Distorting an Adversary's View in Cyber-Physical Systems

Gaurav Kumar Agarwal, Mohammed Karmoose, Suhas Diggavi, Christina Fragouli, Paulo Tabuada
Department of Electrical and Computer Engineering, UCLA, Los Angeles, USA
Email: {gauravagarwal, mkarmoose, suhasdiggavi, christina.fragouli, tabuada}@ucla.edu

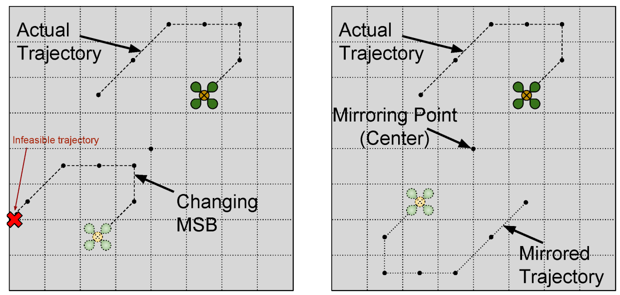
***Abstract*— In Cyber-Physical Systems (CPSs), inference based on communicated data is of critical significance as it can be used to manipulate or damage the control operations by adversaries. This calls for efficient mechanisms for secure transmission of data since control systems are becoming increasingly distributed over larger geographical areas. Distortion based security, recently proposed as one candidate for CPSs security, is not only more appropriate for these applications but also quite frugal in terms of prior requirements on shared keys. In this paper, we propose several distortion-based metrics to protect CPSs communication and show that it is possible to confuse adversaries with just a few bits of pre-shared keys.**

I. Introduction

Wireless networked environments are a natural host for a number of cyber-physical control applications, ranging from autonomous cars and drones, to the Internet-of-Things (IoT), to immersive environments such as augmented reality. It is well recognized that wireless networking is essential to realize the potential of new CPS applications, and is equally well recognized that private and secure exchange of information are necessary and not simply desirable conditions for the CPS ecosystem to thrive. For instance, personal health data in assisted environments, car positions and trajectories, proprietary interests, all need to be protected. This paper introduces a new approach to CPS security, that aims to distort an adversary's view of a control system's states.

Our starting observation is that information security measures (cryptographic and information theoretic secrecy), are not well matched to CPS applications as they impose unnecessary requirements, such as protecting all the raw data, and thus can cause high operational costs. Cryptographic methods rely on computational complexity: they require short keys, but high complexity at the communicating nodes (that can be simple sensors in some cases), and can impose a significant overhead on short packet transmissions, therefore increasing delay [1], [2], [3], [4]. Information theoretic methods rely on keys: they have low complexity and do not add packet overhead, but require the communicating nodes to share large keys - every communication link needs to use a shared secret key (for a one-time pad) of length equal to the entropy (effectively length) of the transmitted data [5]. These costs accumulate rapidly given that large CPS applications can have dense communication patterns.

Instead, we propose a lightweight approach, that uses small amounts of key and low complexity operations, and builds around a distortion measure. To illustrate[1], consider the following simple example of a drone flying motion, depicted in Fig. 1a. The drone starts at any position, and moves between adjacent points in the grid. It regularly communicates its location to a legitimate receiver, Bob; a passive eavesdropper, Eve, wishes to infer the drone's locations, and can perfectly overhear all the transmissions the drone makes. We assume the drone and Bob share just one bit of key, that is secret from Eve, and ask: what is the best use we can make of the key?



(a) (b)

Fig. 1: Example of drone motion: (a) protection of the most significant bit (b) mirroring based scheme.

Using the one bit of shared key to protect the most significant bit (MSB) is not a good solution. As shown in Fig. 1a, the adversary can discover the fake trajectory after a few time steps since this scheme leads to trajectories that do not adhere to the dynamics or environment constraints. At this point, it can learn the real trajectory by flipping back the MSB (we assume that the used scheme is known to everyone). Similar attacks can be made if we use a one-time pad [5] using the same keys over time: as time progresses, more fake trajectories can be discovered and discarded.

Conventional entropy measures also fail to provide insights on how to use the key. For instance, assume we label the 64 squares in Fig. 1a sequentially row per row, and consider two cases: in case I, Eve learns that the drone is in one of the neighboring squares $\{1, 2\}$, each with probability 1/2. For case II, Eve knows that the drone is in one of the squares $\{1, 64\}$, again each with probability 1/2. Both cases are equivalent from an information security perspective since in both cases Eve's uncertainty is a set of size 2 equiprobable elements and hence its entropy is 1. However, the security

The work was partially supported by NSF grant #1740047, by the Army Research Laboratory under Cooperative Agreement W911NF-17-2-0196, and by the UC-NL grant LFR-18-548554

[1] Although we illustrate our approach for a specific simple example, it extends to protecting general system states.

risks in both situations are different. For example, if Eve aims to take a photo of the drone, in the first case she knows where to turn her camera (squares 1 and 2 are close by) while in the second case, she does not (squares 1 and 64 are far apart).

Instead, we propose to use an Euclidean distance distortion measure: how far (in Euclidean distance) is Eve's estimate from the actual location. We then propose encoding/decoding schemes which utilize the shared key to maximize this distance. We first consider an "average" distortion measure. Note that if Eve had not received any of the drone transmissions, then the best (adversarial) estimate of the drone's location at any given time is the center point of the confined region in Fig. 1a. Therefore, a good encryption scheme would strive to maintain Eve's estimate to be as close to the center point as possible; and we achieve the maximum possible distortion, if, after overhearing the drone's transmissions, Eve's best estimate still remains the center point.

The following scheme can achieve this maximum distortion by using exactly one bit of shared secret key. When encoding, the drone either sends its actual trajectory, or a "mirrored" version of it, depending on the value of the secret key. The mirrored trajectory is obtained by reflecting the actual trajectory across a mirroring point in space; in this example, the mirroring point is the center point as shown in Fig. 1b. Since Eve does not know the value of the shared key, its best estimate of the drone's location - after receiving the drone's transmissions - would be the average location given the trajectory and its mirrored version, which is exactly the center point. Our results in Section III extend this idea of mirroring to dynamical systems in higher dimensional spaces, and theoretically analyze the performance in terms of average distortion for a larger variety of distributions (with certain symmetry conditions).

