Adversarial Tradeoffs in Linear Inverse Problems and Robust State Estimation

Bruce Lee*, Thomas Zhang*, Hamed Hassani, and Nikolai Matni

Department of Electrial and Systems Engineering, University of Pennsylvania

Abstract

Adversarially robust training has been shown to reduce the susceptibility of learned models to targeted input data perturbations. However, it has also been observed that such adversarially robust models suffer a degradation in accuracy when applied to unperturbed data sets, leading to a robustness-accuracy tradeoff. In this paper, we provide sharp and interpretable characterizations of such robustness-accuracy tradeoffs for linear inverse problems. In particular, we provide an algorithm to find the optimal adversarial perturbation given data, and develop tight upper and lower bounds on the adversarial loss in terms of the standard (non-adversarial) loss and the spectral properties of the resulting estimator. Further, motivated by the use of adversarial training in reinforcement learning, we define and analyze the *adversarially robust Kalman Filtering problem.* We apply a refined version of our general theory to this problem, and provide the first characterization of robustness-accuracy tradeoffs in a setting where the data is generated by a dynamical system. In doing so, we show a natural connection between a filter's robustness to adversarial perturbation and underlying control theoretic properties of the system being observered, namely the spectral properties of its observability gramian.

1 Introduction

It has been demonstrated across various application areas that contemporary learned models, despite their impressive nominal performance, can be extremely susceptible to small, adversarially designed input perturbations (Carlini and Wagner, 2016; Goodfellow et al., 2014; Szegedy et al., 2013; Huang et al., 2017). In order to mitigate the effects of such attacks, various adversarially robust training algorithms (Carlini and Wagner, 2016, 2017; Madry et al., 2017; Xie et al., 2020; Deka et al., 2020) have been developed. However, it was soon noticed that while adversarial training could be used to improve model robustness, it often came with a corresponding decrease in accuracy on nominal (unperturbed) data. Further, various simplified theoretical models (Tsipras et al., 2018; Zhang et al., 2019; Nakkiran, 2019; Raghunathan et al., 2019; Chen et al., 2020; Javanmard et al., 2020), have been used to explain this phenomena, and to argue that such *robustness-accuracy tradeoffs* are unavoidable.

In this paper, we continue the study of such robustness-accuracy tradeoffs, with a focus on linear inverse problems. Such problems are ubiquitous in machine learning and engineering, with application areas spanning medical imaging (Ribes and Schmitt, 2008), systems biology (Engl et al., 2009),

*Equal contribution

econometrics (Carrasco et al., 2007), recommender systems (Recht et al., 2010), communications (Candes and Tao, 2005), and others. We show that due to the convexity of both the nominal and adversarial linear inverse problems, robustness-accuracy tradeoffs can be sharply and interpretably characterized in terms of the spectral properties of a given linear model. We then define and analyze an *adversarially robust Kalman Filtering problem* by applying a refined version of our general tools. Motivated by applications of adversarial robustness in the reinforcement literature (Lutter et al., 2021; Pinto et al., 2017; Mandlekar et al., 2017), we provide the first theoretical analysis of robustness-accuracy tradeoffs in a setting where data is generated by a *dynamical system*, and in doing so establish connections to natural control theoretic properties of the underlying system.

Our specific contributions can be summarized as follows:

- We propose a simple and computationally efficient algorithm that provably finds the maximal ℓ^2 norm-bounded adversarial perturbation for a given linear model and data-set. This allows us to efficiently compute and explore the Pareto-optimal robustness-accuracy tradeoff curve.
- We derive interpretable lower and upper bounds on the gap between the adversarial and standard (unperturbed) risks in terms of the spectral properties of a given linear model. We also show that our bounds are tight in the one-dimensional setting, recovering the results of (Javanmard et al., 2020), and for systems with full column rank.
- We define and analyze an *adversarially robust Kalman Filtering* problem, and show that lower and upper bounds on the gap between the adversarial and standard (unperturbed) risk can be controlled in terms of the spectral properties of the *observability gramian* (Zhou and Doyle, 1998) of the underlying system.
- We empirically demonstrate through extensive numerical experiments that our results qualitatively and quantitatively predict robustness-accuracy tradeoffs in linear inverse problems as a function of underlying spectral properties of the optimal nominal solution.

The rest of this paper is organized as follows: in Section 2, we present the general linear inverse problem along with an algorithm that solves the problem in an adversarial setting. We then derive tight lower and upper bounds on the adversarial risk, providing an interpretable characterization of the optimal tradeoff between robustness and accuracy. In Section 3, we pose the adversarially robust Kalman Filtering problem as a linear inverse problem, and refine the general bounds developed in the previous section, revealing that for the setting where data is generated by a dynamical system, robustness-accuracy tradeoffs are dictated by natural control theoretic properties of the underlying system (namely, the observability gramian). In Section 4, we provide empirical evidence to support the trends predicted by our bounds, and we end with conclusions and a discussion of future work in Section 5.

1.1 Related Work

Our work makes connections between adversarial robustnesss and robust estimation and control. We now provide a brief overview of work most directly related to ours from these two areas.

Robustness-accuracy tradeoffs: We draw inspiration from recent work offering theoretical characterizations of robustnesss-accuracy tradeoffs: Tsipras et al. (2018) and Zhang et al. (2019) posit that high standard accuracy is fundamentally at odds with high robust accuracy by considering

classification problems, whereas Nakkiran (2019) suggests an alternative explanation that classifiers that are simultaneously robust and accurate are complex, and may not be contained in current function classes. However, Raghunathan et al. (2019) shows that the tradeoff is not due to optimization or representation issues by showing that such tradeoffs exist even for a problem with a convex loss where the optimal predictor achieves 100% standard and robust accuracy. In contrast to previous work, we provide sharp and interpretable characterizations of the robustness-accuracy tradeoffs that may arise in regression problems, albeit restricted to linear models. Most closely related to our work, Javanmard et al. (2020) derive a formula for the exact tradeoff between standard and robust accuracy in the linear regression setting, which can be viewed as a 1-d linear inverse problem. We generalize these results to the matrix-valued setting, and further apply these tools to the adversarially robust Kalman Filtering problem, wherein data is generated by a dynamical system.

Robust estimation and control: Robustness in estimation and control has traditionally been studied from a worst-case induced gain perspective (Hassibi et al., 1999; Zhou and Doyle, 1998). When perturbations are restricted to be ℓ^2-bounded, this gives rise to so-called $\mathcal{H}_\infty$ estimation and control problems. Although widely known, and celebrated for their applications in robust control, $\mathcal{H}_\infty$-based methods are often overly conservative. This conservatism can be reduced by using mixed $\mathcal{H}_2/\mathcal{H}_\infty$ methods (Khargonekar et al., 1996), which blend gaussian disturbance assumptions with worst case disturbances. While such an approach is related to the adversarially robust Kalman Filtering problem that we pose, we note that it decouples the adversarial and stochastic inputs during design, leading to a fundamentally different tradeoff. We leave characterizing a connection between traditional and adversarial robustness to future work. More recently, Al Makdah et al. (2020) have considered the robustness-accuracy tradeoff in data-driven perception-based control. However, the adversary in their paper attacks the noise distribution, and thus the robustness under consideration is distributional, whereas our adversary maximally attacks each measurement instance, which is more aligned to the standard adversarial perturbations considered in machine learning contexts.

2 The Linear Inverse Problem

We consider the following linear inverse problem, where given data $x \in \mathbb{R}^n$, measurements $y \in \mathbb{R}^p$ are generated by

$$y = A_\star x + w \tag{1}$$

where $w \in \mathbb{R}^p$ is a noise vector and $A_\star$ is an unknown ground-truth matrix. Our goal is to estimate the ground truth matrix $A_\star$ through an estimator $\hat{A}$.

In the following subsection, we introduce the fundamental tradeoff between accuracy of the resulting estimator $\hat{A}$ under an assumed noise distribution, and the robustness of this estimator to adversarial attacks. In §2.2, we show how to solve for the adversarial attacks, which provides a means of computing the performance of the estimator in the face of adversarial attacks. §2.3 provides upper and lower bounds on the gap between the robust performance and the nominal performance, which allow for prediction of the severity of the tradeoffs based on properties of the underlying problem.

2.1 Optimal Robustness-Accuracy Tradeoffs

When the noise vector w in (1) is assumed to be stochastic, then the minimum mean square estimator can be found by solving the following stochastic optimization problem:

$$\begin{aligned}\hat{A} &= \underset{A\in\mathbb{R}^{p\times n}}{\operatorname{argmin}}\ \mathbb{E}_{x,w}\left[\|y-Ax\|_2^2\right] \\ &=: \underset{A\in\mathbb{R}^{p\times n}}{\operatorname{argmin}}\ \mathrm{SR}(A).\end{aligned} \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean norm, and we use $\mathrm{SR}(A)$ to denote the *standard risk of* A, i.e., $\mathrm{SR}(A)$ is the mean squared error incurred by a matrix A. We then consider the same problem but with the inclusion of an ℓ^2-norm bounded adversary:

$$\begin{aligned}\hat{A}(\varepsilon) &= \underset{A\in\mathbb{R}^{p\times n}}{\operatorname{argmin}}\ \mathbb{E}_{x,w}\left[\max_{\|\delta\|_2\leq\varepsilon}\|y-A(x+\delta)\|_2^2\right] \\ &=: \underset{A\in\mathbb{R}^{p\times n}}{\operatorname{argmin}}\ \mathrm{AR}(A),\end{aligned} \tag{3}$$

where $\mathrm{AR}(A)$ denotes the *adversarial risk of* A.

Our objective in this section is to precisely characterize the optimal robustness-accuracy tradeoff curve defined by the adversarial and standard risks. We refer to the set of all points $(\mathrm{SR}(A), \mathrm{AR}(A))$ over all $A \in \mathbb{R}^{p\times n}$ as the $(\mathrm{SR}, \mathrm{AR})$ region. The optimal tradeoff between standard and adversarial risks is characterized via the so-called Pareto boundary of this region, which we denote $\{(\mathrm{SR}(A_\lambda), \mathrm{AR}(A_\lambda)) : \lambda \geq 0\}$. Using standard results in multi-objective optimization, A_λ are computed by solving the regularized optimization problem

$$A_\lambda := \underset{A}{\operatorname{argmin}}\ \mathrm{SR}(A) + \lambda\,\mathrm{AR}(A). \tag{4}$$

Varying the regularization parameter λ in problem (4) thus allows us to characterize the aforementioned Pareto boundary by interpolating between the solution to the standard (i.e., A_0) and adversarial (i.e., A_∞) linear inverse problems. Our first contribution is to show that the solution A_λ to the regularized optimization problem (4) can be computed efficiently using stochastic optimization.

In what follows, we assume that the noise $w \sim \mathcal{N}(0, \Sigma_w)$, $\Sigma_w \succ 0$, and that the data $x \sim \mathcal{N}(0, \Sigma_x)$, $\Sigma_x \succ 0$. It is simple to verify that under these assumptions, the standard risk $\mathrm{SR}(A)$ is given by

$$\begin{aligned}\mathrm{SR}(A) &= \mathbb{E}_{x,w}\left[\|(y-A_\star x)+(A_\star-A)x\|_2^2\right] \\ &= \mathrm{tr}(\Sigma_w) + \mathrm{tr}\left((A-A_\star)\Sigma_x(A-A_\star)^\top\right)\end{aligned} \tag{5}$$

However, due to the inner maximization in the adversarial risk, no such closed-form expression exists for $\mathrm{AR}(A)$. We show next that despite the non-convexity of the the inner maximization problem, it can be solved efficiently. This allows us to apply stochastic gradient descent to solve $\min_A \mathrm{AR}(A)$. In particular, note that $\mathrm{AR}(A)$ is the linear combination of a point-wise supremum of convex functions in A, and hence convex in A itself (Boyd and Vandenberghe, 2004). Next observe that we can draw samples of $x \sim \mathcal{N}(0, \Sigma_x)$ and $w \sim \mathcal{N}(0, \Sigma_w)$, and apply the solution to the inner maximization problem to solve for realizations of $\max_{\|\delta\|_2\leq\epsilon}\|y-A(x+\delta)\|_2^2$. Taking the gradient of these realizations with respect to A provides a stochastic descent direction. As the overall expression is convex in A, stochastic gradient descent with an appropriately decaying stepsize converges to the optimal solution (Bottou et al., 2018).

