From Parameter Estimation to Dispersion of Nonstationary Gauss-Markov Processes

Peida Tian, *Student Member, IEEE,* Victoria Kostina, *Member, IEEE*

***Abstract*—This paper provides a precise error analysis for the maximum likelihood estimate $\hat{a}(\boldsymbol{u})$ of the parameter a given samples $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ drawn from a nonstationary Gauss-Markov process $U_i = aU_{i-1} + Z_i,\ i \geq 1$, where $a > 1$, $U_0 = 0$, and Z_i's are independent Gaussian random variables with zero mean and variance σ^2. We show a tight nonasymptotic exponentially decaying bound on the tail probability of the estimation error. Unlike previous works, our bound is tight already for a sample size of the order of hundreds. We apply the new estimation bound to find the dispersion for lossy compression of nonstationary Gauss-Markov sources. We show that the dispersion is given by the same integral formula derived in our previous work [1] for the (asymptotically) stationary Gauss-Markov sources, i.e., $|a| < 1$. New ideas in the nonstationary case include a deeper understanding of the scaling of the maximum eigenvalue of the covariance matrix of the source sequence, and new techniques in the derivation of our estimation error bound.**

***Index Terms*—Parameter estimation, maximum likelihood estimator, unstable processes, finite blocklength analysis, lossy compression, sources with memory, rate-distortion theory, covering in stochastic processes.**

I. Introduction

In parameter estimation, given a sample $\boldsymbol{x}$ drawn according to a distribution $f(\theta_0; \boldsymbol{x})$ in the family $\{f(\theta; \boldsymbol{x}) : \theta \in \Theta\}$, where the set Θ is known and θ_0 is unknown but nonrandom, the goal is to design a good estimate for the underlying parameter θ_0. Many problems in science and engineering can be formulated in the form of parameter estimation, including binary hypothesis testing in statistics (with $|\Theta| = 2$) and system identification in the study of dynamical systems [2]. In this paper, we investigate the use of parameter estimation in the nonasymptotic analysis for lossy compression of nonstationary scalar Gauss-Markov sources.

A scalar Gauss-Markov process $\{U_i\}_{i=1}^{\infty}$ is a random process defined as

$$U_i = aU_{i-1} + Z_i, \quad i \geq 1, \tag{1}$$

where $U_0 = 0$ and Z_i's are independent Gaussian random variables with zero mean and variance σ^2, $Z_i \sim \mathcal{N}(0, \sigma^2)$. We assume without loss of generality that $a \geq 0$ [1]. We make the distinctions between the following three cases: the (asymptotically) stationary case refers to $0 < a < 1$ in (1); the unit-root case to $a = 1$ [2]; and the nonstationary case to $a > 1$. This paper mostly focuses on the nonstationary case.

Gauss-Markov processes have been extensively studied by researchers from many different fields. In the statistical analysis of time series, the Gauss-Markov process is a special case of the autoregressive–moving-average (ARMA) models [3, Chap. 3]. In economics, the process (1) with $a = 1$ is used to model the stochastic structure of the velocity of money [4], see also [5, Sec. 5]. In information theory, Burg's maximum entropy theorem [6, Th. 12.6.1] states that a p-th order Gauss-Markov process attains the maximum entropy rate among all stochastic processes under the autocorrelation constraints $\mathbb{E}[X_i X_{i+k}] = \alpha_k,\ k = 0, \ldots, p,\ \forall i$.

Our primary motivation for studying the Gauss-Markov process is to understand the role of memory in nonasymptotic rate-distortion theory. The Gauss-Markov process (1) is one of the simplest models for information sources with memory. The *rate-distortion function* (RDF) [7] captures the rate-distortion tradeoff when the coding length tends to infinity. The RDF is known in a few cases including the Gauss-Markov process [8]. The central question in nonasymptotic rate-distortion theory is to characterize the rate-distortion tradeoff when the coding length is constrained to be finite, and the *dispersion* is the main quantity of interest. The dispersion of stationary memoryless sources was found in [9], [10]. The dispersion of information sources with memory is largely unknown. Our previous work [1] found the dispersion of the stationary Gauss-Markov source. One of the key ideas in [1] is to construct a typical set based on $\hat{a}(\boldsymbol{u})$, the maximum likelihood estimate (MLE) of a given samples $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$. For a typical $\boldsymbol{u}$, $\hat{a}(\boldsymbol{u})$ is close to a.

The MLE $\hat{a}(\boldsymbol{u})$ of the parameter a is given by [1, App. F-A]

$$\hat{a}(\boldsymbol{u}) = \frac{\sum_{i=1}^{n-1} u_i u_{i+1}}{\sum_{i=1}^{n-1} u_i^2}. \tag{2}$$

For $0 < a < 1$, our previous work [1, Th. 5] provided a tight bound on the tail probability of the estimation error $\hat{a}(\boldsymbol{U}) - a$. Using different tools, for $a > 1$, this paper derives an exponentially decaying upper bound on the tail probability of $\hat{a}(\boldsymbol{U}) - a$. This result complements the large body of works [11–15] studying various aspects of the MLE $\hat{a}(\boldsymbol{u})$. Our bound is nonasymptotic and is tighter than existing bounds, see Fig. 1 in Section III-A below for a comparison. As an application of the error bound, we find the dispersion for the nonstationary Gauss-Markov source. Although the dispersion is represented

P. Tian and V. Kostina are with the Department of Electrical Engineering, California Institute of Technology. (e-mail: {ptian, vkostina}@caltech.edu). This research was supported in part by the National Science Foundation (NSF) under Grant CCF-1751356. A preliminary version of this paper was accepted for publication in the IEEE International Symposium on Information Theory, Paris, France, July 2019.

[1] Otherwise, we instead consider the random process $\{U_i'\}_{i=1}^{\infty}$ defined by the invertible mapping $U_i' \triangleq (-1)^i U_i$ which satisfies $U_i' = (-a)U_{i-1}' + (-1)^i Z_i$, where $(-1)^i Z_i$'s are also independent zero-mean Gaussian random variables with variance σ^2.

[2] Technically, the unit-root case is also nonstationary.

by the same formula as the one we derived for the stationary case [1, Eq. (57)], the analyses of the two scenarios differ significantly. In fact, after the RDF of the stationary Gauss-Markov source was derived [16] (see also [17, Th. 4.5.3]), it still took several decades to completely understand the RDF of the nonstationary one [8, 18, 19], see the detailed discussions in Section IV below.

Notations: For $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, 2, ..., n\}$. We use the standard $O(\cdot), o(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ and $\omega(\cdot)$ notations to characterize functions according to their asymptotic growth rates. Namely, let $f(n)$ and $g(n)$ be two functions of n, then $f(n) = O(g(n))$ means there exists a constant $M > 0$ and $n_0 \in \mathbb{N}$ such that $|f(n)| \le M|g(n)|$ for any $n \ge n_0$; $f(n) = o(g(n))$ means $\lim_{n\to\infty} f(n)/g(n) = 0$; $f(n) = \Theta(g(n))$ means there exist positive constants c_1, c_2 and $n_0 \in \mathbb{N}$ such that $c_1 g(n) \le f(n) \le c_2 g(n)$ for any $n \ge n_0$; $f(n) = \Omega(g(n))$ if and only if $g(n) = O(f(n))$; and $f(n) = \omega(g(n))$ if and only if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = +\infty$. For a matrix M, we denote by $\mathsf{M}^\top$ its transpose, by $\|\mathsf{M}\|$ its operator norm (the largest singular value) and by $\mu_1(\mathsf{M}) \le \ldots \le \mu_n(\mathsf{M})$ its eigenvalues listed in non-decreasing order. We use $\mathcal{S}^c$ to denote the complement of a set $\mathcal{S}$. All logarithms and exponentials are base e.

II. Previous Works

A. Parameter Estimation

The MLE $\hat{a}(\boldsymbol{u})$ of the parameter a given samples $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ drawn from the Gauss-Markov source (1) is given in (2). This paper derives a nonasymptotic fine-grained large deviations analysis of the estimation error. Given an error threshold $\eta > 0$, define the error exponents P_n^+ and P_n^- as

$$P_n^+ \triangleq -\frac{1}{n} \log \mathbb{P}\left[\hat{a}(\boldsymbol{U}) - a > \eta\right], \tag{3}$$

$$P_n^- \triangleq -\frac{1}{n} \log \mathbb{P}\left[\hat{a}(\boldsymbol{U}) - a < -\eta\right]. \tag{4}$$

We also define P_n as

$$P_n \triangleq -\frac{1}{n} \log \mathbb{P}\left[|\hat{a}(\boldsymbol{U}) - a| > \eta\right]. \tag{5}$$

The estimator $\hat{a}(\boldsymbol{U})$ in (2) has been extensively studied in the statistics [11, 12] and economics [20, 21] communities. An estimator $\hat{a}(\boldsymbol{U})$ is said to be *weakly consistent* if the estimation error $\hat{a}(\boldsymbol{U}) - a$ converges to 0 in probability, and *strongly consistent* if $\hat{a}(\boldsymbol{U}) - a$ converges to 0 almost surely, as n tends to infinity. Mann and Wald [20] and Rubin [21] showed that the estimator $\hat{a}(\boldsymbol{U})$ is weakly consistent for any $a \in \mathbb{R}$. Decades later, Rissanen and Caines [12] proved the strong consistency of the maximum likelihood estimator for the general stationary vector Gaussian ARMA processes, which implies that the estimator $\hat{a}(\boldsymbol{U})$ is strongly consistent for $0 < a < 1$. To better understand how the error $\hat{a}(\boldsymbol{U}) - a$ scales as n tends to infinity, researchers turned to study the limiting distribution of the normalized estimation error $h(n)(\hat{a}(\boldsymbol{U}) - a)$ for a careful choice of the standardizing function $h(n)$:

$$h(n) \triangleq \begin{cases} \sqrt{\frac{n}{1-a^2}}, & |a| < 1, \\ \frac{n}{\sqrt{2}}, & |a| = 1, \\ \frac{|a|^n}{a^2-1}, & |a| > 1. \end{cases} \tag{6}$$

Mann and Wald [20] and White [11] showed that the distribution of the normalized estimation error $h(n)(\hat{a}(\boldsymbol{U}) - a)$ converges to $\mathcal{N}(0, 1)$ for $|a| < 1$; to the Cauchy distribution with the probability density function $\frac{1}{\pi(1+x^2)}$ for $|a| > 1$; and for $|a| = 1$, to the distribution of $\frac{B^2(1)-1}{2\int_0^1 B^2(t)\, dt}$, where $\{B(t) : t \in [0, 1]\}$ is a Brownian motion.

Generalizations of the above results in several directions have also been investigated. In [20, Sec. 4], the maximum likelihood estimator for the p-th order stationary autoregressive processes with Z_i's being i.i.d. zero-mean and bounded moments random variables (not necessarily Gaussian) was shown to be weakly consistent, and the scaled estimation errors $\sqrt{n}(\hat{a}_j - a_j)$ for $j = 1, \ldots, p$ were shown to converge in distribution to the Gaussian random variables as n tends to infinity. Anderson [22, Sec. 3] studied the limiting distribution of the maximum likelihood estimator for a nonstationary vector version of the process (1). Chan and Wei [23] studied the performance of the estimation error when a is not a constant but approaches to 1 from below in the order of $1/n$.

Another line of work closely related to this paper is the large deviation principle (LDP) [24, Ch. 1.2] on $\hat{a}(\boldsymbol{U}) - a$ [13, 14]. Bercu et al. [13] showed that for $0 < a < 1$, the estimation error $\hat{a}(\boldsymbol{U}) - a$ satisfies

$$\lim_{n\to\infty} P_n = I_{\text{s}}(\eta), \tag{7}$$

where the rate function $I_{\text{s}}(\eta)$ is given in [13, Prop. 8]. For $a > 1$, Worms [14, Thm. 1] proved that

$$\liminf_{n\to\infty} P_n \ge I_{\text{ns}}(\eta), \tag{8}$$

where $I_{\text{ns}}(\eta)$ is specified in [14, Th. 1] as the optimal value of an optimization problem. A bound similar to (8) for the unit-root case was also presented in [14, Th. 1].

The problem of estimating the parameter a from a block of outcomes of the Gauss-Markov source (1) is one of the simplest versions in recent studies of machine learning for dynamical systems [15, 25–28]. One objective of those studies is to obtain tight performance bounds on the least-squares estimates of the system parameters $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ from a single input / output trajectory $\{\boldsymbol{w}_i, \boldsymbol{y}_i\}_{i=1}^n$ in the state-space model:

$$\boldsymbol{x}_{i+1} = \mathsf{A}\boldsymbol{x}_i + \mathsf{B}\boldsymbol{w}_i + \boldsymbol{z}_i, \tag{9}$$

$$\boldsymbol{y}_i = \mathsf{C}\boldsymbol{x}_i + \mathsf{D}\boldsymbol{w}_i + \boldsymbol{v}_i, \tag{10}$$

where $\boldsymbol{x}_i, \boldsymbol{w}_i, \boldsymbol{z}_i, \boldsymbol{v}_i$'s are random vectors of certain dimensions and the system parameters $\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}$ are matrices of appropriate dimensions [25, 26]. The Gauss-Markov process in (1) can be written as the state-space model with $\mathsf{B} = \mathsf{D} = 0$, $\mathsf{C} = 1$ and $\boldsymbol{v}_i = 0$. For stable systems, $\|\mathsf{A}\| < 1$, Oymak and Ozay [26, Thm. 3.1] showed that the estimation error in spectral norm is $O(1/\sqrt{n})$ with high probability, where n is the number of samples. For the subclass of the regular unstable systems [28, Def. 3], Faradonbeh et al. [28, Thm. 1] proved that the probability of estimation error exceeding a positive threshold in spectral norm decays exponentially in n. For the Gauss-Markov processes considered in the present paper, Simchowitz et al. [25, Thm. B.1] and Sarkar and Rakhlin [27, Prop. 5.2] presented tail bounds on the estimation error of the MLE.

These studies on the limiting distribution and the LDP of the estimation error are asymptotic. To our knowledge, there are two nonasymptotic lower bounds on P_n^+ and P_n^-. For any $a \in \mathbb{R}$, Rantzer [15, Th. 4] showed that

$$P_n^+ \text{ (and } P_n^-) \geq \frac{1}{2}\log(1+\eta^2). \tag{11}$$

Bercu and Touati [29, Cor. 5.2] proved that

$$P_n^+ \text{ (and } P_n^-) \geq \frac{\eta^2}{2(1+y_\eta)}, \tag{12}$$

where y_η is the unique positive solution to $(1+x)\log(1+x) - x - \eta^2 = 0$ in x. Both bounds (11) and (12) do not capture the dependence on a and n, and are the same for P_n^+ and P_n^-. All the bounds in [15, 25–28] are either optimal only order-wise or involve implicit constants. Our main result on parameter estimation is a tight nonasymptotic lower bound on P_n^+ and P_n^-. For larger a, the lower bound becomes larger, which suggests that unstable systems are easier to estimate than stable ones, an observation consistent with [25]. The proof is inspired by Rantzer [15, Lem. 5], but our result significantly improves (11) and (12), see Fig. 1 for a comparison. Most of our results generalize to the case where Z_i's are i.i.d. sub-Gaussian random variables, see Theorem 4 in Section III-A below.

B. Nonasymptotic Rate-distortion Theory

Given a distortion threshold $d > 0$, an excess-distortion probability $\epsilon \in (0,1)$ and $M \in \mathbb{N}$, an (n, M, d, ϵ) lossy compression code for a random vector $\boldsymbol{U} = (U_1, \ldots, U_n)^\top$ of length n consists of an encoder $\mathsf{f}_n \colon \mathbb{R}^n \to [M]$, and a decoder $\mathsf{g}_n \colon [M] \to \mathbb{R}^n$, such that $\mathbb{P}\left[\mathsf{d}\left(\boldsymbol{U}, \mathsf{g}_n\left(\mathsf{f}_n(\boldsymbol{U})\right)\right) > d\right] \leq \epsilon$, where $\mathsf{d}(\cdot,\cdot)$ is the distortion measure. In this paper, we consider the mean squared error (MSE) distortion: $\forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$,

$$\mathsf{d}(\boldsymbol{u}, \boldsymbol{v}) \triangleq \frac{1}{n}\sum_{i=1}^n (u_i - v_i)^2. \tag{13}$$

The minimum achievable code size and source coding rate are defined respectively by

$$M^\star(n, d, \epsilon) \triangleq \min\left\{M \in \mathbb{N} \colon \exists\ (n, M, d, \epsilon) \text{ code}\right\}, \tag{14}$$

$$R(n, d, \epsilon) \triangleq \frac{1}{n}\log M^\star(n, d, \epsilon). \tag{15}$$

The core problem in the nonasymptotic rate-distortion theory is to characterize $R(n, d, \epsilon)$. For stationary memoryless sources, Ingber and Kochman [9] (finite-alphabet and Gaussian sources) and Kostina and Verdú [10] (abstract sources) showed that the minimum achievable source coding rate satisfies a Gaussian approximation of form (68) in Section III-C below. In this paper, we extend our previous analysis [1, Th. 1] of the stationary Gauss-Markov source to the nonstationary one. One of the key ideas behind this extension is to construct a typical set using the MLE of a, and to use our estimation error bound to probabilistically characterize that set.

