RZ 3339 (# 93385) 04/30/01
Computer Science 22 pages

Research Report

Delay-Constrained Capacity and Probabilistic Codes

Xiao-Yu Hu

IBM Research
Zurich Research Laboratory
8803 Rüschlikon
Switzerland
Email: xhu@zurich.ibm.com

Delay-Constrained Capacity and Probabilistic Codes

Xiao-Yu Hu

IBM Research, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

Abstract

The Noisy Channel Coding Theorem discovered by C. E. Shannon assumes infinite coding latency. The objective of this work is to identify the maximal achievable (transmit) rates over noisy, delay-constrained channels, referred to as (ϵ, n)-capacity C_ϵ^n with ϵ denoting target error probability and n coding latency (viz. block length). We investigate a family of block codes based on a probabilistic construction that approaches delay-constrained capacity closely and provably achieves the Shannon limit over an additive white Gaussian noise (AWGN) channel. We also present an improved construction of a probabilistic code with *correlated* codewords, enhancing its asymptotic distance by introducing a specific amount of correlation between codewords. Analytical results show that, if the correlation coefficients are chosen uniformly to be $-1/(M-1)$, where M denotes the number of codewords, the corresponding probabilistic code is asymptotically (in the sense of block length) the "best-$d_{\min}^{\text{asy.}}$" code.

Keywords: AWGN, Shannon limit, probabilistic codes, coding latency, (ϵ, n)-capacity

1 Introduction

Ideas presented in the special issue [1] on codes and graphs and iterative algorithms have allowed us to approach the Shannon limit of an additive white Gaussian noise (AWGN) channel to within hundredths of a decibel at the expense of very long block lengths.

However, in most applications where the system delay is strictly limited, approaching the Shannon limit becomes problematic. To construct good block codes, the major parameters of interest are the probability of block (word) decoding error p_w, the code block length n, and the rate R. The Shannon Noisy Channel Coding Theorem [2] states that, if R is less than the Shannon limit C, no matter how close they are, surely there exist codes for which the word error probability p_w becomes small exponentially with increasing n. There are, of course, prices to be paid for increasing the block length, one of which is *coding latency*. At the transmitter side, the first information bit in a block of incoming data stream must generally be delayed by n samples (or symbols) before a codeword can be formed, and at the receiver it is the same case that decoding also requires a complete codeword. The block length n is thus referred to as *coding latency*. Note that the coding latency differs from the processing delay in that it is inherent in a coding scheme and can not be reduced by increasing the processing capability.

The Shannon limit assumes infinite coding latency, but in practice any application of interest is more or less constrained with a fixed coding latency that is tolerable or affordable. It is thus tempting to ask following questions: what is the maximal achievable rate under a fixed coding latency n constraint together with a target error probability p_w, and subsequently, what is the optimal coding scheme? These two fundamental issues will be addressed in this paper.

In Section II, we present a modified definition of channel capacity, referred to as (ϵ, n)-capacity C_ϵ^n, in which the error probability of a block code with block length n is required to be no larger than ϵ. It follows immediately from the definition that the Shannon limit C turns out to be an asymptotic version of (ϵ, n)-capacity, namely $C = \lim_{n\to\infty} \lim_{\epsilon\to 0} C_\epsilon^n$. Utilizing the upper and lower bounds on the block error probability of a code as a function of its block length, we derive a universal approach to calculate respective lower and upper bounds on the (ϵ, n)-capacity of an AWGN channel. Numerical results agree with the old folk theorem that random codes are always good for large block lengths [2][7], and more importantly, we find that even for moderate to relatively small block lengths, the average performance of the ensemble of spherical random codes is remarkably close to that of a "best" sphere-packing code, despite the fact that a sphere-packing code does not exist at all. This observation strongly suggests a new philosophy to construct good codes, i.e., the use of probabilistic method to "mimic" the ensemble of spherical random codes, rather than the traditional idea of searching for a *deterministic* "best" code.

Section III deals with a particular construction of a probabilistic code over AWGN channels. Codewords are generated by i.i.d. Gaussian random variables. As such, the codebook is inherently *time-varying* from block to block, namely each time that an information block is transmitted, the old codebook is discarded and a new codebook is generated. Since the Shannon limit is an asymptotic form of (ϵ, n)-capacity, optimal codes in the sense of (ϵ, n)-

capacity should also be capable of achieving it. We show that, if block length n tends to infinite, the probabilistic code is indeed capable of achieving the Shannon limit by means of a suboptimal decoding procedure—typical set decoding.

The optimum detector (minimum probability of block error) for the probabilistic code is investigated in Section IV. The error probability performance of the optimum detector is analyzed and approximately formulated. Further, we argue that the performance of the probabilistic code with independent codewords approaches closely the average performance of the ensemble of spherical random codes for moderate to large block lengths, suggesting that the probabilistic codes approach (ϵ, n)-capacity closely.

In Section V we investigate the distance property of the probabilistic code with independent codewords. Using a linear transformation, we present a new construction of a probabilistic code with correlated codewords, improving its asymptotic distance by the introduction of correlation between codewords. Analytical results show that if the correlation coefficients are chosen uniformly to be $-1/(M-1)$, the corresponding probabilistic code is asymptotically (in the sense of block length) a "best-$d_{\min}^{\text{asy.}}$" code.

2 (ϵ, n)-Capacity

The definition of channel capacity involving coding latency n and a target block error probability ϵ is the following.

Definition 1: Consider an (n, M) code with a block length n, M codewords, and a block error probability not greater than ϵ. $R \geq 0$ is an (ϵ, n)-achievable rate if, for every $\delta \geq 0$, there exists at least such a (n, M) code with rate

$$\frac{\log_2 M}{n} \geq R - \delta.$$

The maximum (ϵ, n)-achievable rate is called the (ϵ, n)-capacity C_ϵ^n. The Shannon limit C is defined as the maximal rate that is (ϵ, n)-achievable for all $0 < \epsilon < 1$ and for all positive integer n. It follows immediately from the definition that

$$C = \lim_{n\to\infty} \lim_{\epsilon\to 0} C_\epsilon^n.$$

For the sake of simplicity, we concentrate on a discrete-time memoryless AWGN channel (as shown in Fig. 1), which usually serves as a basic analyzing tool for all other kinds of non-ideal channels. Before going on, we cite the well-known Shannon channel capacity formula — the supremum of all rates R for which there exists at least one code with vanishing error probability, that is

$$C = \max_{p_X} I(X; Y), \tag{1}$$

where p_X denotes the probability distribution of the real-valued channel input X and Y is the real-valued channel output. This formula holds for any ergodic memoryless channels [3]. Specifically, the Shannon limit of a Gaussian channel with power constraint P and noise

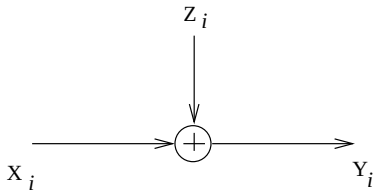

Figure 1: The discrete-time additive white Gaussian noise channel.

variance σ^2 is [1]

$$\begin{aligned} C &= \max_{EX^2 \le P} I(X;Y) \\ &= \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \end{aligned} \tag{2}$$

and the maximum is attained only when $X \sim \mathcal{N}(0,P)$, where $\mathcal{N}(0,P)$ is a zero-mean Gaussian distribution of variance P.

