then, by a well-known folklore argument, the SoS program is equivalent to the corresponding convex program with actual variables $x_i \in \mathbb{R}^d, v_i \in \mathbb{R}^m$, etc. with the same set of constraints. (This is because, by SoS Cauchy Schwarz, replacing the pseudoexpectation $\widetilde{\mathbb{E}}[\cdot]$ with the delta distribution over $\widetilde{\mathbb{E}}[x]$ gives a valid pseudoexpectation with equal or better objective value.) Then the objective is the same as the MAP objective, and as argued below the constraints are satisfied with high probability by the unconstrained MAP solution (Kalman smoother), so our algorithm simply outputs the MAP.

4.2 Feasibility of Oracle Kalman Smoother

In the following section, we show that the output of the oracle Kalman filter, i.e. the algorithm which knows precisely which time steps have been corrupted and runs the offline Kalman filter (Kalman smoother) on the uncorrupted steps to optimally estimate the trajectory, satisfies the constraints of the Program with high probability. In the proof of Lemma 4.3, we show how to do this by reducing to showing that the ground truth x^* is feasible with high probability, which is more straightforward. The key fact which

allows us to do this is knowledge that the posterior is a Gaussian centered at the output of the Kalman filter.

Lemma 4.3. *Let $\{x_i\}$ be the sequence of estimates given by running the Kalman smoother (i.e. offline Kalman filter) on the uncorrupted part of the trajectory, let $a_i = a_i^*$, let $w_i = x_i - Ax_{i-1}$ for all T, let $v_i = y_i - Bx_i$ when $a_i^* = 1$ and otherwise $v_i = 0$. Let $E[\cdot]$ be the expectation with respect to the delta distribution at this point $(x_i, a_i, v_i, w_i)_{i=1}^n$. Then $E[\cdot]$ is feasible for Program 1 with probability at least $1 - \delta$.*

Proof. It is immediate that Constraints 1, 2, and 3 are satisfied. Constraints 4 and 7 only involve a_i^* and we verify them in Lemma 4.6. It remains to check Constraints 5 and 6.

For what follows, suppose a_i^* is fixed. We claim the following two distributions on $\{x_i^*\}$ are equal:

(1) Sample a trajectory $\{x_i^*\}$ from the prior.
(2) Sample a trajectory $\{x_i^0\}$ from the prior, sample observations y_i for times where $a_i^* = 1$ given this trajectory, and sample trajectory $\{x_i^*\}$ from the resulting posterior on $\{x_i^0\}$ given y_i.

The equivalence of these two follows from the following basic fact: given a pair of random variables (X, Y), it's equivalent to sample X from its marginal law directly, or to first sample Y from its marginal law, and then to sample X conditional on Y. In the second case, the observations are the random variable Y and the trajectory is X; the fact implies that Y is sampled from its marginal law, which means that the marginal law of $\{x_i^*\}$ is simply the prior on trajectories. This fact is sometimes called the Nishimori identity.

Recall that the Kalman smoother output is simply the posterior mean $\hat{x}_i = \mathbb{E}[x_i^* \mid \{y_i\}_{i:a_i^*=1}]$ and that the posterior on trajectories is a multivariate Gaussian distribution. By Lemma 4.6, we have that

$$\begin{aligned} \|y_i - Bx_i^*\|^2 &\le O(m\tau^2 + \tau^2 \log(T/\delta)) \\ \|x_i^* - Ax_{i-1}^*\|^2 &\le O(d\sigma^2 + \sigma^2 \log(T/\delta)) \\ \|x_0^*\|^2 &\le R^2(d + O(\log(1/\delta))) \end{aligned} \tag{15}$$

uniformly over i with probability at least $1 - \delta$, then by the law of total probability we know that for $\mathcal{K}$ the feasible set defined by the constraints above in (15),

$$\delta \ge \mathbb{P}[(x^*, y) \notin \mathcal{K}] = \mathbb{E}[\mathbb{P}[(x^*, y) \notin \mathcal{K}|y]]$$

so by Markov's inequality $\mathbb{P}[\mathbb{P}[(x^*, y) \notin \mathcal{K}|y] > 1/3] \le 3\delta$, i.e. $\mathbb{P}[\mathbb{P}[(x^*, y) \in \mathcal{K}|y] \ge 2/3] \ge 1 - 3\delta$, which by Lemma 4.4 implies that $\mathbb{P}[(\mathbb{E}[x^*|y], y) \in \mathcal{K}] \ge 1 - 3\delta$ as well. Adjusting the value of δ by constants proves the result. □

Lemma 4.4. *Suppose that $\mathcal{K}$ is a closed convex set, $Z \sim N(\mu, \Sigma)$ is an arbitrary Gaussian random vector, and $\mathbb{P}[Z \in \mathcal{K}] \ge 0.5$. Then $\mu \in \mathcal{K}$.*

Proof. First we show this when $\mathcal{K}$ is an affine halfspace, i.e. $\mathcal{K} = \{x : \langle a, x\rangle \ge b\}$ for some a and b arbitrary. The assumption gives that $\langle a, Z\rangle \ge b$ with probability greater than 50%; since the marginal law of $\langle a, Z\rangle$ is $N(\langle a, \mu\rangle, a^T\Sigma a)$, and the Gaussian is symmetrical about its mean, it must be that $\langle a, \mu\rangle \ge b$ and so $\mu \in \mathcal{K}$. Now the result follows for arbitrary convex sets by writing them as intersections of affine halfspaces, since the above argument shows that μ will lie in each halfspace (since the probability of lying in each halfspace is at least as large as lying in the intersection), hence in the intersection of the halfspaces. □

Lemma 4.5. *For any $\delta > 0$,*

$$\left\| \sum_{i=0}^{t-1} a_i^* (A^i)^\top B^\top B A^i - (1-\eta) O_t \right\| \le O(\rho^2 \|B\|^2 \sqrt{t \log(d/\delta)})$$

with probability at least $1 - \delta$.

