Writing on Colored Paper

Wei Yu, Arak Sutivong, David Julian
Thomas M. Cover, and Mung Chiang *

March 2, 2002

Abstract

A Gaussian channel, when corrupted by an additive Gaussian interfering signal whose complete sample sequence is known non-causally to the transmitter but not to the receiver, has the same capacity as if the interfering signal were not present. This is true even when the noise and interference are not necessarily stationary or ergodic.

1 Introduction

When an additive white Gaussian channel is corrupted by a Gaussian interfering signal whose entire sample sequence is known non-causally to the transmitter, but not to the receiver, Costa [1] showed that, surprisingly, the capacity of the channel is the same as if the interference were not present. Costa's result is derived under the assumption that both noise and interference are i.i.d. Gaussian processes. This result has since been generalized to cases where the interfering process is an ergodic process with an arbitrary distribution [2], and to cases where the interference process is an arbitrary sequence if a common source of randomness is available at both the transmitter and the receiver [3]. This paper extends Costa's result in a different direction. While retaining the joint Gaussian assumption, we consider the setting where the noise and interfering processes are not necessarily stationary or ergodic. A coding theorem for this general setting is established by looking at a single block of n transmissions as shown in Figure 1:

$$Y^n = X^n + S^n + Z^n, \tag{1}$$

*Address: Electrical Engineering Department, Stanford University, CA 94305, U.S.A. emails: {weiyu, arak, djulian, cover, chiangm}@stanford.edu.

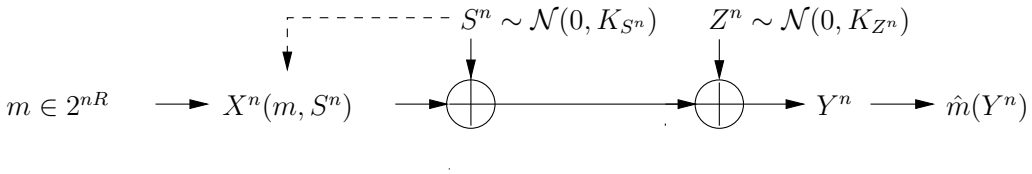

Figure 1: Gaussian channel with interference known at the transmitter

where S^n and Z^n are independent Gaussian sequences with arbitrary finite-dimensional covariance matrices K_{S^n} and K_{Z^n} respectively. The sequence S^n is known to the transmitter but not to the receiver, the sequence Z^n is known neither to the transmitter nor to the receiver. A power constraint P is imposed on the input sequence X^n, i.e. $\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n}(X^n)_i^2] \leq P$. A codeword $X^n(m, S^n)$ is an encoding function that maps a codeword index m and a side information sequence S^n to a block of n transmissions. A codebook of 2^{nR} codewords and a decoding rule $\hat{m}(Y^n)$ can be constructed so that the probability of error $Prob(X^n \neq \hat{X}^n(Y^n))$ goes to zero as the block size n goes to infinity in a way that is independent of the finite-dimensional covariance structure of interference and noise. From this distribution-free probability of error bound, a general coding theorem can be established for the Gaussian channel with non-causal transmitter side information which does not require the usual stationarity and ergodicity assumptions, or the use of a common source of randomness. This general setting also encompasses scenarios where the channel input and output are vector-valued, and scenarios where the channel has memory (that is not necessarily finite).