Evaluating the (ϵ,n)-capacity is not as simple as calculating the Shannon limit, but we can nevertheless employ known analytical results to obtain an upper-bound as well as a lower-bound on the (ϵ,n)-capacity. A classic lower-bound on the error probability for codes of a specific block length is the sphere-packing bound developed by Shannon [4]. This bound has been recently employed as a useful tool to evaluate the "imperfectness" of turbo codes [5] and has also been treated in [6]. The problem posed by Shannon is to estimate, as well as possible, the probability of error for a "best" code of length n containing M codewords, each of power P and perturbed by Gaussian noise of variance σ^2. We denote this minimal or optimal probability of error by $p_w^{\text{opt}}(M,n,\sqrt{P/\sigma^2})$. The sphere-packing bound is equal to the probability that the output sequence Y of an AWGN channel will not be confined to the cone with a solid half-angel θ centralized with respect to the transmitted codeword, which can be expressed in the form

$$\begin{aligned} &p_w^{\text{opt}}(M,n,\sqrt{P/\sigma^2}) \ge Q_{sp}(\theta) = \int_\theta^\pi \frac{(n-1)(\sin\phi)^{n-2}}{2^{n/2}\sqrt{\pi}\Gamma(\frac{n+1}{2})} \cdot \\ &\int_0^\infty r^{n-1} e^{-(r^2+nA^2-2r\sqrt{n}A\cos\phi)/2} dr d\phi, \end{aligned} \tag{3}$$

where A is the squared root of the signal-to-noise ratio (SNR), i.e., $\sqrt{P/\sigma^2}$, $\Gamma(p)$ is the Gamma function $\int_0^\infty t^{p-1}e^{-t}dt$, and θ is the root of the following equation:

$$\int_0^\theta \frac{n-1}{n}\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})\sqrt{\pi}}(\sin\phi)^{n-2}d\phi = 2^{-nR}. \tag{4}$$

[1] In this paper we deal with only real-valued channels; however the results can easily be extended to complex channels, i.e., bandpass signals represented in the equivalent complex baseband. In this case, the factor of 1/2 in (2) does not appear.

For moderate to large n, (3) can be approximated with great accuracy by

$$Q_{sp}(\theta) \approx \frac{[G(\theta)\sin\theta e^{-(A^2-AG(\theta)\cos\theta)/2}]^n}{\sqrt{n\pi}\sqrt{1+G^2(\theta)\sin\theta}[AG(\theta)\sin^2\theta-\cos\theta]}, \tag{5}$$

where $G(\theta) = (1/2)[A\cos\theta + \sqrt{A^2\cos^2\theta+4}]$, and (4) becomes, asymptotically,

$$\frac{\Gamma(\frac{n}{2}+1)(\sin\theta)^{n-1}}{n\Gamma(\frac{n+1}{2})\sqrt{\pi}\cos\theta} = 2^{-nR}. \tag{6}$$

The sphere-packing lower bound on word error probability would be reached with equality only if the code were a *perfect* code for the channel, i.e., if equal-size non-intersecting cones could be drawn around every codeword to completely fill the n-dimensional space. Such a partitioning is clearly possible only for n=1 or 2, if $M > 2$ [4]. It is very plausible intuitively that any realistic code would have a higher probability of error than a sphere-packing code. Recognizing the monotonically increasing error probability with more codewords, the rates specified by the sphere-packing bound can naturally be utilized to *upper-bound* the (ϵ, n)-capacity.

Shannon also computed an upper bound on word error probability by a spherical "random coding" method [4]. The random coding bound gives an expression for the ensemble average word error probability, averaged over the ensemble of all possible spherical codes, where each codeword is selected independently and completely at random, subject to an equal energy constraint. As n grows large enough, an asymptotic formula of the random coding bound turns out to be the sphere-packing bound multiplied by a factor essentially independent of n, that is

$$\begin{aligned} & p_w^{\text{opt}}(M, n, \sqrt{P/\sigma^2}) \le Q_{rc}(\theta) \\ = \; & Q_{sp}(\theta) + 2^{nR}\int_0^\theta \frac{\Gamma(\frac{n}{2}+1)(\sin\phi)^{n-1}}{n\Gamma(\frac{n+1}{2})\pi^{1/2}\cos\phi}\sqrt{\frac{n}{\pi}} \\ & \cdot\frac{[G(\theta)\sin\phi\exp(-\frac{P}{2\sigma^2}+\frac{1}{2}\sqrt{\frac{P}{\sigma^2}}G(\theta)\cos\phi)]^n}{\sqrt{1+G^2(\theta)}\sin^2\phi}d\phi \\ \approx \; & Q_{sp}(\theta)\left(1+\frac{AG(\theta)\sin^2\theta-\cos\theta}{2\cos\theta-AG(\theta)\sin^2\theta}\right). \end{aligned} \tag{7}$$

Since the average error probability over the ensemble of spherical random codes satisfies (7), it is clear that at least one code in the ensemble must have a sufficiently small error probability, i.e. at least one code of block length n meets the target error probability ϵ with a certain rate, which in turn, gives rise to a *lower bound* on the (ϵ, n)-capacity. It is worth emphasizing that, in the case of moderate to large n, the multiplying factor in (7) is just a little over unity; the sphere-packing and the random-coding bounds are close together, thereby yielding a sharp estimate of the (ϵ, n)-capacity.

The significance of the definition of (ϵ, n)-capacity can be seen from the following numerical example. Consider an AWGN channel on which we wish to transmit information with rate

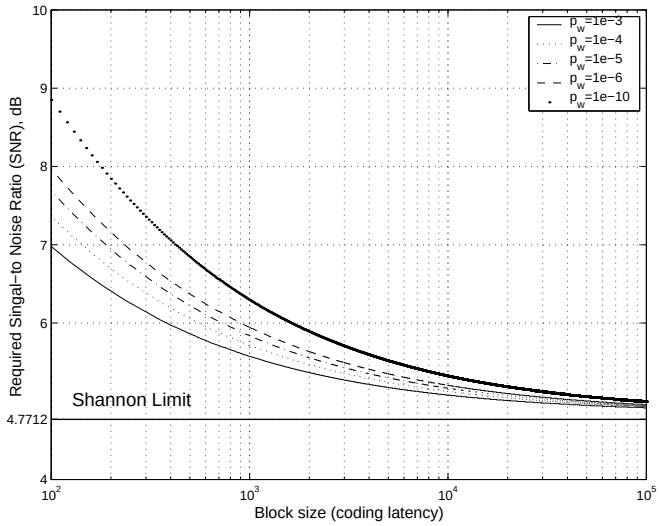

Figure 2: The minimum required signal-to-noise ratio (SNR) by the Shannon sphere-packing bound for codes with varying block length n and rate 1 bit per sample, operating over a continuous-input AWGN channel at $p_w = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-10}$, respectively.

1 bit per sample, for which the minimum SNR specified by the Shannon limit is 3.0. By applying the sphere-packing bound, Fig. 2 shows that, for the same code rate, the minimum threshold for reliable communication (in the sense of achieving a target error probability) is significantly higher than the Shannon limit, provided that the code block length is constrained to a relatively small size. For some real-time applications where large delay is not tolerable, the Shannon limit does not convey much useful information, but the (ϵ, n)-capacity reveals the ultimate limit in such cases instead. It is suggested that, even if a code operates far from the Shannon limit it might perform nearly as well as the best code possible of the same length.

A quantitative overview of the (ϵ, n)-capacity (upper bound) versus the Shannon limit is exhibited in Fig. 3 where an AWGN channel at a SNR of 3.0 is considered. It should not be surprising that, if the code block length is less than 10^4, only the rates significantly lower than the Shannon limit are achievable. For example, codes of block length 100 have a penalty of 0.3 bits per sample with $p_w = 10^{-3}$, and even a penalty of over 0.5 bits per sample with $p_w = 10^{-10}$, as compared with the Shannon limit. The Shannon limit can be approached within 0.05 bits per sample only for block lengths 100,000 and greater. Again, it is evident that, not the Shannon limit, but the (ϵ, n)-capacity should be employed in evaluating a practical coding scheme with finite block lengths.

