MINIMUM ENTROPY PURSUIT: NOISE ANALYSIS

Shirin Jalali[†] *H. Vincent Poor*[⋆]

[†] Nokia - Bell Labs
[⋆]Electrical Engineering Department, Princeton University

ABSTRACT

Universal compressed sensing algorithms recover a "structured" signal from its under-sampled linear measurements, without knowing its distribution. The recently developed minimum entropy pursuit (MEP) optimization suggests a framework for developing universal compressed sensing algorithms. In the noiseless setting, among all signals that satisfy the measurement constraints, MEP seeks the "simplest". In this work, the effect of noise on the performance of the relaxed version of MEP optimization, namely Lagrangian-MEP, is studied. It is proved that the performance the Lagrangian-MEP algorithm is robust to small additive noise.

***Index Terms*—Universal coding, mixing processes, information dimension, compressed sensing**

I. INTRODUCTION

Solving under-determined linear inverse problems has drawn considerable attention in recent years, do to its appearance in many important applications such as high-resolution magnetic resonance imaging (MRI) and high resolution radar imaging. The fundamental problem in all these applications is to estimate a signal $X^n \in \mathbb{R}^n$ from its noisy linear projections $Y^m = AX^n + Z^m$, where m is desired to be much smaller than n. Here $A \in \mathbb{R}^{m\times n}$ denotes the sensing matrix and Z^m denotes the measurement noise.

Clearly, for X^n to be recoverable from $Y^m = AX^n+Z^m$, where $m \ll n$, the signal X^n cannot be any arbitrary signal and needs to be "structured". Luckily, this is the case in the majority of applications. In other words, in most cases the desired signal does not resemble memoryless noise. Starting by sparsity in a transform domain [2], [3], and then moving beyond sparsity to structures such as group-sparsity and low-rankness, in recent years, researchers in the area of compressed sensing have investigated the problem of "structured signal recovery" and its implications.

While many natural signals comply with the simple structures that are studied in the compressed sensing literature, they often follow much more elaborate patterns as well. Taking advantage of such complex patterns can potentially lead to efficient sensing systems that require dramatically fewer measurements m. However, achieving this goal using traditional approaches is challenging. The reason is that, following such methods, for each new type of structure, a new cost function needs to be developed that both enforces that special type of structure and also leads to efficient optimization.

A fundamentally different route towards exploiting complex structures present in a signal is through "universal" coding. In information theory, an algorithm is called "universal", if it is not designed for a specific source distribution and yet it achieves the optimal performance. That is, a universal algorithm, without knowing the source distribution, asymptotically achieves the best performance achievable by Bayesian algorithms that are aware of the source model. Such codes are well-studied in information theory for different applications such as compression [4]–[8], denoising [9], [10] and prediction [11], [12].

In recent years, universal compressed sensing has been the subject of different studies [13]–[18]. In [13], [15] and [16] the authors studied the problem of universal compressed sensing of deterministic signals and proposed a universal recovery algorithm based on Kolmogorov complexity, which is known to be non-computable [19]. In [14] and [17], the authors studied the problem of maximum a posteriori probability (MAP) estimation of X^n from $Y^m = AX^n+Z^m$, where Z^m denotes an additive white Gaussian noise (AWGN). Inspired by [9], considering a universal prior for X^n [19], the authors proposed a universal recovery algorithm, conjecturing that its mean square error (MSE) performance is two times that of an optimal Bayesian estimator.

The problem of universal compressed sensing of stochastic processes was studied in [18], for the case in which there is no measurement noise in the system. For almost lossless recovery of independent and identically distributed (i.i.d.) sources, it was shown that the proposed minimum entropy pursuit (MEP) optimization and its relaxed version, Lagrangian-MEP, achieve the minimum required normalized number of measurements (n/m) [20]. In this paper, we study the performance of the Lagrangian-MEP algorithm, which is in fact is the same algorithm proposed in [14], for the noisy setting. In the small noise regime, we prove the robustness

An extended version of this paper [1] has been submitted to the *IEEE Transaction on Information Theory* and is under review. This research was supported in part by the U. S. National Science Foundation under Grant CCF-1420575.

of the algorithm. We also derive an upper bound on the error for the case in which measurements are distorted by AWGN.