2.2 Solving the Inner Maximization

To the best of our knowledge, no closed-form expression exists for the adversarial risk: indeed, even in the scalar case studied in (Javanmard et al., 2020), it is characterized by a recursive relationship. Further, the techniques used to derive that recursion do not extend to the multi-variable case.

To address this challenge, we show how to efficiently compute solutions to the inner maximization of the adversarial risk $\mathrm{AR}(A)$. We observe that the inner maximization $\max_{\|\delta\|_2 \leq \varepsilon} \|y - A(x+\delta)\|_2^2$ can be expanded and re-written as the following (non-convex) quadratically-constrained quadratic maximization problem:

$$\begin{aligned} \underset{\delta \in \mathbb{R}^n}{\text{maximize}} \quad & \delta^\top A^\top A \delta - 2\delta^\top A^\top b \\ \text{subject to} \quad & \delta^\top \delta \leq \varepsilon^2, \end{aligned} \tag{P}$$

where we set $b := y - Ax$. Let $A = U\Sigma V^\top \in \mathbb{R}^{p \times n}$ be the full singular-value decomposition of A, with $U \in \mathbb{R}^{p \times p}$, $\Sigma \in \mathbb{R}^{p \times n}$, $V \in \mathbb{R}^{n \times n}$, and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{n,p\}})$, $\sigma_1 \geq \cdots \geq \sigma_{\min\{n,p\}} \geq 0$ the nonzero singular values of A. We also denote the columns of U and V by u_i and v_i, respectively. It is known that (P) satisfies strong duality (Boyd and Vandenberghe, 2004) and the optimal primal-dual solutions δ^*, λ^* can be characterized by the KKT conditions:

$$\begin{aligned} 2(\lambda^* I - A^\top A)\delta^* + 2A^\top b &= 0 \\ \lambda^*({\delta^*}^\top \delta^* - \varepsilon^2) &= 0 \\ (\lambda^* I - A^\top A) &\succeq 0. \end{aligned}$$

The KKT conditions can then be leveraged to solve for the optimal dual solution λ^* and subsequently the optimal perturbation δ^*. The full procedure is summarized in Algorithm 1.

Algorithm 1 Inner Maximization Solution

given $A = U\Sigma V^\top \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$, $\varepsilon > 0$
if $\sum_{i:\sigma_i < \sigma_1} \frac{(b^\top u_i)^2 \sigma_i^2}{(\sigma_1^2 - \sigma_i^2)^2} < \varepsilon^2$ **then**
  $c = \sqrt{\varepsilon^2 - \sum_{i:\sigma_i < \sigma_1} \frac{(b^\top u_i)^2 \sigma_i^2}{(\sigma_1^2 - \sigma_i^2)^2}}$
  Set v as any unit vector lying in the null-space of $\left(\sigma_1^2 I - \Sigma^\top \Sigma\right) V^\top$, i.e. $v \in \mathbf{span}\{v_i : \sigma_i = \sigma_1\}$
  $\delta^* = -V(\sigma_1^2 I - \Sigma^\top \Sigma)^\dagger \Sigma^\top U^\top b + cv$
else
  solve $\sum_{i=1}^m \frac{(b^\top u_i)^2 \sigma_i^2}{(\lambda - \sigma_i^2)^2} = \varepsilon^2$ for λ, e.g. by bisection
  $\delta^* = -V(\lambda^* I - \Sigma^\top \Sigma)^{-1} \Sigma^\top U^\top b$
end if
return δ^*

The proof of correctness for Algorithm 1 is detailed in the Appendix.

2.3 Lower and Upper Bounds on $\mathrm{AR}(A) - \mathrm{SR}(A)$

Although no closed-form expression exists for the adversarial risk $\mathrm{AR}(A)$, we show now that interepretable lower and upper bounds on the robustness-accuracy tradeoff, as characterized by the

gap $\mathrm{AR}(A) - \mathrm{SR}(A)$, can be derived. Such bounds predict the severity of the robustness-accuracy traeoff based upon underlying properties of specific linear inverse problems. We further show that these bounds are tight in the sense that they are exact for certain classes of matrices A, and strong in the sense that the lower and upper bounds differ only in higher-order terms with respect to the adversarial budget ε.

Theorem 2.1. *Given any $A \in \mathbb{R}^{p\times n}$, we have the following lower bound on $\mathrm{AR}(A) - \mathrm{SR}(A)$:*

$$\mathrm{AR}(A) - \mathrm{SR}(A) \geq 2\varepsilon\, \mathbb{E}_{x,w}\left[\left\|A^\top(y - Ax)\right\|_2\right] + \varepsilon^2 \lambda_{\min}(A^\top A), \tag{6}$$

and a corresponding upper bound

$$\mathrm{AR}(A) - \mathrm{SR}(A) \leq 2\varepsilon\, \mathbb{E}_{x,w}\left[\left\|A^\top(y - Ax)\right\|_2\right] + \varepsilon^2 \lambda_{\max}(A^\top A), \tag{7}$$

where $\lambda_{\min}(A^\top A)$ and $\lambda_{\max}(A^\top A)$ are the minimum and maximum eigenvalues of $A^\top A$, respectively.

The proof of these bounds relies on turning the inner maximization of the adversarial risk into various equivalent optimization problems, and utilizing the properties of Schur complements and the S-lemma. The full proof is relegated to the appendix.

We note that when $A^\top \in \mathbb{R}^n$, equation (7) recovers the exact characterization of the gap $\mathrm{AR}(A) - \mathrm{SR}(A)$ provided in (Javanmard et al., 2020, Lemma 3.1); thus when $p = 1$, the inequality in equation (7) is in fact an equality. We also note that the upper and lower bounds differ only in the $\mathcal{O}(\varepsilon^2)$ terms. This leads immediately to the following corollary.

Corollary 2.1. *If $p \geq n$ and A has orthogonal columns, then bounds* (6) *and* (7) *match.*

Proof. If $p \geq n$ and A has orthogonal columns, then $\lambda_{\min}(A^\top A) = \lambda_{\max}(A^\top A)$. □

The terms involving the eigenvalues of $A^\top A$ in our bounds also support the intuition that adversarial robustness is a form of implicit regularization, which is visualized in Figure 1. In the one-dimensional linear classification setting, this phenomenon is well-understood (Tsipras et al., 2018; Dobriban et al., 2020), where robustness to adversarial perturbations prevent a robust feature vector from relying on an aggregate of small features.

Leveraging bounds (6) and (7) from Theorem 3.1, we can bound both the susceptibility and robustness of a particular linear model to adversarial perturbations. In particular, the solution to the standard risk problem, which is $A_\star$ using (5), provides key insights into the properties of the problem.

Corollary 2.2. *Assume that $w \sim \mathcal{N}(0, \sigma_w^2 I_p)$. Then*

$$\begin{aligned}
&2\sqrt{\frac{2}{\pi}}\frac{\varepsilon}{\sqrt{p}}\sigma_w \operatorname{tr}\left(\left(A_\star^\top A_\star\right)^{\frac{1}{2}}\right) + \varepsilon^2 \lambda_{\min}(A_\star^\top A_\star) \\
\leq\ & \mathrm{AR}(A_\star) - \mathrm{SR}(A_\star) \\
\leq\ & 2\varepsilon\sigma_w \sqrt{\operatorname{tr}\left(A_\star^\top A_\star\right)} + \varepsilon^2 \lambda_{\max}(A_\star^\top A_\star).
\end{aligned}$$

Corollary 2.2 highlights that the spectral properties of the ground truth model $A_\star$ influences both the susceptibility and robustness of a nominal model to adversarial perturbations. A more refined analysis akin to Corollary 2.2 will form the basis of our analysis for the adversarially robust Kalman Filtering problem in the next section.

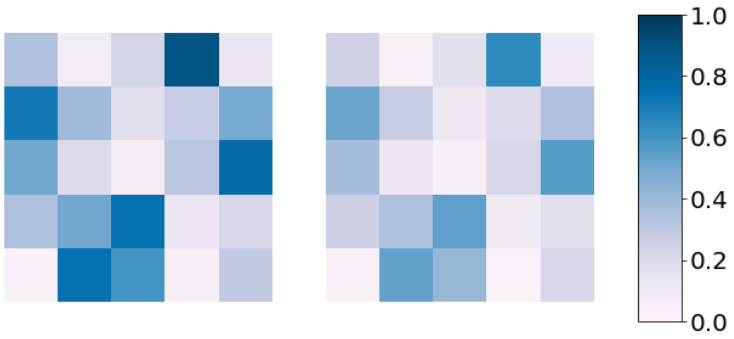


Figure 1: Optimizing for adversarial robustness induces an implicit regularization, which is visualized in this heatmap of a 5×5 nominal solution $A_\star$ (left) and adversarially robust solution $\hat{A}(\varepsilon)$ (right).

3 Adversarially Robust Kalman Filtering

We consider a modification to the standard Kalman Filtering problem (reviewed in §3.1) to incorporate adversarial robustness. We then extend and refine the tools developed in §2.3 to this setting, with the goal of relating control theoretic properties of the underlying linear dynamical system to the robustness-accuracy tradeoffs that it induces.

3.1 State Estimation and Observability

State estimation, which is central to control theory, is one instance of an inverse problem. The goal is to recover an estimate of the underlying state of a dynamical system given noisy partial observations of the system. A particularly interesting example of a state estimator is the Kalman Filter, which is designed for the setting where the underlying dynamical system is linear, and disturbances are gaussian. In particular, consider a linear-time-invariant (LTI) autonomous system with state and measurement disturbances: let $x_t \in \mathbb{R}^n$ be the system state, $w_t \in \mathbb{R}^n$ the process noise, $y_t \in \mathbb{R}^p$ the measurement, and $v_t \in \mathbb{R}^p$ the measurement noise. The initial condition, process noise, and measurement noise are assumed to be i.i.d. zero-mean gussians: $x_0 \sim \mathcal{N}(0, \Sigma_0)$, $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w)$, $v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_v)$. The LTI system is then defined by the following equations:

$$\begin{aligned} x_{t+1} &= Ax_t + w_t \\ y_t &= Cx_t + v_t. \end{aligned} \tag{8}$$

Finite horizon state estimation determines an estimate for the state of the system at time k given some sequence of measurements $y_0, \ldots, y_N$. This problem encompasses smoothing ($k < N$), filtering ($k = N$), and prediction ($k > N$). When the measurement and process noise satisfy the assumptions above, the optimal state estimator is the celebrated Kalman Filter (or smoother/predictor), which produces state estimates that are a linear function of the observations. Therefore, the optimal estimate $\hat{x}_k$ of the state x_k at time k can be written as[1] $\hat{x}_k := LY_N$, where $L \in \mathbb{R}^{n \times p(N+1)}$ is some

[1]We note that state-space representations for the Kalman Filter (see Appendix) also exist Hassibi et al. (1999), but for our purposes it is more convenient to view it as a linear map.

matrix and Y_N is a vector of stacked observations

$$Y_N := \begin{bmatrix} y_0 \\ \vdots \\ y_N \end{bmatrix} \in \mathbb{R}^{p(N+1)}.$$

We similarly define the stacked process and measurement noise vectors as

$$W_N := \begin{bmatrix} w_0 \\ \vdots \\ w_{N-1} \end{bmatrix} \in \mathbb{R}^{nN} \text{ and } V_N := \begin{bmatrix} v_0 \\ \vdots \\ v_N \end{bmatrix} \in \mathbb{R}^{p(N+1)}.$$

Furthermore, suppose $k \leq N$ and let

$$\mathcal{O}_N = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^N \end{bmatrix} \in \mathbb{R}^{p(N+1)\times n},$$

$$\tau_N = \begin{bmatrix} 0 & & & & \\ C & 0 & & & \\ CA & C & & & \\ \vdots & & \ddots & 0 & \\ CA^{N-1} & \dots & & C & 0 \end{bmatrix} \in \mathbb{R}^{p(N+1)\times nN},$$

$$\Gamma_k = \begin{bmatrix} A^{k-1} & A^{k-2} & \dots & I & 0 & \dots & 0 \end{bmatrix} \in \mathbb{R}^{n\times n(N+1)}$$

so that $Y_N = \mathcal{O}_N x_0 + \tau_N W_N$ and $x_k = A^k x_0 + \Gamma_k W_N$. Here $\mathcal{O}_N$ is the referred to as the N-step observability matrix, which is a quantity of interest in our analysis.