III. Main Results

A. Error Exponent Bounds in Parameter Estimation

We first present our nonasymptotic bounds on P_n^+ and P_n^- using two sequences $\{\alpha_\ell\}_{\ell\in\mathbb{N}}$ and $\{\beta_\ell\}_{\ell\in\mathbb{N}}$ defined as follows. Throughout the paper, $\sigma^2 > 0$ and $a > 1$ are fixed constants. For $\eta > 0$ and a parameter $s > 0$, let $\{\alpha_\ell\}_{\ell\in\mathbb{N}}$ be the following sequence

$$\alpha_1 \triangleq \frac{\sigma^2 s^2 - 2\eta s}{2}, \tag{16}$$

$$\alpha_\ell = \frac{[a^2 + 2\sigma^2 s(a+\eta)]\alpha_{\ell-1} + \alpha_1}{1 - 2\sigma^2\alpha_{\ell-1}}, \quad \forall \ell \geq 2. \tag{17}$$

Similarly, let $\{\beta_\ell\}_{\ell\in\mathbb{N}}$ be the following sequence

$$\beta_1 \triangleq \frac{\sigma^2 s^2 - 2\eta s}{2}, \tag{18}$$

$$\beta_\ell = \frac{[a^2 + 2\sigma^2 s(-a+\eta)]\beta_{\ell-1} + \beta_1}{1 - 2\sigma^2\beta_{\ell-1}}, \quad \forall \ell \geq 2. \tag{19}$$

Note the slight difference between (17) and (19). Both sequences depend on η and s. We derive closed-form expressions and analyze the convergence properties of α_ℓ and β_ℓ in Appendix A-B below. For $\eta > 0$ and $n \in \mathbb{N}$, we define the following sets

$$\mathcal{S}_n^+ \triangleq \left\{s \in \mathbb{R} \colon s > 0,\ \alpha_\ell < \frac{1}{2\sigma^2},\ \forall \ell \in [n]\right\}, \tag{20}$$

$$\mathcal{S}_n^- \triangleq \left\{s \in \mathbb{R} \colon s > 0,\ \beta_\ell < \frac{1}{2\sigma^2},\ \forall \ell \in [n]\right\}. \tag{21}$$

Theorem 1. *For any constant $\eta > 0$, the estimator* (2) *satisfies for any $n \geq 2$,*

$$P_n^+ \geq \sup_{s\in\mathcal{S}_n^+} \frac{1}{2n}\sum_{\ell=1}^{n-1} \log\left(1 - 2\sigma^2\alpha_\ell\right), \tag{22}$$

$$P_n^- \geq \sup_{s\in\mathcal{S}_n^-} \frac{1}{2n}\sum_{\ell=1}^{n-1} \log\left(1 - 2\sigma^2\beta_\ell\right), \tag{23}$$

where $\{\alpha_\ell\}_{\ell\in\mathbb{N}}$ and $\{\beta_\ell\}_{\ell\in\mathbb{N}}$ are defined in (17) *and* (19)*, respectively, and $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$ are defined in* (20) *and* (21)*, respectively.*

Proof. Appendix A-A. ∎

The proof of Theorem 1 is a detailed analysis of the Chernoff bound using the tower property of conditional expectations. The proof is motivated by [15, Lem. 5], but our analysis is more accurate and the result is significantly tighter, see Fig. 1 and Fig. 3 for comparisons. Theorem 1 gives the best bound that can be obtained from the Chernoff bound. In view of the Gärtner-Ellis theorem [24, Th. 2.3.6], we conjecture that the bounds (22) and (23) can be reversed in the limit of large n. One recovers Rantzer's lower bound (11) by setting $s = \eta/\sigma^2$ and bounding α_ℓ as $\alpha_\ell \leq \alpha_1$ (due to the monotonicity of α_ℓ, see Appendix A-B below) in Theorem 1. We explicitly state where we diverge from [15, Lem. 5] in the proof in Appendix A-A below.

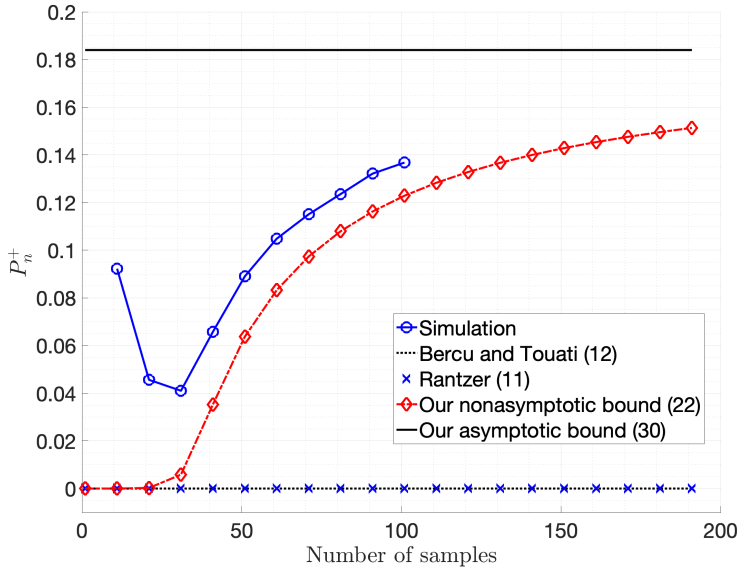

Fig. 1: Numerical simulations for $a = 1.2$ and $\eta = 10^{-3}$.

The exact characterization of $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$ for each n using η is involved. One can see from the definitions (20) and (21) that

$$\mathcal{S}_1^+ = \mathcal{S}_1^- = \left\{ s \in \mathbb{R} \colon 0 < s < \frac{\eta + \sqrt{1+\eta^2}}{\sigma^2} \right\}, \tag{24}$$

and $\mathcal{S}_{n+1}^+ \subseteq \mathcal{S}_n^+$ and $\mathcal{S}_{n+1}^- \subseteq \mathcal{S}_n^-$ for any $n \geq 1$. To obtain the set $\mathcal{S}_{n+1}^+$ from $\mathcal{S}_n^+$, we need to solve $\alpha_{n+1} < \frac{1}{2\sigma^2}$, which is equivalent to solving an inequality involving a polynomial of degree $n+2$ in s. Fig. 2 presents a plot of $\mathcal{S}_n^+$ for $n = 1, ..., 5$. Despite the complexity of the sets $\mathcal{S}_n^+$ and $\mathcal{S}_n^-$, the limits

$$\mathcal{S}_\infty^+ \triangleq \lim_{n\to\infty} \mathcal{S}_n^+ = \bigcap_{n\geq 1} \mathcal{S}_n^+, \tag{25}$$

$$\mathcal{S}_\infty^- \triangleq \lim_{n\to\infty} \mathcal{S}_n^- = \bigcap_{n\geq 1} \mathcal{S}_n^-, \tag{26}$$

can be characterized in terms of the interval

$$\mathcal{I}_\eta \triangleq \left(0, \frac{2\eta}{\sigma^2}\right). \tag{27}$$

Lemma 1. *Fix any constant $\eta > 0$. It holds that*

$$\mathcal{S}_\infty^+ = \mathcal{I}_\eta \cup \left\{ \frac{2\eta}{\sigma^2} \right\}, \tag{28}$$

$$\mathcal{S}_\infty^- \supsetneqq \mathcal{I}_\eta \cup \left\{ \frac{2\eta}{\sigma^2} \right\}. \tag{29}$$

Proof. Appendix A-C. ∎

Using Lemma 1 and taking limits in Theorem 1, we obtain the following result.

Theorem 2. *Fix any constant $\eta > 0$. It holds that*

$$\liminf_{n\to\infty} \; P_n^+ \geq I^+(a,\eta) \triangleq \log(a+2\eta), \tag{30}$$

$$\liminf_{n\to\infty} \; P_n^- \geq I^-(a,\eta), \tag{31}$$

$$\liminf_{n\to\infty} \; P_n \geq I^-(a,\eta), \tag{32}$$

where

$$I^-(a,\eta) \triangleq \begin{cases} \log a, & 0 < \eta \leq \eta_1, \\ \frac{1}{2}\log\frac{2a\eta-(a^2-1)}{1-(\eta-a)^2}, & \eta_1 < \eta < \eta_2, \\ \log(2\eta - a), & \eta \geq \eta_2, \end{cases} \tag{33}$$

where the thresholds η_1 and η_2 are given by

$$\eta_1 \triangleq \frac{a^2-1}{a}, \tag{34}$$

$$\eta_2 \triangleq \frac{3a + \sqrt{a^2+8}}{4}. \tag{35}$$

Proof. We prove (30) and (31) in Appendix A-D. The bound (32) follows from (30) and (31) since

$$\begin{aligned} &\mathbb{P}\left[|\hat{a}(\boldsymbol{U}) - a| > \eta\right] \\ =&\mathbb{P}\left[(\hat{a}(\boldsymbol{U}) - a) > \eta\right] + \mathbb{P}\left[(\hat{a}(\boldsymbol{U}) - a) < -\eta\right] \end{aligned} \tag{36}$$

and

$$\begin{aligned} &\liminf_{n\to\infty} P_n \\ =&\liminf_{n\to\infty} \min\left\{P_n^+, P_n^-\right\} \qquad (37) \\ \geq& I^-(a,\eta). \qquad (38) \end{aligned}$$

∎

Remark 1. The results in (28) and (29), and (30) and (31) indicate the asymmetry between P_n^+ and P_n^-: the set $\mathcal{S}_\infty^-$ contains more elements than $\mathcal{S}_\infty^+$, and $I^+(a,\eta) > I^-(a,\eta)$, which suggests that the maximum likelihood estimator $\hat{a}(\boldsymbol{U})$ is more likely to underestimate a than overestimate it.

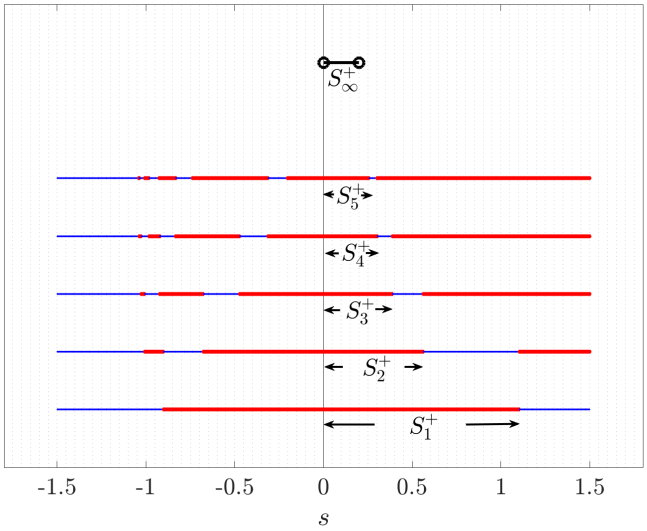

Fig. 2: Numerical computation of the sets $\mathcal{S}_n^+$ for $a = 1.2$ and $\eta = 0.1$. Each horizontal line corresponds to $n = 1, ..., 5$ in the bottom-up order. Within each horizontal line, the red thick parts denote the ranges of s for which $\alpha_n < \frac{1}{2\sigma^2}$, and the blue thin region is where $\alpha_n \geq \frac{1}{2\sigma^2}$. The plot for $\mathcal{S}_n^-$ is similar.

Fig. 3 presents a comparison of (32), Worms' bound (8), Rantzer's bound (11) and Bercu and Touati (12). Worms' bound (8) is tighter than Rantzer's bound (11) when η is small, and looser for large η. Our bound (32) is tighter than both of them for any $\eta > 0$.

If $\eta = \eta_n > 0$ is a sequence decreasing to 0, it is easy to see that Theorem 1 still holds. For Theorem 2 to remain valid, we require that the speed with which η_n decreases to zero is no smaller than $\frac{1}{\sqrt{n}}$, which essentially ensures that the right sides of (22) and (23) still converge to the right sides of (30) and (31), respectively. Let η_n be a positive sequence such that

$$\eta_n = \omega\left(\frac{1}{\sqrt{n}}\right). \tag{39}$$

Theorem 3. *For any $\sigma^2 > 0$ and $a > 1$, let $\eta = \eta_n > 0$ satisfy* (39)*. Then, Theorem 1 holds, and Theorem 2 holds with* (30) *and* (31) *replaced, respectively, by*

$$\liminf_{n\to\infty} P_n^+ \geq \log a, \tag{40}$$

$$\liminf_{n\to\infty} P_n^- \geq \log a. \tag{41}$$

Proof. Appendix A-E. ∎

The following corollary to Theorem 3 is used in Section III-C below to derive the dispersion of nonstationary Gauss-Markov sources.

Corollary 1. *For any $\sigma^2 > 0$ and any $a > 1$, there exists a constant $c \geq \frac{1}{2}\log(a)$ such that for all n large enough,*

$$\mathbb{P}\left[|\hat{a}(\boldsymbol{U}) - a| \geq \sqrt{\frac{\log\log n}{n}}\right] \leq 2e^{-cn}. \tag{42}$$

We now generalize the above results to the case where Z_i's in (1) are zero-mean σ-sub-Gaussian random variables.

Definition 1 (sub-Gaussian random variable, e.g. [30, Def. 2.7])**.** *Fix $\sigma > 0$. A random variable $Z \in \mathbb{R}$ with mean μ is said to be σ-sub-Gaussian with variance proxy σ^2 if its moment-generating function (MGF) satisfies*

$$\mathbb{E}[e^{s(Z-\mu)}] \leq e^{\frac{\sigma^2 s^2}{2}}, \tag{43}$$

for all $s \in \mathbb{R}$.

One important property of σ-sub-Gaussian random variables is the following well-known bound on the MGF of quadratic functions of σ-sub-Gaussian random variables.

Lemma 2 ([15, Prop. 2])**.** *Let Z be a σ-sub-Gaussian random variable with mean μ. Then*

$$\mathbb{E}\exp(sZ^2) \leq \frac{1}{\sqrt{1-2\sigma^2 s}}\exp\left(\frac{s\mu^2}{1-2\sigma^2 s}\right) \tag{44}$$

for any $s < \frac{1}{2\sigma^2}$.

Equality holds in (43) and (44) when Z is Gaussian. In particular, the right side of (44) is the MGF of the noncentral χ^2-distributed random variable Z^2.

Theorem 4 (Generalization to sub-Gaussian case)**.** *Theorems 1–3 and Lemma 1 remain valid for the estimator* (2) *when Z_i's in* (1) *are i.i.d. zero-mean σ-sub-Gaussian random variables.*

Proof. The generalizations of Theorems 1–3 and Lemma 1 from the Gaussian to sub-Gaussian case only require minor changes in the corresponding proofs. See Appendix A-F for the details. ∎

B. Nonasymptotic Rate-distortion Theory

We review some definitions before we can discuss our main results on the dispersion of nonstationary Gauss-Markov processes.

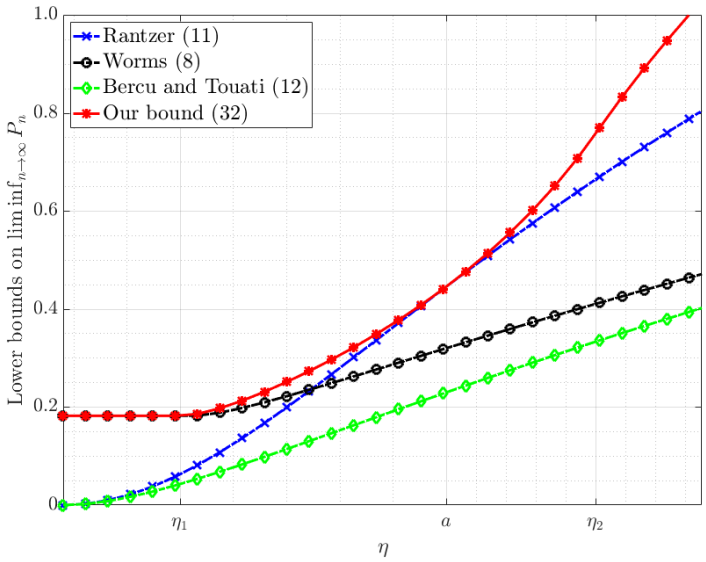

Fig. 3: Comparisons of our lower bound (32) and previous works (8), (11) and (12). $a = 1.2$.