Next, we consider a worst-case-sense distortion-based metric. In other words, our security metric is "in the worst case, how far is Eve's estimate from the actual location?" That is, the adversary's distortion may be different for different time instances and different instances of the actual trajectory, and we are interested in the minimum among these. In Section IV we provide encryption schemes that are suitable for maximizing this distortion metric and show that with 3 bits of shared key per dimension (i.e., 9 for three dimensional motion), our schemes achieve near-perfect worst case distortion. Our main contributions are as follows:

- We define security measures that are based on assessing the distortion: in the average sense over time and over data, and in the minimum sense, providing worst case guarantees at any time and for any particular instances of data.
- For the expected case distortion, we develop a mirroring based scheme which uses exactly one bit of key and can provide maximum possible distortion (equivalent to Eve with no observations) in some cases. We also discuss the cases where it is not optimal and give an analytical characterization of the attained distortion.
- For the worst case distortion, we design a scheme that uses 3 bits of key per dimension and prove it achieves the maximum possible distortion (equivalent to Eve with no observations) when the inputs to the systems are independent from the previous states.

A. Related Work

Secure data communication where the adversary has unlimited computational power is studied from the lens on information theory, most notably by Shannon [5] and Wyner [6]. The study of secure communication from a distortion angle is relatively new and is first studied by Yamamoto [7], where the goal is to maximize the distortion of an eavesdropper's estimate on a message. Schieler and Cuff [8] later showed that, in the limit of an infinite block length (n) code, only $\log(n)$ bits of secret keys are needed to achieve the maximum possible distortion. Schemes for causal encodings (finite block length) were considered in [9] for single shot communication and exponential benefits for each additional bit of shared key were discussed. However, the schemes do not translate to the scenarios where one has to communicate correlated temporal data like the state of a control system.

Secure communication in control systems is studied in [10], [11], [12], [13], [14], [15]. These works either provide distortion only at the steady state or use measures such as differential privacy (does not use keys) and weak information theoretic security; they sometimes also assume that Eve receives different (a subset of the) information than Bob. We allow the adversary to have knowledge about the system dynamics and thus could enable easy generalizations to cases where the adversary might have other side information. We analyze the worst case and expected cases separately and give distortion guarantees in both cases.

B. Notations

For a matrix A, we denote by A' the transpose of A; X and X_a denote random columns, and $X_a^b = [X_a' \, X_{a+1}' \, \cdots \, X_b']'$ for $b \geq a$ and $a, b \in \mathbb{Z}$; for any random vector Y, we denote the mean and covariance matrices of Y by μ_Y and R_Y respectively, thus for example, the mean and the covariance matrix of X_a^b will be denoted by $\mu_{X_a^b}$ and $R_{X_a^b}$ respectively; for a matrix A, we denote by A^r the r-th power of A; $[m] := \{1, 2, \ldots, m\}$ where $m \in \mathbb{Z}^+$.

II. System Model

System Dynamics. we consider the linear dynamical system,

$$X_{t+1} = AX_t + BU_t + w_t, \quad Y_t = CX_t + v_t, \qquad (1)$$

where $X_t \in \mathbb{R}^n$ is the state of the system at time t, $U_t \in \mathbb{R}^m$ is the input to the system at time t, $w_t \in \mathbb{R}^n$ is the process noise, and $v_t \in \mathbb{R}^p$ is the observation noise. We denote $X = X_1^T$, $U = U_1^{T-1}$ and $w = w_1^{T-1}$. Based on the initial state X_1 and target state X_T, the controller computes a sequence of inputs that moves the state from X_1 to X_T.

Communication and Adversary Models. At each time instance the system transmits information about its state to a legitimate receiver, which is referred to as Bob, via a noiseless link. This situation occurs for example when Bob is remotely monitoring the execution of the system as in SCADA systems or in the remote operation of drones. A malicious receiver, referred to as Eve, is assumed to eavesdrop on the communication between the system and Bob and is able to receive all transmitted signals. Eve is assumed to be passive: she does not actively communicate but is interested in learning the underlying system's states from $t = 1$ to T. We assume that the System and Bob have a shared key K which they use to encode/decode the transmitted messages.

Inputs and States Random Process Model. We assume that both receivers are only aware of the system model, the matrices A, B, C and the statistics of noises. Therefore, from the perspective of the receivers, the input and output sequences have random distributions which depend on A, B, C and the statistics of the noise. In addition to the process noise w, the joint distribution $f(X, U, w)$ depends on the initial and target states and the control law of the system. So, even in noiseless systems, X and U possess inherent randomness from a receiver's perspective due to its lack of knowledge about the control law and the initial and target states. In general, the control inputs U can be dependent on the system states X. However, knowing the marginal distribution of U in noiseless systems can specify the marginal distribution of X. This can be shown as $X_2^T = QU + \tilde{Q}W + \hat{Q}X_1$, where

$$Q=\begin{bmatrix} B & 0 & \cdots & 0 \\ AB & B & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A^{T-2}B & \cdots & AB & B \end{bmatrix},\ \tilde{Q}=\begin{bmatrix} I & 0 & \cdots & 0 \\ A & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A^{T-2} & \cdots & A & I \end{bmatrix}$$

and $\hat{Q} = [A' \ (A^2)' \ \cdots \ (A^{T-1})']'$. This implies that for noiseless systems, the marginal distribution of U would imply the marginal distribution of X_2^T for a given initial state X_1 and thus the marginal distribution of X. The mean and covariance matrix of X_2^T become $\mathbb{E}(X_2^T) = \mu_{X_2^T} = Q\mu_U + \hat{Q}\mu_{X_1}$ and $R_{X_2^T} = QR_UQ^T$.

Encoding Model. The system transmits a packet Z_t at each time step t. The t-th transmitted packet can be a function of all previous observations and the shared keys, thus, $Z_t := \mathcal{E}_t(Y_1^t, K)$, where $\mathcal{E}_t$ is the encoding function used at time t. We will denote Z_1^T by Z.

Bob/Eve Models of Decoding. Bob noiselessly receives the transmitted packets from the system, and decodes them using the shared key. Then, using the decoded information, it generates an estimate of the state transmissions of the system at times $t = 1, \cdots, T$. In this work we always require Bob to decode in a lossless manner (i.e., with zero distortion). Formally, $H(X_t|Z_1^t, K) = 0, \ \forall t \in [T]$, where H is the Shannon entropy.

Similarly, Eve also receives all transmissions from the system. However, unlike Bob, she does not have any knowledge about the key K. Therefore, Eve's estimate of X_t is $\hat{X}_t := \phi_t\left(Z_1^T\right), \quad t \in [T]$, where ϕ_t is the decoding function used by Eve at time t.