We are motivated to study this more general setting by the following additive Gaussian noise channel where some side information V about the additive noise W is available to the transmitter, but not to the receiver, as shown in Figure 2. Consider a block-transmission scheme with finite block size n. Let W^n be the additive Gaussian noise sequence distributed as $\mathcal{N}(0, K_{WW})$, where K_{WW} is the covariance matrix. Let V^n be the side information sequence which is Gaussian distributed and correlated with W^n as follows:

$$(W^n, V^n) \sim \mathcal{N}\left(0, \begin{bmatrix} K_{WW} & K_{WV} \\ K_{VW} & K_{VV} \end{bmatrix}\right), \tag{2}$$

where K_{VV} is the covariance matrix of the sequence V^n, and $K_{VW} = K_{WV}^T$ is the cross-covariance matrix between V^n and W^n. When side information V^n is available to the transmitter, the transmitter encodes an index $m \in \{1, \cdots, e^{nR}\}$ by generating a signal $X^n(m, V^n)$ based on m and V^n, which also satisfies a power constraint P, i.e.,

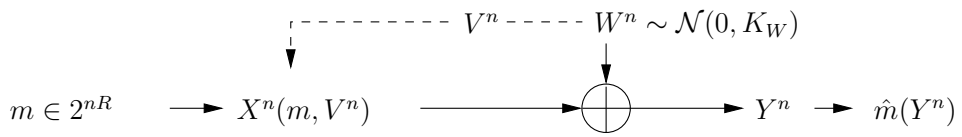

Figure 2: Gaussian channel with partial side information at the transmitter

$\mathbf{E}[\frac{1}{n}\sum_{i=1}^{n} X_i^2] \leq P$. Now, since the minimum mean squared error estimate of W^n given V^n is the conditional mean $\mathbf{E}[W^n|V^n] = K_{WV}K_{VV}^{-1}V^n$ with the corresponding mean squared error of $K_{WW} - K_{WV}K_{VV}^{-1}K_{VW}$, the communication problem depicted in Figure 2 can be modeled by the equivalent problem shown in Figure 1, where $S^n = \mathbf{E}[W^n|V^n]$ and $Z^n = W^n - S^n$. The sequences S^n and Z^n are independent and distributed as $\mathcal{N}(0, K_{WV}K_{VV}^{-1}K_{VW})$ and $\mathcal{N}(0, K_{WW} - K_{WV}K_{VV}^{-1}K_{VW})$ respectively. Solving for the capacity of the channel shown in Figure 2 is equivalent to solving for the capacity of the channel in Figure 1.

An overview of the result of the paper is as follows. In section 2, an asymptotic equipartition theorem for Gaussian random vectors is presented, and the Gaussian channel capacity theorem is restated in a way that does not depend on the finite-dimensional distribution of the noise sequence. In section 3, the distribution-free Gaussian channel capacity theorem is extended to Gaussian channels with side information. The computation of capacity also reveals a connection between the choice of the optimal auxiliary random variable to a Wiener filter. Section 4 contains concluding remarks.

2 Gaussian Channel

We begin the discussion by stating a coding theorem for the ordinary Gaussian channel where the probability of error analysis does not depend on the covariance structure of Gaussian noise. Gaussian processes appear to be special in the sense that an asymptotic equipartition (AEP) theorem holds without the need for stationarity or ergodicity. This was observed in [4], and the proof is based on the fact that the difference between the empirical entropy rate and the true entropy rate for a Gaussian process is a scaled chi-square distribution, whose parameters do not depend on the particular covariance matrices, and therefore converges to zero as the block size goes to infinity in a way that does not rely on the stationarity or ergodicity of the Gaussian process. This result, which is essentially a restatement of Theorem 5 in Cover and Pombra [4], is as follows:

Lemma 1 *Let $(X_{ij})_{i=1\cdots\infty,j=1\cdots i}$ be a zero-mean Gaussian process. Let $p(x_{n1},\cdots,x_{nn})$ be the probability distribution of $X^n=(X_{n1},\cdots,X_{nn})$, and K_{X^n} be its covariance matrix. Let $h(X^n)=\frac{1}{2n}\ln(2\pi e)^n|K_{X^n}|$ be the entropy of X^n. If $|K_{X^n}|>0$, then,*

$$\text{Prob}\left\{\left|-\frac{1}{n}\ln p(x_{n1},\cdots,x_{nn})-h(X^n)\right|\geq\epsilon\right\}\leq e^{-n\left(\epsilon-\frac{1}{2}\ln(1+2\epsilon)\right)}. \tag{3}$$

Proof: See Cover and Pombra [4]. □

Lemma 1 states that if a sequence x^n is generated from a Gaussian process, the probability that x_n is not typical does not depend on the covariance matrix K_{X^n}. This allows a coding theorem for the Gaussian channel to be stated in a way that does not depend on the finite-dimensional covariance structure of the noise process.