Plotted in Figs. 4 and 5 are comparisons of the sphere-packing bound and the spherical random-coding bound with varying block lengths. In particular, information on the upper and lower bounds on the (ϵ, n)-capacity is shown. In this specific setting and for $n \geq 100$, the upper and lower bounds on the (ϵ, n)-capacity are close together enough, thereby delivering precise information concerning the (ϵ, n)-capacity, whereas when $n < 100$, the upper bound and the lower bound are apart and the question of determining (ϵ, n)-capacity for this blocklength

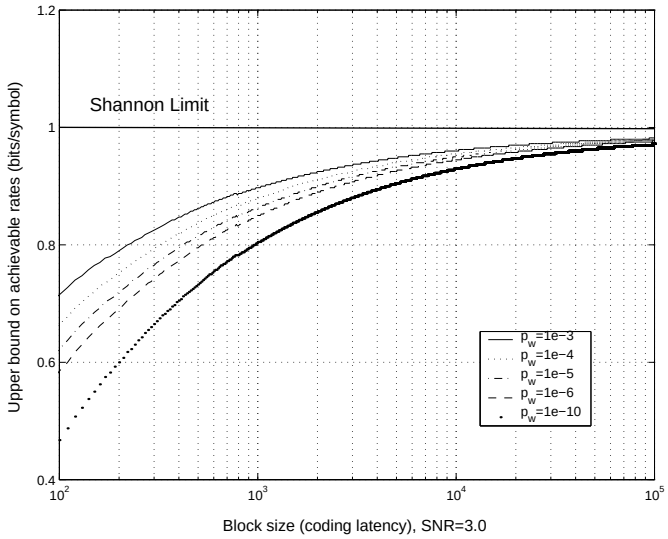

Figure 3: Upper bound on (ϵ, n)-capacity by the sphere-packing bound for codes with varying block length n, operating over a continuous-input AWGN channel at a SNR of 3.0 (4.7712 dB), and $p_w = 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-10}$, respectively.

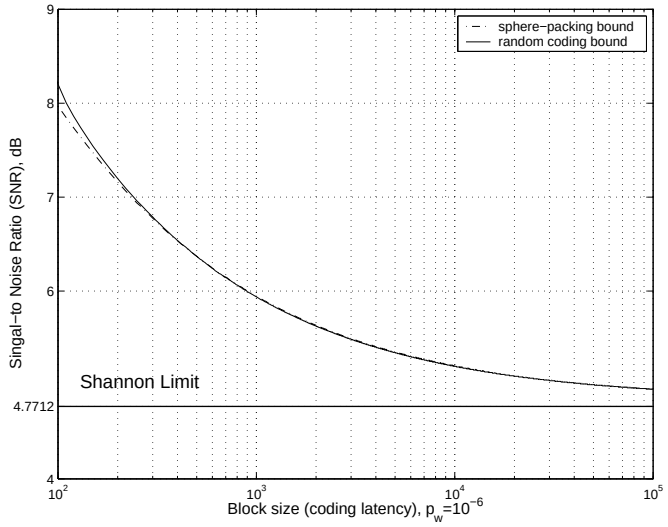

Figure 4: Signal-to-noise ratio by the spherical random-coding bound (as compared with the sphere-packing bound) for codes with varying block length n, operating over a continuous-input AWGN channel at a SNR of 3.0 and $p_w = 10^{-6}$.

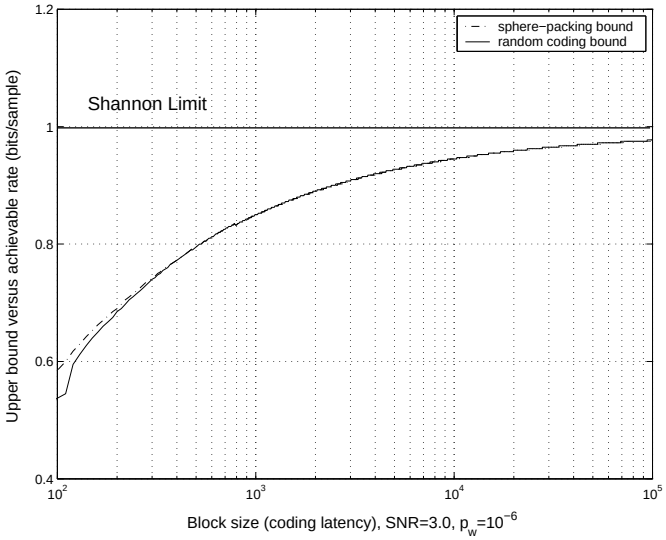

Figure 5: Achievable rates by the spherical random-coding bound (as compared with the sphere-packing bound) for codes with varying block length n, operating over a continuous-input AWGN channel at a SNR of 3.0 and $p_w = 10^{-6}$. Note that the discontinuity in the short blocklength region is caused by numerical difficulty.

region still remains open.

Additional insight into the implications of Figs. 4 and 5 may be obtained by re-examining the definitions of the sphere-packing bound and the spherical random-coding bound. As we know, the performance limit corresponding to the sphere-packing bound would be reached with equality only if the code were a *perfect* spherical code for the continuous-input AWGN, i.e., if equal-sized cones could be drawn around every codeword so as to completely fill the n-dimensional space without intersecting. Actually this is impossible for all $n > 2$ and $M > 2$. On the other hand, we visualize that the spherical random-coding bound is virtually indistinguishable from the sphere-packing bound for block lengths greater than a few hundred. Therefore it is tempting to construct probabilistic codes to "approximate" the ensemble of spherical random codes, rather than search for a deterministic "best" code which does not exist at all.

3 Probabilistic Codes with Independent Codewords

Definition 2: *Probabilistic code with independent codewords*—A (n, M) probabilistic code for a certain channel with power constraint P consists of the following:

- For each encoding block, generate M codewords $X_1^n, X_2^n, \ldots, X_M^n$, that satisfy the power constraint P, i.e., for every codeword

$$\sum_{i=1}^{n} x_{i,w}^2 \leq nP, \qquad w = 1, 2, \ldots, M, \tag{8}$$

where codewords $X_w^n = (x_{1,w}, x_{2,w}, \ldots, x_{n,w})$ are created by independent identically distributed random variables X_w, subject to a common distribution P_X [3] maximizing input-output mutual information $I(X;Y)$, e.g., for an AWGN channel, $P_X \sim \mathcal{N}(0, P)$.

- An encoding function $X : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n$, selecting one codeword from the codebook and passing it through the channel.

- A synchronization scheme between the encoder and decoder that guarantees that the decoder will generate an exact copy of the codebook of encoder for each block. In other words, the receiver has perfect knowledge of the random sources X_w.

- A decoding function
$$g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\}, \tag{9}$$
which is a deterministic rule that assigns an estimate to each possible received vector.

The definition of probabilistic code has the effect of "combined coding and modulation" (baseband), i.e. the encoder feeds its output directly to an AWGN channel. In our notation, the transmission rate is measured as $R = \frac{\log_2 M}{n}$, which can be made readily larger than 1 bit per sample simply by generating a large number of codewords such that $M = 2^{nR}$.

The probabilistic code is inherently time-varying, i.e., the codebook varies from block to block and the codeword with the same index w does not remain the same for different blocks. This scheme differs from the conventional concept wherein a *deterministic* codebook is selected once and used repetitively. The time-varying nature ensures that the channel input resembles a stochastic process with an appropriate distribution, which maximizes the mutual information of channel input and output.