Distortion Metrics. We consider a distortion-based security metric which captures how far (in Euclidean distance) an estimate is from the actual value. More formally, for a given time instance t and a transmitted codeword Z_1^T, we define the following quantity,

$$D(t, Z_1^T) := \mathbb{E}_{X_t|Z_1^T} \left\| X_t - \hat{X}_t \right\|^2 \overset{(a)}{=} \mathrm{tr}\left(R_{X_t|Z_1^T}\right), \quad (2)$$

where (2) captures the distortion incurred by Eve's estimate of X_t. Equality in (a) follows because the best (minimizing) estimates of Eve at time t are, $\hat{X}_t = \phi_t\left(Z_1^T\right) = \mathbb{E}\left[X_t|Z_1^T\right]$.

Note that Bob is required to successfully estimate X_t knowing Z_1^t and the key. Therefore, for a given realization of the key, the encoding function can only map one X_t and that key realization to each value of Z_1^T. Therefore Eve realizes that only trajectories from a strict subset can be the true trajectory for a given Z_1^T: those are the ones which correspond to each key realization. Therefore, the expectation in (2) is in fact taken over the randomness in the key taking into account posterior probabilities given Z_1^T. If Eve does not have observations, the expectation is taken over X_t with prior distribution and will get $D(t, Z_1^T) = \mathrm{tr}(R_{X_t})$.

As $D(t, Z_1^T)$ is a function of time t and the transmitted sequence Z_1^T, we consider two overall distortion metrics: the "average case" distortion (denoted by D_E) where we take expectation over all possible Z_1^T and average out over time; and the "worst case" distortion (denoted by D_W) where we take minimum over all possible Z_1^T and time instances.

$$\text{Average Distortion} \ - \ D_E := \mathbb{E}_{Z_1^T}\left[\frac{1}{T}\sum_{t=1}^{T} D(t, Z_1^T)\right] \quad (3)$$

$$\text{Worst Case Distortion} \ - \ D_W := \min_{Z_1^T}\left[\min_{t\in[T]} D(t, Z_1^T)\right]. \quad (4)$$

It is worth to note that the definitions of D_E and D_W defined in (3) and (4) imply that Eve's state estimation must be associated to a time instance. In other words, making a random/constant estimation of the state hoping that it matches the actual state at some time will lead to high distortion values. Further, D_W can be defined even when there is no prior distribution on X_1^T. However, to compare it to the case when the adversary has no observations, we assume that X_1^T always have a known prior distribution.

Design Goals. Our goal is to choose the encoding and decoding functions, $\mathcal{E}_t$ and ϕ_t, so that Bob can decode loselessly while the distortion is maximized for Eve's estimate. In addition, we seek to achieve this with the minimum amount of shared keys K. In absence of any observation by Eve, these distortions will be,

$$D_E^{\max} = \frac{1}{T}\sum_{t=1}^{T} \mathrm{tr}(R_{X_t}), \quad D_W^{\max} = \min_{t\in[T]} \mathrm{tr}(R_{X_t}).$$

These will serve as upper bounds for our schemes, as

$$D_E = \frac{1}{T}\mathbb{E}_{Z_1^T}\sum_{t=1}^{T}\mathrm{tr}(R_{X_t|Z_1^T}) \overset{(a)}{\leq} \frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}(R_{X_t}) = D_E^{\max}, \tag{5}$$

$$\begin{aligned} D_W &= \min_{Z_1^T}\min_{t\in[T]}\mathrm{tr}(R_{X_t|Z_1^T}) \leq \min_{t\in[T]}\mathbb{E}_{Z_1^T}\left[\mathrm{tr}(R_{X_t|Z_1^T})\right] \\ &\overset{(b)}{\leq} \min_{t\in[T]}\mathrm{tr}(R_{X_t}) = D_W^{\max}, \end{aligned} \tag{6}$$

where (a) and (b) follows by noting that the trace of the conditional covariance matrix is a quadratic (convex) function in Z_1^T and therefore we can use Jensen's inequality.

III. Optimizing Average Distortion D_E

In this section, we assume that the control system in (1) is noise free, that is $v_t = w_t = 0$. In general, an observer would be used to reconstruct or estimate the state from observations. However, to simplify the exposition we assume the state can be directly measured (C is the identity map) although our results can be extended to the case of an arbitrary observable pair (A,C) in (1).

We now discuss our proposed scheme that uses one bit of shared key and show how the achieved distortion compares to the upper bound in (5).

Mirroring Scheme. This scheme works as follows:

$$Z_t = \begin{cases} X_t & \text{if } K = 0, \\ \tilde{X}_t & \text{if } K = 1, \end{cases} \quad \forall t \in [T], \tag{7}$$

where K is the shared bit and $\tilde{X}_t$ is the state vector X_t, mirrored across a particular affine subspace $\mathcal{S}_t$,

$$\mathcal{S}_t = \{x \in \mathbb{R}^n \mid S_t x = b_t\}, \tag{8}$$

where $S_t \in \mathbb{R}^{s_t x n}$ and $b_t \in \mathbb{R}^{s_t}$. Since every affine subspace can be written in terms of orthogonal vectors, we assume that $S_t S_t' = I$. It can be shown that the mirrored point $\tilde{X}_t$ is $(I - 2S_t'S_t)X_t + 2S_t'b_t$.

Example. Consider $X_t \in \mathbb{R}^2$ where $S_t = \frac{1}{\sqrt{2}}[-1 \ 1]$ and $b_t = 0$. Then $\tilde{X}_t$ corresponds to reflecting across a line that passes through the origin with a 45^o angle.

Before performance analysis, we highlight that the encoding/decoding complexity of our scheme is $O(n^2)$. The performance of such scheme is given in the next theorem.

Theorem 3.1 (Proof in Appendix V-A): The mirroring scheme with matrices S_t's and b_t's allows Bob to perfectly estimate X_t's, and the distortion in Eve's estimate is,

$$D_E = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_X\left[\frac{2f_X(\tilde{X})}{f_X(X) + f_X(\tilde{X})}\|S_t X_t - b_t\|^2\right], \tag{9}$$

where $\tilde{X} := [\tilde{X}_1' \ \tilde{X}_2' \ \cdots \ \tilde{X}_T']'$.

Assuming that $f_X(x)$ is known, then Theorem 3.1 provides a closed-form characterization of the achieved average distortion for any mirroring scheme with matrices S_t and b_t. Moreover, under some symmetry conditions on $f_X(x)$, the expression in (9) admits a simplified version which gives insights on the maximum achievable distortion. This is shown in the next corollary.