Definition 1 *Let (X^n,Y^n) be a pair of random sequences. The set of jointly ϵ-typical sequences $x^n=(x_1,\cdots,x_n)$ and $y^n=(y_1,\cdots,y_n)$ with respect to (X^n,Y^n), denoted as $A_\epsilon^n(X^n,Y^n)$ is defined as:*

$$\begin{aligned}A_\epsilon^n(X^n,Y^n) \;=\; \{&(x^n,y^n):\\ &\left|-\frac{1}{n}\ln p(x^n)-h(X^n)\right|\leq\epsilon,\\ &\left|-\frac{1}{n}\ln p(y^n)-h(Y^n)\right|\leq\epsilon,\\ &\left|-\frac{1}{n}\ln p(x^n,y^n)-h(X^n,Y^n)\right|\leq\epsilon\,\},\end{aligned} \tag{4}$$

where $h(X^n)$, $h(Y^n)$ and $h(X^n,Y^n)$ denote the entropy of X^n, Y^n and (X^n,Y^n), respectively, and $p(x^n)$, $p(y^n)$ and $p(x^n,y^n)$ denote the marginal and joint distributions of X^n, Y^n and (X^n,Y^n), respectively.

Lemma 2 *Let X^n and Y^n be jointly Gaussian. The volume of the jointly typical set $A_\epsilon^n(X^n,Y^n)$ satisfies:*

$$\left(1-e^{-n\left(\epsilon-\frac{1}{2}\ln(1+2\epsilon)\right)}\right)e^{h(X^n,Y^n)-n\epsilon}\leq|A_\epsilon^n|\leq e^{h(X^n,Y^n)+n\epsilon}, \tag{5}$$

Also, if (U^n,V^n) are independent random sequences with the same marginal distributions as $p(x^n)$ and $p(y^n)$, then

$$\left(1-e^{-n\left(\epsilon-\frac{1}{2}\ln(1+2\epsilon)\right)}\right)e^{-I(X^n;Y^n)-3n\epsilon}\leq\text{Prob}\left\{(U^n,V^n)\in A_\epsilon^n(X^n,Y^n)\right\}\leq e^{-I(X^n;Y^n)+3n\epsilon}. \tag{6}$$

Proof: This is a straightforward extension of the similar result from Cover and Thomas [5, p.195] and [4]. The upper bound in (5) follows from the following inequality:

$$\begin{aligned} 1 &\geq \int_{A_\epsilon^n} p(x^n, y^n) dx^n dy^n \\ &\geq |A_\epsilon^n| e^{-h(X^n,Y^n)-n\epsilon}. \end{aligned} \tag{7}$$

The lower bound in (5) follows from Lemma 1:

$$\begin{aligned} 1 - e^{-n(\epsilon - \frac{1}{2}\ln(1+2\epsilon))} &\leq \int_{A_\epsilon^n} p(x^n, y^n) dx^n dy^n \\ &\leq |A_\epsilon^n| e^{-h(X^n,Y^n)+n\epsilon}. \end{aligned} \tag{8}$$

The proof of (6) is similar. Since (U^n, V^n) are independent, but with the same marginals as (X^n, Y^n), using the bound on the volume of typical set (5) and the definition of typicality:

$$\begin{aligned} \text{Prob}((U^n, V^n) \in A_\epsilon^n) &= \int_{A_\epsilon^n} p(x^n)p(y^n) dx^n dy^n \\ &\leq e^{h(X^n,Y^n)+n\epsilon} e^{-h(X^n)+n\epsilon} e^{-h(Y^n)+n\epsilon} \\ &\leq e^{-I(X^n;Y^n)+3n\epsilon}. \end{aligned} \tag{9}$$