The organization of the paper is as follows. Section II introduces the notation used in the paper and reviews some related background. In Section III, the MEP optimization and its Lagrangian relaxation are reviewed. Section IV presents the main results of the paper, which are on the performance of the Lagrangian-MEP algorithm in the presence of noise. The proof of Theorem 3 is presented in Appendix A, and the proof of Theorem 4 is given in an extended version of the paper [1].

I-A. Notation

Given $(x_1, \ldots, x_n) \in \mathbb{R}^n$, and $i \leq j \in \{1, \ldots, n\}$, $x_i^j \triangleq (x_i, \ldots, x_j)$. For $i = 1$, x_1^j is simply denoted as x^j. Given $x^n, y^n \in \mathbb{R}^n$, $\|x^n - y^n\|_1 \triangleq \sum_{i=1}^n |x_i - y_i|$.

Given $x \in \mathbb{R}$, let $\lfloor x \rfloor$ denote the largest integer smaller that x. By this definition, $0 \leq x - \lfloor x \rfloor < 1$, and therefore $x - \lfloor x \rfloor = \sum_{i=1}^{\infty} a_i 2^{-i}$, where, for $i = 1, 2, \ldots$, $a_i \in \{0, 1\}$. We define the b-bit quantized version of x, $[x]_b$, as $[x]_b = \lfloor x \rfloor + \sum_{i=1}^{b} a_i 2^{-i}$. For $x^n \in \mathbb{R}^n$, $[x^n]_b = ([x_1]_b, \ldots, [x_n]_b)$. Given a set $\mathcal{X} \subset \mathbb{R}$, let $\mathcal{X}_b \triangleq \{[x]_b : x \in \mathcal{X}\}$. Throughout the paper, for $x \in \mathbb{R}^+$, $\ln x$ and $\log x$ refer to the natural logarithm of x and its logarithm in base 2, respectively.

II. BACKGROUND

In this section, we review some fundamental concepts that will be needed in our presentation and analysis of MEP optimization for universal compressed sensing.

II-A. Conditional empirical entropy

The entropy rate of a stationary process $\mathbf{U} = \{U_i\}_{i=1}^{\infty}$, with finite alphabet $\mathcal{U}$, is defined as $\bar{H}(\mathbf{U}) \triangleq \lim_{n\to\infty} \frac{H(U_1,\ldots,U_n)}{n}$. The entropy rate function $\bar{H}(\cdot)$ is a well-known measure of complexity for finite-alphabet processes. Given U^n generated by the stationary ergodic process $\mathbf{U}$, there are various ways to estimate $\bar{H}(\mathbf{U})$. One such method is using the conditional empirical entropy function. The k-th order empirical distribution induced by $u^n \in \mathcal{U}^n$, $\hat{p}_k(.|u^n)$, is defined as

$$\hat{p}_k(a^k|u^n) = \frac{|\{i : u_{i-k}^{i-1} = a^k, k+1 \leq i \leq n\}|}{n-k},$$

for all $a^k \in \mathcal{U}^k$.

Definition 1 (Conditional empirical entropy)**.** *The k-th order conditional empirical entropy of $u^n \in \mathcal{U}^n$, $\hat{H}_k(u^n)$, is defined as $H(V_{k+1}|V^k)$, where $V^{k+1} \sim \hat{p}_{k+1}(\cdot|u^n)$.*

II-B. Information dimension

As discussed earlier, compressed sensing, i.e., recovering a vector X^n from $Y^m = AX^n + Z^m$, with $m < n$, is only possible if the source is "structured". Unlike discrete-alphabet processes, the entropy function cannot be used to distinguish between structured (low-complexity) and unstructured (high-complexity) continuous-alphabet processes. All such sources have infinite entropy rate. In order to develop a similar measure of complexity for continuous-alphabet processes, in [18], Rényi's notion of information dimension (ID) for random variables and random vectors [21] was extended to stationary processes. In the rest of this section, we briefly review this measure.