Corollary 3.2 (Proof in Appendix V-A): If the mirroring scheme matrices S_t and b_t in Theorem 3.1 are selected such that

$$f_X(X) = f_X(\tilde{X}), \quad \text{for all } X \in \mathbb{R}^{nT}, \tag{10}$$

then (9) becomes,

$$D_E = \frac{1}{T}\sum_{i=1}^{T}\mathrm{tr}\left(S_t R_{X_t} S_t' + (b_t - S_t\mu_{X_t})(b_t - S_t\mu_{X_t})'\right) \tag{11}$$

We can interpret (11) as follows. First, note that condition (10) implies $b_t = S_t\mu_{X_t}$. To see this, note that X and $\tilde{X}$ have the same distribution, hence the same mean. Therefore, we have

$$\begin{aligned} &\mathbb{E}X_t = (I - 2S_t'S_t)\mathbb{E}X_t + 2S_t'b_t \\ &\implies S_t'S_t\mu_{X_t} = S_t'b_t \implies S_tS_t'S_t\mu_{X_t} = S_tS_t'b_t \\ &\overset{(a)}{\implies} S_t\mu_{X_t} = b_t, \end{aligned}$$

where (a) follows because $S_tS_t' = I$. Assuming that condition (10) is met, then the distortion becomes $D_E = \frac{1}{T}\sum_{i=1}^{T}\mathrm{tr}(S_tR_{X_t}S_t')$. The achieved distortion therefore depends on the choice of S_t: if $S_t = I$ then the maximum distortion can be achieved by our mirroring scheme. However, such a choice of S_t may not be a feasible one to ensure that condition (10) is met, as we will see in the following examples. One case for which $S_t = I$ satisfies condition (10) and allows maximum distortion is when X is symmetrically distributed around a point. We show this in the next corollary.

Corollary 3.3: For a random vector $X \in \mathbb{R}^{Tn}$, if there exists a point $v \in \mathbb{R}^{Tn}$ for which $f_X(X) = f_X(2v - X)$, $\forall X \in \mathbb{R}^{Tn}$, then $D_E = \frac{1}{T}\mathrm{tr}(R_X) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{tr}(R_{X_t})$.

Proof: Since X and $2v-X$ have the same distributions, they will have the same mean. This implies that $v = \mu_X$. We then use the following mirroring scheme: $S_t = I$, $b_t = \mu_{X_t}$ for $t \in [1 : T]$. With this, we get $\tilde{X}_t = 2\mu_{X_t} - X_t$, and thus $\tilde{X} = 2\mu_X - X$ where $\tilde{X} := [\tilde{X}_1' \ \tilde{X}_2' \ \cdots \ \tilde{X}_T']'$ and $\mu_X := [{\mu_{X_1}}' \ {\mu_{X_2}}' \ \cdots \ {\mu_{X_T}}']'$. This implies,

$$f_X(X) = f_X(\tilde{X}), \ \forall X \in \mathbb{R}^{nT},$$

so condition (10) is met. Therefore the distortion becomes equal to $D_E^{\max}$. ■

A. Examples

In this section, we show the implications of our results in the context of a few examples.

Example 1. Consider an example where U is distributed as Gaussian with mean μ_U and covariance matrix R_U. Then for a zero initial state, X_2^T is also Gaussian distributed with mean $\mu_{X_2^T} = Q\mu_U$ and variance $R_{X_2^T} = QR_UQ^T$, as we assume noise to be zero. A Gaussian random vector satisfies the conditions in Corollary 3.3, and therefore we can get maximum distortion by setting $b_t = \mu_{X_t}$ and $S_t = I$.

The next example is based on a Markov-based model for the dynamical system. For this example, the following lemma is useful.

Lemma 3.4: Consider the random vectors X_t where the following conditions hold: 1) $f_{X_1}(x_1) = f_{X_1}(2\mu_1 - x_1)$ and 2) $f_{X_t|X_{t-1}}(x_t|x_{t-1}) = f_{X_t|X_{t-1}}(2\mu_t - x_t|2\mu_{t-1} - x_{t-1})$. Then for this case, $f_X(X) = f_X(2\mu - X)$, where $\mu = [\mu_1{}' \ \mu_2{}' \ \cdots \ \mu_T{}']'$. Therefore, by virtue of Corollary 3.3, mirroring schemes can achieve the maximum distortion.

Lemma 3.4 allows us to characterize the performance of the following example.

Example 2. Consider the following random walk mobility model. Let $a \in \mathbb{N}^+$, and X_t be its location at time t, then,

$$\begin{aligned} X_1 &\sim \text{Uni}([-a:a]) \\ X_t|X_{t-1} &\sim \text{Uni}([-a:a] \cap \{X_{t-1}-1, X_{t-1}, X_{t-1}+1\}). \end{aligned}$$

One can see that these distributions satisfy the conditions in Lemma 3.4. Therefore, one can set $b_t = \mu_t = 0$ and $S_t = 1$, which will achieve maximum distortion of D_E.

Example 3. Here we provide a numerical example which shows how our mirroring scheme performs for situations where we do not have an analytical handle on the state distributions. We assume the quadrotor dynamical system provided in [16, (4)]. The quadrotor moves in a 3-dimensional cubed space with a width, length and height of 2 meters, where the origin is the center point of the space. The quadrotor starts its trajectory from an initial point $(-1, y_1, z_1)$ and finishes its trajectory at a target point $(1, y_T, z_T)$ after T time steps, where the points y_1, z_1, y_T, z_T are picked uniformly at random in $[-1, 1]^4$. We assume that $T = 10$ time steps, and that the continuous model in [16, (4)] is discretized with a sample time of $T_s = 0.5$ seconds. We assume that the quadrotor encodes and transmits only the states which contain the location information (first three elements of the state vector X_t). The quadrotor is equipped with an LQR controller which designs the input sequence U_1^{T-1} as the solution of the following problem

$$\begin{array}{ll} \text{minimize} & \|U\|^2 + 10 \left\|X_2^{T-1}\right\|^2 \\ \text{subject to} & X_{t+1} = A^{\text{quad}} X_t + B^{\text{quad}} U_t, \quad \forall t \in [T-1] \\ & X_1 = \begin{bmatrix} -1 & y_1 & z_1 & 0 & \cdots & 0 \end{bmatrix}', \\ & X_T = \begin{bmatrix} 1 & y_T & z_T & 0 & \cdots & 0 \end{bmatrix}', \end{array} \tag{12}$$

where A^{quad} and B^{quad} are the quadrotor's discretized systems matrices. The remaining states of X_1 and X_T are set to zero to allow the drone to hover at the respective locations. We perform numerical simulation of the aforementioned setup: we run 2 millions iterations, where in each iteration a new initial and target points are picked, and the resultant trajectory is recorded. Based on the recorded data, we consider different mirroring schemes and numerically evaluate the attained distortion. To facilitate numerical evaluations, the simulation space is gridded into bins with 0.2 meters of separation, and the location of the drone at each trajectory is approximated to the nearest space bin.