1) Rate-distortion functions: For a random process $\{X_i\}_{i=1}^{\infty}$, the *n-th order rate-distortion function* $\mathbb{R}_{\boldsymbol{X}}(d)$ is defined as

$$\mathbb{R}_{\boldsymbol{X}}(d) \triangleq \inf_{\substack{P_{\boldsymbol{Y}|\boldsymbol{X}}: \\ \mathbb{E}[\mathsf{d}(\boldsymbol{X},\boldsymbol{Y})]\le d}} \frac{1}{n} I(\boldsymbol{X};\boldsymbol{Y}), \tag{45}$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ is the n-dimensional random vector. The *rate-distortion function* $\mathbb{R}_X(d)$ is defined as

$$\mathbb{R}_X(d) \triangleq \limsup_{n\to\infty} \mathbb{R}_{\boldsymbol{X}}(d). \tag{46}$$

Under the average distortion criterion, the informational quantity $\mathbb{R}_X(d)$ has been shown to be equal to the minimum achievable source coding rate as the blocklength n goes to infinity, see [7] for discrete memoryless sources and [31] for general ergodic sources. Gray [8, Th. 2] proved a coding theorem for the Gaussian autoregressive processes, which include the Gauss-Markov source as a special case, under average mean-squared error. Closed-form expressions for $\mathbb{R}_{\boldsymbol{X}}(d)$ and $\mathbb{R}_X(d)$ are known only for a few special random processes including the Gaussian autoregressive processes [8]. Specializing Gray's result [8, Eq. (22)] to our Gauss-Markov source (1), we write down the *n-th order reverse waterfilling* solution for the n-th order rate-distortion function $\mathbb{R}_{\boldsymbol{U}}(d)$:

$$\mathbb{R}_{\boldsymbol{U}}(d) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2} \log \max\left(\mu_i, \frac{\sigma^2}{\theta_n}\right), \tag{47}$$

$$d = \frac{1}{n}\sum_{i=1}^{n} \min\left(\theta_n, \frac{\sigma^2}{\mu_i}\right), \tag{48}$$

where $\theta_n > 0$ is the *water level*, and μ_i's are the eigenvalues of $\mathsf{A}^\top\mathsf{A}$, where A is the following $n \times n$ lower triangular matrix:

$$\mathsf{A}_{ij} = \begin{cases} 1, & i = j, \\ -a, & i = j+1, \\ 0, & \text{otherwise.} \end{cases} \tag{49}$$

The rate-distortion function $\mathbb{R}_U(d)$ of the Gauss-Markov source is given by the *limiting reverse waterfilling*:

$$\mathbb{R}_U(d) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{1}{2} \log \max\left(g(w), \frac{\sigma^2}{\theta}\right) dw, \tag{50}$$

$$d = \frac{1}{2\pi}\int_{-\pi}^{\pi} \min\left(\theta, \frac{\sigma^2}{g(w)}\right) dw, \tag{51}$$

where

$$g(w) \triangleq 1 + a^2 - 2a\cos(w). \tag{52}$$

It is well-known [7] that the rate-distortion function of the Gaussian memoryless source $\{Z_i\}$ is

$$\mathbb{R}_Z(d) = \max\left(0, \frac{1}{2}\log\frac{\sigma^2}{d}\right). \tag{53}$$

See Fig. 4 for a plot of $\mathbb{R}_U(d)$ and $\mathbb{R}_Z(d)$.

2) Critical and maximum distortions: In view of (51), there are two special water levels $\theta_{\min}$ and $\theta_{\max}$, defined as follows:

$$\theta_{\min} \triangleq \min_{w\in[-\pi,\pi]} \frac{\sigma^2}{g(w)} = \frac{\sigma^2}{(a+1)^2} \tag{54}$$

and

$$\theta_{\max} \triangleq \max_{w\in[-\pi,\pi]} \frac{\sigma^2}{g(w)} = \frac{\sigma^2}{(a-1)^2}. \tag{55}$$

The *critical distortion* d_c is defined as the distortion corresponding to water level $\theta_{\min}$. By (51), we have

$$d_c = \theta_{\min} = \frac{\sigma^2}{(a+1)^2}. \tag{56}$$

The *maximum distortion* $d_{\max}$ is defined as the distortion corresponding to water level $\theta_{\max}$. By (51), we have

$$d_{\max} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\sigma^2}{g(w)}\, dw. \tag{57}$$

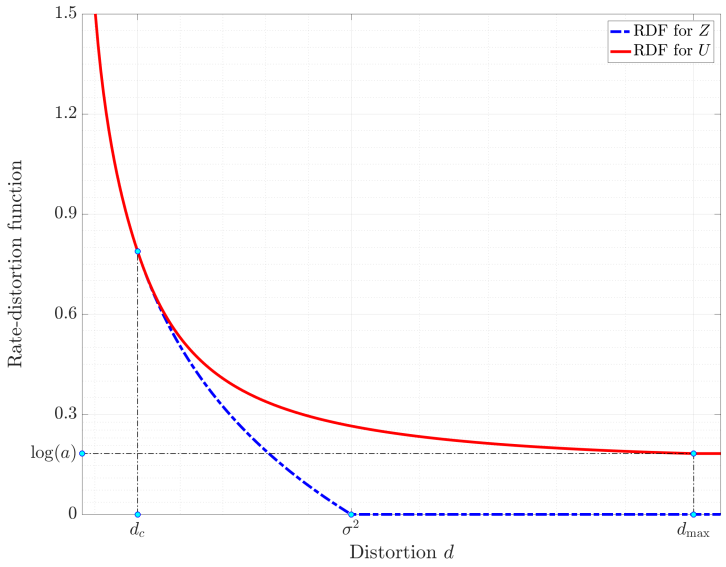

Fig. 4: Rate-distortion functions. $a = 1.2$.

Using similar techniques as in [1, Eq. (169)–(172)], one can obtain

$$d_{\max} = \frac{\sigma^2}{a^2 - 1}. \tag{58}$$

In this paper, we always consider a fixed distortion threshold d such that $0 < d < d_{\max}$.

3) Decorrelation: For the Gauss-Markov source $\{U_i\}_{i=1}^{+\infty}$ in (1), we refer to the transformed random vector

$$\boldsymbol{X} \triangleq \mathsf{S}^\top \boldsymbol{U} \tag{59}$$

as the *decorrelation* of $\boldsymbol{U}$, where $\boldsymbol{U} = (U_1, \ldots, U_n)^\top$ and S is the orthornomal matrix that diagonalizes $(\mathsf{A}^\top \mathsf{A})^{-1}$. The decorrelation $\boldsymbol{X}$ has independent coordinates

$$X_i \sim \mathcal{N}(0, \sigma_i^2), \tag{60}$$

where

$$\sigma_i^2 \triangleq \frac{\sigma^2}{\mu_i} \tag{61}$$

are eigenvalues of the covariance matrix of $\boldsymbol{U}$, or, equivalently, of the covariance matrix of $\boldsymbol{X}$:

$$\Sigma_{\boldsymbol{U}} = \Sigma_{\boldsymbol{X}} = \sigma^2 (\mathsf{A}^\top \mathsf{A})^{-1}. \tag{62}$$

The minimum achievable source coding rates (defined in (15)) for lossy compression of $\boldsymbol{U}$ and $\boldsymbol{X}$ are equal, as are their rate-distortion functions: $\mathbb{R}_{\boldsymbol{U}}(d) = \mathbb{R}_{\boldsymbol{X}}(d)$, see [1, Sec. III.A] for the details.

4) Operational dispersion: In nonasymptotic rate-distortion theory, the *operational dispersion* $V_U(d)$ captures the convergence rate of the minimum achievable source coding rate to the rate-distortion function. Formally, it is defined as

$$V_U(d) \triangleq \lim_{\epsilon \to 0} \limsup_{n \to \infty} n \left(\frac{R(n, d, \epsilon) - \mathbb{R}_U(d)}{Q^{-1}(\epsilon)} \right)^2, \tag{63}$$

where Q^{-1} denotes the inverse Q-function. The main result in the second part of this paper gives $V_U(d)$ for the nonstationary Gauss-Markov source.

5) Informational dispersion: The d-*tilted information*, introduced in [10, Def. 6], is the key random variable that governs the nonasymptotic fundamental limits in the rate-distortion theory. Using (47), (48), (60), and the equivalence between $\boldsymbol{U}$ and its decorrelation $\boldsymbol{X}$ in Section III-B3 above, one can show [1, Eq. (55) and (228)] that the d-tilted information $\jmath_{\boldsymbol{U}}(\boldsymbol{u}, d)$ in $\boldsymbol{u}$ for the Gauss-Markov source satisfies $\jmath_{\boldsymbol{U}}(\boldsymbol{u}, d) = \jmath_{\boldsymbol{X}}(\boldsymbol{x}, d)$, and

$$\jmath_{\boldsymbol{X}}(\boldsymbol{x}, d) = \sum_{i=1}^{n} \frac{\min(\theta_n, \sigma_i^2)}{2\theta_n} \left(\frac{x_i^2}{\sigma_i^2} - 1 \right) + \frac{1}{2} \sum_{i=1}^{n} \log \frac{\max(\theta_n, \sigma_i^2)}{\theta_n}, \tag{64}$$

where $\theta_n > 0$ is given by (48), and $\boldsymbol{x} \triangleq \mathsf{S}^\top \boldsymbol{u}$.

In lossy compression of stationary and memoryless sources, the mean and the variance of the d-tilted information are equal to the rate-distortion function and the dispersion, respectively [10, Th. 12]. This paper provides a natural extension of the above fact to nonstationary Gauss-Markov sources. The main result in this second part of the paper establishes the equality between operational dispersion (63) and the *informational dispersion*, defined as

$$\mathbb{V}_U(d) \triangleq \limsup_{n \to \infty} \frac{1}{n} \mathrm{Var} \left[\jmath_{\boldsymbol{U}}(\boldsymbol{U}, d) \right]. \tag{65}$$

Lemma 3. *The informational dispersion of the nonstationary Gauss-Markov source is given by*

$$\mathbb{V}_U(d) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \min \left[1, \left(\frac{\sigma^2}{\theta g(w)} \right)^2 \right] dw, \tag{66}$$

where $\theta > 0$ is given in (51)*, and g is in* (52)*.*

Proof. Appendix B-A. ∎

It is known [10, Eq. (94)] [9, Sec. IV] that the informational dispersion for the Gaussian memoryless source $\{Z_i\}$ is

$$\mathbb{V}_Z(d) = \frac{1}{2}, \quad \forall d \in (0, \sigma^2). \tag{67}$$

See Fig. 5 for a plot of $\mathbb{V}_U(d)$ and $\mathbb{V}_Z(d)$.

C. Dispersion of the Nonstationary Gauss-Markov Processes

We now state our main results on the dispersion of the nonstationary Gauss-Markov sources.

Theorem 5 (Gaussian approximation)**.** *Consider the Gauss-Markov source* (1) *with $a > 1$. For any fixed excess-distortion probability $\epsilon \in (0,1)$ and distortion threshold $d \in (0, d_{\max})$, the minimum achievable source coding rate $R(n,d,\epsilon)$ admits the following Gaussian approximation:*

$$R(n,d,\epsilon) = \mathbb{R}_U(d) + Q^{-1}(\epsilon)\sqrt{\frac{\mathbb{V}_U(d)}{n}} + o\left(\frac{1}{\sqrt{n}}\right), \tag{68}$$

where Q^{-1} denotes the inverse Q-function; the rate-distortion function $\mathbb{R}_U(d)$ is given in (50)*; the informational dispersion $\mathbb{V}_U(d)$ is given in* (66)*.*

We prove the converse and achievability directions of Theorem 5, stated in the following two theorems, respectively.

Theorem 6 (Converse)**.** *In the setting the Theorem 5, the minimum achievable source coding rate satisfies*

$$R(n,d,\epsilon) \geq \mathbb{R}_U(d) + \sqrt{\frac{\mathbb{V}_U(d)}{n}}Q^{-1}(\epsilon) - \frac{\log n}{2n} + O\left(\frac{1}{n}\right). \tag{69}$$

Proof. The converse proof is similar to that in the asymptotically stationary case in [1, Th. 7]. See Appendix D for the details. ∎

Theorem 7 (Achievability)**.** *In the setting the Theorem 5, the minimum achievable source coding rate satisfies*

$$R(n,d,\epsilon) \leq \mathbb{R}_U(d) + \sqrt{\frac{\mathbb{V}_U(d)}{n}}Q^{-1}(\epsilon) + O\left(\frac{1}{\sqrt{n}\log n}\right). \tag{70}$$

Proof. See the discussions in Section III-D below. ∎

Remark 2. Gray [8, Eq. (24)] showed the following relation between the rate-distortion functions of the Gauss-Markov source and the Gaussian memoryless source:

$$\begin{cases} \mathbb{R}_U(d) = \mathbb{R}_Z(d), & d \in (0, d_c], \\ \mathbb{R}_U(d) > \mathbb{R}_Z(d), & d \in (d_c, d_{\max}), \end{cases} \tag{71}$$

where $\mathbb{R}_Z(d)$ is in (53) above. Using Lemma 3, one can easily show (in the same way as [1, Cor. 1]) that their dispersions are also comparable:

$$\begin{cases} \mathbb{V}_U(d) = \mathbb{V}_Z(d), & d \in (0, d_c], \\ \mathbb{V}_U(d) < \mathbb{V}_Z(d), & d \in (d_c, \sigma^2), \end{cases} \tag{72}$$

where $\mathbb{V}_Z(d)$ is in (67) above. The results (71) and (72) imply that for low distortions $d \in (0, d_c)$, the minimum achievable source coding rate in compressing the Gauss-Markov source and the Gaussian memoryless source are the same up to second-order terms, a phenomenon we observed in the stationary case as well [1, Cor. 1]. See Fig. 4 and Fig. 5 for a visualization of (71) and (72), respectively.

Remark 3. For the red solid curve in Fig. 4, we have

$$\mathbb{R}_U(d_{\max}) = \log a. \tag{73}$$

This result has an interesting connection to the problem of control under communication constraints [32] [33, Th. 1] [34, Prop. 3.1], where it was shown that the minimum rate to asymptotically stabilize a linear, discrete-time, scalar system is also $\log a$, suggesting that stability is unattained with any rate lower than $\log a$ even if an infinite lookahead is allowed. We present two ways to obtain (73). The first one is to directly use (94) in Section IV-A below. For $\theta = \theta_{\max}$, we have $\mathbb{R}_{\mathrm{K}}(d_{\max}) = 0$ in (93), then (73) immediately follows from (94). The second method relies on (50). For $\theta = \theta_{\max}$, observe from (50) that

$$\mathbb{R}_U(d_{\max}) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(g(w))\, dw. \tag{74}$$

Then, computing the integral (74) using Lemma 6 in Appendix B-B below yields (73).

Remark 4. The coordinates of the two special points P_1 and P_2 on the red solid curve in Fig. 5 are given by

$$P_1 = (d_c, 1/2), \quad P_2 = \left(d_{\max}, \frac{(1+a^2)(a-1)}{2(a+1)^3}\right). \tag{75}$$

The derivation for P_2 is the same as that in the stationary case [1, Eq. (61)] except that we need to compute the residue at $1/a$ instead of at a, see [1, App. B-A] for details.

D. Lossy AEP and Parameter Estimation

Central to the proof of Theorem 7 are the random coding bound [10, Cor. 11] and the following second-order refinement of the "lossy AEP" (asymptotic equipartition property) for nonstationary Gauss-Markov sources:

Lemma 4. *For the Gauss-Markov source* (1) *with $a > 1$, let $\boldsymbol{X}$ be the decorrelation of $\boldsymbol{U}$ in* (59)*, let $\boldsymbol{Y}^\star$ be the minimizer of* (45)*, and let $\jmath_{\boldsymbol{X}}(\boldsymbol{X}, d)$ be the* d*-tilted information given by* (64)*. Then,*

$$\mathbb{P}\left[\log \frac{1}{P_{\boldsymbol{Y}^\star}\left(\mathcal{B}\left(\boldsymbol{X}, d\right)\right)} \geq \jmath_{\boldsymbol{X}}(\boldsymbol{X}, d) + p(n)\right] \leq \frac{1}{q(n)}, \tag{76}$$

where

$$p(n) \triangleq c_1(\log n)^{c_2} + c_3 \log n + c_4, \tag{77}$$

$$q(n) \triangleq \Theta(\log n), \tag{78}$$

with constants c_i's, $i = 1, ..., 4$ satisfying $c_1, c_2, c_3 > 0$.