We now introduce some background on the observability of a LTI system and its implications on state estimation.

Definition 3.1. *The LTI system* (8) *is* observable *if when the disturbances are zero, i.e.* $v_t = 0$ *and* $w_t = 0$ *for all* $t \geq 0$*, it is possible to recover the initial state* x_0 *from a sequence of* n *measurements* $y_0, y_1, \dots y_{n-1}$.

Lemma 3.1. *A system of the form* (8) *is* observable *if and only if the* observability matrix

$$\mathcal{O}_{n-1} = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

has rank n.

Throughout the remainder of the paper we make the following assumption:

Assumption 3.1. *The LTI system* (8) *is observable and* $N \geq n - 1$.

As stated, observability is a binary notion that determines whether state estimation is possible. However, observability does not capture the conditioning of the linear inverse problem defining the Kalman Filter. A more refined, non-binary notion of observability can be defined in terms of the *observability gramian* of a system.

Definition 3.2. *The* N-step observability gramian *is defined as* $W_o(N) := \mathcal{O}_N^\top \mathcal{O}_N$. *If the spectral radius of* A *is less than one, then taking the limit as* $N \to \infty$ *results in the* observability gramian $W_o(\infty) = \sum_{t=0}^{\infty} \left(A^t\right)^\top C^\top C A^t$.

Lemma 3.2. *An LTI system* (8) *is observable if and only if* $W_o(\infty)$ *is positive definite.*

The observability gramian provides significantly more information about the difficulty of state estimation than the rank condition in Lemma 3.1. In particular, the ellipsoid $\left\{x | x^\top W_o(\infty) x \leq 1\right\}$ contains the initial states x that lead to measurement signals with ℓ^2 norm bounded by 1 in the absense of process and measurement noise. In particular, letting $x_0 = x$, we have $x_0^\top W_o(\infty) x_0 = \sum_{t=0}^{\infty} x_0^\top \left(A^t\right)^\top C^\top C A^t x_0 = \sum_{t=0}^{\infty} x_t^\top C^\top C x_t = \sum_{t=0}^{\infty} \|y_t\|_2^2$. As such, small singular values of the observability gramian imply that a large subset of the state space leads to relatively small impacts on future measurements. This makes it difficult to use measurements to distinguish states in this region in the presence of process and measurement noise.

3.2 Kalman Filtering and Smoothing

We begin by reviewing relevant results from standard Kalman Filtering and Smoothing. We make the following simplifying assumption for presentation purposes going forward.

Assumption 3.2. $\Sigma_0 = \sigma_0^2 I$, $\Sigma_w = \sigma_w^2 I$, $\Sigma_v = \sigma_v^2 I$. *We further assume that the system matrix* $A = \rho Q$, $\rho \in \mathbb{R}$ *is a scaled orthogonal matrix, such that* ρ *controls the stability of the system.*

A generalization of our subsequent results to generic dynamics A and positive definite covariance matrices can be found in the Appendix: although more notationally cumbersome and difficult to interpret, they nevertheless convey the same overall trends.

Standard State Estimation Under Assumption 3.2, we define the minimum mean square estimator for the state x_k as

$$\hat{L}_k = \underset{L \in \mathbb{R}^{n \times p(N+1)}}{\operatorname{argmin}} \; \mathbb{E}\left[\|x_k - L Y_N\|_2^2\right]. \tag{9}$$

We note that the optimal solution to this problem is precisely the Kalman Filter ($k = N$) or Smoother ($k < N$). We explicitly solve for the minimum mean square estimator $\hat{L}_k$ in the following lemma, which is standard, but included for completeness.

Lemma 3.3. *Suppose* $k \leq N$. *The finite horizon Kalman state estimator is the solution to optimization problem* (9)*, and is given by*

$$\hat{L}_k = \left(\sigma_0^2 A^k \mathcal{O}_N^\top + \sigma_w^2 \Gamma_k \tau_N^\top\right) \left(\sigma_0^2 \mathcal{O}_N \mathcal{O}_N^\top + \sigma_w^2 \tau_N \tau_N^\top + \sigma_v^2 I\right)^{-1}.$$

Adversarially Robust State Estimation We now modify the standard filtering problem (9) to allow adversarial perturbations to enter through sensor measurements.[2] In particular, for some $\varepsilon > 0$, the adversarially robust state estimator is defined by

$$\hat{L}_k(\varepsilon) := \operatorname*{argmin}_{L \in \mathbb{R}^{n \times p(N+1)}} \mathbb{E}\left[\max_{\|\delta\|_2 \leq \varepsilon} \|x_k - L(Y_N + \delta)\|_2^2\right],$$

where the assumptions on the initial condition, process disturbance and measurement disturbance are as in Assumption 3.2.

3.3 Robustness-Accuracy Tradeoffs in Kalman Filtering

The Kalman state estimation problem and adversarial state estimation problems can be viewed as standard and adversarial linear inverse problems by defining

$$\begin{aligned}
\mathrm{SR}(L) &:= \mathbb{E}\left[\|x_k - LY_N\|_2^2\right], \\
\mathrm{AR}(L) &:= \mathbb{E}\left[\max_{\|\delta\|_2 \leq \varepsilon} \|x_k - L(Y_N + \delta)\|_2^2\right].
\end{aligned}$$

As in §2, our goal is to characterize robustness-accuracy trade-offs for this linear inverse problem. As previewed in §2, we show that a refined analysis allows for the the gap $\mathrm{SR}(L) - \mathrm{AR}(L)$ to be bounded in terms of the spectral properties of the observability gramian of the system, establishing a natural connection to the robust control and estimation literature (Hassibi et al., 1999; Zhou and Doyle, 1998). In particular, our results indicate the robustness-accuracy tradeoff is more severe for systems with "uniformly low observability," as characterized by the Frobenius norm of the observability gramian.

To begin, we present a closed form for the standard risk $\mathrm{SR}(L)$.

Lemma 3.4. *For any $L \in \mathbb{R}^{n \times p(N+1)}$, our standard risk is*

$$\mathrm{SR}(L) = \mathbb{E}\left[\|x_k - LY_N\|_2^2\right] = \sigma_0^2 \left\|A^k - L\mathcal{O}_N\right\|_F^2 + \sigma_w^2 \|\Gamma_k - L\tau_N\|_F^2 + \sigma_v^2 \|L\|_F^2.$$

Lemma 3.4 makes clear that the noise terms act as a regularizer: if $\sigma_w^2 = \sigma_v^2 = 0$, then the $\min_L \mathrm{SR}(L) = 0$, and is achieved by $L = A^k(W_o(N))^{-1}\mathcal{O}_N^\top$. This closed-form expression suggests that the spectral properties of $W_o(N)$ may play in important role in the robustness-accuracy tradeoffs satisfied by an LTI system (8). We formalize this intuition next.

3.4 Bounding $\mathrm{AR} - \mathrm{SR}$ for State Estimation

As in §2, we do not have a closed-form for the adversarial risk. Upper and lower bounds on the gap between the adversarial risk and standard risk, however, still highlight the role control theoretic quantities play in robustness-accuracy tradeoffs.

[2] We choose to restrict our attention to adversarial sensor measurements because it is a more direct analog to the traditional adversarial robustness literature, which considers perturbations to image data, and not to the image data-generating distribution (Szegedy et al., 2013; Goodfellow et al., 2014; Carlini and Wagner, 2016).

Theorem 3.1. *For any $L \in \mathbb{R}^{n\times p(N+1)}$, the gap between* $\mathrm{AR}(L)$ *and* $\mathrm{SR}(L)$ *admits the following lower bound:*

$$\mathrm{AR}(L) - \mathrm{SR}(L) \geq 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon}{\sqrt{n}}\sigma_v \left\|L\right\|_F^2 \tag{10}$$

We now turn our attention to studying the tradeoffs enjoyed by the Kalman Filter/Smoother $L = \hat{L}_k$ defined in Lemma 3.3. This choice is made because the Kalman estimator is the optimal estimator in the nominal setting, and is commonly used in practice.

Theorem 3.2. *Suppose that $\hat{L}_k$ is the Kalman estimator from Lemma 3.3. Let us denote $\sigma_\vee^2 = \max\left\{\sigma_0^2, \sigma_w^2\right\}$, $\sigma_\wedge^2 = \min\left\{\sigma_0^2, \sigma_w^2\right\}$ and*

$$r_k(\rho) = \begin{cases} k, & \rho = 1 \\ \frac{1-\rho^{2(k+1)}}{1-\rho^2}, & \rho \neq 1. \end{cases} \tag{11}$$

Then we have the following bound on the gap between AR *and* SR*.*

$$\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k) \geq 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon}{\sqrt{n}}\sigma_v \left\|C\right\|_F^2 \left(\frac{\rho^{2k}\sigma_0^2 + r_k(\rho)\,\sigma_w^2}{(N+1)\sigma_\vee^2 \left\|W_o(N)\right\|_F + \sigma_v^2}\right)^2. \tag{12}$$

We see that the lower bound increases as the Frobenius norm of the observability gramian decreases. This indicates that as observability becomes uniformly low, i.e., if all singular values of $W_o(N)$ are small, then a nominal state estimator $\hat{L}_k$ will have a large gap $\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k)$. Observe that increasing σ_w will increase the lower bound shown above when $\sigma_w \leq \sigma_0$.

We now derive an upper bound on the gap between the standard and adversarial risk for any given L. This bound follows from the upper bound in Theorem 2.1.

Theorem 3.3. *For any $L \in \mathbb{R}^{n\times p(N+1)}$, the following bound holds*

$$\mathrm{AR}(L) - \mathrm{SR}(L) \leq 2\varepsilon \left\|L\right\|_2 \left\|\Sigma^{1/2}\right\|_F + \varepsilon^2 \left\|L\right\|_2^2$$

where $\Sigma^{1/2}$ is the symmetric square root of the covariance of $x_k - LY_N$.

Again, we consider how this upper bound looks for the Kalman estimator $\hat{L}_k$.

Theorem 3.4. *Suppose that $\hat{L}_k$ is the Kalman state estimator from Lemma 3.3. Let $\sigma_\wedge^2$, $\sigma_\vee^2$, and $r_k(\rho)$ be defined as in Theorem 3.2. Then*

$$\begin{aligned}\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k) \leq{}& \varepsilon\left(\frac{\rho^{2k}\sigma_0^2 + r_k(\rho)\,\sigma_w^2}{\sigma_\wedge^2 \lambda_{\min}(W_o(N))^{1/2}}\right) \\ &\times \left[2\sqrt{n}\left(\sigma_\vee^2 + \left(\frac{\sigma_v}{\sigma_\wedge^2\lambda_{\min}(W_o(N))^{1/2}}\right)^2\right)^{1/2} + \varepsilon\left(\frac{1}{\sigma_\wedge^2\lambda_{\min}(W_o(N))^{1/2}}\right)\right].\end{aligned}$$

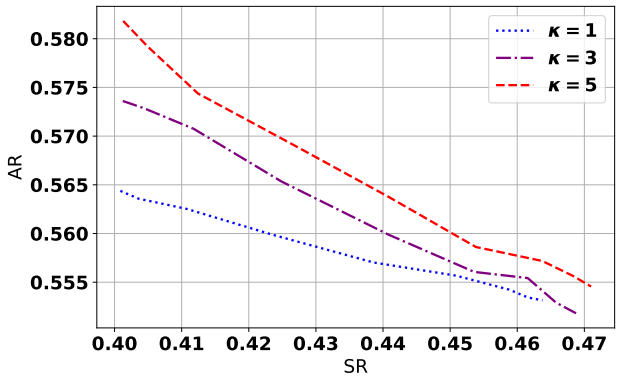


Figure 2: Approximation of Pareto boundaries of $(\mathrm{SR}, \mathrm{AR})$ arising in the general linear inverse problem for normalized matrices of different condition numbers. As the condition number increases, the boundary becomes steeper and moves farther up.