Proof. We apply the Matrix Hoeffding inequality (Lemma 3.8), using that $\|(A^i)^\top B^\top B A^i\| \le \rho^2 \|B\|^2$ by uniform stability. □

Lemma 4.6. *With probability at least $1 - \delta$, the ground truth random variables $(x_i^*, w_i^*, v_i^*, a_i^*)$ satisfies the constraints of Program 1 provided $T = \Omega(\log(2/\delta)/\eta)$.*

Proof. Equality constraints 1, 2, and 3 are satisfied by definition of the process. The remaining inequality constraints follow from a union bound as follows. The bound on Constraint 4 follows from Bernstein's inequality (see e.g. [55]). Constraint 5 follows by standard Gaussian concentration with probability at least $1 - \delta$. The same reasoning applies to Constraints 6 and 8. Constraint 7 follow from Lemma 4.5 applied to every window $0 \le \ell < T/t$. □

4.3 Outer Argument

In this section we reduce the problem of competing with OPT to getting good prediction error on the first iterate of every window.

Lemma 4.7. *Let $\widetilde{\mathbb{E}}[\cdot]$ be the solution to Program 1, assuming it is feasible. Let $\widehat{x}_i \triangleq \widetilde{\mathbb{E}}[x_i]$ and $\widehat{w}_i \triangleq \widetilde{\mathbb{E}}[w_i]$ for every $0 \le i < T$. Provided the event of Lemma 4.3 holds, then*

$$\frac{1}{T}\left(\sum_{i=0}^{T-1} \left(a_i^* \|B\widehat{x}_i - y_i\|^2/\tau^2 + \|\widehat{w}_i\|^2/\sigma^2\right) + \|\widehat{x}_0\|^2/R^2\right) - \mathrm{OPT} \lesssim$$
$$\eta\left(E_{\mathrm{noise}} + \rho^2 \frac{\alpha + \|B\|^2\sqrt{\log(dT/t\delta)/t}}{\tau^2 T/t} \sum_{\ell=0}^{T/t-1} \widetilde{\mathbb{E}}\left[\|x_{\ell t} - x_{\ell t}^*\|^2\right]\right).$$

where E_{noise} is defined in (12).

Before proving this, we will need the following helper lemma.

Lemma 4.8. *Let $\widetilde{\mathbb{E}}[\cdot]$ be the solution to Program 1, assuming it is feasible. Let $\widehat{x}_i \triangleq \widetilde{\mathbb{E}}[x_i]$ and $\widehat{w}_i \triangleq \widetilde{\mathbb{E}}[w_i]$ for every $0 \le i < T$. Provided the event of Lemma 4.3 holds, then*

$$\frac{1}{T}\left(\sum_{i=0}^{T-1} \left(a_i^* \|B\widehat{x}_i - y_i\|^2/\tau^2 + \|\widehat{w}_i\|^2/\sigma^2\right) + \|\hat{x}_0\|^2/R^2\right) - \mathrm{OPT} \le$$
$$\widetilde{\mathbb{E}}\left[\frac{1}{T}\sum_{i=0}^{T-1} (1 - a_i)\|B(x_i - x_i^*)\|^2/\tau^2\right] + O(\eta \cdot (m + \log(T/\delta))).$$

Proof. By Lemma 3.6, for any $0 \le i < T$, $\|B\widehat{x}_i - y_i\|^2 \le \widetilde{\mathbb{E}}[\|Bx_i - y_i\|^2]$ and $\|\widehat{w}_i\|^2 \le \widetilde{\mathbb{E}}[\|w_i\|^2]$, so it suffices to prove that the pseudoexpectation of $\sum_{i=0}^{T-1} \left(a_i^* \|Bx_i - y_i\|^2/\tau^2 + \|w_i\|^2/\sigma^2\right) + \|x_0\|^2/R^2$ is sufficiently bounded using the constraints of Program 1. First, by

splitting up $a_i^* = a_i^* a_i + a_i^*(1-a_i)$, we have

$$\sum_{i=0}^{T-1}\left(a_i^*\|Bx_i-y_i\|^2/\tau^2+\|w_i\|^2/\sigma^2\right)$$
$$=\sum_{i=0}^{T-1}\left(\frac{\|w_i\|^2}{\sigma^2}+a_i^*a_i\frac{\|Bx_i-y_i\|^2}{\tau^2}+a_i^*(1-a_i)\frac{\|Bx_i-y_i\|^2}{\tau^2}\right)$$