Similarly,

$$\begin{aligned} \text{Prob}((U^n, V^n) \in A_\epsilon^n) &= \int_{A_\epsilon^n} p(x^n)p(y^n) dx^n dy^n \\ &\geq \left(1 - e^{-n(\epsilon - \frac{1}{2}\ln(1+2\epsilon))}\right) e^{h(X^n,Y^n)-n\epsilon} e^{-h(X^n)-n\epsilon} e^{-h(Y^n)-n\epsilon} \\ &\geq \left(1 - e^{-n(\epsilon - \frac{1}{2}\ln(1+2\epsilon))}\right) e^{-I(X^n;Y^n)-3n\epsilon}. \end{aligned} \tag{10}$$

□

Theorem 1 *Consider one block of n transmissions over a Gaussian channel $Y^n = X^n + Z^n$, where Z^n is a Gaussian random sequence with covariance matrix K_{Z^n}. Fix $\epsilon > 0$. Let*

$$C_n = \max_{K_{X^n}} \frac{1}{2n} \log \frac{|K_{X^n} + K_{Z^n}|}{|K_{Z^n}|}, \tag{11}$$

where the maximization is over covariance matrices K_{X^n} such that $\frac{1}{n}\text{tr}(K_{X^n}) \leq P(1-\epsilon)$. Suppose that the maximizing K_{X^n} satisfies $|K_{X^n}| > 0$, and that $C_n > 5\epsilon$, then there exists a $(e^{n(C_n-4\epsilon)}, n)$ code over a one-shot use of n transmissions, that satisfies power constraint P and has probability of error $P_e^{(n)} < e^{-n\alpha(\epsilon)}$, where $\alpha(\epsilon)$ does not depend on K_{Z^n}. Thus, C_n is the capacity of the Gaussian channel over one shot of n transmissions.

Proof: For each $n > 0$, construct a codebook $\mathcal{C}_x$ by independently generating $M_n = e^{n(C_n - 4\epsilon)}$ codewords of length n according to a Gaussian distribution $\mathcal{N}(0, K_{X^n})$, where K_{X^n} is the optimal input covariance matrix that maximizes $I(X^n; Y^n)$. Each codeword corresponds to one of $e^{n(C_n-4\epsilon)}$ messages. The encoder sends out the codeword x^n corresponding to the message. The decoder receives $y^n = x^n + z^n$, and decodes the message by looking for $\hat{x}^n$ in the codebook that is jointly ϵ-typical with y^n. An error occurs if $x^n \neq \hat{x}^n$.

We compute the probability of error by averaging over all codewords in all possible codebooks. There are three sources of possible errors. First, the average power of codewords in the codebook may exceed the power constraint. The probability of this happening is computed as follows. Let $X^n(m)_i$ denote the ith sample in mth codeword in the codebook. Let $P_{X^n(m)} = \frac{1}{n}\sum_i X^n(m)_i^2$ be the transmit power of the mth codeword. Now, $\mathbf{E}\left[P_{X^n(m)}\right] = \text{trace}(K_{X^n}) = P(1-\epsilon)$, and

$$\begin{aligned}
\mathbf{var}\left(P_{X^n(m)}\right) &= \mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X^n(m)_i^2 - P(1-\epsilon)\right)^2\right] \\
&\leq \frac{1}{n^2}\sum_{i,j}\mathbf{E}[X^n(m)_i^2 X^n(m)_j^2] + P^2(1-\epsilon)^2 \\
&\leq \frac{1}{n^2}\sum_{i,j}\sqrt{\mathbf{E}[X^n(m)_i^4]}\sqrt{\mathbf{E}[X^n(m)_j^4]} + P^2 \\
&\leq 3(nP)^2 + P \\
&\leq 4(nP)^2, \qquad (12)
\end{aligned}$$