Furthermore, when $\lambda_{\min}(W_o(N))^{1/2} \geq \frac{\sigma_v}{\sigma_\wedge^2}$, *we have*

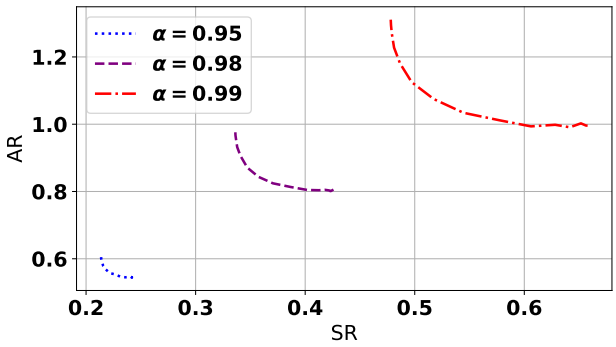
$$\begin{aligned}\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k) \leq \varepsilon \left(\frac{\sigma_\wedge^2 \lambda_{\min}(W_o(N))^{1/2}}{\sigma_\wedge^4 \lambda_{\min}(W_o(N)) + \sigma_v^2} \left(\rho^{2k} \sigma_0^2 + r_k(\rho)\, \sigma_w^2 \right) \right) \\ \times \left[2\sqrt{n} \left(\sigma_\vee^2 + \sigma_v^2 \left(\frac{\sigma_\wedge^2 \lambda_{\min}(W_o(N))^{1/2}}{\sigma_\wedge^4 \lambda_{\min}(W_o(N)) + \sigma_v^2} \right)^2 \right)^{1/2} \right. \\ \left. + \varepsilon \left(\frac{\sigma_\wedge^2 \lambda_{\min}(W_o(N))^{1/2}}{\sigma_\wedge^4 \lambda_{\min}(W_o(N)) + \sigma_v^2} \right) \right].\end{aligned}$$

The upper bound on the gap decreases as the minimum singular value of the observability gramian increases. This indicates that as the observability of the system becomes uniformly strong, the gap between standard and adversarial risk for the nominal Kalman estimator will decrease. Perhaps counter-intuitively, when observability is uniformly poor, i.e., when $\lambda_{\min}(W_o(N))$ is small, increasing the sensor noise σ_v will actually *decrease* the above upper bound, as long as $\lambda_{\min}(W_o(N)) \geq \frac{\sigma_v}{\sigma_\wedge^2}$. This aligns, however, with results demonstrating that injecting artificial noise can improve the robustness of state observers (Doyle and Stein, 1979), and is further consistent with our interpretation of noise as a regularizer following Lemma 3.4.

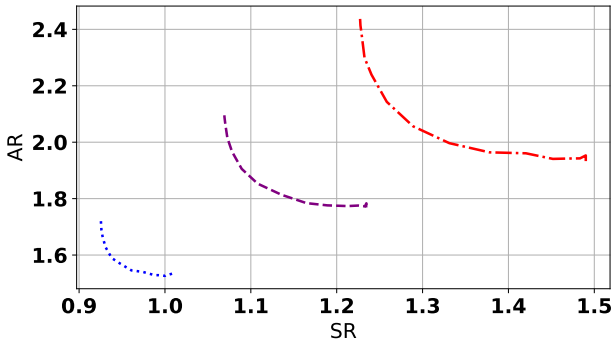
4 Numerical Results

We now demonstrate that the theoretical results shown in the previous two sections predict the tradeoffs arising in linear inverse problems. Additional experiments can be found in the appendix.

In Figure 2, we demonstrate that the tradeoffs curves depend on the condition number of the nominal solution $A_\star$. We approximate the $(\mathrm{SR}, \mathrm{AR})$ Pareto boundary for a linear inverse problem, where $A_\star \in \mathbb{R}^{4\times 4}$ is generated randomly by multiplying a prescribed positive diagonal matrix on the left and right by random orthgonal matrices, such that it has condition number 1, 3, or 5. We also set $\Sigma_x = \Sigma_w = 0.1I$, and provide the adversary a budget of $\varepsilon = 0.25$.



(a) Initial state estimation $k = 0$



(b) Final state estimation $k = N$

Figure 3: Pareto boundaries of (SR, AR) for initial and final state estimation. When α increases, the tradeoff curve becomes larger, enabling the existence of Pareto optimal point with very suboptimal values of SR and AR.

Figure 3 shows tradeoff curves for initial and final state estimation for systems with varying observability for the system $(A, C, \Sigma_0, \Sigma_w, \Sigma_v, N) = \left(\begin{bmatrix} \alpha & \beta \\ -\beta & \alpha \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix}, I, 0.1I, 0.1, 5\right)$ where $\alpha^2+\beta^2 = 1$, where we vary α. The adversary's power is set to $\varepsilon = 0.5$. As α approaches one, the system loses observability, seen by the fact that the minimum eigenvalues of the observability gramian become small. In particular, for $\alpha = 0.95$, $\alpha = 0.98$, $\alpha = 0.99$, the minimum eigenvalues of the observability grammian are given by $1.22, 0.81$ and 0.58 respectively. The results therefore support §3.4, where we showed shrinking the eigenvalues of $W_o(N)$ increases the severity of the tradeoff between SR and AR.

In Figure 4, we demonstrate impact of adversaries on the risk incurred by an estimator optimized for SR versus AR. We consider initial state estimation of a system defined by $(A, C, \Sigma_0, \Sigma_w, \Sigma_v, N) = \left(\begin{bmatrix} 1 & \rho \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & 0 \end{bmatrix}, I, 0.1I, 0.1, 5\right)$, where we vary ρ from 0.1 to $\sqrt{10}$, such that the eigenvalues of $W_o(N)$ decrease as ρ increases. The adversary's power is fixed at $\varepsilon = 0.5$. Evaluating the nominal and robust estimators on this class of systems, we see that the adversarially robust smoother has significantly smaller adversarial risk compared to the nominal Kalman smoother when observability is low. As observability increases, this advantage shrinks. This suggests that adversarially robust state estimation is disproportionately important when observability is low.

5 Conclusion

We analyzed the robustness-accuracy tradeoffs arising in linear inverse problems. We did this in two parts. Firstly, we provided an algorithm to solve for the optimal adversarial perturbation, which can be used to trace out the Pareto boundary. Secondly, we bounded the gap between the adversarial and standard risk in terms of the spectral properties of the underlying linear model. These bounds generalize the robustness-accuracy tradeoffs arising in the classification and linear regression settings. We then specialized these general results to the tradeoffs arising in adversarial state estimation of dynamical systems, where we demonstrated the accuracy decrease of the nominal Kalman estimator under adversarial measurements can be bounded by the eigenvalues

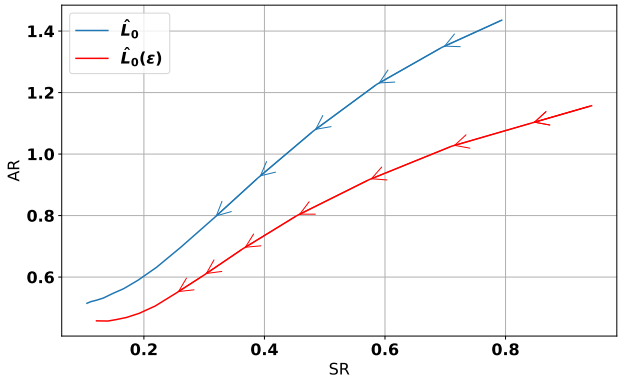


Figure 4: SR vs AR for a nominal Kalman smoother and an adversarially robust smoother, where observability increases in the direction of the arrows. When observability is low, the nominal smoother has lower standard risk than the robust smoother, but a significantly higher adversarial risk. This difference shrinks as observability increases.

of the observability gramian, a control-theoretic object. An interesting avenue of future work is to have an analytic characterization of the solutions A_λ along the Pareto boundary of the linear inverse problem, such that we can not only bound the fragility of the nominal solution, but also the conservatism of adversarially robust solutions. On the control side, we pave the way for numerous avenues for future work including the analysis of robustness-performance tradeoffs in LQR control. It would also be interesting to study the tradeoffs arising in estimation from an infinite horizon perspective.

References

A. A. Al Makdah, V. Katewa, and F. Pasqualetti. Accuracy prevents robustness in perception-based control. In *2020 American Control Conference (ACC)*, pages 3940–3946. IEEE, 2020.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL `https://doi.org/10.1137/16M1080173`.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

N. Carlini and D. Wagner. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*, 2016.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy*, pages 39–57. IEEE, 2017.

M. Carrasco, J.-P. Florens, and E. Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6: 5633–5751, 2007.

L. Chen, Y. Min, M. Zhang, and A. Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pages 1670–1680. PMLR, 2020.

S. A. Deka, D. M. Stipanović, and C. J. Tomlin. Dynamically computing adversarial perturbations for recurrent neural networks. *arXiv preprint arXiv:2009.02874*, 2020.

E. Dobriban, H. Hassani, D. Hong, and A. Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

J. Doyle and G. Stein. Robustness with observers. *IEEE Transactions on Automatic Control*, 24 (4):607–611, 1979. doi: 10.1109/TAC.1979.1102095.

H. W. Engl, C. Flamm, P. Kügler, J. Lu, S. Müller, and P. Schuster. Inverse problems in systems biology. *Inverse Problems*, 25(12):123014, 2009.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

B. Hassibi, T. Kailath, and A. H. Sayed. *Indefinite-quadratic estimation and control: a unified approach to H2 and H∞ theories*. SIAM studies in applied and numerical mathematics, 1999.

S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

A. Javanmard, M. Soltanolkotabi, and H. Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

P. P. Khargonekar, M. A. Rotea, and E. Baeyens. Mixed H2/H∞ filtering. *International Journal of Robust and Nonlinear Control*, 6(4):313–330, 1996.

M. Lutter, S. Mannor, J. Peters, D. Fox, and A. Garg. Robust value iteration for continuous control tasks. *arXiv preprint arXiv:2105.12189*, 2021.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

A. Mandlekar, Y. Zhu, A. Garg, L. Fei-Fei, and S. Savarese. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3932–3939. IEEE, 2017.

P. Nakkiran. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7 (15):510, 2008.

L. Pinto, J. Davidson, R. Sukthankar, and A. Gupta. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pages 2817–2826. PMLR, 2017.