$$\le\sum_{i=0}^{T-1}\Big(\|w_i\|^2/\sigma^2+a_i\|Bx_i-y_i\|^2/\tau^2+2a_i^*(1-a_i)$$
$$\left(\|B(x_i-x_i^*)\|^2/\tau^2+\|v_i\|^2/\tau^2\right)\Big) \quad (16)$$

where in the inequality we used the fact that $a_i^* \le 1$ and that for i satisfying $a_i^* = 1$, $\|Bx_i - y_i\|^2 = \|B(x_i - x_i^*) - v_i\|^2 \le 2\|B(x_i - x_i^*)\|^2 + 2\|v_i\|^2$. Furthermore, note that

$$\sum_{i=0}^{T-1} a_i^*(1-a_i)\|v_i\|^2/\tau^2 \lesssim \eta\,(m+\log(T/\delta))\,T, \quad (17)$$

by Constraints 4 and 5. Putting (16) and (17) together allows us to upper bound the pseudo-expectation of

$$\sum_{i=0}^{T-1}\left(a_i^*\|Bx_i-y_i\|^2/\tau^2+\|w_i\|^2/\sigma^2\right)+\|x_0\|^2/R^2$$

by

$$\mathrm{OPT}+\widetilde{\mathbb{E}}\left[\frac{1}{T}\sum_{i=0}^{T-1}(1-a_i)\|B(x_i-x_i^*)\|^2/\tau^2\right]+O(\eta(m+\log(T/\delta))). \quad (18)$$

where we used the fact that $\widetilde{\mathbb{E}}[\cdot]$ minimizes the objective (14), the fact that the oracle Kalman filter solution is feasible because the event of Lemma 4.3 holds, as well as the fact that $a_i^* \le 1$. □

We now proceed with the proof of Lemma 4.7.

Proof of Lemma 4.7. Lemma 4.8 reduces upper bounding the excess risk achieved by $\{\widehat{x}_i\}, \{\widehat{w}_i\}$ to bounding the main term $\widetilde{\mathbb{E}}[\frac{1}{T}\sum_{i=0}^{T-1}(1-a_i)\|B(x_i-x_i^*)\|^2/\tau^2]$ in (18), which we do now. Using Fact 3.2, for any $i = \ell t + j$ we can write $B(x_i - x_i^*) = BA^j(x_{\ell t} - x_{\ell t}^*) + \sum_{s=0}^{j} BA^{j-s}(w_{\ell t+s} - w_{\ell t+s}^*)$. We thus have

$$\frac{1}{T}\sum_{i=0}^{T-1}(1-a_i)\|B(x_i-x_i^*)\|^2$$
$$=\frac{1}{T}\sum_{\ell,j}(1-a_{\ell t+j})\left\|BA^j(x_{\ell t}-x_{\ell t}^*)+\sum_{s=0}^{j}BA^{j-s}(w_{\ell t+s}-w_{\ell t+s}^*)\right\|^2$$
$$\le\frac{3}{T}\sum_{\ell,j}(1-a_{\ell t+j})\Big(\left\|BA^j(x_{\ell t}-x_{\ell t}^*)\right\|^2+\left\|\sum_{s=0}^{j}BA^{j-s}w_{\ell t+s}\right\|^2$$
$$+\left\|\sum_{s=0}^{j}BA^{j-s}w_{\ell t+s}^*\right\|^2\Big) \quad (19)$$

We can control the two noise terms on the right by noting that for any ℓ, j,

$$\left\|\sum_{s=0}^{j}BA^{j-s}w_{\ell t+s}\right\|^2 \le (j+1)\sum_{s=0}^{j}\|BA^{j-s}w_{\ell t+s}\|^2$$
$$\lesssim t\rho^2\|B\|^2\sigma^2(d+\log(T/\delta)),$$

where in the last step we used Constraint 6. Because the true process noise $\{w_i^*\}$ is part of a feasible solution to Program 1, from Constraint 4 we conclude that

$$\frac{1}{T}\sum_{\ell,j}(1-a_{\ell t+j})\left(\left\|\sum_{s}BA^{j-s}w_{\ell t+s}\right\|^2+\left\|\sum_{s}BA^{j-s}w_{\ell t+s}^*\right\|^2\right)$$
$$\lesssim \eta t\rho^2\|B\|^2\sigma^2(d+\log(T/\delta))$$

For the remaining terms in (19), we invoke Constraint 7 and the bound on $\|O_t\|$ in (10) to get

$$\frac{1}{T}\sum_{\ell,j}(1-a_{\ell t+j})\|BA^j(x_{\ell t}-x_{\ell t}^*)\|^2 \le 1.01\eta\rho^2$$
$$\left(\alpha+O\left(\|B\|^2\sqrt{\log(dT/t\delta)/t}\right)\right)\frac{1}{T/t}\sum_{\ell=0}^{T/t-1}\|x_{\ell t}-x_{\ell t}^*\|^2$$

from which the lemma follows by substituting the two estimates above into (19). □

4.4 Decay of Unobservable Subspace

In this section we show how to bound our prediction error on the first iterate of every window. Towards proving this, the main result of this subsection is to show that our error in estimating these first iterates decays exponentially over time provided that a certain matrix concentration event holds in every window.