Proof. Appendix E-D. ∎

The precise connection between the lossy AEP and the achievability proof is the following. Assume we have shown (76) for $p(n)$ and $q(n)$ satisfying $p(n) = \omega(1)$ and

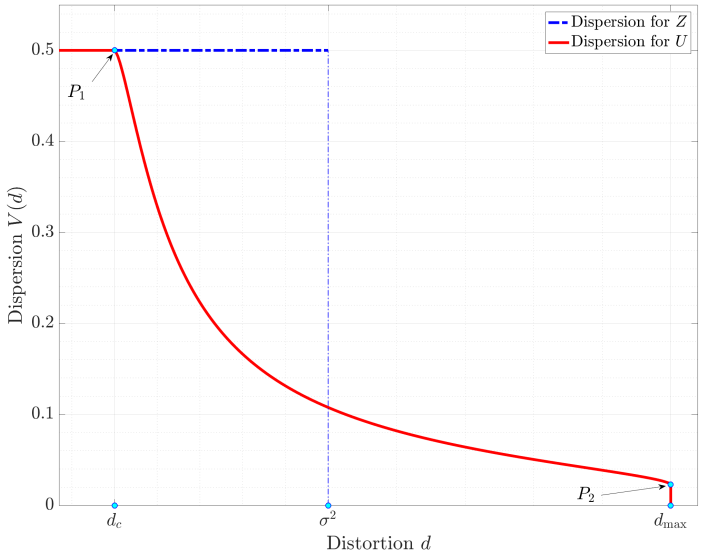

Fig. 5: Dispersions. $a = 1.2$

$q(n) = \omega(1)$. Then, using Propositions 1 and 2 in Appendix C below, one can apply the proof in [1, Sec. V-C] to obtain

$$R(n,d,\epsilon) \le \mathbb{R}_U(d) + \sqrt{\frac{\mathbb{V}_U(d)}{n}}Q^{-1}(\epsilon) + \frac{K_1 \log\log n}{n} + \frac{p(n)}{n} + \frac{K_2}{\sqrt{n}q(n)}, \tag{79}$$

where $K_1 > 0$ is a universal constant and K_2 is a constant depending on ϵ. Hence, in order to obtain the second-order term in (79), we require in (76) that

$$p(n) = o(\sqrt{n}), \tag{80}$$
$$q(n) = \omega(1). \tag{81}$$

Therefore, Theorem 7 follows immediately from (79) with the choices of $p(n)$ and $q(n)$ in (77) and (78), respectively. We have $O(\cdot)$ in (70) since K_2 could be positive or negative. The rest of the paper focuses on the proof of Lemma 4.

The proof of lossy AEP in the form of Lemma 4 is technical even for stationary memoryless sources. A lossy AEP for stationary α-mixing processes was derived in [35, Cor. 17]. The idea in [10, Lem. 2] is to form a typical set of source outcomes (the set F_n in [10, Lem. 4]) using the product of the empirical distributions [10, Eq. (270)]: $P_{\hat{X}} \times \ldots \times P_{\hat{X}}$, where $P_{\hat{X}}(x) \triangleq \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{x_i = x\}$ is the empirical distribution of a given source sequence $\boldsymbol{x}$, and then to show that the inequality inside the bracket in (76) holds for $\boldsymbol{x} \in F_n$. The Gauss-Markov source is not memoryless and it is nonstationary for $a > 1$. To form a typical set of source outcomes, we define the following proxy random variable using the estimator $\hat{a}(\boldsymbol{u})$.

Definition 2 (Proxy random variable)**.** *For each sequence $\boldsymbol{u}$ of length n generated by the Gauss-Markov source, define the proxy random variable $\hat{\boldsymbol{X}}$ as an n-dimensional Gaussian random vector with independent coordinates, each of which follows the distribution $\mathcal{N}(0, \hat{\sigma}_i^2)$ with*

$$\hat{\sigma}_1^2 \triangleq \sigma^2 \hat{a}(\boldsymbol{u})^{2n}, \tag{82}$$
$$\hat{\sigma}_i^2 \triangleq \frac{\sigma^2}{1 + \hat{a}(\boldsymbol{u})^2 - 2\hat{a}(\boldsymbol{u})\cos\frac{i\pi}{n+1}}, \quad i = 2, ..., n. \tag{83}$$

The proxy random variables in Definition 2 differ from that in [1, Eq. (119)] in the behavior of the largest variance $\hat{\sigma}_1^2$.

Remark 5. Since the proxy random variable $\hat{\boldsymbol{X}}$ depends on the realization of $\boldsymbol{U}$, Definition 2 defines the joint distribution of $(\boldsymbol{X}, \hat{\boldsymbol{X}})$, where $\boldsymbol{X}$ is the decorrelation of $\boldsymbol{U}$ in (59).

The following convex optimization problem will be instrumental: for any two random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ with distributions $P_{\boldsymbol{X}}$ and $P_{\boldsymbol{Y}}$, respectively, define

$$\mathbb{R}(\boldsymbol{X}, \boldsymbol{Y}, d) \triangleq \inf_{\substack{P_{\boldsymbol{F}|\boldsymbol{X}}: \\ \mathbb{E}[\mathsf{d}(\boldsymbol{X},\boldsymbol{F})] \le d}} \frac{1}{n} D(P_{\boldsymbol{F}|\boldsymbol{X}} || P_{\boldsymbol{Y}}|P_{\boldsymbol{X}}), \tag{84}$$

where $D(P_{\boldsymbol{F}|\boldsymbol{X}}||P_{\boldsymbol{Y}}|P_{\boldsymbol{X}})$ is the conditional relative entropy. Denote the minimizer of (84) by $\boldsymbol{F}^\star$. See [1, Sec. II-B] for a discussion on various aspects of this optimization problem. For each $\boldsymbol{x}$, define for each $i = 1, ..., n$,

$$m_i(\boldsymbol{x}) \triangleq \mathbb{E}\left[(\hat{F}_i^\star - x_i)^2 | \hat{X}_i = x_i\right], \tag{85}$$

where $\hat{\boldsymbol{F}}^\star = (\hat{F}_1^\star, \ldots, \hat{F}_n^\star)^\top$ is the random variable that achieves $\mathbb{R}(\hat{\boldsymbol{X}}, \boldsymbol{Y}^\star, d)$, and $\boldsymbol{Y}^\star$ is the random variable that achieves $\mathbb{R}_{\boldsymbol{X}}(d)$ in (45). Denote η_n as

$$\eta_n \triangleq \sqrt{\frac{\log\log n}{n}}. \tag{86}$$

The typical set for the Gauss-Markov source is defined as follows.

Definition 3 (Typical set [1, Def. 1])**.** *For any $d \in (0, d_{\max})$, $n \geq 2$ and a constant $p > 0$, define $\mathcal{T}(n,p)$ to be the set of vectors $\boldsymbol{u} \in \mathbb{R}^n$ that satisfy the following conditions:*

$$|\hat{a}(\boldsymbol{u}) - a| \leq \eta_n, \tag{87}$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i^2}{\sigma_i^2} \right) - (2k-1)!! \right| \leq 2, \quad k = 1, 2, 3, \tag{88}$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} m_i(\boldsymbol{x}) - d \right| \leq p\eta_n, \tag{89}$$

where $\boldsymbol{x} = \mathsf{S}^\top \boldsymbol{u}$ is the decorrelation (59)*.*

The typical set in Definition 3 is in the same form as that in the stationary case [1, Def. 2], but the definition of proxy random variables and the analysis are different.

Theorem 8. *For any $d \in (0, d_{\max})$, there exists a constant $p > 0$ such that the probability that the Gauss-Markov source produces a typical sequence satisfies*

$$\mathbb{P}\left[\boldsymbol{U} \in \mathcal{T}(n,p)\right] \geq 1 - \Theta\left(\frac{1}{\log n}\right). \tag{90}$$

Proof. Appendix E-B. ∎

Let $\mathcal{E}$ denote the event inside the square bracket in (76). Then, we prove Lemma 4 by intersecting $\mathcal{E}$ with the typical set $\mathcal{T}(n,p)$ and the complement $\mathcal{T}(n,p)^c$, respectively, and then bounding the probability of the two intersections separately. See Appendix E-D for the details.

IV. Discussion

A. Differences between Stationary and Nonstationary Gauss-Markov Processes

It took several decades [8, 16, 18, 19, 36] to completely understand the difference in rate-distortion functions between stationary and nonstationary Gaussian autoregressive sources. We briefly summarize this subtle difference here to make the point that generalizing results from the stationary case to the nonstationary one is natural but nontrivial.

Since $\det(\mathsf{A}) = 1$, the eigenvalues μ_i's of $\mathsf{A}^\top\mathsf{A}$ satisfy

$$\prod_{i=1}^{n} \mu_i = 1. \tag{91}$$

Using (91), (47) can be equivalently rewritten as

$$\mathbb{R}_{\boldsymbol{U}}(d) = \frac{1}{n} \sum_{i=1}^{n} \max\left(0, \ \frac{1}{2} \log \frac{\sigma_i^2}{\theta_n}\right), \tag{92}$$

where $\theta_n > 0$ is given in (48) and σ_i^2's are given in (61). Both (47) and (92) are valid expressions for the n-th order rate-distortion function $\mathbb{R}_{\boldsymbol{U}}(d)$, regardless of whether the source is stationary or nonstationary. The classical Kolmogorov reverse waterfilling result [16, Eq. (18)], obtained by taking the limit in (92), implies that the rate-distortion function of the *stationary* Gauss-Markov source ($0 < a < 1$) is given by (K stands for Kolmogorov)

$$\mathbb{R}_{\mathrm{K}}(d) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \max\left(0, \ \frac{1}{2} \log \frac{\sigma^2}{\theta g(w)}\right) \ dw, \tag{93}$$

where $\theta > 0$ is given in (51) and $g(w)$ is given in (52). While (50) and (51) are valid for both stationary and nonstationary cases, Hashimoto and Arimoto [18] noticed in 1980 that (93) is not correct for the nonstationary Gaussian autoregressive source. The reason is the different asymptotic behaviors of the eigenvalues μ_i's of $\mathsf{A}^\top\mathsf{A}$ (49) in the stationary and nonstationary cases: while in the stationary case, the spectrum is bounded away from zero, in the nonstationary case, the smallest eigenvalue μ_1 approaches 0, causing a discontinuity. By treating that smallest eigenvalue in a special way, they extended (93) [18, Th. 2] (HA stands for the authors of [18]) [3]

$$\mathbb{R}_{\mathrm{HA}}(d) = \mathbb{R}_{\mathrm{K}}(d) + \log(\max(a, 1)) \tag{94}$$

to both stationary and nonstationary Gauss-Markov sources. In 2008, Gray and Hashimoto [19] showed the equivalence between $\mathbb{R}_{\mathrm{HA}}(d)$ in (94), obtained by taking a limit in (92), and Gray's result $\mathbb{R}_U(d)$ in (50), obtained by taking a limit in (47).

The tool that allows one to take limits in (92) and (47) is the following theorem on the asymptotic eigenvalue distribution of the almost Toeplitz matrix $\mathsf{A}^\top\mathsf{A}$, which is the (rescaled) inverse of the covariance matrix (62). Denote

$$\alpha \triangleq \min_{w \in [-\pi, \pi]} \ g(w) = (a-1)^2, \tag{95}$$

and

$$\beta \triangleq \max_{w \in [-\pi, \pi]} \ g(w) = (a+1)^2. \tag{96}$$

To indicate the dependence on n, we change the notation a little bit to denote $\mu_{n,i}, \ i = 1, \ldots, n$ the eigenvalues of $\mathsf{A}^\top\mathsf{A}$ listed in increasing order. Gray [37, Th. 2.4] generalized the result of Grenander and Szegö [38, Th. in Sec. 5.2] on the asymptotic eigenvalue distribution of Toeplitz forms to that of matrices that are asymptotically equivalent to Toeplitz forms, see [37, Chap. 2.3] for the details. Define

$$\alpha' \triangleq \inf_{n \in \mathbb{N}, \ i \in [n]} \ \mu_{n,i}. \tag{97}$$

Theorem 9 (Gray [8, Eq. (19)], Hashimoto and Arimoto [18, Th. 1])**.** *For any continuous function $F(t)$ over the interval*

$$t \in [\alpha', \ \beta], \tag{98}$$

the eigenvalues $\mu_{n,i}$'s of the rescaled inverse covariance matrix in (62) *of $\boldsymbol{U}$, or, equivalently, $\boldsymbol{X}$, satisfy*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} F(\mu_{n,i}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F\left(g(w)\right) \ dw, \tag{99}$$

where $g(w)$ is defined in (52)*.*

The eigenvalues $\mu_{n,i}$'s behave quite differently in the following three cases, leading the subtle difference in rate-distortion functions.

1) For the stationary case $a \in (0, 1)$, it can be easily shown [1, Eq. (71)] that $\alpha' = \alpha > 0$ and all eigenvalues $\mu_{n,i}$'s lie in between α and β. Kolmogorov's formula (93)

[3] For the general higher-order Gaussian autoregressive source, the correction term needed in (94) depends on the unstable roots of the characteristic polynomial of the source, see [18, Th. 2] for the details.

is obtained by applying Theorem 9 to (92) using the function

$$F_{\mathrm{K}}(t) \triangleq \max\left(0,\ \frac{1}{2}\log\frac{\sigma^2}{\theta t}\right), \tag{100}$$

where $\theta > 0$ is given by (51).

2) For unit-root processes / Wiener processes $a = 1$, closed-form expressions of $\mu_{n,i}$'s are given by Berger [36, Eq. (2)]. Those results imply that the smallest eigenvalue $\mu_{n,1}$ is of order $\Theta\left(\frac{1}{n^2}\right)$ and $\alpha' = \alpha = 0$. Using the same function as in (100), Berger obtained the rate-distortion functions for the Wiener processes $a = 1$ [36, Eq. 4] [4].
3) For the nonstationary case $a > 1$, we have $\alpha' = 0 < \alpha$, the smallest eigenvalue $\mu_{n,1}$ is of order $\Theta(a^{-2n})$ and the largest $n-1$ eigenvalues lie in between α and β. This behavior of eigenvalues was shown by Hashimoto and Arimoto [18, Lemma] for higher-order Gaussian autoregressive sources, and we will show a refined version for the Gauss-Markov source in Lemma 5 below. As pointed out by [18, Th. 1], an application of Theorem 9 using the function (100) fails to yield the correct rate-distortion function for nonstationary sources due to the discontinuity of $F_{\mathrm{K}}(t)$ at 0.
Gray [8, Eq. (22)] and Hashimoto and Arimoto [18] circumvent the above difficulty in two different ways, which lead to (50) and (94), respectively. Gray [8] applied Theorem 9 on (47) using the function

$$F_{\mathrm{G}}(t) = \frac{1}{2}\log\max\left(t,\ \frac{\sigma^2}{\theta}\right), \tag{101}$$

which is indeed continuous at 0, while Hashimoto and Arimoto [18, Th. 2] still use the function $F_{\mathrm{K}}(t)$ but consider $\mu_{n,1}$ and $\mu_{n,i},\ i \geq 2$ separately:

$$\frac{1}{n}\sum_{i=2}^{n} F_{\mathrm{K}}(\mu_{n,i}) + \frac{1}{n}F_{\mathrm{K}}(\mu_{n,1}), \tag{102}$$

which in the limit yields (94) by plugging $\mu_{n,1} = \Theta(a^{-2n})$ into (100).

B. New Results on the Spectrum of the Covariance Matrix

The following result on the scaling of the eigenvalues $\mu_{n,i}$'s refines [18, Lemma].