where the third line follows from the Cauchy-Schwartz inequality, the fourth line follows the fact that the fourth moment of a Gaussian random variable $\mathcal{N}(0, \sigma^2)$ is $3\sigma^4$, and the variance of $X^n(m)_i$ is at most nP. Now, let $M_n = e^{n(C_n-4\epsilon)}$ be the number of codewords in a codebook. Then, the probability that the average power of a codebook exceeds the power constraint can be bounded using the Chebyshev inequality:

$$\begin{aligned}
P_{e,1}^{(n)} &= \text{Prob}\left\{\frac{1}{M_n}\sum_{m=1}^{M_n} P_{X^n(m)} > P\right\} \\
&\leq \text{Prob}\left\{\left|\frac{1}{M_n}\sum_{m=1}^{M_n} P_{X^n(m)} - P(1-\epsilon)\right| > \epsilon P\right\} \\
&\leq \mathbf{var}\left(\frac{1}{M_n}\sum_{m=1}^{M_n} P_{X^n(m)}\right)/(\epsilon P)
\end{aligned}$$

$$\begin{aligned} &\leq \frac{1}{M_n} 4n^2P^2/(\epsilon P) \\ &\leq e^{-n\epsilon}\epsilon^{-1}4n^2P, \end{aligned} \tag{13}$$

where $\mathbf{var}\left(\frac{1}{M_n}\sum_{m=1}^{M_n} P_{X^n(m)}\right) = \frac{1}{M_n}\mathbf{var}(P_{X^n(m)})$ since $P_{X^n(m)}$ and $P_{X^n(k)}$ are independent if $m \neq k$. The last inequality in (13) follows from the assumption that $C_n > 5\epsilon$, so $M_n = e^{n(C_n-4\epsilon)} \geq e^{n\epsilon}$. Note that $P_{1,e}^{(n)}$ goes to zero as n goes to infinity in a way that does not depend on K_{X^n}.

Second, the received signal y^n may not be jointly typical with x^n. By Lemma 1, and the assumption $|K_{X^n}| > 0$, this happens with probability at most $P_{e,2}^{(n)} \leq 3e^{-n(\epsilon-\frac{1}{2}\ln(1+2\epsilon))}$.

Third, there might be other codewords $\hat{x}^n \neq x^n$ that are jointly typical with y^n. By Lemma 2, the probability that an independently chosen codeword is jointly typical with y^n is at most $P_{e,3}^{(n)} \leq e^{-I(X^n;Y^n)+3n\epsilon}$, where $I(X^n;Y^n)$ is derived from the input covariance matrix K_{X^n}. There are $(e^{n(C_n-4\epsilon)}-1)$ other codewords, so the probability that any of these codewords is typical with y^n is at most $(e^{n(C_n-4\epsilon)}-1)P_{e,3}^{(n)}$. Now, K_{X^n} is the covariance matrix that maximizes $I(X^n;Y^n)$. For this K_{X^n}, $nC_n = \frac{1}{2}\log\frac{|K_{X^n}+K_{Z^n}|}{|K_{Z^n}|} = I(X^n;Y^n)$. Thus, $(e^{n(C_n-4\epsilon)}-1)P_{e,3}^{(n)} \leq e^{-n\epsilon}$.

Therefore, the total probability of error, averaged over all codebooks, is bounded by:

$$\begin{aligned} P_e^{(n)} &\leq P_{e,1}^{(n)} + P_{e,2}^{(n)} + (e^{n(C_n-4\epsilon)}-1)P_{e,3}^{(n)} \\ &\leq 4e^{-n\epsilon}\epsilon^{-1}n^2P + 3e^{-n(\epsilon-\frac{1}{2}\ln(1+2\epsilon))} + e^{-n\epsilon}. \end{aligned} \tag{14}$$

This implies that there exists at least one codebook with a probability of error that is at most this. Note that this probability of error bound does not depend on K_{Z^n}.