We begin by describing this event. Let Π denote the projection to the *observable subspace*, that is, to the subspace of $v \in \mathbb{R}^d$ for which $v^\top O_t v \ge \zeta$ for $\zeta \triangleq \frac{\kappa t}{40000\rho^4}$, where the window size t will be optimized at the end of this section. The matrix concentration that we need to hold in every window is the following:

Lemma 4.9. *Suppose $t = \widetilde{\Omega}\left(\kappa^{-2}\rho^{12}\|B\|^4\log(dT/\delta)\right)$. Then with probability at least $1-\delta$ over the randomness of $\{a_i^*\}$, we have that for all windows $0 \le \ell < T/t$, there is a degree-2 SoS proof of the psd inequality*

$$\sum_{i=0}^{t-1}a_{\ell t+i}^*a_{\ell t+i}\Pi(A^i)^\top B^\top BA^i\Pi \succeq \frac{1}{100}\Pi O_t\Pi \quad (20)$$

using the constraints of Program 1.

We leave the proof of Lemma 4.9 to the full version of the paper.

We now turn to showing the main result of this section, namely that provided the event of Lemma 4.9 holds, our prediction error on the first iterate of every window decays exponentially over time.

Lemma 4.10. *Let pseudoexpectation $\widetilde{\mathbb{E}}[\cdot]$ be the solution to Program 1, assuming it is feasible. Provided the event of Lemma 4.9 holds, we have*

$$\widetilde{\mathbb{E}}[\|x_{\ell t}-x_{\ell t}^*\|^2] \le \frac{1}{2}\widetilde{\mathbb{E}}[\|x_{(\ell-1)t}-x_{(\ell-1)t}^*\|^2]+O(\rho^6E_{\mathrm{noise}}t/\kappa),$$

where E_{noise} is defined in (12).

Before proving Lemma 4.10, we first show how to use it to conclude the proof of Theorem 4.1.

Proof of Theorem 4.1. Take t as in (13). By the union bound, the events of Lemma 4.3 and Lemma 4.9 hold with probability at least $1-2\delta$. By summing the conclusion of Lemma 4.10 over the time windows, we get

$$\frac{1}{T/t}\sum_{\ell=0}^{T/t-1}\widetilde{\mathbb{E}}\left[\|x_{\ell t}-x^*_{\ell t}\|^2\right]\le\frac{1}{T/t}\widetilde{\mathbb{E}}\left[\|x_0-x^*_0\|^2\right]+O(\frac{\rho^6 E_{\text{noise}}t}{\kappa}). \tag{21}$$

Recall that $x^*_0\sim\mathcal{N}(0,R^2\cdot\mathrm{Id})$, so by standard concentration, $\|x^*_0\|^2\le R^2(d+O(\log(1/\delta)))$ with probability at least $1-\delta$. $\|x^*_0\|^2$ is similarly bounded by Constraint 8. We can thus bound $\widetilde{\mathbb{E}}\left[\|x_0-x^*_0\|^2\right]$ by $2R^2(d+O(\log(1/\delta)))$, so plugging this into (21) and invoking Lemma 4.7, we conclude the proof of Theorem 4.1. □

We now proceed to the proof of Lemma 4.10. The first step is an averaging argument to show that applying A^t to a vector in the unobservable subspace is guaranteed to decrease its norm.

Lemma 4.11. *For any vector $x\in\mathbb{R}^d$, $\|A^t\Pi^\perp x\|^2\le\frac{1}{40000\rho^2}\|\Pi^\perp x\|^2$.*

Proof. We have that

$$\begin{aligned}\frac{1}{t/s}\sum_{j=0}^{t/s-1}\|A^{js}\Pi^\perp x\|^2&=\frac{1}{t/s}\sum_{j=0}^{t/s-1}x^\top\Pi^\perp(A^{js})^\top A^{js}\Pi^\perp x\\&\le\frac{1}{\kappa t}\sum_{j=0}^{t/s-1}x^\top\Pi^\perp(A^{js})^\top O_sA^{js}\Pi^\perp x\\&=\frac{1}{\kappa t}x^\top\Pi^\perp O_t\Pi^\perp x\le\frac{1}{40000\rho^4}\|\Pi^\perp x\|^2,\end{aligned}$$

where the third step follows by the first part of Fact 3.1 and the last step follows by the definition of $\Pi^\perp$ and $\zeta=\frac{\kappa t}{40000\rho^4}$. By averaging, there exists there some $0\le j<t/s$ for which $\|A^{js}\Pi^\perp x\|^2\le\frac{1}{40000\rho^4}\|\Pi^\perp x\|^2$. The lemma follows by uniform stability. □

We will eventually take x to be the difference between our estimate of an iterate at the beginning of a window and the ground truth. Informally, this will tell us that over the course of a window of size t, the component of the error that started in the unobservable subspace has decayed.

What about the component of the error that started in the *observable* subspace? By uniform stability, it cannot increase by too much, but unlike the unobservable component, it need not decay. This brings us to the win-win argument at the core of the proof of Lemma 4.13: when the observable component does not decay, we can still relate it to the observable component in the previous time window just by uniform stability, and then bound this by a tiny fraction of the unobservable component in the previous time window!