A. Raghunathan, S. M. Xie, F. Yang, J. C. Duchi, and P. Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

A. Ribes and F. Schmitt. Linear inverse problems in imaging. *IEEE Signal Processing Magazine*, 25(4):84–99, 2008.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

C. Xie, M. Tan, B. Gong, A. Yuille, and Q. V. Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020.

H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

K. Zhou and J. C. Doyle. *Essentials of Robust Control*. Prentice-Hall, 1998.

Supplementary Material

A Proofs from Section 2

A.1 Proof of Correctness for Algorithm 1

Recalling the optimization problem

$$\begin{aligned} \underset{\delta \in \mathbb{R}^n}{\text{maximize}} \quad & \delta^\top A^\top A\delta - 2\delta^\top A^\top b \\ \text{subject to} \quad & \delta^\top \delta \le \varepsilon^2, \end{aligned} \tag{P}$$

and the corresponding KKT conditions:

$$\begin{aligned} 2(\lambda^* I - A^\top A)\delta^* + 2A^\top b &= 0 \\ \lambda^*({\delta^*}^\top \delta^* - \varepsilon^2) &= 0 \\ (\lambda^* I - A^\top A) &\succeq 0. \end{aligned}$$

The third condition implies that $\lambda^* \ge \sigma_1^2$. We assume $\sigma_1 > 0$, otherwise the problem is trivial. Then using the SVD of A to re-arrange the first stationarity condition, we get

$$(\lambda^* I - \Sigma^\top \Sigma) V^\top \delta^* = \Sigma^\top U^\top b.$$

Maximizing a convex function over a convex set achieves its maximum on the boundary; it suffices to search over $\delta^\top \delta = \varepsilon^2$. We now consider two cases: $\lambda^* > \sigma_1^2$ and $\lambda^* = \sigma_1^2$. In the first case, we know $\lambda^* I - \Sigma^\top \Sigma$ must be invertible, and thus

$$\begin{aligned} \delta^* &= -V(\lambda^* I - \Sigma^\top \Sigma)^{-1} \Sigma^\top U^\top b \\ \varepsilon^2 = {\delta^*}^\top \delta^* &= b^\top U \Sigma (\lambda^* I - \Sigma^\top \Sigma)^{-2} \Sigma^\top U^\top b \\ &= \sum_{i=1}^{\min\{n,p\}} \frac{(b^\top u_i)^2 \sigma_i^2}{(\lambda^* - \sigma_i^2)^2}, \end{aligned}$$

where u_i are the columns of U. Observe that

$$f(\lambda) := \sum_{i=1}^{\min\{n,p\}} \frac{(b^\top u_i)^2 \sigma_i^2}{(\lambda - \sigma_i^2)^2},$$

is a strictly monotonically decreasing function when $\lambda > \sigma_1^2$, and converges to 0 when $\lambda \to \infty$. This implies there is a unique λ^* such that $f(\lambda^*) = \varepsilon^2$, which can be numerically solved for in various ways, such as bisection.

Now we consider the case where $\lambda^* = \sigma_1^2$. In this case, δ^* will no longer be unique, and will come in the form

$$\delta^* = -V(\sigma_1^2 I - \Sigma^\top \Sigma)^\dagger \Sigma^\top U^\top b + cv,$$

where † denotes the Moore-Penrose pseudoinverse, and v is any unit vector lying in the null-space of $\left(\sigma_1^2 I - \Sigma^2\right) V^\top$, which is precisely characterized in this case by

$$\ker\left(\left(\sigma_1^2 I - \Sigma^2\right) V^\top\right) = \mathbf{span}\left(\left\{v_i : \sigma_i^2 = \sigma_1^2\right\}\right),$$

with v_i denoting the ith column of V. To find the appropriate scaling c, we observe

$$\begin{aligned}\delta^{*\top}\delta^* &= b^\top U\Sigma\left((\sigma_1^2 I - \Sigma^\top\Sigma)^\dagger\right)^2\Sigma^\top U^\top b + c^2\\ &= \sum_{i:\sigma_i<\sigma_1}\frac{(b^\top u_i)^2\sigma_i^2}{(\sigma_1^2-\sigma_i^2)^2} = \varepsilon^2\\ c &= \sqrt{\varepsilon^2 - \sum_{i:\sigma_i<\sigma_1}\frac{(b^\top u_i)^2\sigma_i^2}{(\sigma_1^2-\sigma_i^2)^2}}.\end{aligned}$$

Combining our precise characterization of $\ker\left(\left(\sigma_1^2 I - \Sigma^\top\Sigma\right)V^\top\right)$ using the columns of V, and the formula for c, we can extract an optimal perturbation vector δ^*. Therefore, we have demonstrated that (P), as well as extracting its optimal solution, can be solved to arbitrary precision.

A.2 Proof of Theorem 2.1

First recall the definitions of SR and and AR:

$$\begin{aligned}\mathrm{SR}(A) &= \mathbb{E}\left[\|y - Ax\|_2^2\right]\\ \mathrm{AR}(A) &= \mathbb{E}\left[\max_{\|\delta\|_2\le\varepsilon}\|y - A(x+\delta)\|_2^2\right].\end{aligned}$$

Given fixed y, x, let us define the quantity $d(A) := y - Ax$. Writing out the inner maximization of AR we have:

$$\max_{\|\delta\|_2\le\varepsilon}\|y - A(x+\delta)\|_2^2 = \max_{\|\delta\|_2\le\varepsilon}\|d(A) - A\delta\|_2^2.$$

Observe that this is equivalent to the problem

$$\begin{aligned}\underset{s}{\text{minimize}}\quad & s\\ \text{subject to}\quad & s - \|d(A) - A\delta\|_2^2 \ge 0 \text{ for all } \delta^\top\delta \le \varepsilon^2.\end{aligned}\tag{P1}$$

We now recall the S-lemma for quadratic functions.

Lemma A.1 (S-lemma)**.** *Given quadratic functions $p(x), q(x) : \mathbb{R}^n \to \mathbb{R}$, suppose there exists x such that $p(x) > 0$. Then,*

$$p(x) \ge 0 \implies q(x) \ge 0 \text{ for all } x$$

if and only if

$$\exists t \ge 0 \text{ such that } q(x) \ge tp(x) \text{ for all } x.$$

Using this lemma, we set $p(\delta) = \varepsilon^2 - \delta^\top\delta$, $q(\delta) = s - \|d(A) - A\delta\|_2^2 \ge 0$. We observe that trivially, there exists $\delta = 0$ such that $p(\delta) > 0$. Now given feasible s for (P1), we observe that by our constraints, any δ such that $p(\delta) \ge 0$ immediately implies $q(\delta) \ge 0$. By the S-lemma, this is equivalent to the existence of some $t \ge 0$ such that

$$q(\delta) - tp(\delta) = s - \|d(A) - A\delta\|_2^2 - t(\varepsilon^2 - \delta^\top\delta) \ge 0$$

for all δ. Therefore, we can re-write the optimization problem (P1) into

$$\begin{aligned} \underset{s}{\text{minimize}} \quad & s \\ \text{subject to} \quad & \exists t \geq 0 \text{ s.t. } s - \|d(A) - A\delta\|_2^2 - t(\varepsilon^2 - \delta^\top \delta) \geq 0 \text{ for all } \delta. \end{aligned} \tag{P2}$$

Re-arranging the terms in the quadratic expression, we get:

$$\begin{aligned} s - \|d(A) - A\delta\|_2^2 - t(\varepsilon^2 - \delta^\top \delta) &= s - \left(\delta^\top A^\top A\delta - 2d(A)^\top A\delta + \|d(A)\|_2^2\right) - t(\varepsilon^2 - \delta^\top \delta) \\ &= \delta^\top \left(tI - A^\top A\right)\delta + 2d(A)^\top A\delta + \left(s - t\varepsilon^2 - \|d(A)\|_2^2\right). \end{aligned}$$

Now we recall a property of Schur complements.

Lemma A.2 (Schur Complement)**.** *Given $p(x) = x^\top Px + b^\top x + c$, we have*

$$\begin{aligned} p(x) \geq 0 \; \forall \; x \iff & \begin{bmatrix} P & b \\ b^\top & c \end{bmatrix} \succeq 0 \\ \iff & P \succeq 0, \; c - b^\top P^\dagger b \geq 0. \end{aligned}$$

Applying this to (P2), we see the constraints can be re-written

$$\begin{aligned} & \exists t \geq 0 \text{ s.t. } s - \|d(A) - A\delta\|_2^2 - t(\varepsilon^2 - \delta^\top \delta) \geq 0 \text{ for all } \delta \\ \iff \quad & \exists t \geq 0, \; tI - A^\top A \succeq 0, \; s - t\varepsilon^2 - \|d(A)\|_2^2 - d(A)^\top A(tI - A^\top A)^\dagger A^\top d(A) \geq 0 \\ \iff \quad & \exists t \geq \lambda_{\max}(A^\top A), \; s - t\varepsilon^2 - \|d(A)\|_2^2 - d(A)^\top A(tI - A^\top A)^\dagger A^\top d(A) \geq 0. \end{aligned}$$

Therefore, we get the optimization problem

$$\begin{aligned} \underset{s,t}{\text{minimize}} \quad & s \\ \text{subject to} \quad & t \geq \lambda_{\max}(A^\top A) \\ & s - t\varepsilon^2 - \|d(A)\|_2^2 - d(A)^\top A(tI - A^\top A)^\dagger A^\top d(A) \geq 0. \end{aligned}$$

However, this is clearly equivalent and has the same optimal value as the following problem

$$\begin{aligned} \underset{t}{\text{minimize}} \quad & t\varepsilon^2 + \|d(A)\|_2^2 + d(A)^\top A(tI - A^\top A)^\dagger A^\top d(A) \\ \text{subject to} \quad & t \geq \lambda_{\max}(A^\top A). \end{aligned}$$

Notice that so far we are simply considering equivalent formulations to the original optimization. The ensuing step is where the lower and upper bounds 6 and 7 arise. Recall the Neumann series, where since $t \geq \lambda_{\max}(A^\top A)$, we have

$$\begin{aligned} (tI - A^\top A)^{-1} &= \frac{1}{t}\left(I - \frac{1}{t}A^\top A\right)^{-1} \\ &= \frac{1}{t}\left(I + \frac{1}{t}A^\top A + \frac{1}{t^2}\left(A^\top A\right)^2 + \cdots\right). \end{aligned}$$

From the Neumann series, we see that we can upper and lower bound the inverse using geometric series of the largest and smallest eigenvalues of $A^\top A$, respectively,

$$\frac{1}{t - \lambda_{\min}(A^\top A)} I \preceq (tI - A^\top A)^{-1} \preceq \frac{1}{t - \lambda_{\max}(A^\top A)} I$$

From now on, we will deal with the inverse, since instead of the pseudo-inverse we can take the infimum of the above problem, which is bounded from below. Let us consider the lower bound first–the upper bound follows using the exact same techniques. We have

$$\begin{aligned} t\varepsilon^2 + \|d(A)\|_2^2 + d(A)^\top A(tI - A^\top A)^{-1}A^\top d(A) &\geq t\varepsilon^2 + \|d(A)\|_2^2 + d(A)^\top A\left(\frac{1}{t-\lambda_{\min}(A^\top A)}I\right)A^\top d(A) \\ &= t\varepsilon^2 + \|d(A)\|_2^2 + \frac{1}{t-\lambda_{\min}(A^\top A)}\left\|A^\top d(A)\right\|_2^2. \end{aligned}$$

Therefore, we have

$$\max_{\|\delta\|_2\leq\varepsilon}\|x_k - A(Y_T+\delta)\|_2^2 \geq \|d(A)\|_2^2 + \min_{t>\lambda_{\max}(A^\top A)} t\varepsilon^2 + \frac{1}{t-\lambda_{\min}(A^\top A)}\left\|A^\top d(A)\right\|_2^2.$$

We now make a second relaxation:

$$\begin{aligned} \|d(A)\|_2^2 + \min_{t>\lambda_{\max}(A^\top A)} t\varepsilon^2 + \frac{1}{t}\left\|A^\top d(A)\right\|_2^2 &\geq \|d(A)\|_2^2 + \min_{t\geq 0} t\varepsilon^2 + \frac{1}{t-\lambda_{\min}(A^\top A)}\left\|A^\top d(A)\right\|_2^2 \\ &= \|d(A)\|_2^2 + 2\varepsilon\left\|A^\top d(A)\right\|_2 + \varepsilon^2\lambda_{\min}(A^\top A), \end{aligned}$$

which we get by deriving the unconstrained minimizer $t^* = \frac{\left\|A^\top d(A)\right\|_2}{\varepsilon} + \lambda_{\min}(A^\top A)$. Now putting expectations on both sides of the inequality, we get

$$\mathrm{AR}(A) \geq \mathrm{SR}(A) + 2\varepsilon\,\mathbb{E}\left[\left\|A^\top(x_k - AY_T)\right\|_2\right] + \varepsilon^2\lambda_{\min}(A^\top A).$$