Definition 2 (ID of a stationary process)**.** *The k-th order upper ID of a stationary process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$ is defined as*

$$\bar{d}_k(\mathbf{X}) = \limsup_{b\to\infty} \frac{H([X_{k+1}]_b|[X^k]_b)}{b}.$$

Similarly, the k-th order lower ID of $\mathbf{X}$ is defined as $\underline{d}_k(\mathbf{X}) = \liminf_{b\to\infty} \frac{H([X_{k+1}]_b|[X^k]_b)}{b}$. The upper (lower) ID of $\mathbf{X}$, $\bar{d}_o(\mathbf{X})$ ($\underline{d}_o(\mathbf{X})$), is defined as $\bar{d}_o(\mathbf{X}) = \lim_{k\to\infty} \bar{d}_k(\mathbf{X})$ ($\underline{d}_o(\mathbf{X}) = \lim_{k\to\infty} \underline{d}_k(\mathbf{X})$), when this limit exists. If $\bar{d}_o(\mathbf{X}) = \underline{d}_o(\mathbf{X})$, then the ID of $\mathbf{X}$, $d_o(\mathbf{X})$, is defined as $d_o(\mathbf{X}) = \bar{d}_o(\mathbf{X}) = \underline{d}_o(\mathbf{X})$.

As argued in [18], the ID of a process is a measure of its structuredness and has close connections to the problem of compressed sensing. (Refer to [1] for the evaluation of the IDs of several structured processes.)

II-C. Mixing processes

As discussed above, the ID of a stationary continuous-valued process $\mathbf{X} = \{X_n\}_{-\infty}^{\infty}$ serves as a measure of complexity for *processes*. However, in order to develop a universal compressed sensing algorithm, we need to be able to somehow distinguish between unstructured and structured *sequences*. Therefore, similar to finite-alphabet sources, an estimator of the ID of a stationary process is a reasonable candidate for a general measure of complexity for continuous-alphabet sequences. To be able to develop such an estimator, [18] imposed a mixing condition on the source process, which requires sufficiently-spaced future and past of the process to be almost independent of each other.

Given stationary process $\mathbf{X}$ and $j \leq k$, let $\mathcal{F}_j^k$ denote the σ-field of events generated by X_j^k. Then, the function $\psi^* : \mathbb{N} \to \mathbb{R}^+$ is defined as

$$\psi^*(g) = \sup \frac{\mathrm{P}(\mathcal{A} \cap \mathcal{B})}{\mathrm{P}(\mathcal{A})\,\mathrm{P}(\mathcal{B})},$$

where the supremum is taken over all events $\mathcal{A} \in \mathcal{F}_{-\infty}^j$ and $\mathcal{B} \in \mathcal{F}_{j+g}^{\infty}$, such that $P(\mathcal{A}) > 0$ and $\mathrm{P}(\mathcal{B}) > 0$.[1]

Definition 3 (ψ^*-mixing processes)**.** *A stationary process $\mathbf{X}$ is called ψ^*-mixing, if $\lim_{g\to\infty} \psi^*(g) = 1$.*

Some examples of ψ^*-mixing include memoryless sources, aperiodic finite-alphabet Markov chains and moving

[1]For more information on ψ^*-mixing, and its connection to other mixing conditions, the reader is referred to [22].

averages of i.i.d. processes. The following result proved in [18] shows that the k-th order empirical distribution of the quantized version of a sequence X^n generated by a ψ^*-mixing process $\mathbf{X}$ converges to the k-th order distribution of the quantized process $[\mathbf{X}]_b = \{[X_i]_b\}$. This result is an important tool that is used in all of the following theorems on the performance of MEP optimization.