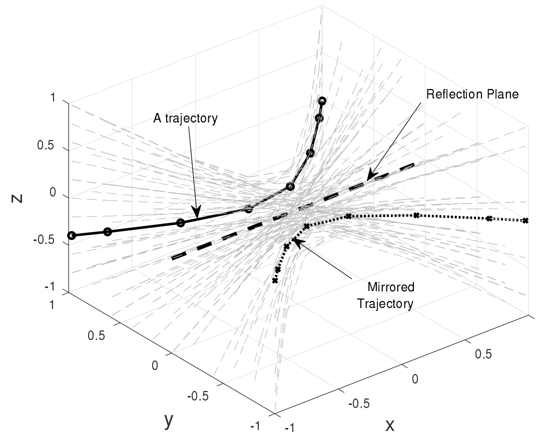


Fig. 2: An illustration of some trajectories. The reflection plane is shown as a dashed-black line. One trajectory (solid-black) is shown along with its mirrored image (dotted-black).

Figure 2 shows some of the drone trajectories obtained from our numerical simulation. It is clear that not all trajectories are equiprobable, and therefore the distribution of X_t is not uniform across all bins in space. However, the computation of $\mathbb{E}X_t$ shows the expected value of the position to be the origin. Moreover, since the motion of the drone is mainly progressive in the positive x-axis direction, reflection across the origin results in mirrored trajectories that are progressing in the opposite direction, and therefore are identified to be fake automatically. Therefore, mirroring across a point here is useless: the numerically computed distortion for this scheme is equal to zero.

Next we consider mirroring across the reflection plane shown in Figure 2, where $b_t = 0$ and $S_t = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. As can be seen from the figure, the reflection plane is indeed an axis of symmetry for the distribution of the drones trajectories, and therefore is expected to provide high distortion values. We numerically evaluate the attained distortion using the scheme by using equation (9), which evaluates to $D_E = 0.3971$. This is slightly less than $D_E^{\max} = 0.3979$.

Remark. Optimizing D_E requires the use of one bit of key. However, it may be insufficient for some application to allow the adversary to confuse the actual state between only two possibilities. For such scenarios, our scheme can be extended to larger uncertainty sets by using larger keys, along with symmetric rotations and mirroring.

IV. Optimizing The Worst Case Distortion D_W

In this section, we present an encryption scheme that attempts to maximize the worst case distortion for Eve. The main idea is to obfuscate the initial state in a such a way that Eve, even if she optimally uses her knowledge about the dynamics and her observations, her best estimate is close to the maximal distortion. We start by studying the problem of distorting the transmission of a single random variable in Theorems 4.2 and 4.3. These results then form the basis for maximizing the worst case distortion of a trajectory, as described in Theorem 4.4.

A. Building Step: Scalar Case

Consider the case where the system wants to communicate a single scalar random variable X to Bob by transmitting Z. The worst case distortion D_W for Eve will be $D_W = \min_Z \text{Var}(X|Z)$. Note that if Eve does not overhear Z, Eve will use as her estimate the minimum mean square estimate (the mean value), and thus experience a worst case distortion equal to the variance of X.

We first assume that $X \sim \mathcal{N}(0,1)$, and thus, the worst case distortion can not be larger than 1 by (6). We next develop our scheme progressively, from simple to more sophisticated steps. We will also use the following lemma.

Lemma 4.1: The variance of two real numbers a and b with probabilities p_a and p_b is given by $p_a p_b (a-b)^2$.

Mirroring or Shifting. Reflecting around the origin (as we did for optimizing the average case distortion in Section III) does not work well when X takes small values: indeed $\text{Var}(X|Z)$ is $Pr(X=Z|Z)(Pr(X=-Z|Z))(Z-(-Z))^2$ using Lemma 4.1 and as the plot in Fig. 3a shows has a worst case value that goes to zero as X approaches zero. To avoid this, we could try to use a "shifting" scheme where we add a constant θ to X whenever the shared key bit is one; but now this scheme does not perform well for large values of X, as X increases $\text{Var}(X|Z)$ goes to zero as in Fig. 3b.

Shifting+Mirroring. We here combine shifting and mirroring, in order to achieve a good performance for both small and large values of X. We start from the case where we have $k=1$ bit of key and then go to the case $k \geq 1$.

- $k=1$. We select a $\theta_1 \in \mathbb{R}$ that determines a window size (θ_1 is public and known by Eve). The encoding function is

$$Z = \mathcal{E}(X,K) = \begin{cases} X & \text{if } K=0 \\ -X & \text{if } K=1,\ |X|>\theta_1 \\ X+\theta_1 & \text{if } K=1,\ -\theta_1 \leq X < 0 \\ X-\theta_1 & \text{if } K=1,\ 0 \leq X < \theta_1 \end{cases}$$

We note that there is one particular value of X, $X=\theta_1$, which we do not transmit. Since this is of zero probability measure, it can be safely ignored. Given Z, there are two possibilities for X:

$$X \in \begin{cases} \{Z,-Z\} & \text{if } |Z|>\theta_1 \\ \{Z,Z+\theta_1\} & \text{if } -\theta_1 \leq Z < 0 \\ \{Z,Z-\theta_1\} & \text{if } 0 \leq Z < \theta_1. \end{cases}$$

Using the fact that $X \sim \mathcal{N}(0,1)$, we can calculate the posterior probabilities $Pr(X|Z)$ and use Lemma 4.1 to compute $\text{Var}(X|Z)$. Fig. 3c plots $\text{Var}(X|Z)$ for $\theta = 1.76$. The worst case distortion in this case becomes 0.4477, which is the best we can hope for if we have only one bit of shared key. This follows because for any mapping from X to Z, a transmitted symbol Z can have at most two pre-images (as Bob needs to reliably decode with one bit of key), and if one of these is $X=0$, then no matter what the second one is, the distortion corresponding to Z will be at most 0.4477. Equality occurs when the second pre-image of Z is ± 1.76, which is what our scheme also maps 0 to (using $\theta = 1.76$).