Lemma 5. *Fix $a > 1$. For any $i = 2, \ldots, n$, the eigenvalues of $\mathsf{A}^\top\mathsf{A}$ are bounded as*

$$\xi_{n-1,i-1} \leq \mu_{n,i} \leq \xi_{n,i}, \tag{103}$$

where

$$\xi_{n,i} \triangleq 1 + a^2 - 2a\cos\left(\frac{i\pi}{n+1}\right). \tag{104}$$

The smallest eigenvalue is bounded as

$$2\log a + \frac{c_2}{n} \geq -\frac{1}{n}\log\mu_{n,1} \geq 2\log a - \frac{c_1}{n}, \tag{105}$$

where $c_1 > 0$ and c_2 are constants given by

$$c_1 = 2\log(a+1) + \frac{a\pi}{a^2-1}, \tag{106}$$

$$c_2 = 2\log\frac{a}{a^2-1} + \frac{2a\pi}{a^2-1}. \tag{107}$$

Proof. Appendix B-C. ∎

Remark 6. The constant c_1 in (106) is positive, while c_2 in (107) can be positive, zero or negative, depending on the value of $a > 1$. Lemma 5 indicates that a^{-2n} is a good approximation to $\mu_{n,1}$. Using (104) and (103), we deduce that for $i = 2, \ldots, n$,

$$\mu_{n,i} \in [\alpha, \beta]. \tag{108}$$

Based on Lemma 5, we obtain a nonasymptotic version of Theorem 9, which is useful in the analysis of the dispersion, in particular, in Proposition 1 in Appendix C-A below.

Theorem 10. *Fix any $a > 1$. For any bounded, L-Lipschitz and nondecreasing function (or nonincreasing function) $F(t)$ over the interval* (98) *and any $n \geq 1$, the eigenvalues $\mu_{n,i}$'s of $\mathsf{A}^\top\mathsf{A}$ satisfy*

$$\left|\frac{1}{n}\sum_{i=1}^{n} F(\mu_{n,i}) - \frac{1}{2\pi}\int_{-\pi}^{\pi} F\left(g(w)\right)\ dw\right| \leq \frac{C_L}{n}, \tag{109}$$

where $g(w)$ is defined in (52) *and $C_L > 0$ is a constant that depends on L and the maximum absolute value of F.*

Proof. Appendix B-D. ∎

V. Conclusion

In this paper, we obtain a nonasymptotic bound (Theorem 1) on the estimation error of the maximum likelihood estimator of the parameter a of the nonstationary scalar Gauss-Markov process. An asymptotic bound (Theorem 2) follows immediately. Numerical simulations in Fig. 1 confirm the tightness of our estimation error bounds compared to previous works. As an application of the estimation error bound (Corollary 1), we find the dispersion for lossy compression of the nonstationary Gauss-Markov sources (Theorem 5). Future research directions include generalizing the error exponent bounds in this paper, applicable to identification of scalar dynamical systems, to vector systems, and finding the dispersion of the Wiener process.

Appendix A

A. Proof of Theorem 1

Proof. We present the proof of (22). The proof of (23) is similar, which we omit here. For any $n \geq 2$, denote by $\mathcal{F}_n$ the σ-algebra generated by $Z_1, \ldots, Z_n$. For any $s > 0$, $\eta > 0$, and $n \geq 2$, we denote the following random variable

$$W_n \triangleq \exp\left\{s\sum_{i=1}^{n-1}(U_i Z_{i+1} - \eta U_i^2)\right\}. \tag{110}$$

By the Chernoff bound, we have

$$\mathbb{P}\left[\hat{a}(\boldsymbol{U}) - a \geq \eta\right] \leq \inf_{s>0}\ \mathbb{E}\left[W_n\right]. \tag{111}$$

[4] To be precise, although the rate-distortion function for the Wiener process is correct in [36, Eq. 4], the proof there is not rigorous since in this case $\alpha' = \alpha = 0$ but $F_{\mathrm{K}}(t)$ is not continuous at $t = 0$ as pointed out in [19, Eq. (23)]. Therefore, the limit leading to [36, Eq. 4] needs extra justifications.

To compute $\mathbb{E}[W_n]$, we first condition on $\mathcal{F}_{n-1}$. Since Z_n is the only term in W_n that does not belong to $\mathcal{F}_{n-1}$, we have

$$\begin{aligned}&\mathbb{E}[W_n]\\ =&\mathbb{E}\left\{W_{n-1}\cdot\mathbb{E}[\exp(s(U_{n-1}Z_n-\eta U_{n-1}^2))|\mathcal{F}_{n-1}]\right\} \quad (112)\\ =&\mathbb{E}\left[W_{n-1}\cdot\exp(\alpha_1 U_{n-1}^2)\right], \quad (113)\end{aligned}$$

where α_1 is the deterministic function of s and η defined in (16), and (113) follows from the moment generating function of Z_n. To obtain a recursion, we condition on $\mathcal{F}_{n-2}$. Since U_{n-1}^2 and $U_{n-2}Z_{n-1}$ are the only two terms in $W_{n-1}\cdot\exp(\alpha_1 U_{n-1}^2)$ that do not belong to $\mathcal{F}_{n-2}$, we use the relation $U_{n-1}=aU_{n-2}+Z_{n-1}$ and we complete squares in Z_{n-1} to obtain

$$\begin{aligned}&W_{n-1}\cdot\exp(\alpha_1 U_{n-1}^2)\\ =&W_{n-2}\cdot\exp\left\{\alpha_1\left(Z_{n-1}+(a+\frac{s}{2\alpha_1})U_{n-2}\right)^2+\right.\\ &\left.(a^2\alpha_1-s\eta)U_{n-2}^2-\alpha_1\left(a+\frac{s}{2\alpha_1}\right)^2U_{n-2}^2\right\}. \quad (114)\end{aligned}$$

Furthermore, using the formula for the moment generating function of the noncentral χ^2-distributed random variable

$$\left(Z_{n-1}+\left(a+\frac{s}{2\alpha_1}\right)U_{n-2}\right)^2 \quad (115)$$

with 1 degree of freedom, we obtain

$$\begin{aligned}&\mathbb{E}\left[W_{n-1}\cdot\exp(\alpha_1 U_{n-1}^2)\right]\\ =&\frac{1}{\sqrt{1-2\sigma^2\alpha_1}}\mathbb{E}\left[W_{n-2}\cdot\exp(\alpha_2 U_{n-2}^2)\right]. \quad (116)\end{aligned}$$

This is where our method diverges from Rantzer [15, Lem. 5], who chooses $s=\frac{\eta}{\sigma^2}$ and bounds $\alpha_2\le\alpha_1$ (due to property (A4) in Appendix A-B below) in (116). Instead, by conditioning on $\mathcal{F}_{n-3}$ in (116) and repeating the above recursion, we compute $\mathbb{E}[W_n]$ exactly using the sequence $\{\alpha_\ell\}$:

$$\mathbb{E}[W_n]=\exp\left\{-\frac{1}{2}\sum_{\ell=1}^{n-1}\log(1-2\sigma^2\alpha_\ell)\right\}. \quad (117)$$

If $s\notin\mathcal{S}_n^+$ then $\mathbb{E}[W_n]=+\infty$. Therefore,

$$\inf_{s>0}\ \mathbb{E}[W_n]=\inf_{s\in\mathcal{S}_n^+}\ \mathbb{E}[W_n]. \quad (118)$$

■

B. Properties of the Sequences α_ℓ and β_ℓ

We derive several important elementary properties about the sequences α_ℓ and β_ℓ. First, we consider α_ℓ. We find the two fixed points $r_1<r_2$ of the recursive relation (17) by solving the following quadratic equation in x:

$$2\sigma^2x^2+[a^2+2\sigma^2s(a+\eta)-1]x+\alpha_1=0. \quad (119)$$

(A1) For any $s>0$ and $\eta>0$, (119) has two roots $r_1<r_2$, and $r_1<0$. The two roots r_1 and r_2 are given by

$$r_1=\frac{-[a^2+2\sigma^2(a+\eta)s-1]-\sqrt{\Delta}}{4\sigma^2}, \quad (120)$$

$$r_2=\frac{-[a^2+2\sigma^2(a+\eta)s-1]+\sqrt{\Delta}}{4\sigma^2}, \quad (121)$$

where Δ denotes the discriminant of (119), given by

$$\begin{aligned}\Delta=&4\sigma^4[(a+\eta)^2-1]s^2+\\ &4\sigma^2[(a+\eta)(a^2-1)+2\eta]s+(a^2-1)^2. \quad (122)\end{aligned}$$

Proof. This is verified by noting that the discriminant Δ satisfies

$$\Delta>(a^2-1)^2>0, \quad (123)$$

where $a>1$ is used. Then, it is easy to see from (120) that $r_1<0$. ■

(A2) For $\frac{2\eta}{\sigma^2}\ne s>0$ and $\eta>0$, the sequence $\left\{\frac{\alpha_\ell-r_1}{\alpha_\ell-r_2}\right\}_{\ell\in\mathbb{N}}$ is a geometric sequence with common ratio

$$q\triangleq\frac{[a^2+2\sigma^2s(a+\eta)]+2\sigma^2r_1}{[a^2+2\sigma^2s(a+\eta)]+2\sigma^2r_2}. \quad (124)$$

Furthermore,

$$q\in(0,1), \quad (125)$$

and it follows immediately that

$$\alpha_\ell=r_1+\frac{(r_1-r_2)\frac{\alpha_1-r_1}{\alpha_1-r_2}q^{\ell-1}}{1-\frac{\alpha_1-r_1}{\alpha_1-r_2}q^{\ell-1}}, \quad (126)$$

$$=r_2+\frac{r_2-r_1}{\frac{\alpha_1-r_1}{\alpha_1-r_2}q^{\ell-1}-1}. \quad (127)$$

Proof. Using the recursion (17) and the fact that r_1 and r_2 are the fixed points of (17), one can verify that $\left\{\frac{\alpha_\ell-r_1}{\alpha_\ell-r_2}\right\}_{\ell\in\mathbb{N}}$ is a geometric sequence with common ratio q given by (124). The relation (125) is verified by direct computations using (120) and (121). ■

(A3) For any $\frac{2\eta}{\sigma^2}\ne s>0$ and $\eta>0$, we have

$$\lim_{\ell\to\infty}\alpha_\ell=r_1. \quad (128)$$

For $s=\frac{2\eta}{\sigma^2}$, we have $\alpha_\ell=0=r_2>r_1,\ \forall\ell\ge1$.

Proof. The limit (128) follows from (125) and (126). Plugging $s=\frac{2\eta}{\sigma^2}$ into (16) yields $\alpha_1=0$, which implies $\alpha_\ell=0$ for $\ell\ge1$ by (17). ■

(A4) For any $s\in\mathcal{I}_\eta$, we have $\alpha_\ell<0$, which decreases to r_1 geometrically. For $s>\frac{2\eta}{\sigma^2}$, (128) still holds, but the convergence is not monotone: there exists an $\ell^\star\ge1$ such that $\alpha_\ell>0$ and increases to $\alpha_{\ell^\star}$ for $1\le\ell\le\ell^\star$; and $\alpha_\ell<0$ and increases to r_1 for $\ell>\ell^\star$.

Proof. Due to (127), the monotonicity of α_ℓ depends on the signs of r_2-r_1 and $\frac{\alpha_1-r_1}{\alpha_1-r_2}$. Note that $r_2-r_1>0$ by (A1). Plugging $x=\alpha_1$ into (119), we have

$$(\alpha_1-r_1)(\alpha_1-r_2)=(a+\sigma^2s)^2\alpha_1. \quad (129)$$

Since for $s\in\mathcal{I}_\eta$ we have $\alpha_1<0$ by (16), (129) implies that $\frac{\alpha_1-r_1}{\alpha_1-r_2}<0$ for any $s\in\mathcal{I}_\eta$. This immediately implies that α_ℓ decreases to r_1 due to (126) and (127). Therefore, $\alpha_\ell\le\alpha_1<0,\ \forall\ell\ge1$. For any $s>\frac{2\eta}{\sigma^2}$, we have $\alpha_1>0$ and $\frac{\alpha_1-r_1}{\alpha_1-r_2}>0$. In fact, since $r_1<0$, we have $\alpha_1>r_2$, which implies $\frac{\alpha_1-r_1}{\alpha_1-r_2}>1$. Therefore, the conclusion follows from (127). ■

(A5) For any $\eta>0$, the root r_1 in (120) is a decreasing function in $s>0$.

Proof. Direct computation using (120) and (122). Here we rely on $a > 1$. ■

The sequence β_ℓ is analyzed similarly, although it is slightly more involved than α_ℓ. We only consider $0 < s \leq \frac{2\eta}{\sigma^2}$ in the rest of this section. We find the two fixed points $t_1 < t_2$ of the recursive relation (19) by solving the following quadratic equation in x:

$$2\sigma^2 x^2 + [a^2 + 2\sigma^2 s(-a+\eta) - 1]x + \beta_1 = 0. \tag{130}$$

(B1) For $s = \frac{2\eta}{\sigma^2}$, we have $\beta_\ell = 0,\ \forall \ell \geq 1$. For any $\eta > 0$ and $s \in \mathcal{I}_\eta$, (130) has two distinct roots $t_1 < 0 < t_2$, given by

$$t_1 = \frac{-[a^2 + 2\sigma^2 s(-a+\eta) - 1] - \sqrt{\Gamma}}{4\sigma^2}, \tag{131}$$

$$t_2 = \frac{-[a^2 + 2\sigma^2 s(-a+\eta) - 1] + \sqrt{\Gamma}}{4\sigma^2}, \tag{132}$$

where the discriminant Γ of (130) is

$$\Gamma = 4\sigma^4[(-a+\eta)^2 - 1]s^2 + 4\sigma^2[(-a+\eta)(a^2-1) + 2\eta]s + (a^2-1)^2. \tag{133}$$

Proof. We verify that $\Gamma > 0$ for any $\eta > 0$ and $s \in \mathcal{I}_\eta$. The reason that $\Gamma > 0$ is not as obvious as (123) is due to the subtle difference between (122) and (133) in the negative sign of a. Note that Γ in (133) is a quadratic equation in s and the discriminant of Γ is given by (with some elementary manipulations)

$$\gamma = 16\sigma^4(2a\eta - a^2 + 1)^2 \geq 0. \tag{134}$$

Hence, in general, (133) has two roots (distinct when $\eta \neq \frac{a^2-1}{2a}$) and Γ could be positive or negative. However, an analysis of two cases $(-a+\eta)^2 - 1 \geq 0$ and $(-a+\eta)^2 - 1 < 0$ reveals that $\Gamma > 0$ for any $\eta > 0$ and $s \in \mathcal{I}_\eta$. Therefore, (130) has two distinct roots $t_1 < t_2$ given in (131) and (132) above. From (130), we have $t_1 t_2 = \frac{\beta_1}{2\sigma^2}$, which is negative for $s \in \mathcal{I}_\eta$. Therefore, we have $t_1 < 0 < t_2$. ■

(B2) For any $\eta > 0$ and $s \in \mathcal{I}_\eta$, the sequence $\left\{\frac{\beta_\ell - t_1}{\beta_\ell - t_2}\right\}$ is a geometric sequence with common ratio

$$p \triangleq \frac{[a^2 + 2\sigma^2 s(-a+\eta)] + 2\sigma^2 t_1}{[a^2 + 2\sigma^2 s(-a+\eta)] + 2\sigma^2 t_2}. \tag{135}$$

In addition, for any $\eta > 0$ and $s \in \mathcal{I}_\eta$, we also have

$$p \in (0, 1). \tag{136}$$

It follows immediately that

$$\beta_\ell = t_1 + \frac{(t_1 - t_2)\frac{\beta_1 - t_1}{\beta_1 - t_2} p^{\ell-1}}{1 - \frac{\beta_1 - t_1}{\beta_1 - t_2} p^{\ell-1}}, \tag{137}$$

$$= t_2 + \frac{t_2 - t_1}{\frac{\beta_1 - t_1}{\beta_1 - t_2} p^{\ell-1} - 1}. \tag{138}$$

Proof. The derivation is similar to that of (A2) above for α_ℓ. ■

(B3) For any $\eta > 0$ and $s \in \mathcal{I}_\eta$, we have $\beta_\ell \leq \beta_1 < 0$ and β_ℓ decreases to t_1 geometrically:

$$\lim_{\ell \to \infty} \beta_\ell = t_1. \tag{139}$$

Proof. This can be verified using (137) and (138) by noticing that $t_2 - t_1 > 0$ and for $s \in \mathcal{I}_\eta$,

$$(\beta_1 - t_1)(\beta_2 - t_2) = (a - \sigma^2 s)^2 \beta_1 < 0. \tag{140}$$

■

(B4) For any constant $a > 1$, recall the two thresholds η_1 and η_2, defined in (34) and (35) in Section III-A, respectively. Then,

a) When $0 < \eta \leq \eta_1$, the root t_1 in (131) is an increasing function in $s \in \mathcal{I}_\eta$.
b) When $\eta \geq \eta_2$, t_1 is a decreasing function in $s \in \mathcal{I}_\eta$.
c) When $\eta_1 < \eta < \eta_2$, t_1 is a decreasing function in $s \in (0, s^\star)$; and an increasing function in $s \in \left(s^\star, \frac{2\eta}{\sigma^2}\right)$, where $s^\star$ is the unique solution in the interval $\mathcal{I}_\eta$ to

$$\frac{dt_1}{ds}\Big|_{s=s^\star} = 0, \tag{141}$$

and $s^\star$ is given by

$$s^\star \triangleq \frac{a\eta(\eta - \eta_1)}{\sigma^2(1 - (\eta - a)^2)}. \tag{142}$$