The probabilistic code should not be confused with the standard method of proof of coding theorems based on a *random-coding argument.* Whereas a probabilistic code constitutes a communication technique, a random-coding argument is a proof technique often used to establish the existence of a (single) deterministic code which yields good performance on a specific channel without actually constructing the code. This is done by introducing a probability mass function (pmf) on an ensemble of codes, computing the corresponding average performance over such an ensemble, and then invoking the argument to show that if this average performance is good, then there must exist at least one code in the ensemble with good performance. In contrast, a probabilistic code constitutes a communication technique, the implementation of which requires the availability of a common source of randomness at the transmitter and receiver.

Armed with this formal definition of probabilistic codes, it is now straightforward to prove that a probabilistic code is a capacity-achieving code.

Theorem 1: The Shannon limit of a Gaussian channel with power constraint P and noise variance σ^2,
$$C = \frac{1}{2}\log_2(1 + \frac{P}{\sigma^2})$$
can be attained by a probabilistic code.

Proof: We will use the same decoding rule as in [8], namely, joint typicality decoding.

1. *Generation of probabilistic codebook.* For every encoding block, we generate codewords with each element i.i.d. according to a normal distribution with variance $P-\epsilon$. For large n, we have $\frac{1}{n}\sum x_{i,w}^2 \to P-\epsilon$, then the probability that a codeword does not satisfy the power constraint is quite small. Assume $x_{i,w}, i = 1, 2, \ldots, n, w = 1, 2, \ldots, 2^{nR}$ be i.i.d. subject to $\mathcal{N}(0, P-\epsilon)$, which form codewords $X_1^n, X_2^n, \ldots, X_{2^{nR}}^n \in \mathcal{R}^n$. Thus for each particular encoding block, we have 2^{nR} codewords as the *rows* of a matrix

$$\begin{bmatrix} x_{1,1} & x_{2,1} & \cdots & x_{n,1} \\ x_{1,2} & x_{2,2} & \cdots & x_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1,2^{nR}} & x_{2,2^{nR}} & \cdots & x_{n,2^{nR}}. \end{bmatrix}_{2^{nR}\times n}$$

2. *Encoding.* To send the message index w, the transmitter sends the wth codeword X_w^n in the codebook.

3. *Decoding.* The receiver first regenerates the same codebook as that of the transmitter of the corresponding block, then searches the list of codewords $\{X_w^n\}$ for one that is jointly typical with the received vector. If there is one and only one such codeword, the receiver declares it to be the transmitted codeword. Otherwise the receiver declares an error. The receiver also declares an error if the chosen codeword does not satisfy the power constraint.

4. *Probability of error.* By the symmetry of the code construction, the probability of error p_w does not depend on the particular index that was sent. Without loss of generality, assume that codeword 1 was sent. Thus $Y^n = X_1^n + Z^n$, where Z^n is the channel noise vector.

Define the following events:

$$E_0 = \left\{ \frac{1}{n}\sum_{i=1}^{n} x_{i,1}^2 > P \right\}.$$

and

$$E_i = \{(X_i^n, Y^n) \quad \text{is in joint typical set } A_\epsilon^{(n)}\}.$$

An error occurs if E_0 occurs (the power constraint is violated) or E_1^c occurs (the transmitted codeword and the received sequence are not jointly typical) or $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (an incorrect codeword is jointly typical with the received sequence). Hence the word error probability can be expressed as

$$\begin{aligned} p_w &= \Pr(E|W=1) = \Pr(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}) \\ &\le \Pr(E_0) + \Pr(E_1^c) + \sum_{i=2}^{2^{nR}} \Pr(E_i) \end{aligned}$$

by the union bound for probabilities. Applying the law of large numbers, $\Pr(E_0) \to 0$ as n tends to infinity. Now, by the joint asymptotic equipartition property (AEP), $\Pr(E_1^c) \to 0$, and hence

$$\Pr(E_1^c) \leq \epsilon \quad \text{for sufficiently large } n.$$

By the code generation process we know that X_1^n and X_i^n are independent, Y^n and X_i^n, $\forall i \neq 1$, are also independent. Hence, the probability that X_i^n and Y^n will be jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP [8]. Thus

$$\begin{aligned}
p_w &= \Pr(E|W=1) = \Pr(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}) \\
&\leq \Pr(E_0) + \Pr(E_1^c) + \sum_{i=2}^{2^{nR}} \Pr(E_i) \\
&\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} \Pr(E_i) \\
&\leq 2\epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\
&= 2\epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\
&\leq 2\epsilon + 2^{-n(I(X;Y)-R-3\epsilon)} \\
&\leq 3\epsilon
\end{aligned}$$

for sufficiently large n and $R < I(X;Y) - 3\epsilon$ which, together with (2), concludes the proof.

4 The Optimum Decoder

In the previous section, we constructed time-varying probabilistic codes and showed that they are able to achieve the Shannon limit asymptotically via *typical set decoding*. Although typical set decoding is asymptotically optimal and conceptually easy to analyze, its drawback is two-fold: firstly it is not optimal in the sense of minimizing the probability of error; secondly, it is merely a statistical term and somewhat cumbersome to obtain error probability performance.

The optimum procedure to minimize the probability of error is the maximum *a posteriori* probability (MAP) decoding, i.e., the receiver should choose the *a posteriori* most likely message. Using Bayes' rule, the posterior probabilities may be expressed as

$$p(X_w^n|Y^n) = \frac{p(Y^n|X_w^n)p(X_w^n)}{p(Y^n)}, \tag{10}$$

where $p(Y^n|X_w^n)$ is the conditional probability density function (pdf) of the observed vector given X_w^n, and $p(X_w^n)$ is the *a priori* probability of the wth codeword being transmitted. Assume that the 2^{nR} codewords are equally probable *a priori*, i.e., $p(X_w^n) = 2^{-nR}$ for all w. Furthermore, note that the denominator in (10) is independent of which codeword is transmitted. Consequently, the decision rule based on finding the codeword that maximizes

$p(X_w^n|Y^n)$ is equivalent to finding the codeword that maximizes $p(Y^n|X_w^n)$. It is evident that a detector based on the MAP criterion and one that is based on the maximum likelihood criterion make the same decisions as long as *a prior* probabilities $p(X_w^n)$ are all equal.

In the case of an AWGN channel, the logarithm likelihood function of $p(Y^n|X_w^n)$ is given by

$$\ln p(Y^n|X_w^n) = -\frac{1}{2}n\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(y_k - x_{k,w})^2. \tag{11}$$

Hence, the maximum of $p(Y^n|X_w^n)$ over X_w^n is equivalent to finding the index w that minimizes the Euclidean distance

$$D(Y^n, X_w^n) = \sum_{k=1}^{n}(y_k - x_{k,w})^2. \tag{12}$$

We call the normalized version of $D(Y^n, X_w^n)$ by distance metrics $d(w)$, i.e.,

$$d(w) = \frac{1}{n}D(Y^n, X_w^n)$$

for $w = 1, 2, \ldots, 2^{nR}$. This decision rule is usually referred to as the *minimum distance criterion.* It should be mentioned that each codeword in the probabilistic code is characterized by a discrete stochastic process that maximizes the input-output mutual information. Although the average power remains the same, the energies of particular codewords may not necessarily be the same. Hence, a correlation detector is no longer optimal for the probabilistic code.