- $k \geq 1$. For $K \in \{0,1\}^k$, we use the following encoding:

$$\begin{aligned} Z &= \mathcal{E}(X,K) \qquad (13)\\ &= \begin{cases} \begin{cases} X & \text{if } K \leq 2^{k-1} \\ -X & \text{if } K > 2^{k-1} \end{cases} & |X| > \theta_k \\ X + K\frac{2\theta_k}{2^k} \mod [-\theta_k, \theta_k) & X \in [-\theta_k, \theta_k), \end{cases} \end{aligned}$$

where the optimal value of the constant θ_k depends on the number k of keys we have, K is the decimal equivalent of a binary string of length k, and $r \mod [a,b) = r - i(b-a)$ is such that i is an integer and $r - i(b-a) \in [a,b)$ for $r,a,b \in \mathbb{R}$. Intuitively, if $|X| > \theta_k$ then for half of the keys, we reflect across origin and for other half we do nothing; if $|X| < \theta_k$, we divide this window of size $2\theta_k$ into 2^k equal size windows and shift a point from one window to another by jumping K (in decimal) windows. An example for $k=2$ is shown in Fig. 4a for the key values $K=11$ and $K=10$. Fig. 4b plots D_W as a function of the number of keys k. Using $k=3$ and $\theta_3 = 4.84$ we achieve $D_W = 0.9998$ which is very close to 1, the best we can hope for.

Theorem 4.2: A Gaussian random variable with mean μ and variance σ^2 can be near perfectly (~ 0.9998 times the perfect distortion) distorted in worst case settings by just using three bits of shared keys.

Proof: Generate the random variable $V \sim \mathcal{N}(0,1)$ as $V = (X-\mu)/\sigma$ and encrypt it using $k=3$ key bits and the previously described scheme. For $c = 0.9998$ we have

$$\begin{aligned} D_W &= \min_Z \text{Var}(X|Z) = \min_Z \text{Var}(\sigma V + \mu|Z) \\ &= \sigma^2 \min_Z \text{Var}(V|Z) = c\sigma^2. \end{aligned}$$

■

Remark: We optimized the parameter θ_k of our scheme assuming Gaussian distribution. For other distributions, the optimal choice of θ_k and the corresponding worst case distortion would be different.

B. Vector Case and Time Series

Theorem 4.3 (Proof in Appendix V-B): For a Gaussian random vector $X \in \mathbb{R}^n$ with mean μ and a diagonal covariance matrix Σ we can achieve D_W within 0.9998 of the optimal by using $3n$ bits of shared keys.

This theorem uses our 3-bit encryption for each element in the vector. Assume now that this vector captures the probability distribution of the initial state of dynamical system; by encrypting this state we can guarantee the following.

Theorem 4.4 (Complete Proof in Appendix V-C): Using $3n$ bits of shared keys we can achieve D_W within 0.9998 of the optimal for the dynamical systems (1) with $C = I$, $v_t = 0$, singular values of A more than 1, and initial state $X_1 \sim \mathcal{N}(\mu, \Sigma)$, where Σ is diagonal covariance matrix, and U_t and w_t are independent of X_t.

Remark: Although the independence assumption on the inputs is rather restrictive, the result serves as a stepping stone towards understanding general cases.

Proof: The system transmits $Z_1 = f(Y_1, K) = f(X_1, K)$ where f is the encoding in Theorem 4.3, and

$$\begin{aligned} Z_{t+1} &= AZ_t + (Y_{t+1} - AY_t) \\ &= AZ_t + BU_t + w_t, \quad \forall t \in [T-1]. \end{aligned}$$

Bob can decode X_1 using Z_1 and K. Then:

$$\begin{aligned} \hat{X}_{t+1} &= Z_{t+1} - AZ_t + A\hat{X}_t \\ &= (AZ_t + BU_t + w_t) - AZ_t + A\hat{X}_t \\ &= AX_t + BU_t + w_t = X_{t+1}, \quad \forall t \in [T-1]. \end{aligned}$$

Eve's distortion is calculated in the Appendix V-C. ■

Complexity: $\mathcal{O}(n^2)$ per time instance for both encoding and decoding.

REFERENCES

[1] J. Wan, A. B. Lopez, and M. A. Al Faruque, "Exploiting wireless channel randomness to generate keys for automotive cyber-physical system security," in *Proceedings of the 7th International Conference on Cyber-Physical Systems*. IEEE Press, 2016, p. 13.

[2] B. Zan, M. Gruteser, and F. Hu, "Key agreement algorithms for vehicular communication networks based on reciprocity and diversity theorems," *IEEE Transactions on Vehicular Technology*, vol. 62, pp. 4020–4027, 2013.

[3] W. Trappe, R. Howard, and R. S. Moore, "Low-energy security: Limits and opportunities in the internet of things," *IEEE Security Privacy*, vol. 13, no. 1, pp. 14–21, Jan 2015.

[4] P. Koopman and C. Szilagyi, "Integrity in embedded control networks," *IEEE Security Privacy*, vol. 11, no. 3, pp. 61–63, May 2013.

[5] C. E. Shannon, "Communication theory of secrecy systems," *Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[6] A. D. Wyner, "The wire-tap channel," *The bell system technical journal*, vol. 54, no. 8, pp. 1355–1387, 1975.

[7] H. Yamamoto, "A rate-distortion problem for a communication system with a secondary decoder to be hindered," *IEEE Trans. Inf. Theory*, vol. 34, no. 4, pp. 835–842, 1988.

[8] C. Schieler and P. Cuff, "Rate-distortion theory for secrecy systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 12, pp. 7584–7605, 2014.

[9] C. Y. Tsai, G. K. Agarwal, C. Fragouli, and S. Diggavi, "A distortion based approach for protecting inferences," in *2017 IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 1913–1917.

[10] A. Tsiamis, K. Gatsis, and G. J. Pappas, "State-secrecy codes for networked linear systems," *CoRR*, vol. abs/1709.04530, 2017.

[11] ——, "State estimation with secrecy against eavesdroppers," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 8385 – 8392, 2017, 20th IFAC World Congress. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896317321559

[12] T. Tanaka, M. Skoglund, H. Sandberg, and K. H. Johansson, "Directed information and privacy loss in cloud-based control," in *2017 American Control Conference (ACC)*, May 2017, pp. 1666–1672.

[13] D. van Dooren, S. Schiessl, A. Molin, J. Gross, and K. H. Johansson, "Safety analysis for controller handover in mobile systems," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 10 090 – 10 095, 2017, 20th IFAC World Congress. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896317324035

[14] W. A. Malik, N. C. Martins, and A. Swami, *LQ Control under Security Constraints*. Heidelberg: Springer International Publishing, 2013, pp. 101–120. [Online]. Available: https://doi.org/10.1007/978-3-319-01159-2_6

[15] J. Corts, G. E. Dullerud, S. Han, J. L. Ny, S. Mitra, and G. J. Pappas, "Differential privacy in control and network systems," in *2016 IEEE 55th Conference on Decision and Control (CDC)*, Dec 2016, pp. 4252–4272.