Theorem 1. *Consider a ψ^*-mixing process $\mathbf{X}$, with continuous alphabet $\mathcal{X}$. Then, for any $\epsilon > 0$, there exists $g \in \mathbb{N}$, depending only on ϵ, such that for any $n > 6(k+g)/\epsilon + k$,*

$$\mathrm{P}(\|\hat{p}_k(\cdot|[X^n]_b) - \mu_k\|_1 \geq \epsilon) \leq 2^{c\epsilon^2/8}(k+g)n^{|\mathcal{Z}|^k} 2^{-\frac{nc\epsilon^2}{8(k+g)}},$$

where $c = 1/(2\ln 2)$. Here, for $a^k \in \mathcal{X}_b^k$, $\mu_k(a^k) = \mathrm{P}([X^k]_b = a^k)$.

III. MINIMUM ENTROPY PURSUIT

A universal compressed sensing decoder estimates X^n from observations $Y^m = AX^n$, without knowing the source distribution. The (upper) ID of a process $\mathbf{X}$, $\bar{d}_o(\mathbf{X}) = \lim_{k\to\infty} \limsup_{b\to\infty} H([X_{k+1}]_b|[X^k]_b)/b$, captures the level of structuredness of process $\mathbf{X}$. This suggests that, given an individual sequence $x^n \in \mathcal{X}^n$, $\hat{H}_k([x^n]_b)/b$ might serve as a good candidate to measure the complexity of x^n. Based on this intuition and Occam's razor, MEP optimization proposed in [18] estimates X^n as the sequence that satisfies the measurements constraints and among all such sequences minimizes $\hat{H}_k([x^n]_b)/b$. In other words,

$$\hat{X}^n_{\mathrm{MEP}} = \underset{x^n:\, Ax^n = Y^m}{\arg\min}\ \hat{H}_k([x^n]_b).$$

Note that this is a very challenging optimization problem to solve, since it is requires minimizing a discrete cost function over continuous variables. In order to derive a more manageable optimization, [18] considered the Lagrangian relaxed version of MEP, in which the optimization is now over a discrete set. More precisely,

$$\hat{X}^n_{\mathrm{L-MEP}} = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \Big(\hat{H}_k(u^n) + \frac{\lambda}{n^2}\|Au^n - Y^m\|_2^2\Big). \quad (1)$$

To evaluate the performance of this algorithm, there are three parameters that need to be specified: i) k (memory parameter), ii) b (quantization level) and iii) λ (regularization coefficient).

Theorem 2. *Consider a ψ^*-mixing stationary process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$, with $\mathcal{X} = [0,1]$ and upper information dimension $\bar{d}_o(X)$. Choose $r > 1$ and $\delta > 0$, and let $b = b_n = \lceil r \log\log n \rceil$, $k = k_n = o(\frac{\log n}{\log\log n})$, $\lambda = \lambda_n = (\log n)^{2r}$ and $m = m_n \geq (1+\delta)\bar{d}_o(X)n$. Also, let the entries of $A \in \mathbb{R}^{m\times n}$ be drawn i.i.d. $\mathcal{N}(0,1)$. Given X^n generated by the source $\mathbf{X}$ and $Y^m = AX^n$, let $\hat{X}^n_{\mathrm{MEP}} = \hat{X}^n_{\mathrm{MEP}}(Y^m, A)$ denote the solution of* (1)*. Then, as $n \to \infty$,*

$$\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n_{\mathrm{MEP}}\|_2 \xrightarrow{\mathrm{P}} 0.$$

For i.i.d. sources with a mixed discrete and continuous distribution, Theorem 2 proves that in the noiseless setting there is no loss in the performance due to universality. In other words, in such a setting, asymptotically, Lagrangian-MEP is successful as long as m/n exceeds the ID of the source, which is the fundamental limit of m/n in Bayesian compressed sensing [20], [23].

It turns out that (1) is the same optimization derived in [17] for MAP estimation of X^n from $Y^m = AX^n + Z^m$, using a universal prior on X^n and an AWGN Z^m.