Suppose the codeword X_1^n is transmitted. Then the 2^{nR} decision variables $d(w)$ are

$$\begin{cases} d(1) &= \frac{1}{n}\sum_{k=1}^{n} z_k^2 \\ d(2) &= \frac{1}{n}\sum_{k=1}^{n}(x_{k,1} + z_k - x_{k,2})^2 \\ \vdots & \vdots \\ d(2^{nR}) &= \frac{1}{n}\sum_{k=1}^{n}(x_{k,1} + z_k - x_{k,2^{nR}})^2 \end{cases} \tag{13}$$

where z_k is the channel noise. The decision is made in favor of the codeword X_w^n having the least decision value $d(w)$ among the whole set of all codewords. As all codewords are mutually i.i.d. Gaussian variables and independent of the channel noise, the decision variable $d(1)$ [2] has the normalized chi-square distribution with n degrees of freedom (see Appendix B)

$$f_1(y) = \frac{1}{(2\sigma^2)^{n/2}\Gamma(n/2)}n^{n/2}y^{n/2-1}e^{-ny/2\sigma^2}, \quad y \geq 0 \tag{14}$$

and all other decision variables $d(w), w \neq 1$ yield the normalized chi-square distribution of

$$f_w(y) = \frac{1}{[2(\sigma^2 + 2P)]^{n/2}\Gamma(n/2)}n^{n/2}y^{n/2-1}e^{-ny/2(\sigma^2+2P)}, \quad y \geq 0. \tag{15}$$

[2] Again assume the codeword 1 was sent.

It is interesting to note that each decision variable in (13) is actually an estimate of the variance of the Gaussian process — the correct branch has the minimum variance of σ^2 and all other branches have variances of $\sigma^2 + 2P$. By the law of large numbers, as block length n goes to infinity, these estimates tend to be increasingly accurate and thus the probability of making an error vanishes.

It is mathematically convenient to first derive the probability that the detector makes a correct decision. This is the probability that $d(1)$ is greater than each of the other $2^{nR} - 1$ decision variables $d(2), d(3), \ldots, d(2^{nR})$. This probability may be expressed as

$$p_c = \int_0^\infty \Pr[d(2) > r, d(3) > r, \cdots, d(2^{nR}) > r | d(1) = r] f_1(r) dr, \tag{16}$$

where $\Pr[d(2) > r, d(3) > r, \ldots, d(2^{nR}) > r | d(1) = r]$ denotes the joint probability that $d(2)$, $d(3)$, ..., $d(2^{nR})$ are all greater than r, where $r \geq 0$. This joint probability is then averaged over all r. Unfortunately, the values of d_w, $w = 1, 2, \ldots, 2^{nR}$, are not statistically independent, and the evaluation of the influence of correlations is rather complicated, even impossible. Therefore, we resort to an approximation expression in which the correlations in decision variables are neglected.[3] One approach is to factor the joint probability into a product of $2^{nR} - 1$ marginal probabilities, yielding

$$p_c \approx \int_0^\infty [\Pr(d(w) > r | d(1) = r)]^{2^{nR}-1} f_1(r) dr. \tag{17}$$

Thus the probability of word error is

$$p_w = 1 - p_c. \tag{18}$$

Under the condition that $d(1) = r$, the decision variable $d(w)$ has the normalized chi-square distribution of

$$\frac{1}{[2(r+2P)]^{n/2}\Gamma(n/2)} n^{n/2} y^{n/2-1} e^{-ny/2(r+2P)}, \quad y \geq 0.$$

Thus the conditional probability $\Pr[d(w) > r | d(1) = r]$ yields

$$\Pr[d(w) > r | d(1) = r] = 1 - \int_0^r \frac{1}{[2(r+2P)]^{n/2}\Gamma(n/2)} n^{n/2} y^{n/2-1} e^{-ny/2(r+2P)} dy. \tag{19}$$

Without loss of generality, we assume that n is an even integer, the integral can be expressed in closed form by repeated integration by parts, which yields

$$\Pr[d(w) > r | d(1) = r] = e^{-nr/2(2P+r)} \sum_{k=0}^{n/2-1} \frac{1}{k!} \left[\frac{nr}{2(2P+r)}\right]^k. \tag{20}$$

Substituting (20) into (17) and (18), the probability of word error can be expressed as

[3]The correlations in decision variables due to the channel noise $\{z_k\}$ vanish asymptotically as the signal-to-noise ratio increases. Hence, for the high-SNR regime, this approximation has only a minor impact on accuracy.

$$p_w = 1 - \int_0^\infty \frac{1}{(2\sigma^2)^{n/2}\Gamma(n/2)} n^{n/2} r^{n/2-1} e^{-nr/2\sigma^2} \left\{ e^{-nr/2(2P+r)} \sum_{k=0}^{n/2-1} \frac{1}{k!} \left[\frac{nr}{2(2P+r)} \right]^k \right\}^{2^{nR}-1} dr. \tag{21}$$

Denoting $m = \frac{n}{2}$ and $M = 2^{nR}$, (21) can be simplified as

$$p_w = 1 - \int_0^\infty \frac{1}{\sigma^{2m}\Gamma(m)} m^m r^{m-1} e^{-mr/\sigma^2} \left\{ e^{-mr/(2P+r)} \sum_{k=0}^{m-1} \frac{1}{k!} \left[\frac{mr}{(2P+r)} \right]^k \right\}^{M-1} dr. \tag{22}$$

For moderate to large block lengths n, the evaluation of (22) becomes rather difficult due to the existence of the factorial of a large number. Hence we derive an asymptotic formula for probabilistic codes, which is based on the connection between probabilistic codes and spherical random codes. Note that the only difference between the probabilistic code and the ensemble of spherical random codes lies in the power constraint. In Shannon's spherical random codes, each codeword is required to lie exactly on the surface of the sphere of radius $\sqrt{nP}$, but in the probabilistic code, all codewords are required to have a power of P in a probabilistic manner. As stated by the law of large numbers, as the block length n tends to infinity, the probability that a codeword deviates from this power constraint can be made arbitrarily small. Thus one may presume that the asymptotic performance of both codes is nearly the same. In fact, the probabilistic code can be made to be an instance of the ensemble of spherical random codes by expanding one sample in the block length. Suppose we have a probabilistic code with each block of length n, and all codewords satisfy the power constraint $\sum_{k=1}^{n} x_{k,w}^2 \leq nP$. To each codeword, add a further element of such value that the $n+1$ dimensional codeword lies exactly on the $n+1$ sphere surface. Specifically, the added $n+1$ element will have the value

$$x_{n+1,w} = \sqrt{(n+1)P - \sum_{k=1}^{n} x_{k,w}^2} \quad . \tag{23}$$

This yields a snapshot of spherical random codes with 2^{nR} codewords of block length $n+1$, and the overall transmit rate becomes $\frac{n}{n+1}R$. Asymptotically as n tends to infinity, this rate loss is negligible. This justification enables us to use (7) to calculate the asymptotic performance of the probabilistic code based on the ensemble of spherical random codes, namely

$$p_w \approx Q_{sp}(\theta) \left(1 + \frac{AG(\theta)\sin^2\theta - \cos\theta}{2\cos\theta - AG(\theta)\sin^2\theta} \right), \tag{24}$$

where θ is defined in (4). Considering the fact that the performance of a sphere-packing code is indistinguishable from that of the ensemble of spherical random codes for moderate to large block lengths, it is evident that a probabilistic code with independent codewords approaches the (ϵ, n)-capacity closely in this blocklength region.

5 Probabilistic Codes with Correlated Codewords

5.1 Distance Analysis

Recognizing the fact that the codewords $\{X_w^n\}$ are independent of the channel noise, the normalized decision variables in (13) can be rewritten as

$$\begin{cases} d(1) &= \frac{1}{n}\sum_{k=1}^{n} z_k^2 \\ d(2) &= \frac{1}{n}\sum_{k=1}^{n}(x_{k,1}-x_{k,2})^2 + \frac{1}{n}\sum_{k=1}^{n} z_k^2 \\ \vdots & \vdots \\ d(2^{nR}) &= \frac{1}{n}\sum_{k=1}^{n}(x_{k,1}-x_{k,2^{nR}})^2 + \frac{1}{n}\sum_{k=1}^{n} z_k^2 \end{cases} \tag{25}$$

Clearly the probability of making an error is closely related to the normalized Euclidean distances of the codewords, i.e. the set

$$d_{i,j} = \frac{1}{n}\sum_{k=1}^{n}(x_{k,i}-x_{k,j})^2, \quad \text{for all } i \neq j. \tag{26}$$

Because all codewords in a probabilistic code with independent codewords are generated via Gaussian random variables in a time-varying manner, the normalized Euclidean distances between the codewords are essentially random variables subject to the normalized chi-square distribution

$$f_D(d) = \frac{1}{(4P)^{n/2}\Gamma(n/2)} n^{n/2} d^{n/2-1} e^{-nd/4P}, \quad d \geq 0 \tag{27}$$

whose cumulative distribution functions are illustrated in Fig. 6 for various n.