[16] V. Kumar and N. Michael, "Opportunities and challenges with autonomous micro aerial vehicles," *The International Journal of Robotics Research*, vol. 31, no. 11, pp. 1279–1291, 2012.

V. APPENDICES

A. Proof of Theorem 3.1 and Corollary 3.2

We start by computing $R_{X_t|Z_1^T}$. Note that given a sequence of transmitted symbol Z_1^T there are two possible values of sequence of message symbols X_1^T which are $X_1^T = Z_1^T$ and $X_1^T = \tilde{Z}_1^T$, where $\tilde{Z}_t$ is the image of Z_t across the affine subspace given by $S_t x = b_t$.

With this, the posterior probability of $X_t = Z_t$ given Z_1^T i.e., $Pr(X_t = Z_t|Z_1^T)$ will be equal to $Pr(X_1^T = Z_1^T|Z_1^T) := p_Z$. We note that $p_Z = \frac{f(Z)}{f(Z)+f(\tilde{Z})}$, where $\tilde{Z} := [\tilde{Z}_1' \; \tilde{Z}_2' \; \cdots \; \tilde{Z}_T']'$. Then $\mathbb{E}(X_t|Z_1^T)$,

$$= p_Z Z_t + (1-p_Z)(\tilde{Z}_t) = Z_t + 2(1-p_Z)S_t'(b_t - S_t Z_t).$$

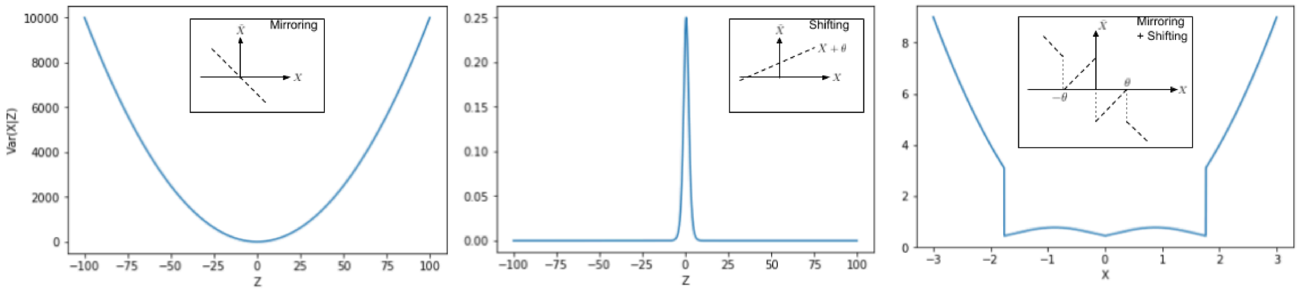
$$\begin{aligned} R_{X_t|Z_1^T} &= \mathbb{E}_{X_t|Z_1^T}\left[\left(X_t - \mathbb{E}(X_t|Z_1^T)\right)\left(X_t - \mathbb{E}(X_t|Z_1^T)\right)'\right] \\ &= p_Z\left(4(1-p_Z)^2\left(S_t'(b_t - S_tZ_t)(b_t - S_tZ_t)'S_t\right)\right) \\ &+ (1-p_Z)\left(4{p_Z}^2\left(S_t'(b_t - S_tZ_t)(b_t - S_tZ_t)'S_t\right)\right) \\ &= \underbrace{4p_Z(1-p_Z)}_{\eta(Z)} S_t'(b_t - S_tZ_t)(b_t - S_tZ_t)'S_t. \end{aligned}$$

$$\begin{aligned} D_E &= \mathbb{E}_Z \frac{1}{T}\sum_{t=1}^{T} \text{tr}\left(R_{X_t|Z_1^T}\right) \\ &= \mathbb{E}_Z \frac{1}{T}\sum_{t=1}^{T} \text{tr}\left(\eta(Z)S_t'(b_t - S_tZ_t)(b_t - S_tZ_t)'S_t\right) \end{aligned}$$

$$\begin{aligned} &= \mathbb{E}_Z \frac{1}{T}\sum_{t=1}^{T} \eta(Z)\text{tr}\left(S_t'(b_t - S_tZ_t)(b_t - S_tZ_t)'S_t\right) \\ &= \frac{1}{T}\mathbb{E}_Z\left[\sum_{t=1}^{T} \eta(Z)\|S_tZ_t - b_t\|^2\right] \\ &= \frac{1}{T}\mathbb{E}_Z\left[\sum_{t=1}^{T} 4p_Z(1-p_Z)\|S_tZ_t - b_t\|^2\right] \\ &= \frac{1}{T}\mathbb{E}_Z\left[\sum_{t=1}^{T} 4\frac{f_X(Z)f_X(\tilde{Z})}{(f_X(Z)+f_X(\tilde{Z}))^2}\|S_tZ_t - b_t\|^2\right]. \end{aligned}$$

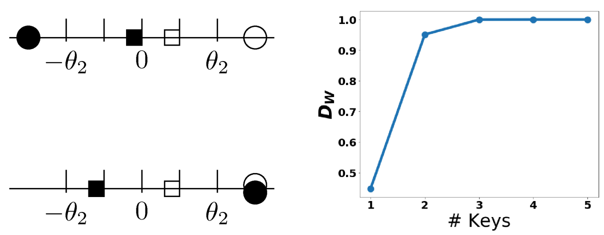
Now, Z_1^T is the transmitted symbols if $X_1^T = Z_1^T$ and key was zero or if $\{X_t = \tilde{Z}_t, \; \forall t \in [T]\}$ and key was one. So $f_Z(Z) = \frac{f_X(Z)+f_X(\tilde{Z})}{2}$. Thus D_E,

$$\begin{aligned} &= \frac{1}{T}\mathbb{E}_Z\left[\sum_{t=1}^{T} 4\frac{f_X(Z)f_X(\tilde{Z})}{(f_X(Z)+f_X(\tilde{Z}))^2}\|S_tZ_t - b_t\|^2\right] \\ &= \frac{1}{T}\int f_Z(Z)\left[\sum_{t=1}^{T} \frac{4f_X(Z)f_X(\tilde{Z})}{(f_X(Z)+f_X(\tilde{Z}))^2}\|S_tZ_t - b_t\|^2\right]dZ \\ &= \frac{1}{T}\int\left[\sum_{t=1}^{T} \frac{2f_X(Z)f_X(\tilde{Z})}{f_X(Z)+f_X(\tilde{Z})}\|S_tZ_t - b_t\|^2\right]\; dZ \\ &= \frac{1}{T}\mathbb{E}_X\left[\sum_{t=1}^{T} \frac{2f_X(\tilde{X})}{f_X(X)+f_X(\tilde{X})}\|S_tX_t - b_t\|^2\right], \end{aligned}$$



(a) (b) (c)

Fig. 3: Var$(X|Z)$ Vs Z (a) for mirroring based scheme; $D_W = 0$, (b) for shift based scheme; $D_W = 0$, (c) for mirroring+shift based scheme with $\theta_1 = 1.76$; $D_W = 0.4477$.