Lemma 4.12. *With probability at least $1-\delta$, the following holds true for for every window $0\le\ell<T/t$, for $q\triangleq x_{\ell t}-x^*_{\ell t}$, all for all $i<t$. There is a degree-4 SoS proof from the constraints in Program 1 that*

$$a^*_{\ell t+i}a_{\ell t+i}\|BA^i\Pi q\|^2\le 4a^*_{\ell t+i}a_{\ell t+i}\|BA^i\Pi^\perp q\|^2+O(E_{\text{noise}}),$$

where recall that E_{noise} is defined in (12).

Proof. Without loss of generality we can assume $\ell=0$. For any $i<t$, we have the following sequence of inequalities in degree-4 SoS

$$\begin{aligned}&a_ia^*_i\left\|BA^i(\Pi+\Pi^\perp)q\right\|^2=a_ia^*_i\left\|BA^iq\right\|^2\\&=a_ia^*_i\left\|(y_i-Bx^*_i)-(y_i-Bx_i)+B(x^*_i-A^ix^*_0)-B(x_i-A^ix_0)\right\|^2\\&\le 3\|v^*_i\|^2+3\|v_i\|^2+3\left\|\sum_{s=1}^{i}BA^{i-s}(w_s-w^*_s)\right\|^2\le E_{\text{noise}},\end{aligned}$$

where we used the constraints and event of Lemma 4.3 in the last step (which holds with probability at least $1-\delta$). So the lemma follows by applying Fact 3.7 to $\varepsilon\triangleq E_{\text{noise}}$, $v_1=a^*_ia_iBA^i\Pi q$, and $v_2=a^*_ia_iBA^i\Pi^\perp q$. □

Lemma 4.13. *Let pseudoexpectation $\widetilde{\mathbb{E}}[\cdot]$ be the solution to Program 1, assuming it is feasible. Provided the events of Lemma 4.9 and Lemma 4.12 hold, then at least one of the following holds for every window $0\le\ell<T/t$ for $q\triangleq x_{\ell t}-x^*_{\ell t}$:*

(1) (Observable component decays)

$$\widetilde{\mathbb{E}}\left[\|A^t\Pi q\|^2\right]\le\frac{1}{10}\widetilde{\mathbb{E}}\left[\|\Pi q\|^2\right].$$

(2) (Observable error bounded by unobservable error)

$$\widetilde{\mathbb{E}}\left[\|\Pi q\|^2\right]\le\frac{1}{10\rho^2}\widetilde{\mathbb{E}}\left[\|\Pi^\perp q\|^2\right]+O\left(E_{\text{noise}}\,\rho^4/\kappa\right).$$

Proof. Without loss of generality we can assume $\ell=0$. In addition to the lower bound of (20), we also have a degree-2 SoS proof of the upper bound

$$\sum_{i=0}^{T-1}a^*_ia_i\Pi^\perp(A^i)^TB^TBA^i\Pi^\perp\preceq\Pi^\perp O_t\Pi^\perp\preceq\zeta\cdot I, \tag{22}$$

where in the first step we used that $a_ia^*_i\le 1$ by Constraint 1 and in the second step we used the definition of Π.

For convenience, define $q\triangleq x_0-x^*_0$. We proceed by casework on whether there is a gap between $\widetilde{\mathbb{E}}\left[(\Pi q)^\top O_t(\Pi q)\right]$ and $\widetilde{\mathbb{E}}[\zeta\|\Pi q\|^2]$:

Case 1: $\widetilde{\mathbb{E}}\left[\|\Pi q\|^2\right]\ge\widetilde{\mathbb{E}}\left[\frac{1}{4000\rho^2\zeta}\sum_{i=0}^{t-1}\|BA^i\Pi q\|^2\right]$.

The analysis for this case is very similar to the analysis in Lemma 4.11. We have

$$\begin{aligned}\widetilde{\mathbb{E}}\left[\|\Pi q\|^2\right]&\ge\widetilde{\mathbb{E}}\left[\frac{1}{4000\rho^2\zeta}\sum_{i=0}^{t-1}\|BA^i\Pi q\|^2\right]\\&=\widetilde{\mathbb{E}}\left[\frac{1}{4000\rho^2\zeta}\sum_{j=0}^{t/s-1}q^\top\Pi A^{js\top}O_sA^{js}\Pi q\right]\\&\ge\widetilde{\mathbb{E}}\left[\frac{10\rho^2 s}{t}\sum_{j=0}^{t/s-1}\|A^{js}\Pi q\|^2\right],\end{aligned}$$

where in the last step we used the definition of ζ and the assumption that $\lambda_{\min}(O_s)\ge\kappa s$. Rearranging, we obtain

$$\frac{1}{10\rho^2}\widetilde{\mathbb{E}}[\|\Pi q\|^2]\ge\frac{1}{t/s}\sum_{j=0}^{t/s-1}\widetilde{\mathbb{E}}\left[\|A^{js}\Pi q\|^2\right].$$

Therefore, there exists some index $0 \le j < t/s$ for which

$$\widetilde{\mathbb{E}}\big[\|A^{js}\Pi q\|^2\big] \le \frac{1}{10\rho^2}\widetilde{\mathbb{E}}\big[\|\Pi q\|^2\big].$$

By uniform stability, we obtain the first desired outcome in the lemma statement.

Case 2: $\widetilde{\mathbb{E}}\big[\|\Pi q\|^2\big] \le \widetilde{\mathbb{E}}\Big[\frac{1}{4000\rho^2\zeta}\sum_{i=0}^{t-1}\|BA^i\Pi q\|^2\Big]$.