As n goes to infinity, we obtain the asymptotic distance set [4] $d_{i,j}^{\text{asy.}}$ as

$$d_{i,j}^{\text{asy.}} \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(x_{k,i}-x_{k,j})^2, \quad i \neq j. \tag{28}$$

Under the assumption of independent codewords treated in the previous section, the asymptotic distance between codeword X_i^n and codeword X_j^n becomes

$$\begin{aligned} d_{i,j}^{\text{asy.}} &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n}(x_{k,i}-x_{k,j})^2 \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_{k,i}^2 + \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_{k,j}^2 - \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 2x_{k,i}x_{k,j} \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_{k,i}^2 + \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} x_{k,j}^2 \\ &= 2P, \end{aligned} \tag{29}$$

[4] By definition, the asymptotic distance denotes mathematical "expectation" over block length. Thus the asymptotic distance can also be interpreted as expectation distance.

which holds for all $i \neq j$. Note that in (29) the third term $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 2x_{k,i}x_{k,j}$ falls to zero under independent conditions. We observe that in a probabilistic code with independent codewords, the asymptotic distance set $d_{i,j}^{\text{asy.}}$ is uniform with a unique value of $2P$ that is precisely twice the transmit power.

As $d_{i,j}$ is an average sum of independent random variables, we can then use the Chernoff bound to give an exponential bound on the probability that $d_{i,j}$ will deviate from its mean $2P$ by more than a fraction ϵ. A stronger bound can be obtained if the elements of codewords are amplitude-limited, or say, under peak-power constraint.

Lemma 1: Let $\{x_{k,i}\}$ be peak-power constrained, i.e., $\forall\ x_{k,i}^2 < E_p$, then

$$\Pr\{|d_{i,j} - 2P| > \epsilon\} \leq e^{-\frac{1}{2E_p}\epsilon^2 n}.$$

Proof: See the Appendix of [9].

The above Lemma provides a straightforward justification of the probabilistic construction. As the block length n is sufficiently large, the performance of an instance of probabilistic codes should fall within the ϵ region of that of the average ensemble with probability exponentially close to 1. The concentration of the performance on the expectation grows exponentially with the code block length. It thus appears that any effort to search for a deterministic code that outperforms this probabilistic construction would turn out to be in vain if the block length is large enough, taking into account the fact that the performance of the ensemble of spherical random codes is indistinguishable from that of a best sphere-packing code.

If the block length is relatively small, there exists a non-negligible probability that some probabilistic constructions yield good codes, and some yield bad codes. Through repetitively independent trials, the probability of selecting a bad code can be made arbitrarily small in an exponential manner as the number of independent trials increases.

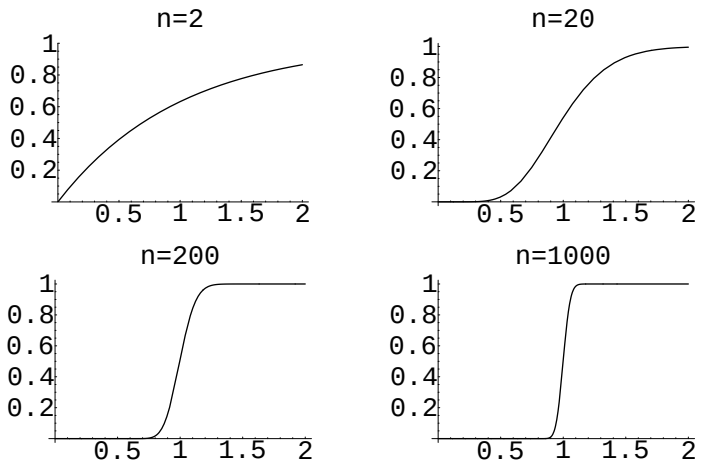

Figure 6: Cumulative distribution function (cdf) of the normalized Euclidean distance with varying block length n, where $P = 0.5$. As n increases, its cdf becomes a step function meaning that the probability mass concentrates on its expectation $2P$.

5.2 Best Asymptotic $d_{\min}^{\text{asy.}}$ Codes

The asymptotic distance can be made even larger to improve the error probability performance. From the above derivation we know that, if the ith codeword $\{x_{k,i}\}$ has a certain correlation with the jth codeword $\{x_{k,j}\}$, it is also possible for the third term in (29) to contribute a positive value to the asymptotic distance. In this subsection we will present a new construction of probabilistic code with correlated codewords, which has the property of max-min distance.

Let us assume that $X_i, i = 1, 2, \ldots, M$, are i.i.d. Gaussian random variables, which are used to generate independent codewords $X_1^n, X_2^n, \ldots, X_M^n$. These Gaussian random variables are mutually independent and with zero-mean, variances of $P_1, P_2, \ldots, P_M$ whose values will be determined later. Let $\mathbf{M}$ denote the $M \times M$ covariance matrix which is diagonal, i.e. $\mathbf{M} = \text{diag}(P_1, P_2, \cdots, P_M)$. Let X denote the $M \times 1$ column vector consisting of random variables $X_i, i = 1, 2, \ldots, M$. The joint pdf of the Gaussian random vector X is defined as

$$p(X) = \frac{1}{(2\pi)^{M/2}(\det \mathbf{M})^{1/2}} \exp[-\frac{1}{2} X' \mathbf{M}^{-1} X], \tag{30}$$

where $\mathbf{M}^{-1}$ denotes the inverse of $\mathbf{M}$ and X' the transpose of X. Now let us consider a linear transformation of M independently Gaussian random variables X

$$Y = \mathbf{A}X, \tag{31}$$

where Y is an $M \times 1$ column vector, $\mathbf{A}$ is a nonsingular orthogonal matrix ($\mathbf{A}' = \mathbf{A}^{-1}$). The Jacobian of this transformation is $J = 1/\det(\mathbf{A})$. As $X = \mathbf{A}^{-1}Y$, we may substitute for X in (30) and thus obtain the joint pdf of Y in the form

$$\begin{aligned} p(Y) &= \frac{1}{(2\pi)^{M/2}(\det \mathbf{M})^{1/2} \det \mathbf{A}} \exp[-\frac{1}{2}(\mathbf{A}^{-1}Y)' \mathbf{M}^{-1}(\mathbf{A}^{-1}Y)] \\ &= \frac{1}{(2\pi)^{M/2}(\det \mathbf{Q})^{1/2}} \exp[-\frac{1}{2} Y' \mathbf{Q}^{-1} Y], \end{aligned}$$

where the covariance matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{A}\mathbf{M}\mathbf{A}'. \tag{32}$$

Thus we come to the following lemma:

Lemma 2: A set of correlated Gaussian random variables Y can be obtained via linear transformation from a set of statistically independent Gaussian random variables X.

Equations (27), (28) and (29) suggest a method to generate a probabilistic code with correlated codewords via linear transformation from a probabilistic code with independent codewords, which is summarized as follows.