Finally, a converse can be proved in the usual manner [5]. So, C_n is the capacity of the Gaussian channel over a one-shot of n transmissions. □

Theorem 1 shows that for a Gaussian channel, a codebook can be constructed for each block-size n in such a way that the probability of error bound is independent of the covariance structure of Z^n. Thus, the noise process may have memory, or it may be non-stationary or non-ergodic. As long as the noise is Gaussian, the probability of error can be made arbitrarily small as the block-size goes to infinity irrespective of the finite-dimensional distribution of the noise process. This is true even though C_n may fluctuate arbitrarily. However, the condition $|K_{X^n}| > 0$ must be met. This condition ensures that X^n has enough degrees of freedom to drive the error probability to zero

exponentially. In fact, this assumption can be further relaxed. Let $f(n)$ be the number of positive eigenvalues of K_{X^n}. Then it is not difficult to see that the error probability is now bounded by $P_e^{(n)} \leq e^{-\alpha(\epsilon)f(n)}$. Thus, as long as $f(n) \to \infty$ as $n \to \infty$, the probability of error goes to zero.

3 Writing on Colored Paper

3.1 Dirty Paper

We wish to develop a similar distribution-free result for the Gaussian channel with non-causal side information at the transmitter. We will begin with a brief review of "Writing on dirty paper" [1]. Costa considered the communication problem over a Gaussian channel $Y^n = X^n + S^n + Z^n$ with the interference signal S^n known non-causally at the transmitter but not at the receiver. S^n and Z^n are assumed to be independent and identically Gaussian distributed with variances Q and N respectively. The transmitted signal X^n has a power constraint P. This communication channel is named "dirty-paper" because S^n can be thought of as dirt on a piece of paper. The transmitter tries to encode messages by writing on top of the dirt, knowing its location and intensity, while the receiver decodes the message without prior knowledge of the dirt. Costa considered the case where S^n and Z^n are both white Gaussian processes. The goal of this paper is to extend the result to dirt and noise processes that are colored, thus the title "Writing on colored paper".

The capacity of memoryless channels $p(y|x,s)$ with non-causal side information S available at the transmitter is given by Gel'fand and Pinsker [6] and Heegard and El Gamal [7]:

$$C = \max_{p(u,x|s)} \{I(U;Y) - I(U;S)\}, \tag{15}$$

where the maximization is over all joint distributions of the form $p(s)p(u,x|s)p(y|x,s)$[1], where U is an auxiliary random variable representing the codebook. The key to finding the capacity is to identify an appropriate U. The insight of Costa [1] is the following: despite the fact that the encoding of X^n necessarily depends on both the message and S^n, the distribution of X^n can be chosen to be independent of S^n. Costa showed that the optimal auxiliary variable U^n takes the form $U^n = X^n + \alpha S^n$, where X^n and S^n are independent Gaussian distributed, and the optimal α is $P/(P+N)$. This choice of

[1]The joint distribution can be further restricted to the form $p(s)p(u|s)p(x|u,s)$. Moreover, $p(x|u,s)$ can be taken as a deterministic function $x = f(u,s)$ without loss of capacity [6].

auxiliary variable U^n gives $C = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$, which is also the channel capacity when the interference S^n is also known at the receiver. Interestingly, neither capacity nor the optimal α depend on the variance of S.