In this case we invoke (20) to obtain

$$\begin{aligned}\widetilde{\mathbb{E}}\big[\|\Pi q\|^2\big] &\le \widetilde{\mathbb{E}}\left[\frac{1}{4000\rho^2\zeta}\sum_{i=0}^{t-1}\|BA^i\Pi q\|^2\right]\\ &\le \widetilde{\mathbb{E}}\left[\frac{1}{40\rho^2\zeta}\sum_{i=0}^{t-1}a_i^*a_i\|BA^i\Pi q\|^2\right].\end{aligned}$$

Recall from Lemma 4.12 that we have a degree-4 SoS proof of

$$a_i^*a_i\|BA^i\Pi q\|^2 \le 4a_i^*a_i\|BA^i\Pi^\perp q\|^2 + O(E_{\text{noise}}).$$

Summing this inequality over $i < t$ and taking pseudo-expectations, we get

$$\widetilde{\mathbb{E}}\left[\sum_{i=0}^{t-1}a_i^*a_i\|BA^i\Pi q\|^2\right] \le \widetilde{\mathbb{E}}\left[4\sum_{i=0}^{t-1}a_i^*a_i\|BA^i\Pi^\perp q\|^2\right] + O(E_{\text{noise}}t).$$

Substituting this back into the main bound (**??**), we get

$$\begin{aligned}\widetilde{\mathbb{E}}\big[\|\Pi q\|^2\big] &\le \widetilde{\mathbb{E}}\left[\frac{1}{10\rho^2\zeta}\sum_{i=0}^{t-1}a_i^*a_i\|BA^i\Pi^\perp q\|^2\right] + O\left(\frac{E_{\text{noise}}t}{30\zeta}\right)\\ &\le \widetilde{\mathbb{E}}\left[\frac{1}{10\rho^2}\|\Pi^\perp q\|^2\right] + O\left(\frac{E_{\text{noise}}t}{30\zeta}\right),\end{aligned}$$

where in the last step we used (22). Unpacking the definition of ζ, we arrive at the second desired bound. □

We are now ready to prove Lemma 4.10:

Proof of Lemma 4.10. By the SoS triangle inequality,

$$\begin{aligned}&\|x_{\ell t} - x^*_{\ell t}\|^2\\ &\quad\le 2\|A^t(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2 + O(t^2\sigma^2(d+\log(T/\delta)))\\ &\le 4\|A^t\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2 + 4\|A^t\Pi^\perp(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\\ &\qquad + O(t^2\sigma^2(d+\log(T/\delta)))\\ &\le 4\|A^t\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2 + (1/10000)\|\Pi^\perp(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\\ &\qquad + O(t^2\sigma^2(d+\log(T/\delta)))\end{aligned}$$

where we used Constraint 6 and triangle inequality in the first inequality, SoS triangle inequality in the second inequality, and Lemma 4.11 in the third inequality.

Now based on Lemma 4.13 applied to $\ell - 1$, we consider the following two cases:

Case 1. : Observable component decays, that is, we have

$$\widetilde{\mathbb{E}}\Big[\|A^t\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big] \le \frac{1}{10}\widetilde{\mathbb{E}}\Big[\|\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big].$$

Then we can argue that the error at time ℓt is a small fraction of the error at time $(\ell-1)t$ because both the observable and unobservable components of the error at time $(\ell-1)t$ have decayed over t steps. Formally:

$$\begin{aligned}&\widetilde{\mathbb{E}}[\|x_{\ell t} - x^*_{\ell t}\|^2]\\ &\le \widetilde{\mathbb{E}}\Big[4\|A^t\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big]\\ &\quad + \widetilde{\mathbb{E}}\Big[10^{-4}\|\Pi^\perp(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big] + O(t^2\sigma^2(d+\log(T/\delta)))\\ &\le \widetilde{\mathbb{E}}\Big[(2/5)\|\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big]\\ &\quad + \widetilde{\mathbb{E}}\Big[10^{-4}\|\Pi^\perp(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big] + O(t^2\sigma^2(d+\log(T/\delta)))\\ &\le (1/2)\widetilde{\mathbb{E}}\Big[\|x_{(\ell-1)t} - x^*_{(\ell-1)t}\|^2\Big] + O(t^2\sigma^2(d+\log(T/\delta))).\end{aligned}$$

where in the last step we used the Pythagorean Theorem.

Case 2. : Observable error bounded by unobservable error, that is

$$\widetilde{\mathbb{E}}\big[\|\Pi q\|^2\big] \le \frac{1}{10\rho^2}\widetilde{\mathbb{E}}\big[\|\Pi^\perp q\|^2\big] + O(E_{\text{noise}}\rho^4/\kappa) \tag{23}$$

where $q \triangleq x_{(\ell-1)t} - x^*_{(\ell-1)t}$. Then we can argue that the error at time ℓt is a small fraction of the error at time $(\ell-1)t$ as follows. As in Case 1, the unobservable error at time $(\ell-1)t$ has decayed. As discussed above, the observable error at time ℓt might even be bigger than the observable error at time $(\ell-1)t$, but it can't be much bigger because of uniform stability. On the other hand, the latter is bounded by a small fraction of the *unobservable* error at time $(\ell-1)t$. This lets us conclude that the overall error at time ℓt is bounded even by the unobservable error at time $(\ell-1)t$. Formally,