Definition 3: *Probabilistic code with correlated codewords*

- At each time index j, $j = 1, 2, \ldots, n$, generate a random (column) vector $X^j = (X_1^j, X_2^j, \cdots, X_M^j)'$, where X_i^j's are instances of i.i.d. Gaussian variables whose covariance matrix $\mathbf{M}$ equals $\text{diag}(P_1, P_2, \cdots, P_M)$.

- At each time index j, $j = 1, 2, \ldots, n$, compute column vector Y^j via linear transformation $Y^j = \mathbf{A}X^j$.

- Finally the codewords are formed by the rows of the M-by-n matrix $(Y^1, Y^2, \cdots, Y^n)$.

By application of basic matrix theory, matrix $\mathbf{A}$ consists of rows that are the eigenvectors of the covariance matrix $\mathbf{Q}$, and $\mathbf{M}$ is a diagonal matrix with elements equal to the eigenvalues of $\mathbf{Q}$. To satisfy the average energy constraint, P should be chosen as the main diagonal elements of the matrix $\mathbf{Q}$, meaning that each codeword has an average power of P. To maintain the symmetry, we impose a constraint of equal correlation coefficients in $\mathbf{Q}$, which gives

$$\mathbf{Q} = \begin{bmatrix} P & \rho P & \cdots & \rho P \\ \rho P & P & \cdots & \rho P \\ \vdots & \vdots & \ddots & \vdots \\ \rho P & \rho P & \cdots & P \end{bmatrix}_{M \times M}, \tag{33}$$

where $-1 \leq \rho \leq 1$. Hence, the asymptotic distance in (29) yields

$$\begin{aligned} d_{i,j}^{\text{asy.}} &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} y_{k,i}^2 + \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} y_{k,j}^2 - \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} 2 y_{k,i} y_{k,j} \\ &= P + P - 2\rho P \\ &= 2(1-\rho)P. \end{aligned} \tag{34}$$

It is thus evident that, in order to maximize the asymptotic distance, we need to make ρ as small as possible. However, there exists a tight lower bound on ρ that induces the covariance matrix $\mathbf{Q}$ to be positive.

Lemma 3: The covariance matrix $\mathbf{Q}$ is positive if and only if ρ is greater than $-1/(M-1)$.

Proof:

$$\begin{aligned} \det(\lambda \mathbf{I} - \mathbf{Q}) &= \det \begin{bmatrix} \lambda - P & -\rho P & \cdots & -\rho P \\ -\rho P & \lambda - P & \cdots & -\rho P \\ \vdots & \vdots & \ddots & \vdots \\ -\rho P & -\rho P & \cdots & \lambda - P \end{bmatrix}_{M \times M} \\ &= [\lambda - (1 + (M-1)\rho)P][\lambda - (1-\rho)P]^{M-1}, \end{aligned}$$

thus the matrix $\mathbf{Q}$ has $M-1$ duplicate eigenvalues of $(1-\rho)P$ and one of $[1+(M-1)\rho]P$. In order to guarantee that $\mathbf{Q}$ is positive, we obtain

$$\begin{cases} (1-\rho)P & > 0 \\ [1+(M-1)\rho]P & > 0, \end{cases} \tag{35}$$

which simplifies to $\rho > -1/(M-1)$. This completes the proof.

Since M is a diagonal matrix with elements equal to the eigenvalues of $\mathbf{Q}$, we obtain

Lemma 4: The diagonal matrix $\mathbf{M} = \mathrm{diag}(P_1, P_2, \ldots, P_M)$ should be

$$\mathrm{diag}\{\underbrace{(1-\rho)P, \ldots, (1-\rho)P}_{M-1}, [1+(M-1)\rho]P\},$$

where P is the average transmit power. As ρ tends to $-1/(M-1)$, we have

$$\lim_{\rho \to -1/(M-1)} \mathbf{M} = \mathrm{diag}\{\underbrace{\frac{MP}{M-1}, \ldots, \frac{MP}{M-1}}_{M-1}, 0\}.$$

Proof: The proof is straightforward from the above theorem.

Substituting the minimum (in the limiting case) value of ρ into (34), the asymptotic distance becomes a constant

$$d_{i,j}^{\text{asy.}} = \frac{2MP}{M-1} \tag{36}$$

for all $i \neq j, \in 1, 2, \ldots, M$. It should be pointed out that, with increasing n, the improvement due to the introduction of correlated codewords becomes negligible while the additional complexity of linear transformation becomes prohibitive. However, it is interesting to note that the probabilistic code with correlated codewords of $\rho = -\frac{1}{M-1}$ is asymptotically "best-$d_{\min}^{\text{asy.}}$" code, i.e. the best code for continuous-input AWGN channel.

Theorem 2: The probabilistic code with correlated codewords of $\rho = -\frac{1}{M-1}$ achieves asymptotically the maximum-minimum distance.

Proof: First we derive an upper bound on the *average* asymptotic distance over all distinct codewords $\{x_{k,j}\}$.

$$\begin{aligned}
\bar{D} &= \frac{1}{M(M-1)} \sum_{i \neq j} \lim_{n \to \infty} \left[\frac{1}{n} \sum_{k=1}^{n} (x_{k,i} - x_{k,j})^2 \right] \\
&= \frac{1}{M(M-1)} \sum_{i,j} \lim_{n \to \infty} \left[\frac{1}{n} \sum_{k=1}^{n} (x_{k,i} - x_{k,j})^2 \right] \\
&= \frac{1}{M(M-1)} \sum_{i,j} \lim_{n \to \infty} \left[\frac{1}{n} \sum_{k=1}^{n} (x_{k,i}^2 - 2x_{k,i}x_{k,j} + x_{k,j}^2) \right] \\
&= \frac{1}{M(M-1)} \lim_{n \to \infty} \frac{1}{n} \left(\sum_{i,j,k} x_{k,i}^2 - 2 \sum_{k} \sum_{i,j} x_{k,i}x_{k,j} + \sum_{i,j,k} x_{k,j}^2 \right) \\
&= \frac{1}{M(M-1)} \lim_{n \to \infty} \frac{1}{n} \left(2M \sum_{i,k} x_{k,i}^2 - 2 \sum_{k} \left(\sum_{i} x_{k,i} \right)^2 \right) \\
&\leq \frac{1}{M(M-1)} \lim_{n \to \infty} \left(2M \sum_{i,k} x_{k,i}^2 \right) \\
&= \frac{2MP}{M-1}.
\end{aligned}$$

Maximum-minimum distance among a finite set must be less than or equal to the average distance, otherwise this would lead to a paradox. It is thus clear that the probabilistic code with correlated codewords of $\rho = -\frac{1}{M-1}$ is asymptotically (as n goes to infinity) "best-$d_{\min}^{\text{asy.}}$" code with a *uniform* distance of

$$d_{i,j}^{\text{asy.}} = \bar{D} = \frac{2MP}{M-1}$$

for all $i \neq j$. This completes the proof.

When $M = 2$, this bound in (37) is $4P$ and is simply equal to the distance for antipodal signaling. As $n \to \infty$, and thus $M \to \infty$ for any fixed R, the bound approaches $2P$ quickly, which equals the asymptotic distance of a probabilistic code with independent codewords. The implication is that, for large n or M, only very little performance improvement is gained by the use of correlated codewords. The performance improvement vanishes exponentially as the block length grows. It reveals to us that, for moderate to large block lengths, the probabilistic code with independent codewords performs almost as well as the best code.

6 Conclusion

In this paper we have identified the maximal achievable rates over noisy, delay-constrained AWGN channels, referred to as (ϵ, n)-capacity C_ϵ^n with ϵ denoting target error probability and n coding latency (viz. block length). We have also investigated a family of block codes based on a probabilistic construction that approaches closely the delay-constrained capacity and provably achieves the Shannon limit over an AWGN channel. The decoding complexity of probabilistic codes with independent or correlated codewords grows exponentially with block lengths, while they are the right codes capable of approaching the (ϵ, n)-capacity closely except for very small block lengths. The theoretical characterization of the (ϵ, n)-capacity and its corresponding probabilistic codes offers insights into how the optimal block codes look like and what is the ultimate limit under a critical coding latency constraint. The (ϵ, n)-capacity can be used as a natural criterion against how good a practical coding scheme is with a finite block length.