IV. ROBUSTNESS OF MEP TO NOISE

In almost all practical situations measurements are contaminated by noise. Therefore, it is important to study the performance of the Lagrangian-MEP algorithm in the presence of the measurement noise. In this section, we explore the effect of noise on the Lagrangian-MEP algorithm. We consider two types of noise: i) small noise, where the noise power per measurement goes to zero, ii) normal noise, where the noise is AWGN and has a constant power. In the small noise regime, we prove that the performance of the Lagrangian-MEP algorithm is robust to measurement noise. For the normal noise regime, given the noise power, we characterize the trade-off between the number of measurements and the reconstruction error per symbol.

Assume that instead of AX^n, the decoder observes $Y^m = AX^n + Z^m$, where Z^m denotes the noise in the measurement system. The decoder employs the Lagrangian-MEP algorithm to recover X^n, i.e.,

$$\hat{X}^n = \underset{u^n \in \mathcal{X}_b^n}{\arg\min} \Big(\hat{H}_k(u^n) + \frac{\lambda}{m}\|Au^n - Y^m\|_2^2\Big). \quad (2)$$

By comparing (1) and (2), it can be observed that in (2) the coefficient multiplied by $\|Au^n - Y^m\|_2^2$ is changed from $\frac{\lambda}{n^2}$ to $\frac{\lambda}{m}$. The reason for this modification is that throughout this section, the entries of the matrix A are assumed to be i.i.d. $\mathcal{N}(0, \frac{1}{n})$, instead of $\mathcal{N}(0,1)$. While in the noiseless setting the variance of the entries of A does not have an impact on the performance, in the noisy setting, this power affects the signal to noise ratio (SNR) experienced by the measurements. Drawing the entries of A i.i.d. $\mathcal{N}(0, \frac{1}{n})$ ensures a fixed SNR per measurement that does not grow with n.

The following theorem asserts that Lagrangian-MEP is robust to measurement noise, and as long as the ℓ_2 norm of the noise vector is small enough, the algorithm recovers the source vector from the same number of measurements, despite receiving noisy observations.

Theorem 3. *Consider X^n generated by a ψ^*-mixing stationary process $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$, with $\mathcal{X} = [0,1]$ and upper ID $\bar{d}_o(\mathbf{X})$. Assume that we observe $Y^m = AX^n + Z^m$, where $A \in \mathbb{R}^{m\times n}$ is i.i.d. $\mathcal{N}(0, \frac{1}{n})$, and there exists a deterministic sequence c_m such that $\lim_{m\to\infty} \mathrm{P}(\frac{1}{\sqrt{m}}\|Z^m\|_2 > c_m) = 0$,*

and $c_m = O(\frac{1}{(\log m)^r})$. *Given* $r > 1$ *and* $\delta > 0$, *let* $b = b_n = \lceil r \log\log n \rceil$, $k = k_n = o(\frac{\log n}{\log\log n})$, $\lambda = \lambda_n = (\log m)^{2r}$ *and* $m = m_n \geq (1+\delta)\bar{d}_o(\mathbf{X})n$. *Further let* $\hat{X}^n_{\mathrm{L-MEP}} = \hat{X}^n_{\mathrm{L-MEP}}(Y^m, A)$ *denote the solution of* (2). *Then,* $\frac{1}{\sqrt{n}}\|X^n - \hat{X}^n\|_2 \xrightarrow{\mathrm{P}} 0$, *as* $n \to \infty$.

The following result considers the case in which the measurements are distorted by i.i.d. Gaussian noise. It suggests an alternative choice of the coefficient λ, which depends on the noise power and the source complexity. For this choice of the parameter λ, it characterizes the trade-off between the number of measurements and the per-symbol reconstruction error.