$$\begin{aligned}&\widetilde{\mathbb{E}}[\|x_{\ell t} - x^*_{\ell t}\|^2]\\ &\le \widetilde{\mathbb{E}}\Big[4\|A^t\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big]\\ &\quad + \widetilde{\mathbb{E}}\Big[10^{-4}\|\Pi^\perp(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\Big] + O(t^2\sigma^2(d+\log(T/\delta)))\\ &\le 4\rho^2\|\Pi(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2 + 10^{-4}\|\Pi^\perp(x_{(\ell-1)t} - x^*_{(\ell-1)t})\|^2\\ &\quad + O(t^2\sigma^2(d+\log(T/\delta)))\\ &\le \frac{1}{2}\widetilde{\mathbb{E}}\big[\|q\|^2\big] + O(\rho^6E_{\text{noise}}t/\kappa).\end{aligned}$$

where $C < 1$ is an absolute constant and in the last step we used (23) and absorbed $O(t^2\sigma^2(d+\log(T/\delta)))$ into $O(\rho^6E_{\text{noise}}t/\kappa)$. Since we showed the desired conclusion in both cases, the proof is complete. □

4.5 Confidence Band Recovery

Here we note that as a consequence of Theorem 4.1, our estimate $\{\widetilde{\mathbb{E}}[x_i]\}$ of the trajectory is actually pointwise $O(\log T)$-close to the true trajectory at all time steps, except for a $o(1)$ proportion of time close to time zero. This will be useful in the full version of the paper when we use this as a warm start for a second sum-of-squares relaxation that will achieve excess risk *doubly logarithmic* in T. A proof is given in the full version of the paper.

Corollary 4.14. *Let pseudoexpectation $\widetilde{\mathbb{E}}[\cdot]$ be the solution to Program 1, assuming it is feasible. Then provided the event of Lemma 4.9 holds, for all $0 \le i < T$ the estimates $\{\widetilde{\mathbb{E}}[x_i]\}$ satisfy*

$$\|\widetilde{\mathbb{E}}[x_i] - x_i^*\| \lesssim \frac{\rho}{2^{\ell(i)/2}}R\left(\sqrt{d} + \sqrt{\log(1/\delta)}\right) + O(\rho^4E_{\text{noise}}^{1/2}t^{1/2}/\kappa^{1/2})$$

REFERENCES

[1] Brian DO Anderson and John B Moore. 2007. *Optimal control: linear quadratic methods*. Courier Corporation.

[2] Brian DO Anderson and John B Moore. 2012. *Optimal filtering*. Courier Corporation.

[3] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. 2020. Robustly Learning Mixtures of k Arbitrary Gaussians. arXiv:2012.02119 [cs.DS]

[4] Ainesh Bakshi and Pravesh Kothari. 2020. Outlier-Robust Clustering of Non-Spherical Mixtures. *arXiv preprint arXiv:2005.02970* (2020).

[5] Ainesh Bakshi and Adarsh Prasad. 2021. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 102–115.

[6] Jess Banks, Sidhanth Mohanty, and Prasad Raghavendra. 2021. Local statistics, semidefinite programming, and community detection. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 1298–1316.

[7] Boaz Barak, Jonathan A Kelner, and David Steurer. 2014. Rounding sum-of-squares relaxations. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. 31–40.

[8] Boaz Barak and David Steurer. 2014. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236* (2014).

[9] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. 2017. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. 47–60.

[10] Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. 2020. Online and Distribution-Free Robustness: Regression and Contextual Bandits with Huber Contamination. *arXiv preprint arXiv:2010.04157* (2020).

[11] Yeshwanth Cherapanamjeri, Efe Aras, Nilesh Tripuraneni, Michael I Jordan, Nicolas Flammarion, and Peter L Bartlett. 2020. Optimal Robust Linear Regression in Nearly Linear Time. *arXiv preprint arXiv:2007.08137* (2020).

[12] Geoffrey Chinot et al. 2020. ERM and RERM are optimal estimators for regression problems when malicious outliers corrupt the labels. *Electronic Journal of Statistics* 14, 2 (2020), 3563–3605.

[13] Arnak Dalalyan and Philip Thompson. 2019. Outlier-robust estimation of a sparse linear model using l1-penalized Huber's M-estimator. In *Advances in Neural Information Processing Systems*. 13188–13198.

[14] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. 2020. Robustly Learning any Clusterable Mixture of Gaussians. *arXiv preprint arXiv:2005.06417* (2020).

[15] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2017. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 999–1008.

[16] Ilias Diakonikolas, Gautam Kamath, Daniel M Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. 2018. Robustly learning a gaussian: Getting optimal error, efficiently. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2683–2702.

[17] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. 2018. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1047–1060.

[18] Erwin Enrique Fetzer and PM Anderson. 1975. Observability in the state estimation of power systems. *IEEE transactions on power Apparatus and Systems* 94, 6 (1975), 1981–1988.

[19] M. Grötschel, L. Lovász, and A. Schrijver. 1981. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica* 1, 2 (01 Jun 1981), 169–197. https://doi.org/10.1007/BF02579273

[20] Didier Henrion and Andrea Garulli. 2005. *Positive polynomials in control*. Vol. 312. Springer Science & Business Media.