Proof. Using (131) and (133), we compute the derivatives of t_1 as follows:

$$\frac{dt_1}{ds} = -\frac{\eta - a}{2} - \frac{1}{\sqrt{\Gamma}}\left\{\sigma^2[(-a+\eta)^2 - 1]s + \frac{1}{2}[(-a+\eta)(a^2-1) + 2\eta]\right\}, \tag{143}$$

$$\frac{d^2 t_1}{ds^2} = \frac{\sigma^2(2a\eta - a^2 + 1)^2}{\Gamma^{\frac{3}{2}}} \geq 0. \tag{144}$$

To simplify notations, denote by $L(s)$ the first derivative:

$$L(s) \triangleq \frac{dt_1}{ds}(s). \tag{145}$$

From (143), we have

$$L(0) = \frac{-a^2(\eta - \eta_1)}{a^2 - 1}, \tag{146}$$

and

$$L\left(\frac{2\eta}{\sigma^2}\right) = \begin{cases} \frac{-2(2\eta - a)(\eta - \eta_2)(\eta - \eta_2')}{(a-2\eta)^2 - 1}, & \eta \in \left(0, \frac{a-1}{2}\right) \cup \left(\frac{a+1}{2}, +\infty\right) \\ \frac{\eta}{1 - (a-2\eta)^2}, & \eta \in \left(\frac{a-1}{2}, \frac{a+1}{2}\right), \end{cases} \tag{147}$$

where η_2' is given by

$$\eta_2' \triangleq \frac{3a - \sqrt{a^2 + 8}}{4}. \tag{148}$$

Since $L(s)$ is an increasing function in s due to (144), to determine the monotonicity of t_1, we only need to consider the following three cases.

a) When $L(0) \geq 0$, or equivalently, $0 < \eta \leq \eta_1$, we have $L(s) \geq 0$ for any $s \in \mathcal{I}_\eta$. Hence, t_1 is an increasing function in s.

b) When $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$, we have $L(s) \leq 0$ for any $s \in \mathcal{I}_\eta$. Hence, t_1 is a decreasing function in s. We now

show that $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$ is equivalent to $\eta \geq \eta_2$. When $\eta \in \left(\frac{a-1}{2}, \frac{a+1}{2}\right)$, we have $L\left(\frac{2\eta}{\sigma^2}\right) > 0$ by (147) and $\eta > 0$. When $\eta \in \left(0, \frac{a-1}{2}\right) \cup \left(\frac{a+1}{2}, +\infty\right)$, it is easy to see from (147) that $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$ is equivalent to $\eta \in [\eta_2', a/2] \cup [\eta_2, +\infty)$. Hence, the equivalent condition for $L\left(\frac{2\eta}{\sigma^2}\right) \leq 0$ is $\eta \in [\eta_2, +\infty)$.

c) When $L(0) < 0$ and $L\left(\frac{2\eta}{\sigma^2}\right) > 0$, or equivalently, $\eta \in (\eta_1, \eta_2)$, solving (141) using (143) yields (142). Since $L(s)$ is monotonically increasing due to (144), we know that $s^\star$ given by (142) is the unique solution to (141) in $\mathcal{I}_\eta$, and $L(s) \leq 0$ for $s \in (0, s^\star]$ and $L(s) > 0$ for $s \in (s^\star, 2\eta/\sigma^2)$. ■

C. Proof of Lemma 1

Proof. We first prove $\mathcal{I}_\eta \cup \{\frac{2\eta}{\sigma^2}\} \subseteq \mathcal{S}_\infty^+$. Property (A4) above in Appendix A-B implies that for any $0 < s \leq \frac{2\eta}{\sigma^2}$, we have $\alpha_\ell \leq 0 < \frac{1}{2\sigma^2}$. Hence $\mathcal{I}_\eta \cup \{\frac{2\eta}{\sigma^2}\} \subset \mathcal{S}_n^+$ for any $n \geq 1$. To show the other direction $\mathcal{S}_\infty^+ \subseteq \mathcal{I}_\eta \cup \{\frac{2\eta}{\sigma^2}\}$, it suffices to show that for any $s > \frac{2\eta}{\sigma^2}$, there exists $n \in \mathbb{N}$ such that $\alpha_n \geq \frac{1}{2\sigma^2}$. Let $\ell^\star$ be the integer defined in property (A4) above. Then, $\ell^\star$ satisfies the following two conditions

$$\frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell^\star - 1} \geq 1, \tag{149}$$

$$\frac{\alpha_1 - r_1}{\alpha_1 - r_2} q^{\ell^\star} < 1. \tag{150}$$

We show that $\alpha_{\ell^\star} \geq \frac{1}{2\sigma^2}$, which would complete the proof. Due to $r_2 - r_1 > 0$, using (127) and (150), we have

$$\alpha_{\ell^\star} \geq r_2 + \frac{r_2 - r_1}{\frac{1}{q} - 1} \tag{151}$$

$$= \frac{r_2 - r_1 q}{1 - q} \tag{152}$$

$$= \frac{1}{2\sigma^2}, \tag{153}$$

where (153) [5] is by plugging (120), (121) and (124) into (152).

To show (29), for any $0 < s \leq \frac{2\eta}{\sigma^2}$, we have $\beta_\ell \leq 0 < \frac{1}{2\sigma^2}$, $\forall \ell \geq 1$, hence $\mathcal{I}_\eta \cup \{\frac{2\eta}{\sigma^2}\} \subseteq \mathcal{S}_\infty^-$. The other direction cannot hold since we can find many counterexamples, e.g., $a = 1.2$, $\sigma^2 = 1$, $\eta = 0.15$ and $s = 0.35 > \frac{2\eta}{\sigma^2}$, where the sequence β_ℓ increases monotonically to $t_1 \approx 0.0411 < \frac{1}{2\sigma^2}$. Hence, in this case, $0.35 \in \mathcal{S}_\infty^-$ but $0.35 \notin \left(0, \frac{2\eta}{\sigma^2}\right]$. ■

D. Proof of Theorem 2

Proof. Theorem 1 and Lemma 1 imply that for any $s \in \mathcal{I}_\eta$, we have

$$\liminf_{n\to\infty} P_n^+ \geq \lim_{n\to\infty} \frac{1}{2n} \sum_{\ell=1}^{n-1} \log(1 - 2\sigma^2 \alpha_\ell), \tag{154}$$

where recall that α_ℓ depends on s. By (128), the continuity of the function $x \mapsto \log(1-x)$ and the Cesàro mean convergence, we have

$$\liminf_{n\to\infty} P_n^+ \geq \frac{1}{2} \log(1 - 2\sigma^2 r_1), \tag{155}$$

[5] It is pretty amazing that (153) is in fact an equality.

where r_1 is given in (120). Since (155) holds for any $s \in \mathcal{I}_\eta$, using Property (A5) in Appendix A-B above and supremizing (155) over $s \in \mathcal{I}_\eta$, we obtain (30). That is, plugging $s = \frac{2\eta}{\sigma^2}$ into (155) yields (30).

Similarly, to show (31), using Property (B3) in Appendix A-B above, we have

$$\liminf_{n\to\infty} P_n^- \geq \sup_{s\in\mathcal{I}_\eta} \frac{1}{2} \log(1 - 2\sigma^2 t_1). \tag{156}$$

Then, by Property (B4) in Appendix A-B above, the supermizer s' in (156) is given by

$$s' = \begin{cases} 0, & 0 < \eta \leq \eta_1 \\ s^\star, & \eta_1 < \eta < \eta_2 \\ \frac{2\eta}{\sigma^2}, & \eta \geq \eta_2, \end{cases} \tag{157}$$

where $s^\star$ is given by (142). Plugging (157) into (156) yields (31). ■

E. Proof of Theorem 3

Proof. For any sequence η_n decreasing to 0, the proof of Theorem 1 in Appendix A-A above remains valid. We present the proof of (40), and omit that of (41), which is similar. In this regime, for each $n \geq 1$, the proof of Lemma 1 in Appendix A-C implies that

$$\left(0, \frac{2\eta_n}{\sigma^2}\right) \subset \mathcal{S}_n^+, \tag{158}$$

therefore, in (22), we choose

$$s = s_n = \frac{\eta_n}{\sigma^2} \in \mathcal{S}_n^+. \tag{159}$$

First, using (120), (121) (124) and the choice (159), we determine the asymptotic behavior of quantities involved in determining α_ℓ in (126) and (127), summarized in TABLE I.

α_1	r_1	r_2	$r_2 - r_1$	q	$-\frac{\alpha_1 - r_1}{\alpha_1 - r_2}$
$-\Theta(\eta_n^2)$	$-\Theta(1)$	$\Theta(\eta_n^2)$	$\Theta(1)$	$\Theta(1)$	$\Theta(1/\eta_n^2)$

TABLE I: Order dependence in η_n of the quantities involved in determining α_ℓ in (126) and (127).

We make two remarks before proceeding further. It can be easily verified from (124) that the common ratio q is a constant belonging to $(0, 1)$ and

$$\lim_{\eta\to 0} q = \frac{1}{a^2} \in (0, 1). \tag{160}$$

Hence, for all large n, q is bounded by positive constants between 0 and 1. Besides, from (120), we have

$$\lim_{\eta\to 0} r_1 = -\frac{a^2 - 1}{2\sigma^2}. \tag{161}$$

Second, from (126), (22) and the choice (159), we have

$$P_n^+ \geq \frac{n-1}{2n} \log(1 - 2\sigma^2 r_1) + \frac{1}{2n} \sum_{\ell=1}^{n-1} \log\left(1 - \frac{2\sigma^2(r_2 - r_1)}{1 - 2\sigma^2 r_1} \cdot \frac{\left(-\frac{\alpha_1 - r_1}{\alpha_1 - r_2}\right) q^{\ell-1}}{1 + \left(-\frac{\alpha_1 - r_1}{\alpha_1 - r_2}\right) q^{\ell-1}}\right). \tag{162}$$

Using the inequality $\log(1-x) \geq \frac{x}{x-1},\ \forall x \in (0,1)$, we have

$$\begin{aligned} & P_n^+ \\ \geq & \frac{n-1}{2n}\log(1-2\sigma^2 r_1) + \\ & \frac{1}{2n}\sum_{\ell=1}^{n-1} \frac{-1}{\frac{1-2\sigma^2 r_2}{2\sigma^2(r_2-r_1)} + \frac{1-2\sigma^2 r_1}{2\sigma^2(r_2-r_1)} \cdot \frac{1}{\left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right)q^{\ell-1}}}. \end{aligned} \tag{163}$$

Since $1-2\sigma^2 r_2 > 0$ due to (121), we can further bound P_n^+ as

$$\begin{aligned} P_n^+ \geq & \frac{n-1}{2n}\log(1-2\sigma^2 r_1) - \\ & \frac{1}{n}\left(\sum_{\ell=1}^{n-1} q^{\ell-1}\right)\frac{2\sigma^2(r_2-r_1)}{1-2\sigma^2 r_1} \cdot \left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right) && (164) \\ \geq & \frac{n-1}{2n}\log(1-2\sigma^2 r_1) - \\ & \frac{1}{n}\frac{2\sigma^2(r_2-r_1)}{(1-2\sigma^2 r_1)(1-q)} \cdot \left(-\frac{\alpha_1-r_1}{\alpha_1-r_2}\right) && (165) \\ = & \frac{n-1}{2n}\log(1-2\sigma^2 r_1) - \frac{1}{n\Theta(\eta_n^2)}, && (166) \end{aligned}$$

where in the last step we used the results in TABLE I. Due to the assumption (39) on η_n and (161), we obtain (40). ■

F. Proof of Theorem 4

Proof. We point out the proof changes in generalizing to the sub-Gaussian case. There are two changes to be made in the proof of Theorem 1 in Appendix A-A above, the equality from (112) to (113) is replaced by $\leq$ since Z_n is σ-sub-Gaussian; the equality in (116) is replaced by $\leq$ due to Lemma 2. The rest of the proof for Theorem 1 remains the same for the sub-Gaussian case. Since Lemma 1 and Theorem 2, 3 depend only on the properties of the sequences α_ℓ and β_ℓ and not on the distribution of Z_n's as long as Theorem 1 holds, their proofs remain exactly the same for the sub-Gaussian case. ■

APPENDIX B

A. Proof of Lemma 3

Proof. Taking variances of both sides of (64), we obtain

$$\mathbb{V}_U(d) = \limsup_{n\to\infty}\ \frac{1}{2n}\sum_{i=1}^{n}\min\left[1,\ \left(\frac{\sigma_i^2}{\theta_n}\right)^2\right]. \tag{167}$$

Note that $\lim_{n\to\infty}\theta_n = \theta$, where $\theta > 0$ is the water level given by (51). Applying Theorem 9 to (167) with the function

$$F(t) \triangleq \frac{1}{2}\min\left[1,\ \left(\frac{\sigma^2}{\theta t}\right)^2\right], \tag{168}$$

we obtain (66). Note that the function $F(t)$ in (168) is continuous at $t = 0$. ■

B. An Integral

Lemma 6. *For any constant $r \in [-1,1]$, it holds that*

$$\int_{-\pi}^{\pi}\log(1-r\cos(w))\ dw = 4\pi\log\frac{\sqrt{1+r}+\sqrt{1-r}}{2}. \tag{169}$$

Proof. Denote

$$I(r) \triangleq \int_{-\pi}^{\pi}\log(1-r\cos(w))\ dw. \tag{170}$$

By Leibniz's rule for differentiation under the integral sign, we have

$$\begin{aligned} \frac{\mathrm{d}I(r)}{\mathrm{d}r} &= \int_{-\pi}^{\pi}\frac{\partial}{\partial r}\log(1-r\cos(w))\ dw && (171) \\ &= -2\cdot\int_0^{\pi}\frac{\cos w}{1-r\cos w}\ dw. && (172) \end{aligned}$$

With the change of variable $u = \tan\left(\frac{w}{2}\right)$ and partial-fraction decomposition, we obtain the closed-form solution to the integral in (172):

$$\frac{\mathrm{d}I(r)}{\mathrm{d}r} = \frac{2\pi}{r} - \frac{2\pi}{r\sqrt{1-r^2}}. \tag{173}$$

It can be easily verified by directly taking derivatives that the right-side of (169) is indeed the antiderivative of (173). ■

C. Proof of Lemma 5

Proof. The bound (103) is obtained by partitioning $\mathsf{A}^\top\mathsf{A}$ into its leading principal submatrix of order $n-1$ and then applying the Cauchy interlacing theorem to that partition, see [1, Lem. 1] for details. To obtain (105), observe from (91)

$$\mu_{n,1} = \left(\prod_{i=2}^{n}\mu_{n,i}\right)^{-1}. \tag{174}$$

Combining (174) and (103) yields

$$L_n \geq -\frac{1}{n}\log\mu_{n,1} \geq R_n, \tag{175}$$

where

$$L_n \triangleq \frac{1}{n}\sum_{i=2}^{n}\log\xi_{n,i} \quad \text{and} \quad R_n \triangleq \frac{1}{n}\sum_{i=1}^{n-1}\log\xi_{n-1,i}. \tag{176}$$

Plugging (104) into (176) and then taking the limit, we obtain

$$\begin{aligned} \lim_{n\to\infty} L_n &= \lim_{n\to\infty} R_n \\ &= \frac{1}{\pi}\int_0^{\pi}\log(1+a^2-2a\cos(w))\ dw && (177) \\ &= 2\log a, && (178) \end{aligned}$$

where the last equality is due to Lemma 6 in Appendix B-B above. In the rest of the proof, we obtain the following refinement of (178): for any $n \geq 1$,

$$\begin{aligned} R_n &\geq 2\log a - \frac{c_1}{n}, && (179) \\ L_n &\leq 2\log a + \frac{c_2}{n}, && (180) \end{aligned}$$

where c_1 and c_2 are the constants given by (106) and (107) in Lemma 5, respectively. Then, (105) will follow directly from (175), (179) and (180).