Theorem 4. *Consider a* ψ^**-mixing stationary process* $\mathbf{X} = \{X_i\}_{i=1}^{\infty}$, *with* $\mathcal{X} = [0,1]$ *and upper ID* $\bar{d}_o(\mathbf{X})$. *Given* $\delta > 0$, $w > 0$, $\tau \in (0,1)$ *and* $r > 1$, *let* $b = b_n = \lceil r \log\log n \rceil$, $k = k_n = o(\frac{\log n}{\log\log n})$, $\lambda = \lambda_n = \frac{w\bar{d}_o(\mathbf{X})b_n}{\sigma^2}$, *and* $m = m_n \geq \frac{4}{\tau^2 \log \mathrm{e}}(1+w+\delta)\bar{d}_o(\mathbf{X})b_n n$. *We observe* $Y^m = AX^n + Z^m$, *where* X^n *is generated by the source* $\mathbf{X}$, *the entries of* A *are i.i.d.* $\mathcal{N}(0, \frac{1}{n})$ *and* Z_i, $i = 1, \ldots, m$, *are i.i.d.* $\mathcal{N}(0, \sigma^2)$. *Let* $\hat{X}^n_{\mathrm{L-MEP}} = \hat{X}^n_{\mathrm{L-MEP}}(Y^m, A)$ *denote the solution of* (1). *Then,*

$$\mathrm{P}\Big(\frac{1}{\sqrt{n\sigma^2}}\|X^n - \hat{X}^n_{\mathrm{L-MEP}}\|_2 > \frac{2}{1-\tau}\sqrt{\frac{2(1+\frac{\delta}{8w})(1+w+\delta)\bar{d}_o n}{m}} + \sqrt{\frac{1+\frac{\delta}{2}}{(1-\tau)w}} + \delta\Big)$$

converges to zero, as n *grows to infinity.*

V. DISCUSSION

Theorems 3 and 4 characterize the performance of the Lagrangian-MEP algorithm in the presence of noise. Theorem 3 asserts that Lagrangian-MEP is a *robust* universal compressed sensing algorithm for almost lossless recovery, where the noise is small. Theorem 4 provides a probabilistic upper bound on the average per-symbol reconstruction error, but it does not prove the optimality or sub-optimality of the Lagrangian-MEP optimization, in the noisy setting.

APPENDIX
PROOF OF THEOREM 3

Throughout the proof, for ease of notation, $\bar{d}_o(\mathbf{X})$ and $\hat{X}^n_{\mathrm{L-MEP}}$ are denoted by $\bar{d}_o$ and $\hat{X}^n$, respectively. Let $q^n \triangleq X^n - [X^n]_b$, $\epsilon > 0$, $\tau > 0$ and $\mathcal{C}_n \triangleq \{[x^n]_b : \frac{1}{nb}\ell_{\mathrm{LZ}}([x^n]_b) \leq \bar{d}_o + 3\epsilon\}$, where $\ell_{\mathrm{LZ}}(u^n)$ denotes the length of the compressed version of u^n using the Lempel-Ziv compression algorithm [4]. Let $\sigma_{\max}(A)$ denote the maximum singular value of matrix A. Define events $\mathcal{E}_1 \triangleq \Big\{\sigma_{\max}(A) \leq 1 + 2\sqrt{\frac{m}{n}}\Big\}$, $\mathcal{E}_2 \triangleq \{\frac{1}{b}\hat{H}_k([X^n]_b) \leq \bar{d}_o + \epsilon\}$, $\mathcal{E}_3 \triangleq \{\|A(u^n - [X^n]_b)\|_2 \geq \|u^n - [X^n]_b\|_2\sqrt{\frac{(1-\tau)m}{n}} : \forall u^n \in \mathcal{C}_n\}$, and $\mathcal{E}_4 \triangleq \{\frac{1}{\sqrt{m}}\|Z^m\|_2 \leq c_m\}$.