[21] Samuel Hopkins. 2018. *STATISTICAL INFERENCE AND THE SUM OF SQUARES METHOD*. Ph. D. Dissertation. Cornell University.

[22] Samuel B Hopkins and Jerry Li. 2018. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1021–1034.

[23] Peter J Huber. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* (1964), 73–101.

[24] Peter J Huber. 1973. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics* (1973), 799–821.

[25] T. Kailath, A.H. Sayed, and B. Hassibi. 2000. *Linear Estimation*. Prentice Hall. https://books.google.com/books?id=zNJFAQAAIAAJ

[26] Rudolf Kalman. 1959. On the general theory of control systems. *IRE Transactions on Automatic Control* 4, 3 (1959), 110–110.

[27] Rudolf E. Kálmán and Richard S. Bucy. 1961. New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering* 83 (1961), 95–108.

[28] Christopher D Karlgaard. 2015. Nonlinear regression Huber–Kalman filtering and fixed-interval smoothing. *Journal of guidance, control, and dynamics* 38, 2 (2015), 322–330.

[29] Adam Klivans, Pravesh K Kothari, and Raghu Meka. 2018. Efficient Algorithms for Outlier-Robust Regression. In *Conference On Learning Theory*. 1420–1430.

[30] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. 2018. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 1035–1046.

[31] Kevin A Lai, Anup B Rao, and Santosh Vempala. 2016. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 665–674.

[32] Jean B. Lasserre. 2001. *New Positive Semidefinite Relaxations for Nonconvex Quadratic Programs*. Springer US, Boston, MA, 319–331. https://doi.org/10.1007/978-1-4613-0279-7_18

[33] Allen Liu and Ankur Moitra. 2021. Settling the robust learnability of mixtures of gaussians. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*. 518–531.

[34] John Mattingley and Stephen Boyd. 2010. Real-time convex optimization in signal processing. *IEEE Signal processing magazine* 27, 3 (2010), 50–61.

[35] Shahar Mendelson. 2018. Learning without concentration for general loss functions. *Probability Theory and Related Fields* 171, 1 (2018), 459–502.

[36] Yilin Mo and Bruno Sinopoli. 2008. A characterization of the critical value for Kalman filtering with intermittent observations. In *2008 47th IEEE Conference on Decision and Control*. IEEE, 2692–2697.

[37] Yilin Mo and Bruno Sinopoli. 2011. Kalman filtering with intermittent observations: Tail distribution and critical value. *IEEE Trans. Automat. Control* 57, 3 (2011), 677–689.

[38] PC Müller and HI Weber. 1972. Analysis and optimization of certain qualities of controllability and observability for linear dynamical systems. *Automatica* 8, 3 (1972), 237–246.

[39] Yurii Nesterov. 2000. *Squared Functional Systems and Optimization Problems*. Springer US, Boston, MA, 405–440. https://doi.org/10.1007/978-1-4757-3216-0_17

[40] Ryan O'Donnell. 2014. *Analysis of boolean functions*. Cambridge University Press.

[41] Ryan O'Donnell and Yuan Zhou. 2013. Approximability and proof complexity. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 1537–1556.

[42] Pablo A Parrilo. 2000. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. Ph. D. Dissertation. California Institute of Technology.

[43] H Poor. 1980. On robust Wiener filtering. *IEEE Trans. Automat. Control* 25, 3 (1980), 531–536.

[44] Stephen Prajna, Antonis Papachristodoulou, Peter Seiler, and Pablo A Parrilo. 2005. SOSTOOLS and its control applications. In *Positive polynomials in control*. Springer, 273–292.

[45] Michael Roth, Emre Özkan, and Fredrik Gustafsson. 2013. A Student's t filter for heavy tailed process and measurement noise. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5770–5774.

[46] Irvin C Schick. 1989. *Robust recursive estimation of the state of a discrete-time stochastic linear dynamic system in the presence of heavy-tailed observation noise*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[47] Irvin C Schick and Sanjoy K Mitter. 1994. Robust recursive estimation in the presence of heavy-tailed observation noise. *The Annals of Statistics* (1994), 1045–1080.

[48] N.Z. Shor. 1987. Quadratic optimization problems. *Soviet Journal of Computer and Systems Sciences* 25 (11 1987).

[49] Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. 2018. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*. PMLR, 439–473.

[50] Max Simchowitz, Karan Singh, and Elad Hazan. 2020. Improper learning for non-stochastic control. In *Conference on Learning Theory*. PMLR, 3320–3436.

[51] Bruno Sinopoli, Luca Schenato, Massimo Franceschetti, Kameshwar Poolla, Michael I Jordan, and Shankar S Sastry. 2004. Kalman filtering with intermittent observations. *IEEE transactions on Automatic Control* 49, 9 (2004), 1453–1464.

[52] Didier Sornette and Kayo Ide. 2001. The Kalman–Lévy filter. *Physica D: Nonlinear Phenomena* 151, 2-4 (2001), 142–174.

[53] Joel A Tropp. 2012. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics* 12, 4 (2012), 389–434.

[54] Anastasios Tsiamis and George Pappas. 2020. Online learning of the kalman filter with logarithmic regret. *arXiv preprint arXiv:2002.05141* (2020).

[55] Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.