■

A.3 Proof of Corollary 2.2

We note that the upper and lower bounds boil down to upper and lower bounding $\mathbb{E}\left[\left\|A_\star^\top(y - A_\star x)\right\|_2\right]$. We first observe $A_\star^\top(y - A_\star x) = A_\star^\top w \sim \mathcal{N}(0, \sigma_w^2 A_\star^\top A_\star)$. The upper bound therefore comes from an application of Jensen's inequality:

$$\begin{aligned} \mathbb{E}\left[\left\|A_\star^\top w\right\|_2\right] &= \mathbb{E}\left[\sqrt{\|A_\star^\top w\|_2^2}\right] \\ &\leq \sqrt{\mathbb{E}\left[\|A_\star^\top w\|_2^2\right]} \\ &= \sqrt{\mathrm{tr}\,(A_\star^\top A_\star)}. \end{aligned}$$

For the lower bound, we first consider the full SVD of $A_\star = U\Sigma V^\top$, where U, V are orthogonal matrices. Therefore, we have

$$\begin{aligned} \left\|A^\top w\right\|_2^2 &= \left\|\Sigma U^\top w\right\|_2^2 \\ &= \sum_{i=1}^{p}\sigma_i^2 w_i^2. \end{aligned}$$

Using the ℓ^2-ℓ^1 equivalence of norms, we have

$$\sqrt{\sum_{i=1} \sigma_i^2 w_i^2} \geq \frac{1}{\sqrt{p}} \sum_{i=1}^{p} \sigma_i |w_i|.$$

Putting this together, we have

$$\mathbb{E}\left[\left\|A_\star^\top w\right\|\right] \geq \frac{1}{\sqrt{p}} \mathbb{E}\left[\sum_{i=1}^{p} \sigma_i |w_i|\right] = \frac{1}{\sqrt{p}} \mathbb{E}\left[|w_1|\right] \sum_{i=1}^{p} \sigma_i,$$

where $w_1 \sim \mathcal{N}(0,1)$. The quantity $\mathbb{E}\left[|w_1|\right]$ is the mean of a folded standard normal, and we may express $\sum_{i=1}^{p} \sigma_i = \operatorname{tr}\left((A_\star^\top A_\star)^{1/2}\right)$, which yields us the desired lower bound. ■

B Kalman Filtering State Space Solution

Consider the setting defined in §3.1, i.e. we have a dynamical system which progresses according to

$$\begin{aligned} x_{t+1} &= Ax_t + w_t \\ y_t &= Cx_t + v_t \end{aligned}$$

with $x_0 \sim \mathcal{N}(0, \Sigma_0)$, $w_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_w)$, $v_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_v)$. Consider estimating state x_k given measurements $y_0, \ldots, y_k$. The minimum mean square estimator is given by

$$\hat{x}_k = \min_z \mathbb{E}\left[\|z - x_k\|_2^2 \,|y_0, \ldots, y_k\right].$$

This can be written as an integral

$$\min_z \int_{\mathbb{R}^n} \|z - x_k\|_2^2 \, f(x_k|y_0, \ldots, y_k) dx_k$$

where f denotes the conditional density of x_k given the measurements. The objective is convex in z, and thus we can find the minimizer by setting the gradient with respect to z to zero. In particular, we have

$$\begin{aligned} \frac{d}{dz} \int_{\mathbb{R}^n} \|z - x_k\|_2^2 \, f(x_k|y_0, \ldots, y_k) &= \int_{\mathbb{R}^n} \frac{d}{dz} \|z - x_k\|_2^2 \, f(x_k|y_0, \ldots, y_k) dx_k \\ &= \int_{\mathbb{R}^n} 2(z - x_k) f(x_k|y_0, \ldots, y_k) dx_k \\ &= 0 \end{aligned}$$

where dominated convergence theorem permits the exchange of integration and differentiation. Therefore, the state estimate may be expressed

$$\hat{x}_k = \int_{\mathbb{R}^n} x_k f(x_k|y_0, \ldots, y_k) dx_k = \mathbb{E}\left[x_k|y_0, \ldots y_k\right].$$

Thus we can determine the state estimates $\hat{x}_k$ as the mean of the conditional distribution $f(x_k|y_0, \ldots y_k)$. This may be computed recursively. In particular, let $x_k|y_{0:k}$ be the random variable with probability

density function $f(x_k|y_0, \dots y_k)$. Then for all k, we have that $x_k|y_{0:k} = \mathcal{N}(\hat{x}_k^+, P_k^+)$ where

$$\begin{aligned} x_0^- &= 0 \\ P_0^- &= \Sigma_0 \\ \hat{x}_k^+ &= x_k^- + P_k^- C^\top \left(CP_k^- C^\top + \Sigma_v\right)^{-1} \left(y_k - C\hat{x}_k^-\right) \\ P_k^+ &= P_k^- - P_k^- C^\top \left(CP_k^- C^\top + \Sigma_v\right) CP_k^- \\ \hat{x}_{k+1}^- &= A\hat{x}_k^+ \\ P_{k+1}^- &= A^\top P_k^+ A + \Sigma_w \end{aligned} \tag{13}$$

To see that this is the case, recall that $x_0 \sim \mathcal{N}(0, \Sigma_0)$, by assumption. Now suppose that $x_k|y_{0:k-1} \sim \mathcal{N}(\hat{x}_k^-, P_k^-)$. Observe that

$$\begin{bmatrix} x_k|y_{0:k-1} \\ y_k \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \hat{x}_k^- \\ C\hat{x}_k^- \end{bmatrix}, \begin{bmatrix} P_k^- & P_k^- C^\top \\ CP_k^- & CP_k^- C^\top + \Sigma_v \end{bmatrix}\right)$$

Thus

$$\begin{aligned} x_k|y_{0:k} &\sim \mathcal{N}\left(x_k^- + P_k^- C^\top \left(CP_k^- C^\top + \Sigma_v\right)^{-1} \left(y_k - C\hat{x}_k^-\right), P_k^- - P_k^- C^\top \left(CP_k^- C^\top + \Sigma_v\right) CP_k^-\right) \\ &= \mathcal{N}(x_k^+, P_k^+) \end{aligned}$$

Now, given $x_k|y_{0:k} \sim \mathcal{N}(x_k^+, P_k^+)$, observe that $x_{k+1}|y_{0:k} = Ax_k|y_{0:k} + w_k \sim \mathcal{N}(A^\top x_k^+, A^\top P_k^+ A + \Sigma_w) = \mathcal{N}(x_{k+1}^-, P_k^-)$.

Using the equations in (13), we can write the Kalman filter as a state space system with inputs y_t. In particular, if P_k^- and P_k^+ are defined as in (13), we can let $K_k = P_k^- C^\top \left(CP_k^- C^\top + \Sigma_v\right)^{-1}$. Then we may express our state estimates using the following time varying system.

$$\hat{x}_{k+1} = (A - K_k CA)\hat{x}_k + K_k y_k.$$

C General Statements and Proofs from Section 3

C.1 Lemma 3.3

Lemma 3.3: Suppose $k \leq N$. The finite horizon Kalman state estimator is the solution to optimization problem (9), and is given by

$$\hat{L}_k = \left(A^k \Sigma_0 \mathcal{O}_N{}^\top + \Gamma_k \Sigma_w \tau_N{}^\top\right) \left(\mathcal{O}_N \Sigma_0 \mathcal{O}_N{}^\top + \tau_N \left(I_N \otimes \Sigma_w\right) \tau_N{}^\top + \left(I_{N+1} \otimes \Sigma_v\right)\right)^{-1}.$$

Proof: We know $\mathrm{SR}(L)$ is convex in L, thus we may take the derivative of $\mathrm{SR}(L)$ with respect to L and set it to 0 to solve for $\hat{L}_k$. Matrix derivatives can be found in Petersen et al. (2008).

C.2 Proof of Lemma 3.4

Proof: This follows simply by expanding the norm inside the expectation, and noticing that since x_0, W_N, V_N are defined to be zero-mean Gaussian random vectors, their cross terms vanish. More

precisely, we have

$$\begin{aligned}
\mathbb{E}\left[\|x_k - LY_N\|_2^2\right] &= \mathbb{E}\left[\operatorname{tr}\left((x_k - LY_N)(x_k - LY_N)^\top\right)\right] \\
&= \mathbb{E}\Bigg[\operatorname{tr}\Bigg(\left(A^k - L\mathcal{O}_N\right) x_0 x_0^\top \left(A^k - L\mathcal{O}_N\right)^\top + \left(\Gamma_k - L\tau_N\right) W_N W_N^\top (\Gamma_k - L\tau_N)^\top \\
&\quad + LV_N V_N^\top L^\top\Bigg)\Bigg] + 0 \\
&= \left\|\left(A^k - L\mathcal{O}_N\right)\Sigma_0^{1/2}\right\|_F^2 + \left\|\left(\Gamma_k - L\tau_N\right)\left(I_N \otimes \Sigma_w\right)^{1/2}\right\|_F^2 + \left\|L\left(I_{N+1} \otimes \Sigma_v\right)^{1/2}\right\|_F^2.
\end{aligned}$$

■

C.3 Theorem 3.1

Theorem 3.1. *For any $L \in \mathbb{R}^{n \times p(N+1)}$, the gap between $\mathrm{AR}(L)$ and $\mathrm{SR}(L)$ admits the following lower bound:*

$$\begin{aligned}
\mathrm{AR}(L) - \mathrm{SR}(L) &\geq 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon}{\sqrt{n}} \operatorname{tr}\left(\left(L^\top\left(S\Sigma_0 S^\top + T\left(I_N \otimes \Sigma_w\right)T^\top + L\left(I_{N+1} \otimes \Sigma_v\right)L^\top\right)L\right)^{1/2}\right) \\
&\geq 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon}{\sqrt{n}}\sigma_{\min}(\Sigma_v)^{1/2}\|L\|_F^2 \qquad (14)
\end{aligned}$$

where $S := A^k - L\mathcal{O}_N$, $T := \Gamma_k - L\tau_N$.

Proof: Applying the lower bound (6), we have

$$\mathrm{AR}(L) \geq \mathrm{SR}(L) + 2\varepsilon\,\mathbb{E}\left[\left\|L^\top(x_k - LY_T)\right\|_2\right].$$

Then to derive the lower bound (14), we observe that the random vector

$$\begin{aligned}
z &= L^\top(x_k - LY_T) \\
&= L^\top\left(Sx_0 - TW_T + LV_T\right),
\end{aligned}$$

is a zero-mean gaussian with covariance

$$\Sigma = L^\top\left(S\Sigma_0 S^\top + T\Sigma_w T^\top + L\Sigma_v L^\top\right)L.$$

We can also write $z = \Sigma^{1/2}w$ where $w \sim \mathcal{N}(0, I)$. Consider the diagonalization of $\Sigma^{1/2} = VSV^\top$. Then

$$\begin{aligned}
\mathbb{E}\left[\|z\|_2\right] &= \mathbb{E}\left[\left\|VSV^\top w\right\|_2\right] \\
&= \mathbb{E}\left[\left\|\sum_i s_i w_i v_i\right\|_2\right],
\end{aligned}$$

where v_i is the ith row of V and s_i is the ith singular value of $\Sigma^{1/2}$. We have that

$$\left\|\sum_i s_i w_i v_i\right\|_2^2 = \sum_i s_i^2 w_i^2 v_i^\top v_i = \sum_i s_i^2 w_i^2$$