Since $\hat{X}^n$ is the minimizer of the cost function in (1), we have

$$\hat{H}_k(\hat{X}^n) + \frac{\lambda}{m}\|A\hat{X}^n - Y^m\|_2^2 \leq \hat{H}_k([X^n]_b) + \frac{\lambda}{m}\|Aq^n + Z^m\|_2^2.$$

But $\|Aq^n + Z^m\|_2^2 \leq \|Aq^n\|_2^2 + \|Z^m\|_2^2 + 2\|Aq^n\|_2\|Z^m\|_2 \leq (\sigma_{\max}(A))^2\|q^n\|_2^2 + \|Z^m\|_2^2 + 2\sigma_{\max}(A)\|q^n\|_2\|Z^m\|_2$. Since $\|q^n\|_2 \leq \sqrt{n}2^{-b}$ and $m \leq n$, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, we get

$$\begin{aligned}
&\hat{H}_k(\hat{X}^n) + \frac{\lambda}{m}\|A\hat{X}^n - Y^m\|_2^2 \\
&\leq \hat{H}_k([X^n]_b) + \frac{\lambda}{m}\Big((1 + 2\sqrt{\frac{m}{n}})^2 n 2^{-2b} + \\
&\qquad\qquad mc_m^2 + 2\sqrt{m}c_m(1 + 2\sqrt{\frac{m}{n}})\sqrt{n}2^{-b}\Big) \\
&\leq b(\bar{d}_o + \epsilon) + \lambda\Big(\frac{9n}{m}(2^{-2b}) + c_m^2 + 6c_m\sqrt{\frac{n}{m}}2^{-b}\Big).
\end{aligned} \tag{A.3}$$

Dividing both sides of (A.3) by b, it follows that

$$\begin{aligned}
&\frac{1}{b}\hat{H}_k(\hat{X}^n) + \frac{\lambda}{bm}\|A\hat{X}^n - Y^m\|_2^2 \\
&\leq \bar{d}_o + \epsilon + \frac{\lambda}{b}\Big(\frac{9n}{m}(2^{-2b}) + c_m^2 + 6c_m\sqrt{\frac{n}{m}}2^{-b}\Big).
\end{aligned} \tag{A.4}$$

By the theorem's assumption, $\lambda = \lambda_n = (\log m)^{2r}$, $b = b_n = \lceil r \log\log n \rceil$ and $m = m_n \geq (1+\delta)\bar{d}_o n$. For this choice of the parameters, $\frac{\lambda n}{b2^{2b}m} \leq \frac{1}{(1+\delta)\bar{d}_o r \log\log n}$, $\frac{\lambda c_m^2}{b} \leq \frac{((\log m)^r c_m)^2}{r\log\log n}$, and finally $\frac{\lambda c_m}{b2^b}\sqrt{\frac{n}{m}} \leq \frac{c_m(\log m)^r}{r(\log\log n)\sqrt{(1+\delta)\bar{d}_o}}$. Therefore, since $c_m = O(\frac{1}{(\log m)^r})$, for all n large enough, $\frac{\lambda}{b}(\frac{9n}{m}(2^{-2b}) + c_m^2 + 6c_m\sqrt{\frac{n}{m}}2^{-b}) < \epsilon$. Hence, from (A.4), for all n large enough, conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_4$, $\hat{H}_k(\hat{X}^n) \leq \bar{d}_o(1+2\epsilon)$ and $\frac{\lambda}{m}\|A\hat{X}^n - Y^m\|_2^2 \leq \bar{d}_o(1+2\epsilon)$. On the other hand, by the triangle inequality, $\|A\hat{X}^n - Y^m\|_2 \geq \|A(\hat{X}^n - X^n)\|_2 - \|Z^m\|_2$, and hence $\|A(\hat{X}^n - X^n)\|_2 \leq \|A\hat{X}^n - Y^m\|_2 + \|Z^m\|_2$. Therefore,

$$\begin{aligned}
\|A(\hat{X}^n - X^n)\|_2 &\leq \sqrt{\frac{(\bar{d}_o + 2\epsilon)bm}{\lambda}} + c_m\sqrt{m} \\
&\leq \sqrt{\frac{bm}{\lambda}}\left(\sqrt{\bar{d}_o + 2\epsilon} + \sqrt{\frac{c_m^2(\log m)^{2r}}{r\log\log n}}\right).
\end{aligned}$$