The proofs of the refinements (179) and (180) are similar, and both are based on the elementary relations between Riemann sums and their corresponding integrals. We present the proof of (179), and omit that of (180). Note that the function $h(w) \triangleq \frac{1}{\pi}\log(1+a^2-2a\cos(w))$ is an increasing function in $w \in [0,\pi]$, and its derivative is bounded above by $M_1 \triangleq \frac{2a}{\pi(a^2-1)}$ for any fixed $a>1$. Therefore, from (104) and (176), we have

$$\left| R_n + \frac{1}{n}\log(a+1)^2 - \frac{1}{\pi}\int_0^\pi \log(g(w))\ dw \right| \le \frac{M_1\pi^2}{2n}, \tag{181}$$

and (179) follows immediately. ∎

D. Proof of Theorem 10

Proof. From Lemma 5, we know that $\alpha' = 0 < \alpha$ (recall (95) and (97)). Since $g(w)$ is an even function, we have

$$I \triangleq \frac{1}{2\pi}\int_{-\pi}^{\pi} F(g(w))\ dw \tag{182}$$

$$= \frac{1}{\pi}\int_0^\pi F(g(w))\ dw. \tag{183}$$

Denote the maximum absolute value of F over the interval (98) by $T>0$. It is easy to check that the function $F(g(w))$ is $2aL$-Lipschitz since $F(\cdot)$ is L-Lipschitz and the derivative of $g(w)$ is bounded by $2a$. For the following Riemann sum

$$S_n \triangleq \frac{1}{n}\sum_{i=1}^n F\left(g\left(\frac{i\pi}{n}\right)\right), \tag{184}$$

the Lipschitz property implies that

$$|S_n - I| \le \frac{2aL}{\pi n}. \tag{185}$$

For $i \ge 2$, rewrite (104) and (103) as

$$g\left(\frac{(i-1)\pi}{n}\right) \le \mu_{n,i} \le g\left(\frac{i\pi}{n+1}\right). \tag{186}$$

Suppose, according to the assumption of Theorem 10, that $F(t)$ is a non-decreasing function (the non-increasing case is analyzed similarly). Denote the sum in (109) as

$$Q_n \triangleq \frac{1}{n}\sum_{i=1}^n F(\mu_{n,i}). \tag{187}$$

Then, separating $F(\mu_{n,1})$ from Q_n and applying (186), we have

$$Q_n \ge S_n - \frac{2T}{n}, \tag{188}$$

$$Q_n \le \frac{n+1}{n}S_{n+1} + \frac{3T}{n}. \tag{189}$$

Therefore, there is a constant $C_L>0$ depending on L and T such that (109) holds. ∎

APPENDIX C

In the rest of the paper, we frequently use the following notations. For any given distortion threshold $d>0$, let $\theta>0$ be the water level corresponding to d in the limiting reverse waterfilling (51). For each $n\ge 1$, let θ_n be the water level corresponding to d in the n-th order reverse waterfilling (48), and let d_n be the distortion associated to the water level θ in the n-th order reverse waterfilling (48). For clarity, we explicitly write down

$$d = \frac{1}{n}\sum_{i=1}^n \min(\theta_n,\ \sigma_i^2), \tag{190}$$

$$d_n = \frac{1}{n}\sum_{i=1}^n \min(\theta,\ \sigma_i^2), \tag{191}$$

where σ_i^2's are given in (61). Note the d and θ are constants independent of n, while d_n and θ_n are functions of n. There is no direct reverse waterfilling relation between d_n and θ_n. It is easy to see that

$$\lim_{n\to\infty}\theta_n = \theta, \tag{192}$$

and

$$\lim_{n\to\infty} d_n = d. \tag{193}$$

A. Expectation and Variance of the d-tilted Information

Proposition 1. *For any $d\in(0,d_{\max})$ and $n\ge 1$, let d_n be defined in* (191) *in Appendix C above. Then, the expectation and variance of the* d*-tilted information $\jmath_{\boldsymbol{U}}(\boldsymbol{U},d_n)$ satisfy*

$$\left|\frac{1}{n}\mathbb{E}\left[\jmath_{\boldsymbol{U}}(\boldsymbol{U},d_n)\right] - \mathbb{R}_U(d)\right| \le \frac{C_1}{n}, \tag{194}$$

$$\left|\frac{1}{n}\mathbb{V}\left[\jmath_{\boldsymbol{U}}(\boldsymbol{U},d_n)\right] - \mathbb{V}_U(d)\right| \le \frac{C_2}{n}, \tag{195}$$

where $\mathbb{R}_U(d)$ and $\mathbb{V}_U(d)$ are the rate-distortion function given in (50) *and the informational dispersion given in* (66)*, respectively, and C_1 and C_2 are positive constants.*

Proof. Similar to (64), one can obtain

$$\jmath_{\boldsymbol{U}}(\boldsymbol{U},d_n) = \sum_{i=1}^n \frac{\min(\theta,\ \sigma_i^2)}{2\theta}\left(\frac{X_i^2}{\sigma_i^2}-1\right) + \frac{1}{2}\sum_{i=1}^n \log\frac{\max(\theta,\ \sigma_i^2)}{\theta}, \tag{196}$$

where $\boldsymbol{X} = (X_1,\ldots,X_n)^\top$ is the decorrelation of $\boldsymbol{U}$ defined in (59). Using (60) and taking expectations and variances of both sides of (196), we have

$$\frac{1}{n}\mathbb{E}\left[\jmath_{\boldsymbol{U}}(\boldsymbol{U},d_n)\right] = \frac{1}{2n}\sum_{i=1}^n \log\max\left(1\ ,\frac{\sigma_i^2}{\theta}\right), \tag{197}$$

$$\frac{1}{n}\mathrm{Var}\left[\jmath_{\boldsymbol{U}}(\boldsymbol{U},d_n)\right] = \frac{1}{2n}\sum_{i=1}^n \min\left(1\ ,\frac{\sigma_i^4}{\theta^2}\right). \tag{198}$$

Applying Theorem 10 to (197) with the function $F_{\mathrm{G}}(t)$ defined in (101) yields (194). Similarly, applying Theorem 10 to (198) with the function (168) yields (195). ∎

Proposition 1 and its proof are similar to those of [1, Eq. (95)–(96)]. The difference is that we apply Theorem 10, which is the nonstationary version of [1, Th. 4], to a different function in (197).

B. Approximation of the d-tilted Information

The following proposition approximates the d-tilted information $\jmath_{\boldsymbol{U}}(\boldsymbol{U}, d)$ by $\jmath_{\boldsymbol{U}}(\boldsymbol{U}, d_n)$. The proof in [1, App. D-B] works through for the nonstationary case as well.

Proposition 2. *For any $d \in (0, d_{\max})$, there exists a constant $t > 0$ (depending on d only) such that*

$$\mathbb{P}\left[|\jmath_{\boldsymbol{U}}(\boldsymbol{U}, d) - \jmath_{\boldsymbol{U}}(\boldsymbol{U}, d_n)| > t\right] \le \frac{1}{n}, \tag{199}$$

where d_n is defined in (191)*.*

Appendix D
Converse Proof

The following lemma is a direct application of the general converse result by Kostina and Verdú [10] to the Gauss-Markov source.

Lemma 7 ([10, Th. 7])**.** *Fix $d \in (0, d_{\max})$. Any (n, M, d, ϵ) code must satisfy*

$$\epsilon \ge \sup_{\gamma \ge 0} \left\{\mathbb{P}\left[\jmath_{\boldsymbol{X}}(\boldsymbol{X}, d) \ge \log M + \gamma\right] - \exp(-\gamma)\right\}, \tag{200}$$

where $\boldsymbol{X} = (X_1, \ldots, X_n)^\top$ is the decorrelation (59) *of the Gauss-Markov source $\boldsymbol{U}$.*

Proof of Theorem 6. With Lemma 7 and the established Propositions 1 and 2, the proof is the same as the converse proof in the asymptotically stationary case [1, Th. 7, Eq. (97)–(109)]. ■

Appendix E
Lossy AEP

A. Notations

For the optimization problem $\mathbb{R}(\boldsymbol{X}, \boldsymbol{Y}, d)$ in (84), the *generalized tilted information* is given by [10, Eq. (28)]

$$\Lambda_{\boldsymbol{Y}}(\boldsymbol{x}, \delta, d) \triangleq -\delta n d - \log \mathbb{E}\left[\exp(-n\delta \mathsf{d}(\boldsymbol{x}, \boldsymbol{Y}))\right], \tag{201}$$

where $\boldsymbol{x} \in \mathbb{R}$, $\delta > 0$ and $d \in (0, d_{\max})$. For properties of the generalized tilted information, see [10, App. D]. For clarity, we list the notations used throughout this section:

1) $\boldsymbol{X}$ denotes the decorrelation of $\boldsymbol{U}$ defined in (59);
2) $\hat{\boldsymbol{X}}$ is the proxy random variable of $\boldsymbol{X}$ defined in Definition 2 in Section III-D above;
3) For any $\boldsymbol{Y}$, the random vector $\boldsymbol{F}^\star$ achieves $\mathbb{R}(\boldsymbol{X}, \boldsymbol{Y}, d)$ in (84);
4) For $\boldsymbol{Y}^\star$ that achieves $\mathbb{R}_{\boldsymbol{X}}(d)$ in (45), $\hat{\boldsymbol{F}}^\star$ is the random vector that achieves $\mathbb{R}\left(\hat{\boldsymbol{X}}, \boldsymbol{Y}^\star, d\right)$;
5) We denote $\lambda^\star$ the negative slope of $\mathbb{R}_{\boldsymbol{X}}(d)$:

$$\lambda^\star \triangleq -\mathbb{R}'_{\boldsymbol{X}}(d). \tag{202}$$

Given any source outcome $\boldsymbol{u}$, let $\boldsymbol{x}$ be the decorrelation of $\boldsymbol{u}$. Define $\lambda(\boldsymbol{x})$ as

$$\lambda(\boldsymbol{x}) \triangleq -\mathbb{R}'(\hat{\boldsymbol{X}}, \boldsymbol{Y}^\star, d). \tag{203}$$

6) Comparing the definitions of d-tilted information and the generalized tilted information, one can see that [1, Eq. (18)]

$$\jmath_{\boldsymbol{X}}(\boldsymbol{x}, d) = \Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{x}, \lambda^\star, d). \tag{204}$$

7) Recalling (60) and applying the reverse waterfilling result [6, Th. 10.3.3], we know that the coordinates of $\boldsymbol{Y}^\star$ are independent and satisfy

$$Y_i^\star \sim \mathcal{N}(0,\ \nu_i^2), \tag{205}$$

where

$$\nu_i^2 \triangleq \max(0,\ \sigma_i^2 - \theta_n), \tag{206}$$

with $\theta_n > 0$ given in (190).

B. Proof of Theorem 8

The proof is similar to [1, Th. 12], and we point out the differences between the two.

1) Our Corollary 1 implies that the condition (87) is violated with probability at most $\Theta\left(e^{-cn}\right)$ for a constant $c > \frac{1}{2}\log(a)$. This is much stronger than the bound $\Theta\left(\frac{1}{\text{poly}\log n}\right)$ in the stationary case [1, Th. 6].
2) The Berry-Esseen theorem implies that the condition (88) is violated with probability at most $\Theta\left(\frac{1}{\sqrt{n}}\right)$. This is the same as the stationary case [1, Eq. (279)–(280)].
3) Similar to [1, Eq. (281)–(299), (305)–(313)] (with minor distinctions discussed next), we can show that there exists a constant $p > 0$ such that the condition (89) is violated with probability at most $\Theta\left(\frac{1}{\log n}\right)$. The relations [1, Eq. (281)–(291)] remains the same for the nonstationary case except that the justification of [1, Eq. (286)] needs to be modified since in the nonstationary case, [1, Eq. (307), (308), (313)] hold only for $i = 2, ..., n$ but not for $i = 1$ due to Lemma 5. In order to justify [1, Eq. (286)], we separately consider the case when $i = 1$ and $i \ge 2$ in [1, Eq. (305)–(313)]. Since $\nu_1^2 = \sigma_1^2 - \theta_n = \Theta\left(a^{2n}\right)$ and $\lambda(\boldsymbol{x}) = \Theta(1)$, in the nonstationary case, we replace [1, Eq. (305)–(306)] by

$$\frac{1}{n}\sum_{i=1}^n \bar{m}_i(x_i) - \frac{1}{n}\sum_{i=1}^n m_i(\boldsymbol{x}) \tag{207}$$

$$= \frac{1}{n}\sum_{i=2}^n \frac{2\nu_i^4(\lambda(\boldsymbol{x}) - \lambda^\star)}{(1 + 2\lambda(\boldsymbol{x})\nu_i^2)(1 + 2\lambda^\star\nu_i^2)} + \frac{1}{n}\sum_{i=2}^n \frac{2x_i^2\nu_i^2(2 + 2\lambda(\boldsymbol{x})\nu_i^2 + 2\lambda^\star\nu_i^2)(\lambda(\boldsymbol{x}) - \lambda^\star)}{(1 + 2\lambda(\boldsymbol{x})\nu_i^2)^2(1 + 2\lambda^\star\nu_i^2)^2} + O\left(\frac{1}{n}\right), \tag{208}$$

and restrict the summation in [1, Eq. (309)–(313)] to be $i = 2, ..., n$. Combining this modified version of [1, Eq. (309)–(313)] and (208) implies [1, Eq. (286)] for the nonstationary case. In addition, due to 1) and 2) above, the right side of [1, Eq. (293)] is strengthened to be $\Theta\left(\frac{1}{\sqrt{n}}\right)$,

and hence the right side of [1, Eq. (297)] is strengthened to be $\Theta\left(\frac{1}{\log n}\right)$ by choosing the constant p such that

$$p \geq A_1'' + \sqrt{2}\beta, \tag{209}$$

where $A_1'' > 0$ is the constant in [1, Eq. (286)] and β defined in [1, Eq. (288)] is bounded. The reason that β is bounded is that [1, Eq. (288)] has a limit when n tends to infinity, due to [1, Eq. (285)] and Theorem 9.

4) In [1, Eq. (311)–(312)], we used the following bound: for any $\boldsymbol{u} \in \mathcal{T}(n,p)$ and $\boldsymbol{x} = \mathsf{S}^\top \boldsymbol{u}$, it holds that

$$|\lambda(\boldsymbol{x}) - \lambda^\star| \leq B_1 \eta_n, \tag{210}$$

where $B_1 > 0$ is a constant and η_n is in (86). In the stationary case, the bound (210) follows from [1, Lem. 6 and Eq. (128)]. In the nonstationary case, due to Lemma 5, [1, Eq. (128)] holds only for $i = 2, ..., n$, while for $i = 1$, we have

$$\left|\hat{\sigma}_1^2(\boldsymbol{x}) - \sigma_1^2\right| \leq O\left(\rho_1^{2n}\right), \tag{211}$$

where $\rho_1 \in (a, a^2)$ is a constant. The bound (211) follows from Lemma 5 and Corollary 1. Then, to justify (210) in the nonstationary case, we again separately consider $i = 1$ and $i \geq 2$ in [1, Proof of Lem. 6]. Specifically, we rewrite [1, Eq. (222)] as

$$\begin{aligned}
&\lambda(\boldsymbol{x}) \\
&= \left[\Theta(a^{-4n}) + \sum_{i \geq 2: \hat{\sigma}_i^2 > \hat{\theta}_n} \frac{1}{(1+2\lambda(\boldsymbol{x})\nu_i^2)^2}\right] \Bigg/ \\
&\quad \left[O\left(\rho_2^{2n}\right) + \sum_{i \geq 2: \hat{\sigma}_i^2 > \hat{\theta}_n} \frac{2(\hat{\sigma}_i^2 - \sigma_i^2 + \theta_n)}{(1+2\lambda(\boldsymbol{x})\nu_i^2)^2}\right],
\end{aligned} \tag{212}$$

where $\rho_2 \triangleq \frac{\rho_1}{a^2} \in (0,1)$ and $\hat{\theta}_n > 0$ is the water level matched to d via the n-th order reverse waterfilling (48) over $\{\hat{\sigma}_i^2\}_{i=1}^n$. It is easy to see from (212) that

$$\begin{aligned}
&\lambda(\boldsymbol{x}) \\
&= O(\rho_3^n) + \left[\sum_{i \geq 2: \hat{\sigma}_i^2 > \hat{\theta}_n} \frac{1}{(1+2\lambda(\boldsymbol{x})\nu_i^2)^2}\right] \Bigg/ \\
&\quad \left[\sum_{i \geq 2: \hat{\sigma}_i^2 > \hat{\theta}_n} \frac{2(\hat{\sigma}_i^2 - \sigma_i^2 + \theta_n)}{(1+2\lambda(\boldsymbol{x})\nu_i^2)^2}\right],
\end{aligned} \tag{213}$$

where $\rho_3 \in (0,1)$ is a constant. Then, similar to [1, Eq. (223)–(227)], plugging the bound [1, Eq. (128)], which holds only for $i = 2, ..., n$, into (213) yields (210).