We have by equivalence of norms, $\sqrt{\sum_{i=1}^n x_i^2} \geq n^{-1/2} \sum_{i=1}^n |x_i|$. Therefore,

$$\sqrt{\sum_{i=1}^n s_i^2 w_i^2} \geq \frac{1}{\sqrt{n}} \sum_{i=1}^n s_i |w_i|,$$

and thus

$$\mathbb{E}\left[\|z\|_2\right] \geq \frac{1}{\sqrt{n}} \mathbb{E}\left[\sum_{i=1}^n s_i |w|_i\right] = \frac{1}{\sqrt{n}} \mathbb{E}\left[|w|\right] \sum_{i=1}^n s_i$$

where $w \sim \mathcal{N}(0,1)$. The quantity $\mathbb{E}\left[|w|\right]$ is the expected value of a folded standard normal, which is $\sqrt{\frac{2}{\pi}}$, while $\sum_{i=1}^n s_i = \operatorname{tr}\left(\Sigma^{1/2}\right)$. Putting this together, we have that

$$\mathrm{AR}(L) - \mathrm{SR}(L) \geq 2\sqrt{\frac{2}{\pi}} \frac{\varepsilon}{\sqrt{n}} \operatorname{tr}\left(\left(L^\top \left(S\Sigma_0 S^\top + T\left(I_N \otimes \Sigma_w\right) T^\top + L\left(I_{N+1} \otimes \Sigma_v\right) L^\top\right) L\right)^{1/2}\right).$$

From the above bound, we may now derive a cruder lower bound from which we can observe a dependence on the singular values of the observability grammian, $W_o(N)$. In particular, begin with the expression above, and note that the terms involving Σ_0 and Σ_w are positive definite to achieve a lower bound in terms of L:

$$\begin{aligned}
&2\sqrt{\frac{2}{\pi}} \frac{\varepsilon}{\sqrt{n}} \operatorname{tr}\left(\left(L^\top \left(S\Sigma_0 S^\top + T\left(I_N \otimes \Sigma_w\right) T^\top + L\left(I_{N+1} \otimes \Sigma_v\right) L^\top\right) L\right)^{1/2}\right) \\
&\geq 2\sqrt{\frac{2}{\pi}} \frac{\varepsilon}{\sqrt{n}} \sigma_{\min}\left(\Sigma_v\right)^{1/2} \operatorname{tr}\left(\left(L^\top L L^\top L\right)^{1/2}\right) \\
&= 2\sqrt{\frac{2}{\pi}} \frac{\varepsilon}{\sqrt{n}} \sigma_{\min}\left(\Sigma_v\right)^{1/2} \|L\|_F^2 .
\end{aligned}$$

which completes the proof of inequality (14). ■

Introducing additional notation to express the Kalman estimator will be helpful in subsequent sections. Let

$$\bar{\Sigma} := \begin{bmatrix} \Sigma_0 & \\ & I_N \otimes \Sigma_w \end{bmatrix} \tag{15}$$

$$H := \begin{bmatrix} I & & & \\ A & I & & \\ \vdots & & \ddots & \\ A^N & A^{N-1} & \dots & I \end{bmatrix} \bar{\Sigma}^{1/2}$$

$$H_k := E_k^\top H = \begin{bmatrix} A^k & A^{k-1} & \dots & I & 0 & \dots & 0 \end{bmatrix} \bar{\Sigma}^{1/2}$$

$$M := H^\top \left(C^\top \otimes I\right) = \bar{\Sigma}^{1/2} \begin{bmatrix} \mathcal{O}_N^\top \\ \left(\mathcal{Z}\mathcal{O}_N\right)^\top \\ \vdots \\ \left(\mathcal{Z}^N \mathcal{O}_N\right)^\top \end{bmatrix},$$

where $\mathcal{Z} \in \mathbb{R}^{p(N+1) \times p(N+1)}$ is a block downshift operator, with blocks of size m. With this notation, the Kalman estimator given in Lemma 3.3 may be rewritten more compactly as

$$\hat{L}_k = H_k M \left(M^\top M + \Sigma_v\right)^{-1}.$$

C.4 Theorem 3.2

Theorem 3.2. *Suppose that $\hat{L}_k$ is the Kalman estimator from Lemma 3.3. Then we have the following bound on the gap between* AR *and* SR*.*

$$\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k) \geq 2\sqrt{\frac{2}{\pi}}\frac{\varepsilon}{\sqrt{n}}\sigma_{\min}\left(\Sigma_v\right)\|C\|_F^2 \left(\frac{\sigma_{\min}\left(A^k\Sigma_0 {A^k}^\top + \sum_{i=1}^k A^{k-i}\Sigma_w {A^{k-i}}^\top\right)}{(N+1)\left\|\bar{\Sigma}\right\|_2^2 \|W_o(N)\|_F + \|\Sigma_v\|_2}\right)^2 .$$

Proof: We begin by writing the Kalman estimator using the notation defined in (15)

$$L_k = H_k M\left(M^\top M + I_{N+1}\otimes\Sigma_v\right)^{-1}.$$

Suppose the rank of M is m. Then the singular value decomposition of M can be taken to be

$$U\begin{bmatrix} S & 0 \\ 0 & 0\end{bmatrix}V^\top = M$$

where $S = \mathrm{diag}\left(\begin{bmatrix} s_1 & s_2 & \dots s_m\end{bmatrix}\right)$ with $s_1 \geq s_2 \geq \cdots \geq s_m \geq 0$, while $U \in \mathbb{R}^{n(N+1)\times n(N+1)}$ and $V \in \mathbb{R}^{p(N+1)\times p(N+1)}$. We can now lower bound the Frobenius norm of L_k as follows.

$$\|L_k\|_F^2 \geq \|H_k M\|_F^2 \,\sigma_{\min}\left\{\left(M^\top M + \Sigma_v I\right)^{-1}\right\}^2 .$$

Note that $\sigma_{\min}\left\{\left(M^\top M + \Sigma_v I\right)^{-1}\right\} \geq \sigma_{\min}\left\{\left(M^\top M + \|\Sigma_v\|_2 I\right)^{-1}\right\}$. Therefore

$$\begin{aligned}\|L_k\|_F^2 &\geq \sigma_{\min}\left\{\left(\begin{bmatrix} S^2 & \\ & 0\end{bmatrix} + \|\Sigma_v\|_2 I\right)^{-2}\right\}\|H_k M\|_F^2 \\ &= \sigma_{\min}\left\{\left(\begin{bmatrix} S^2 & \\ & 0\end{bmatrix} + \|\Sigma_v\| I\right)^{-2}\right\}\left\|H_k H^\top\left(C^\top\otimes I\right)\right\|_F^2 \\ = \sigma_{\min}&\left\{\left(\begin{bmatrix} S^2 & \\ & 0\end{bmatrix} + \|\Sigma_v\|_2 I\right)^{-2}\right\}\left\|\begin{bmatrix} A^k\Sigma_0 C^\top & \dots & A^k\Sigma_0\left(A^N\right)^\top C^\top + \sum_{i=1}^k A^{k-i}\Sigma_w\left(A^{N-i}\right)^\top C^\top\end{bmatrix}\right\|_F^2\end{aligned} \tag{16}$$

Now observe that

$$\sigma_{\min}\left\{\left(\begin{bmatrix} S^2 & \\ & 0\end{bmatrix} + \|\Sigma_v\| I\right)^{-2}\right\} = \frac{1}{\left(s_1^2 + \|\Sigma_v\|_2\right)^2} \tag{17}$$

Also note that $s_1 = \|M\|_2 \leq \left\|\bar{\Sigma}^{1/2}\right\|_2 \left\|\bar{\Sigma}^{-1/2}M\right\|_F$. We have that

$$\left\|\bar{\Sigma}^{-1/2}M\right\|_F^2 = \left\|\begin{bmatrix} \mathcal{O}_N^\top \\ \left(\mathcal{Z}\mathcal{O}_N\right)^\top \\ \vdots \\ \left(\mathcal{Z}^N\mathcal{O}_N\right)^\top\end{bmatrix}\right\|_F^2 \leq \sum_{i=0}^N \left\|\mathcal{Z}^i\mathcal{O}_N\right\|_F^2 \leq (N+1)\left\|\mathcal{O}_N\right\|_F^2 .$$

Then

$$s_1 \leq \left\|\bar{\Sigma}^{1/2}\right\|_2 \sqrt{N+1}\,\|\mathcal{O}_N\|_F \tag{18}$$

When $k \geq 0$, we have

$$\begin{aligned}
\left\| \begin{bmatrix} A^k \Sigma_0 C^\top & \dots & A^k \Sigma_0 \left(A^N\right)^\top C^\top + \sum_{i=1}^k A^{k-i} \Sigma_w \left(A^{T-i}\right)^\top C^\top \end{bmatrix} \right\|_F^2 \\
&\geq \left\| \left(A^k \Sigma_0 \left(A^k\right)^\top + \sum_{i=1}^k A^{k-i} \Sigma_w \left(A^{k-i}\right)^\top \right) C^\top \right\|_F^2 \\
&\geq \sigma_{\min} \left(A^k \Sigma_0 \left(A^k\right)^\top + \sum_{i=1}^k A^{k-i} \Sigma_w \left(A^{k-i}\right)^\top \right)^2 \|C\|_F^2 .
\end{aligned}$$

In conjunction with (16), (17), and (18), this leads to (12). ■

C.5 Proof of Theorem 3.3

Theorem 3.3. *For any $L \in \mathbb{R}^{n \times p(N+1)}$, the following bound holds*

$$\mathrm{AR}(L) - \mathrm{SR}(L) \leq 2\varepsilon \|L\|_2 \left\| \Sigma^{1/2} \right\|_F + \varepsilon^2 \|L\|_2^2$$

where $\Sigma^{1/2}$ is the symmetric square root of the covariance of $x_k - LY_N$.

Proof: By Theorem 2.1,

$$\mathrm{AR}(L) - \mathrm{SR}(L) \leq 2\varepsilon \mathbb{E}\left[\|L(x_t - LY_N)\|_2\right] + \varepsilon^2 \lambda_{\max}\left(L^\top L\right) \leq 2\varepsilon \|L\|_2 \mathbb{E}\left[\|x_t - LY_N\|_2\right] + \varepsilon^2 \|L\|_2^2$$

We can upper bound $\mathbb{E}\left[\|x_t - LY_N\|_2\right]$ by bounding the expectation of the euclidean norm of a normal random variable. In particular, let w be a n dimensional standard normal random variable, so that $\mathbb{E}\left[\|x_t - LY_N\|_2\right] = \mathbb{E}\left[\left\|\Sigma^{1/2} w\right\|_2\right]$, where Σ is defined as the covariance of $x_k - LY_N$, and $\Sigma^{1/2}$ is its symmetric square root. Let $US^{1/2}U^\top := \Sigma^{1/2}$ be the eigenvalue decomposition of $\Sigma^{1/2}$ so that

$$\left\| \Sigma^{1/2} z \right\|_2 = \left\| U S^{1/2} U^\top w \right\|_2$$

Now define $z = U^\top w$. We have that $z \sim N(0, I)$. Then the above quantity equals $\sqrt{\left\|US^{1/2}w\right\|_2^2}$. Jensen's inequality tells us that

$$\mathbb{E}\left[\sqrt{\left\|US^{1/2}w\right\|_2^2}\right] \leq \sqrt{\mathbb{E}\left[\left\|US^{1/2}w\right\|_2^2\right]} = \sqrt{\mathbb{E}\left[w^\top S^{1/2} w\right]} = \left\|S^{1/2}\right\|_F = \left\|\Sigma^{1/2}\right\|_F ,$$

from which the theorem follows. ■

C.6 Theorem 3.4

Theorem 3.4. *Suppose that $\hat{L}_k$ is the Kalman state estimator given by Lemma 3.3. Then the gap between $\mathrm{AR}(\hat{L}_k)$ and $\mathrm{SR}(\hat{L}_k)$ is upper bounded by*

$$\begin{aligned}
\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k) \geq \epsilon \left(\frac{\left\| A^k \Sigma_0 \left(A^k\right)^\top + \sum_{i=1}^k A^{k-i} \Sigma_w \left(A^{k-i}\right)_2^\top \right\|_2}{\sigma_{\min}(W_o(N))^{1/2} \sigma_{\min}\left(\bar{\Sigma}\right)} \right) \\
\times \left(2\sqrt{n} \left(\left\|\bar{\Sigma}\right\|_2 + \left(\frac{\sqrt{\|\Sigma_v\|_2}}{\sigma_{\min}(W_o(N))^{1/2} \sigma_{\min}\left(\bar{\Sigma}\right)} \right)^2 \right)^{1/2} + \epsilon \left(\frac{1}{\sigma_{\min}(W_o(N))^{1/2} \sigma_{\min}\left(\bar{\Sigma}\right)} \right) \right) .
\end{aligned}$$