Appendix A: Evaluation of $\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})}$

It would be very difficult to evaluate $\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})}$ for a large block length n by direct calculation. In this appendix, we present an alternative method to evaluate it approximately. By Stirling's formula, we have

$$\Gamma(z) \approx e^{-z} z^{z-\frac{1}{2}} (2\pi)^{\frac{1}{2}} \left[1 + \frac{1}{12z} + \frac{1}{288z^2} - \frac{139}{51840z^3} - \frac{571}{2488320z^4} + \cdots\right]$$

for $z \to \infty$. Substituting this formula into $\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})}$, and after some algebra, we obtain

$$\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})} \approx \frac{(\frac{n}{2}+1)^{1/2}(1+\frac{1}{n+1}^{n/2})}{e^{1/2}} \frac{[1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{2488320z^4}]_{z=\frac{n}{2}+1}}{[1+\frac{1}{12z}+\frac{1}{288z^2}-\frac{139}{51840z^3}-\frac{571}{2488320z^4}]_{z=\frac{n+1}{2}}}.$$

If n is sufficiently large, we obtain a simplified asymptotic version

$$\frac{\Gamma(\frac{n}{2}+1)}{\Gamma(\frac{n+1}{2})} \approx \sqrt{\frac{n}{2}}.$$

Appendix B: Normalized Chi-Square Distribution

Let X be Gaussian distributed with zero mean and variance σ^2. Setting $Y = X^2$, we obtain the pdf of Y in the form

$$p_Y(y) = \frac{1}{\sqrt{2\pi y}\sigma} e^{-y/2\sigma^2}, \quad y \geq 0. \tag{37}$$

The cdf of Y is

$$\begin{aligned} F_Y(y) &= \int_0^y p_Y(u)du \\ &= \frac{1}{\sqrt{2\pi y}} \int_0^y \frac{1}{\sqrt{u}} e^{-u/2\sigma^2} du, \end{aligned} \tag{38}$$

which cannot be expressed in closed form. The characteristic function, however, can be determined in closed form. It is [10]

$$\phi(jv) = \frac{1}{(1-j2v\sigma^2)^{1/2}}. \tag{39}$$

Now, suppose that the random variable Y is defined as

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \tag{40}$$

where the X_i, $i = 1, 2, \cdots, n$, are statistically independent and identically distributed Gaussian random variables with zero mean and variance σ^2. As a consequence of the statistical independence of the X_i, the characteristic function of Y is

$$\phi_Y(jv) = \frac{1}{(1-j2v\sigma^2/n)^{n/2}}. \tag{41}$$

The inverse transform of this characteristic function yields the pdf

$$p_Y(y) = \frac{1}{(2\sigma^2)^{n/2}\Gamma(n/2)} n^{n/2} y^{n/2-1} e^{-ny/2\sigma^2}, \quad y \geq 0, \tag{42}$$

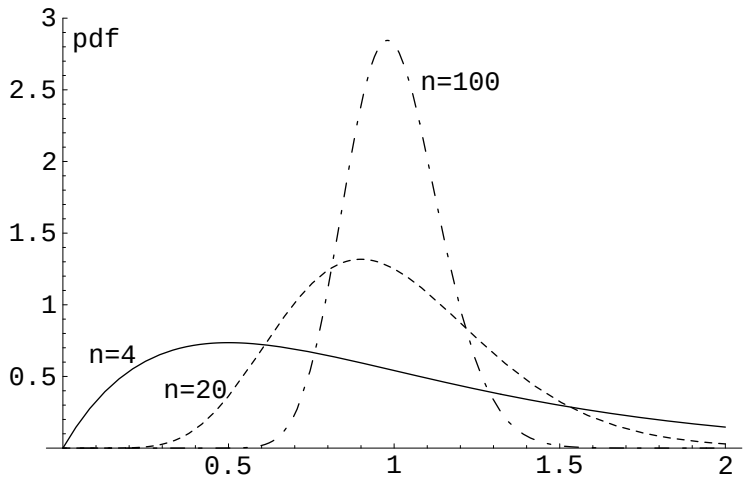

Figure 7: The pdf of a normalized chi-square-distributed random variable for several degrees of freedom with $\sigma^2 = 1$.

where $\Gamma(p)$ is the gamma function defined as

$$\begin{aligned}
\Gamma(p) &= \int_0^\infty t^{p-1} e^{-t} dt. \quad p > 0 \\
\Gamma(p) &= (p-1)!, \quad p \text{ an interger}, \; p > 0 \\
\Gamma(\frac{1}{2}) &= \sqrt{\pi}, \quad \Gamma(\frac{3}{2}) = \frac{1}{2}\sqrt{\pi}.
\end{aligned}$$

This pdf, which is a generalization of chi-square distribution, is called a normalized chi-square pdf with n degrees of freedom. It is illustrated in Fig. 7.

The first two moments of Y are

$$\begin{aligned}
E(Y) &= \sigma^2 \\
E(Y^2) &= \frac{2\sigma^4}{n} + \sigma^4 \\
\sigma_y^2 &= \frac{2\sigma^4}{n}.
\end{aligned}$$

The cdf of Y is

$$F_Y(y) = \int_0^y \frac{1}{(2\sigma^2)^{n/2}\Gamma(n/2)} n^{n/2} u^{n/2-1} e^{-nu/2\sigma^2} du, \quad y \geq 0. \tag{43}$$

When n is even, this integral can be expressed in closed form. Specifically, let $m = \frac{1}{2}$, where m is an integer. Then, by repeated integration by parts, we obtain

$$F_Y(y) = 1 - e^{-my/\sigma^2} \sum_{k=0}^{m-1} \frac{1}{k!} \left(\frac{my}{\sigma^2}\right)^k, \quad y \geq 0. \tag{44}$$

Substituting N, $(N + 2P)$, $2P$ for σ^2 in (42) yields (14), (15), and (27), respectively.

References

[1] "Special issue on codes and graphs and iterative algorithms," *IEEE Trans. Inform. Theory,* vol. 47, pp. 493-849, Feb. 2001.

[2] C.E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423 and pp. 623-656, July and Oct. 1948.

[3] S. Shamai (Shitz), and S. Verdu, "The empirical distribution of good codes," *IEEE Trans. Inform. Theory,* vol. 43, pp. 836-846, May 1997.

[4] C.E. Shannon, "Probability of error for optimal codes in a Gaussian channel," *Bell Syst. Tech. J.,* vol. 38, pp. 611-656, May 1959.

[5] S. Dolinar, D. Divsalar, and F. Pollara, "Code performance as a function of block size," *TMO Progress Report 42-133,* Jet Propulsion Laboratory, Pasadena, California, pp. 1-23, May 1998.

[6] S. J. MacMullan, and O. M. Collins, "A comparision of known codes, random codes, and the best codes," *IEEE Trans. Inform. Theory,* vol. 44, pp. 3009-3022, Nov. 1998.

[7] S. Verdu, "Fifty years of Shannon theory," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2057-2077, Oct. 1998.

[8] T.M. Cover and J.A. Thomas, *Elements of information theory,* New York: Wiley, 1991.

[9] N. Alon, J. Spencer, and P. Erdos, *The Probabilistic Method*, New York: Wiley, 1992.

[10] J. G. Proakis, *Digital Communications,* McGraw-Hill, third edition, 1995.