(a) (b)

Fig. 4: (a) Transparent shapes represent true values and solid shapes represent their respective mapping when two bit key is 11 and 10 respectively. (b) D_W, as a function of number of keys for optimal choice of θ_k.

which proves (9). Again, if we can choose S_t's and b_t's such that,

$$f_X(X) = f_X(\tilde{X}), \ \forall X \in \mathbb{R}^{nT},$$

the distortion D_E becomes,

$$\frac{1}{T}\mathbb{E}_X \sum_{t=1}^{T} \|S_t X_t - b_t\|^2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{X_t}\|S_t X_t - b_t\|^2$$
$$= \frac{1}{T}\sum_{t=1}^{T} \text{tr}\left(S_t R_{X_t} S_t' + (b_t - S_t\mu_{X_t})(b_t - S_t\mu_{X_t})'\right),$$

which proves (11).

B. *Proof for Theorem 4.3*

Let the shared key K is $(K_1, K_2, \ldots, K_n)$ where all K_i's are i.i.d. and uniformly distributed in $\{0,1\}^3$. Let us also assume that $X = (X^{(1)}, X^{(2)}, \ldots, X^{(n)})$, where each $X^{(i)} \in \mathbb{R}$. Similar to the scheme for scalar case, we create a random vector $V = (V^{(1)}, \ldots, V^{(n)})$ where,

$$V^{(i)} = (X^{(i)} - \mu^{(i)})/\sqrt{\Sigma_{ii}},$$

and encode $V^{(i)}$ using key K_i as in the case of a scalar for all $i \in [n]$. Thus, the distortion D_W will be,

$$= \min_Z \text{tr}(R_{X|Z}) = \min_Z \sum_{i=1}^{n} \text{Var}(X^{(i)}|Z)$$
$$= \min_Z \sum_{i=1}^{n} (\Sigma_{ii})\text{Var}(V^{(i)}|Z) = \sum_{i=1}^{n} (\Sigma_{ii}) \min_Z \text{Var}(V^{(i)}|Z)$$
$$= \sum_{i=1}^{n} (\Sigma_{ii}) \min_{Z^{(i)}} \text{Var}(V^{(i)}|Z^{(i)}) = c\sum_{i=1}^{n} (\Sigma_{ii}) = c \ \text{tr}(\Sigma),$$

where $c = 0.9998$. And since tr(Σ) is the expected distortion even when the adversary has no observations, and as we can not beat this by (6), this is optimal.

C. *Proof for Theorem 4.4*

Distortion at the adversary's end. Based on the coding scheme we can see that the adversary get $BU_t + w_t$ by just subtracting AZ_t from Z_{t+1} for $t \in [1 : T-1]$. So the adversary's information is given by following set:

$$E_{\text{info}} = \{Z_1, BU_t + w_t, \ t \in [1 : T-1]\}$$
$$= \{f(X_1, K), BU_t + w_t, \ t \in [1 : T-1]\}.$$

Thus, $D(t, Z_1^T) = D(t, E_{\text{info}}) = \text{tr}(R_{X_t|E_{\text{info}}})$.

Let's first compute $D(t = 1, Z_1^T)$,

$$D(t=1, Z_1^T) = \text{tr}(R_{X_1|E_{\text{info}}}) \overset{(a)}{=} \text{tr}(R_{X_1|f(X_1,K)}) \overset{(b)}{=} c \ \text{tr}(\Sigma),$$

where (a) is because U_t and w_t are independent on X_t and (b) is due to the encoding used in Theorem 4.3 with $c = 0.9998$.

Now, for other time instances we can use induction to prove that we will have worst case distortion at least $\text{tr}(\Sigma)$.

$$
\begin{aligned}
&D(t+1, Z_1^T) = \text{tr}(R_{X_{t+1}|E_{\text{info}}}) = \text{tr}(R_{(AX_t+BU_t+w_t)|E_{\text{info}}}) \\
&= \text{tr}(R_{(AX_t)|E_{\text{info}}}) = \text{tr}(AR_{X_t|E_{\text{info}}}A') = \text{tr}(A'AR_{X_t|E_{\text{info}}}) \\
&\overset{(a)}{=} \text{tr}(V\Lambda V' R_{X_t|E_{\text{info}}}) = \text{tr}(\Lambda V' R_{X_t|E_{\text{info}}} V) \\
&\overset{(b)}{=} \sum_{i=1}^{n} \lambda_i d_i (V' R_{X_t|E_{\text{info}}} V) \overset{(c)}{\geq} \sum_{i\in[n]} d_i (V' R_{X_t|E_{\text{info}}} V) \\
&\overset{(d)}{=} \sum_{i\in[n]} \nu_i (V' R_{X_t|E_{\text{info}}} V) \overset{(e)}{=} \sum_{i\in[n]} \nu_i (R_{X_t|E_{\text{info}}}) \\
&= \text{tr}(R_{X_t|E_{\text{info}}}) \overset{(f)}{\geq} c\ \text{tr}(\Sigma),
\end{aligned}
$$

where in (a), we do eigenvalue decomposition of $A'A$ which is a positive definite matrix and thus will have non negative eigenvalues; in (b) $d_i(V'R_{X_t|E_{\text{info}}}V)$ is the i-th diagonal entry of $V'R_{X_t|E_{\text{info}}}V$; (c) is true because $V'R_{X_t|E_{\text{info}}}V$ is a positive definite matrix and all the diagonal entries of a positive semi definite matrix are non-negative and because of our assumption that singular values of A, i.e. the square root of eigenvalues of $A'A$ are all more than one; (d) is because summation of eigenvalues is equal to the sum of all the diagonal entries for any square matrix, where $\nu_i(V'R_{X_t|E_{\text{info}}}V)$ is the i-th eigenvalue of $V'R_{X_t|E_{\text{info}}}V$; (e) is because a unitary transform preserve the eigenvalues. Finally (f) is because of the induction step.