Furthermore, if $\lambda_{\min}(W_o(N))^{1/2} \geq \sigma_v/\sigma_{\min}\left(\bar{\Sigma}\right)$*, and defining* $\kappa = \frac{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)}{\lambda_{\min}(W_o(N))\sigma_{\min}\left(\bar{\Sigma}\right)^2+\sigma_v}$*, we get the bound*

$$\begin{aligned}\mathrm{AR}(\hat{L}_k) - \mathrm{SR}(\hat{L}_k) \leq \varepsilon &\left(\kappa\sqrt{\left\|A^k\Sigma_0\left(A^k\right)^\top + \sum_{i=1}^{k} A^{k-i}\Sigma_w\left(A^{k-i}\right)_2^\top\right\|_2}\right) \\ &\times \left(2\sqrt{n}\left(\sigma_{\max}(\bar{\Sigma})^2 + \sigma_{\min}(\Sigma_v)\kappa^2\right)^{1/2} + \varepsilon\kappa\right)\end{aligned}$$

Proof: By Theorem 3.3, upper bounding the gap between $\mathrm{AR}(L_k)$ and $\mathrm{SR}(L_k)$ reduces to upper bounding $\left\|\Sigma^{1/2}\right\|_F$ and $\|L_k\|_2$. First consider $\left\|\Sigma^{1/2}\right\|_F$. Equivalence of norms tells us that

$$\left\|\Sigma^{1/2}\right\|_F \leq \sqrt{n}\left\|\Sigma^{1/2}\right\|_2 = \sqrt{n}\left\|\Sigma\right\|_2^{1/2}. \tag{19}$$

Recalling the notation defined in (15), the Kalman filter may be expressed as $L_k = H_k M\left(M^\top M + \Sigma_v\right)^{-1}$. We may also express Σ in terms of this notation: $\Sigma = \left(H_k - L_k M\right)\bar{\Sigma}\left(H_k - L_k M\right)^\top + L_k\left(I \otimes \Sigma_v\right){L_k}^\top$. To upper bound the spectral radius of this, we can leverage triangle inequality and submultiplicativity

$$\begin{aligned}\|\Sigma\|_2 &\leq \left\|\left(H_k - L_k M\right)\bar{\Sigma}\left(H_k - L_k M\right)^\top\right\|_2 + \left\|L_k\left(I \otimes \Sigma_v\right){L_k}^\top\right\|_2 \\ &\leq \left\|\bar{\Sigma}\right\|_2\left\|H_k - L_k M\right\|_2^2 + \left\|\Sigma_v\right\|_2\left\|L_k\right\|_2^2.\end{aligned}$$

Note that $H_k - L_k M = H_k - H_k M\left(M^\top M + \Sigma_v\right)^{-1} M = H_k\left(I - M\left(M^\top M + \Sigma_v\right)^{-1} M^\top\right)$. Then by submultiplicativity,

$$\left\|H_k\left(I - M\left(M^\top M + \Sigma_v\right)^{-1} M^\top\right)\right\|_2 \leq \|H_k\|_2\left\|I - M\left(M^\top M + \Sigma_v\right)^{-1} M^\top\right\|_2 \leq \|H_k\|_2.$$

We can further upper bound $\|H_k\|_2$ in terms of system properties. In particular, we have

$$\|H_k\|_2 = \sqrt{\|H_k\|_2^2} = \sqrt{\left\|H_k H_k^\top\right\|_2} = \sqrt{\left\|A^k\Sigma_0\left(A^k\right)^\top + \sum_{i=1}^{k} A^{k-i}\Sigma_w\left(A^{k-i}\right)_2^\top\right\|_2}.$$

Thus

$$\|\Sigma\|_2 \leq \left\|\bar{\Sigma}\right\|_2\|H_k\|_2^2 + \|\Sigma_v\|_2\|L\|_2^2 \leq \left\|\bar{\Sigma}\right\|_2\left\|A^k\Sigma_0\left(A^k\right)^\top + \sum_{i=1}^{k} A^{k-i}\Sigma_w\left(A^{k-i}\right)_2^\top\right\|_2 + \|\Sigma_v\|_2\|L\|_2^2. \tag{20}$$

Next we obtain a bound on $\|L_k\|_2$. As in the proof of Theorem 3.2, we will assign $m := \mathrm{rank}(M)$ and take the singular value decomposition of M to be $U\begin{bmatrix}S & 0\\ 0 & 0\end{bmatrix}V^\top$ where $S = \mathrm{diag}\left(\begin{bmatrix}s_1 & \dots & s_m\end{bmatrix}\right)$

with $s_1 \geq s_2 \geq \cdots \geq s_m \geq 0$, while $U \in \mathbb{R}^{n(N+1)\times n(N+1)}$ and $V \in \mathbb{R}^{p(N+1)\times p(N+1)}$.

$$\begin{aligned}
\|L_k\|_2 &= \left\| H_k M \left(M^\top M + \Sigma_v \right)^{-1} \right\|_2 \leq \|H_k\|_2 \left\| M \left(M^\top M + \sigma_{\min}\left(\Sigma_v\right) \right)^{-1} \right\|_2 \\
&= \|H_k\|_2 \left\| U \begin{bmatrix} S & \\ & 0 \end{bmatrix} V^\top \left(V \left(\begin{bmatrix} S^2 & \\ & 0 \end{bmatrix} + \sigma_{\min}(\Sigma_v) \right) V^\top \right)^{-1} \right\| \\
&= \|H_k\|_2 \left\| U \begin{bmatrix} S & \\ & 0 \end{bmatrix} \left(\begin{bmatrix} S^2 & \\ & 0 \end{bmatrix} + \sigma_{\min}(\Sigma_v) \right)^{-1} V^\top \right\|_2 \\
&= \|H_k\|_2 \left\| S(S^2 + \sigma_{\min}(\Sigma_v))^{-1} \right\|_2 \\
&\leq \|H_k\|_2 \max_{1\leq k\leq m} \frac{s_k}{s_k^2 + \sigma_{\min}(\Sigma_v)}.
\end{aligned}$$

A simple bound on the last maximization would be

$$\max_{1\leq k\leq m} \frac{s_k}{s_k^2 + \sigma_{\min}(\Sigma_v)} \leq \max_{1\leq k\leq m} \frac{s_k}{s_k^2} \leq \frac{1}{s_m} \tag{21}$$

Note that $s_m \geq \lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)$, so $\frac{1}{s_m} \leq \frac{1}{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)}$. Then

$$\|L_k\|_2 \leq \frac{\|H_k\|_2}{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)} \leq \frac{\sqrt{\left\| A^k \Sigma_0 \left(A^k\right)^\top + \sum_{i=1}^k A^{k-i}\Sigma_w \left(A^{k-i}\right)_2^\top \right\|_2}}{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)}. \tag{22}$$

Then the first half of the theorem follows by combining the result of Theorem 3.3 with (19), (20) and (22).

However, if $\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right) \geq \sigma_v$, then the maximum (21) is attained at

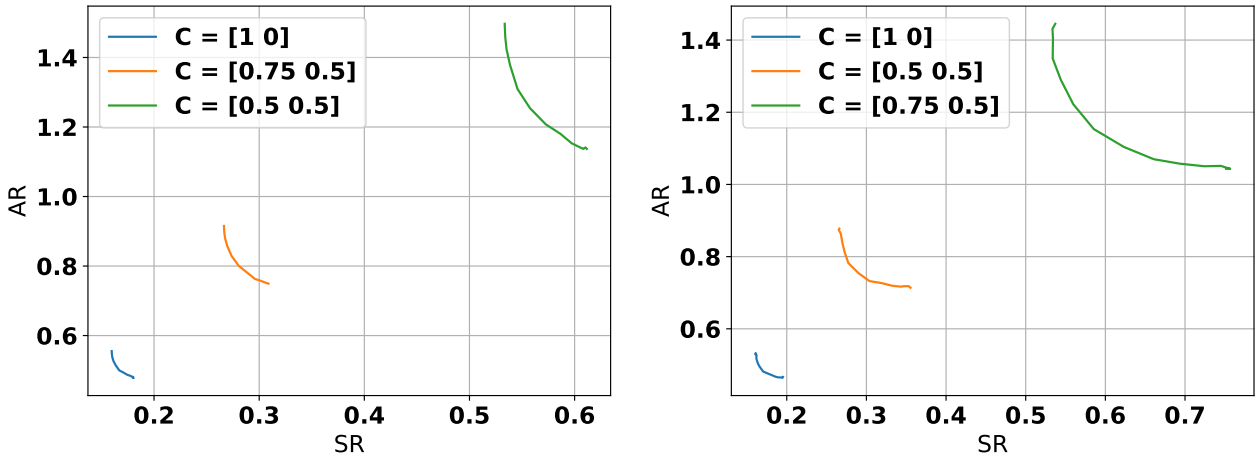
$$\begin{aligned}
\max_{1\leq k\leq m} \frac{s_k}{s_k^2 + \sigma_{\min}(\Sigma_v)} &= \frac{s_m}{s_m^2 + \sigma_v} \\
&\leq \frac{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)}{\lambda_{\min}(W_o(N))\sigma_{\min}\left(\bar{\Sigma}\right)^2 + \sigma_v}.
\end{aligned}$$

Therefore, when $\lambda_{\min}(W_o(N))^{1/2} \geq \sigma_v/\sigma_{\min}\left(\bar{\Sigma}\right)$, we have the more precise bound

$$\begin{aligned}
\|L_k\|_2 &\leq \|H_k\|_2 \frac{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)}{\lambda_{\min}(W_o(N))\sigma_{\min}\left(\bar{\Sigma}\right)^2 + \sigma_v} \\
&\leq \sqrt{\left\| A^k \Sigma_0 \left(A^k\right)^\top + \sum_{i=1}^k A^{k-i}\Sigma_w \left(A^{k-i}\right)_2^\top \right\|_2} \frac{\lambda_{\min}(W_o(N))^{1/2}\sigma_{\min}\left(\bar{\Sigma}\right)}{\lambda_{\min}(W_o(N))\sigma_{\min}\left(\bar{\Sigma}\right)^2 + \sigma_v},
\end{aligned}$$

which leads to the second half of the theorem.

■



(a) Linear estimator

(b) Neural Network

Figure 5: Pareto boundaries of $(\mathrm{SR}, \mathrm{AR})$ for initial state estimation for a variety of measurement matrices C by both a linear state estimator and a neural network (Lower bound on AR plotted for NN). As the first entry of C decreases, the tradeoff curve becomes more severe. The trade-offs are not alleviated by a nonlinear estimator.

D Additional Experiments

In Figure 5, we demonstrate that the fundamental tradeoffs are not overcome by using nonlinear state estimators. In particular, we consider the system and adversary defined by $(A, \Sigma_0, \Sigma_w, \Sigma_v, N) = \left(\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, I, 0.1I, 0.1, 5\right)$, and C as in the legend of the plots. We solve for the Pareto boundary for the linear estimator as in Figure 3. We also parametrize a two layer network with 10 neurons per layer, and perform a SGD procedure similar to that used in the linear case, with the exception being that we do not solve the exact adversary corresponding to each data point, but rather apply 100 steps of gradient ascent to find the adversarial perturbation. Once the neural network is trained, this same approach to find the adversary is used again to estimate the adversarial risk of the resultant estimator. Thus the tradeoff curves shown for the neural network use an under approximation to the true values of AR achieved. As such, it appears that the neural network is getting roughly the same tradeoff curves as the linear estimator. This may suggest that linear estimators are optimal, even for minimizing the adversarial risk.