But again since $c_m = O(\frac{1}{(\log m)^r})$, $\sqrt{\frac{c_m^2(\log m)^{2r}}{r\log\log n}}$ can be made arbitrarily small, for all n large enough. On the other hand conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_4$, for all n large enough, $\|A(\hat{X}^n - [X^n]_b)\|_2 \geq \|\hat{X}^n - [X^n]_b\|_2\sqrt{\frac{(1-\tau)m}{n}}$. Combining this lower bound by the just derived upper bound on $\|A(\hat{X}^n - X^n)\|_2$ establishes the desired result. The remaining step is to show that $\mathrm{P}(\mathcal{E}_1 \cap \ldots \cap \mathcal{E}_4) \to 1$, as $n \to \infty$. This can be done following steps similar to those used in the proof of Theorem 8 in [1].

[1] S. Jalali and H. V. Poor, "Universal compressed sensing of Markov sources," *arXiv preprint arXiv:1406.7807*, 2014.

[2] D.L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[3] E. J Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[4] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Trans. Inf. Theory*, vol. 23, no. 3, pp. 337–343, 1977.

[5] D. J. Sakrison, "The rate of a class of random processes," *IEEE Trans. Inf. Theory*, vol. 16, pp. 10–16, Jan. 1970.

[6] J. Ziv, "Coding of sources with unknown statistics part II: Distortion relative to a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 18, pp. 389–394, May 1972.

[7] D. L. Neuhoff, R. M. Gray, and L. D. Davisson, "Fixed rate universal block source coding with a fidelity criterion," *IEEE Trans. Inf. Theory*, vol. 21, pp. 511–523, May 1972.

[8] D. L. Neuhoff and P. L. Shields, "Fixed-rate universal codes for Markov sources," *IEEE Trans. Inf. Theory*, vol. 24, pp. 360–367, May 1978.

[9] D. Donoho, "The Kolmogorov sampler," Tech. Rep. 2002-04, Stanford University, Jan. 2002.

[10] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú, and M. Weinberger, "Universal discrete denoising: Known channel," *IEEE Trans. Inf. Theory*, vol. 51, no. 1, pp. 5–28, 2005.

[11] M. Feder, N. Merhav, and M. Gutman, "Universal prediction for individual sequences," *IEEE Trans. Inf. Theory*, vol. 38, no. 4, pp. 1258–1270, 1992.

[12] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[13] S. Jalali and A. Maleki, "Minimum complexity pursuit," in *Proc. 49th Annual Allerton Conf. on Commun., Control, and Comp.*, Sep. 2011, pp. 1764–1770.

[14] D. Baron and M. F. Duarte, "Universal MAP estimation in compressed sensing," in *Proc. 49th Annual Allerton Conf. on Commun., Control, and Comp.*, Sep. 2011.

[15] S. Jalali, A. Maleki, and R. Baraniuk, "Minimum complexity pursuit: Stability analysis," in *Proc. IEEE Int. Symp. Inform. Theory*, Jul. 2012, pp. 1857–1861.

[16] S. Jalali, A. Maleki, and R.G. Baraniuk, "Minimum complexity pursuit for universal compressed sensing," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 2253–2268, April 2014.

[17] J. Zhu, D. Baron, and M. F. Duarte, "Recovery from linear measurements with complexity-matching universal signal estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 6, pp. 1512–1527, Mar. 2015.

[18] S. Jalali and H. V. Poor, "Universal compressed sensing," in *Proc. IEEE Int. Symp. Inform. Theory*, Jul. 2016, pp. 2369–2373.

[19] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, New York, second edition, 2006.

[20] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, vol. 56, no. 8, pp. 3721 –3748, Aug. 2010.

[21] A. Rényi, "On the dimension and entropy of probability distributions," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 1-2, pp. 193–215, 1959.

[22] R. C. Bradley, "Basic properties of strong mixing conditions. a survey and some open questions," *Probability Surveys*, vol. 2, no. 2, pp. 107–144, 2005.

[23] Y. Wu and S. Verdú, "Optimal phase transitions in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6241–6263, Oct. 2012.