With these modifications above, the proof of Theorem 8 follows in a similar way as [1, Th. 12]. ∎

C. Auxiliary Lemmas

Lemma 8 (Lower bound on the probability of distortion balls)**.** *Fix $d \in (0, d_{\max})$. For any n large enough and any $\boldsymbol{u} \in \mathcal{T}(n,p)$ defined in Definition 3 in Section III-D above, and γ defined in* (242) *below, it holds that*

$$\mathbb{P}\left[d - \gamma \leq \mathsf{d}\left(\boldsymbol{x}, \hat{\boldsymbol{F}}^\star\right) \leq d \,|\hat{\boldsymbol{X}} = \boldsymbol{x}\right] \geq \frac{K_1}{\sqrt{n}}, \tag{214}$$

where $K_1 > 0$ is a constant and $\hat{\boldsymbol{F}}^\star$ is in 3) in Appendix E-A above.

Proof. Appendix E-E. ∎

Lemma 9. *Fix $d \in (0, d_{\max})$ and $\epsilon \in (0,1)$. There exists constants C and $K_2 > 0$ such that for all n large enough,*

$$\begin{aligned}
&\mathbb{P}\left[\Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{X}, \lambda(\boldsymbol{X}), d) \leq \Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{X}, \lambda^\star, d) + C\log n\right] \\
\geq& 1 - \frac{K_2}{\sqrt{n}},
\end{aligned} \tag{215}$$

where $\lambda^\star$ and $\lambda(\boldsymbol{X})$ are defined in (202) *and* (203)*, respectively.*

Proof. The proof of Lemma 9 is the same as [1, Eq. (314)–(333)] except that we strengthen the right side of [1, Eq. (322)] to be $\Theta(e^{-cn})$ for a constant $c > \frac{1}{2}\log(a)$ due to Corollary 1. ∎

D. Proof of Lemma 4

After we have proved Lemmas 8 and 9, the proof of Lemma 4 is almost the same as that in the stationary case [1, Eq. (270)-(278)]. For completeness, we sketch the proof here. We weaken the bound [10, Lem. 1] by setting $P_{\hat{X}}$ as $P_{\hat{\boldsymbol{X}}}$ and P_Y as $P_{\boldsymbol{Y}^\star}$ to obtain that for any $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned}
&\log \frac{1}{P_{\boldsymbol{Y}^\star}(\mathcal{B}_d(\boldsymbol{x}))} \\
&\leq \inf_{\gamma > 0} \Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{x}, \lambda(\boldsymbol{x}), d) + \lambda(\boldsymbol{x}) n\gamma - \\
&\qquad \log \mathbb{P}\left[d - \gamma \leq \mathsf{d}\left(\boldsymbol{x}, \hat{\boldsymbol{F}}^\star\right) \leq d | \hat{\boldsymbol{X}} = \boldsymbol{x}\right],
\end{aligned} \tag{216}$$

where $\lambda(\boldsymbol{x})$ is in (203). Let $\mathcal{E}$ denote the event inside the square brackets in (76). Then,

$$\begin{aligned}
&\mathbb{P}[\mathcal{E}] \\
=& \mathbb{P}[\mathcal{E} \cap \mathcal{T}(n,p)] + \mathbb{P}[\mathcal{E} \cap \mathcal{T}(n,p)^c] && (217) \\
\leq& \mathbb{P}\Big[\Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{X}, \lambda(\boldsymbol{X}), d) \geq \Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{X}, \lambda^\star, d) + p(n) - \lambda(\boldsymbol{X}) n\gamma - \\
&\qquad \frac{1}{2}\log n + \log K_1,\ \mathcal{T}(n,p)\Big] + \mathbb{P}[\mathcal{T}(n,p)^c] && (218) \\
\leq& \mathbb{P}\Big[\Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{X}, \lambda(\boldsymbol{X}), d) \geq \Lambda_{\boldsymbol{Y}^\star}(\boldsymbol{X}, \lambda^\star, d) + C\log n\Big] + \\
&\qquad \mathbb{P}[\mathcal{T}(n,p)^c] && (219) \\
\leq& \frac{1}{q(n)}, && (220)
\end{aligned}$$

where

- (218) is due to (216) and Lemma 8;
- From (218) to (219), we used the fact that for $\boldsymbol{u} \in \mathcal{T}(n,p)$, $\lambda(\boldsymbol{x})$ can be bounded by

$$\left|\lambda(\boldsymbol{x}) - \frac{1}{2\theta}\right| \leq B_1, \tag{221}$$

where $B_1 > 0$ is a constant and $\theta > 0$ is given by (51). The bound (221) is obtained by the same argument as that

in the stationary case [1, Eq. (273)]; γ is chosen in (242) below; the constants c_i's, $i = 1, ...4$ in (77) are chosen as

$$c_1 = B_1 + \frac{1}{2\theta}, \tag{222}$$

$$c_2 = B_4, \tag{223}$$

$$c_3 = C + \frac{1}{2}, \tag{224}$$

$$c_4 = -\log K_1, \tag{225}$$

where $B_4 > 0$ is given in (241) below and K_1 and C are the same constants in Lemmas 8 and 9, respectively.

- (220) is due to Lemma 9 and Theorem 8.

■

E. Proof of Lemma 8

Proof. The proof is similar to the stationary case [1, Lem. 10]. We streamline the proof and point out the differences. Conditioned on $\hat{\boldsymbol{X}} = \boldsymbol{x}$, the random variable

$$\mathsf{d}\left(\boldsymbol{x}, \hat{\boldsymbol{F}}^\star\right) = \frac{1}{n}\sum_{i=1}^n \left(\hat{F}_i^\star - x_i\right)^2 \tag{226}$$

follows a noncentral χ^2-distribution with (at most) n degrees of freedom, since it is shown in [1, Eq. (282) and Lem. 4] that conditioned on $\hat{\boldsymbol{X}} = \boldsymbol{x}$, the distribution of the random variable $\hat{F}_i^\star - x_i$ is given by

$$\mathcal{N}\left(\frac{-x_i}{1+2\lambda(\boldsymbol{x})\nu_i^2}, \frac{\nu_i^2}{1+2\lambda(\boldsymbol{x})\nu_i^2}\right), \tag{227}$$

where ν_i^2's are given in (206). Then, the conditional expectation is given by

$$\mathbb{E}\left[\mathsf{d}\left(\boldsymbol{x}, \hat{\boldsymbol{F}}^\star\right) | \hat{\boldsymbol{X}} = \boldsymbol{x}\right] = \frac{1}{n}\sum_{i=1}^n m_i(\boldsymbol{x}), \tag{228}$$

where $m_i(\boldsymbol{x})$ is defined in (85) in Section III-C above. In view of (226), (228) and (89), we expect that $\mathsf{d}\left(\boldsymbol{x}, \hat{\boldsymbol{F}}^\star\right)$ concentrates around d conditioned on $\hat{\boldsymbol{X}} = \boldsymbol{x}$ for $\boldsymbol{u} \in \mathcal{T}(n,p)$. Note that the proof of Theorem 8 related to (89) is different from the one in the stationary case, see Appendix E-B above for the details. To simplify notations, we denote the variances as

$$V_i(\boldsymbol{x}) \triangleq \text{Var}\left[\left(\hat{F}_i^\star - x_i\right)^2 | \hat{\boldsymbol{X}} = \boldsymbol{x}\right], \tag{229}$$

$$V(\boldsymbol{x}) \triangleq \sqrt{\frac{1}{n}\sum_{i=1}^n V_i(\boldsymbol{x})}. \tag{230}$$

Due to (227) and (89), we see $(\hat{F}_i - x_i)^2$'s have finite second- and third- order absolute moments. That is, we have

$$V(\boldsymbol{x}) = \Theta(1), \tag{231}$$

for $\boldsymbol{u} \in \mathcal{T}(n,p)$. Therefore, we can apply the Berry-Esseen theorem. Hence,

$$\begin{aligned}
&\mathbb{P}\left[d - \gamma \le \mathsf{d}\left(\boldsymbol{x}, \hat{\boldsymbol{F}}^\star\right) \le d \,|\hat{\boldsymbol{X}} = \boldsymbol{x}\right] \\
=&\mathbb{P}\Bigg[\frac{n(d-\gamma) - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})} \\
&\le \frac{1}{\sqrt{n}V(\boldsymbol{x})}\sum_{i=1}^n \left[\left(\hat{F}_i^\star - x_i\right)^2 - m_i(\boldsymbol{x})\right] \\
&\le \frac{nd - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})} \,|\hat{\boldsymbol{X}} = \boldsymbol{x}\Bigg] \qquad (232) \\
\ge&\Phi\left(\frac{nd - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})}\right) \\
&- \Phi\left(\frac{n(d-\gamma) - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})}\right) - \frac{2B_1}{\sqrt{n}} \qquad (233) \\
=&\frac{\sqrt{n}\gamma}{V(\boldsymbol{x})}\Phi'(\xi) - \frac{2B_1}{\sqrt{n}}, \qquad (234)
\end{aligned}$$

where

- (233) follows from the Berry-Esseen theorem; $B_1 > 0$ is a constant, and

$$\Phi(t) \triangleq \frac{1}{\sqrt{2\pi}}\int_{-\infty}^t e^{-\frac{\tau^2}{2}}\, d\tau \tag{235}$$

is the cumulative distribution function of the standard Gaussian distribution;

- (234) is due to the mean value theorem and

$$\Phi'(t) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}; \tag{236}$$

- In (234), ξ satisfies

$$\frac{n(d-\gamma) - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})} \le \xi \le \frac{nd - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})}. \tag{237}$$

By (89) and (231), we see that there is a constant $B_2 > 0$ such that

$$\left|\frac{nd - \sum_{i=1}^n m_i(\boldsymbol{x})}{\sqrt{n}V(\boldsymbol{x})}\right| \le B_2\sqrt{\log\log n}. \tag{238}$$

Hence, as long as γ in (237) satisfies

$$\gamma \le O(\eta_n), \tag{239}$$

where η_n is defined in (86), there exists a constant $B_3 > 0$ such that

$$|\xi| \le B_3\sqrt{\log\log n}. \tag{240}$$

Let $B_4 > 0$ be a constant such that

$$B_4 \ge \frac{B_3^2}{2} + 1, \tag{241}$$

and choose γ as

$$\gamma \triangleq \frac{(\log n)^{B_4}}{n}, \tag{242}$$

which satisfies (239). Then, plugging the bounds (231), (240), (241) and (242) into (234), we conclude that there exists a constant $K_1 > 0$ such that (234) is further bounded from below by $\frac{K_1}{\sqrt{n}}$. ■

REFERENCES

[1] P. Tian and V. Kostina, "The dispersion of the Gauss-Markov source," *IEEE Transactions on Information Theory*, 2019, to appear.

[2] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, New Jersey: P T R Prentice Hall, 1987.

[3] J. D. Hamilton, *Time Series Analysis*. Princeton, New Jersey: Princeton University Press, Jan. 1994.

[4] J. P. Gould and C. R. Nelson, "The stochastic structure of the velocity of money," *The American Economic Review*, vol. 64, no. 3, pp. 405–418, Jun. 1974.

[5] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366a, pp. 427–431, Jun. 1979.

[6] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Nov. 2012.

[7] C. E. Shannon, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 1, pp. 142–163, Mar. 1959.

[8] R. M. Gray, "Information rates of autoregressive processes," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 412–421, Jul. 1970.

[9] A. Ingber and Y. Kochman, "The dispersion of lossy source coding," in *Proceedings of 2011 IEEE Data Compression Conference*, Snowbird, UT, USA, Mar. 2011, pp. 53–62.

[10] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 58, no. 6, pp. 3309–3338, Jun. 2012.

[11] J. S. White, "The limiting distribution of the serial correlation coefficient in the explosive case," *The Annals of Mathematical Statistics*, pp. 1188–1197, Dec. 1958.

[12] J. Rissanen and P. Caines, "The strong consistency of maximum likelihood estimators for ARMA processes," *The Annals of Statistics*, pp. 297–315, Mar. 1979.

[13] B. Bercu, F. Gamboa, and A. Rouault, "Large deviations for quadratic forms of stationary Gaussian processes," *Stochastic Processes and their Applications*, vol. 71, no. 1, pp. 75–90, Oct. 1997.

[14] J. Worms, "Large and moderate deviations upper bounds for the Gaussian autoregressive process," *Statistics & probability letters*, vol. 51, no. 3, pp. 235–243, Feb. 2001.

[15] A. Rantzer, "Concentration bounds for single parameter adaptive control," in *Proceedings of 2018 IEEE Annual American Control Conference*, Milwaukee, WI, USA, Jun. 2018, pp. 1862–1866.

[16] A. Kolmogorov, "On the Shannon theory of information transmission in the case of continuous signals," *IRE Transactions on Information Theory*, vol. 2, no. 4, pp. 102–108, Dec. 1956.

[17] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, New Jersey: Prentice Hall, 1971.

[18] T. Hashimoto and S. Arimoto, "On the rate-distortion function for the nonstationary Gaussian autoregressive process," *IEEE Transactions on Information Theory*, vol. 26, no. 4, pp. 478–480, Jul. 1980.

[19] R. M. Gray and T. Hashimoto, "A note on rate-distortion functions for nonstationary Gaussian autoregressive processes," *IEEE Transactions on Information Theory*, vol. 54, no. 3, pp. 1319–1322, Feb. 2008.

[20] H. B. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica, Journal of the Econometric Society*, pp. 173–220, Jul. 1943.

[21] H. Rubin, "Consistency of maximum likelihood estimates in the explosive case," *Statistical Inference in Dynamic Economic Models*, pp. 356–364, 1950.

[22] T. W. Anderson, "On asymptotic distributions of estimates of parameters of stochastic difference equations," *The Annals of Mathematical Statistics*, pp. 676–687, Sep. 1959.

[23] N. H. Chan and C.-Z. Wei, "Asymptotic inference for nearly nonstationary AR(1) processes," *The Annals of Statistics*, pp. 1050–1063, Sep. 1987.

[24] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Berlin: Springer-Verlag, 2010.

[25] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht, "Learning without mixing: Towards a sharp analysis of linear system identification," *arXiv preprint arXiv:1802.08334*, May 2018.

[26] S. Oymak and N. Ozay, "Non-asymptotic identification of LTI systems from a single trajectory," *arXiv preprint arXiv:1806.05722*, Jun. 2018.

[27] T. Sarkar and A. Rakhlin, "How fast can linear dynamical systems be learned?" *arXiv preprint arXiv:1812.01251*, Dec. 2018.

[28] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Finite time identification in unstable linear systems," *Automatica*, vol. 96, pp. 342–353, Oct. 2018.

[29] B. Bercu and A. Touati, "Exponential inequalities for self-normalized martingales with applications," *The Annals of Applied Probability*, vol. 18, no. 5, pp. 1848–1869, 2008.

[30] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge: Cambridge University Press, 2019, vol. 48.

[31] T. Goblick, "A coding theorem for time-discrete analog data sources," *IEEE Transactions on Information Theory*, vol. 15, no. 3, pp. 401–407, May 1969.

[32] W. S. Wong and R. W. Brockett, "Systems with finite communication bandwidth constraints–II: Stabilization with limited information feedback," *IEEE Transactions on Automatic Control*, vol. 44, no. 5, pp. 1049–1053, May 1999.

[33] J. Baillieul, "Feedback designs for controlling device arrays with communication channel bandwidth constraints," in *Proceedings of 1999 ARO Workshop on Smart Structures*, Pennsylvania State University, State College, PA, USA, Aug. 1999, pp. 48–55.

[34] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on Automatic Control*, vol. 49, no. 7, pp. 1056–1068, Jul. 2004.

[35] A. Dembo and L. Kontoyiannis, "Source coding, large deviations, and approximate pattern matching," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1590–1615, Jun. 2002.

[36] T. Berger, "Information rates of Wiener processes," *IEEE Transactions on Information Theory*, vol. 16, no. 2, pp. 134–139, Mar. 1970.

[37] R. M. Gray, "Toeplitz and Circulant Matrices: A Review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[38] U. Grenander and G. Szegö, *Toeplitz Forms and their Applications*. New York: Chelsea Publishing Company, 1984.