An outline of the transmission strategy using the auxiliary variable U is as follows. First, generate $e^{n(I(U;Y)-\epsilon)}$ i.i.d. sequences u^n according to its marginal distribution $\mathcal{N}(0, P+\alpha^2 Q)$, and distribute them at random into $e^{n(C-2\epsilon)}$ bins. To encode a message $m \in \{1,\cdots,e^{n(C-2\epsilon)}\}$ given s^n, look in bin m for a sequence u^n such that the (u^n, s^n) pair is jointly typical, or equivalently find a sequence u^n from bin m such that $u^n - \alpha s^n$ is uncorrelated with s^n. The transmitter then sends $x^n = u^n - \alpha s^n$. The decoder looks for the unique sequence $\hat{u}^n$ such that $(\hat{u}^n, y^n)$ is jointly typical and lets $\hat{m}$ be the index of the bin containing $\hat{u}^n$.

3.2 Colored Paper

Consider now the more general additive Gaussian noise channel $Y_i = X_i + S_i + Z_i, i = 1,\cdots,\infty$, where $\{Z_i\}$ and $\{S_i\}$ are independent Gaussian processes with arbitrary finite-dimensional covariance matrices, and thus are not necessarily stationary or ergodic. The complete sample sequence S^n is known non-causally to the transmitter, but not to the receiver. Let an $(e^{n(C_n-\epsilon)}, n)$-code be a codebook of size $e^{n(C_n-\epsilon)}$ over a single block of n transmissions. Let $P_e^{(n)}$ be the average probability of error. We wish to construct a sequence of $(e^{n(C_n-\epsilon)}, n)$ codes for each finite block-size n, such that $P_e^{(n)}$ goes to zero as n goes to infinity, even as K_{S^n} and K_{Z^n} vary and C_n fluctuates arbitrarily. Again, the achievability result is based on the distribution-free Gaussian AEP.

Definition 2 *Consider one block of n transmissions over a Gaussian channel $Y^n = X^n + S^n + Z^n$, where S^n and Z^n are arbitrary Gaussian sequences, S^n is completely known at the transmitter only, and Z^n is known neither at the transmitter nor at the receiver. An (M_n, n) code consists of the following:*

1. *An index set $\{1,\cdots,M_n\}$.*
2. *An encoding function $X^n(m, S^n) : \{1,\cdots,M_n\} \times \mathcal{S}^n \to \mathcal{X}^n$.*
3. *A decoding function $\hat{m}(Y^n) : \mathcal{Y}^n \to \{1,\cdots,M_n\}$.*

The (M_n, n) code is said to satisfy a power constraint P if

$$\frac{1}{M_n}\sum_{m=1}^{M_n} \mathbf{E}_{S^n}[X^n(m,S^n)^T X^n(m,S^n)] \le nP, \tag{16}$$

and the probability of error $P_e^{(n)}$ *of the* (M_n, n) *code is defined to be:*

$$P_e^{(n)} = \frac{1}{M_n} \sum_{m=1}^{M_n} \mathbf{E}_{S^n} \left[\text{Prob}\{\hat{m}(Y^n) \neq m | X^n(m, S^n)\} \right], \tag{17}$$

where both expectations are over the interference sequence S^n*, as codewords are functions of* S^n*.*

Theorem 2 *Fix* $\epsilon > 0$*. Consider one block of* n *transmissions over a Gaussian channel* $Y^n = X^n + S^n + Z^n$*, where* S^n *and* Z^n *are independent Gaussian sequences with covariance* K_{S^n} *and* K_{Z^n} *respectively,* S^n *is known non-causally at the transmitter but not at the receiver, and* $|K_{S^n}| > 0$*. Let* $p(x^n|u^n, s^n)p(u^n|s^n)p(s^n)$ *be a joint Gaussian distribution such that* $|K_{U^n}| > 0$*,* $|K_{X^n}| > 0$*, and* $\frac{1}{n}\text{tr}(K_{x^n}) \leq P(1-\epsilon)$*. Let* $R_n = \frac{1}{n}(I(U^n; Y^n) - I(U^n; S^n))$*, and suppose that* $R_n > 6\epsilon$*. Then, there exists a* $(e^{n(R_n - 5\epsilon)}, n)$ *code satisfying a power constraint* P *and has a probability of error* $P_e^{(n)} \leq e^{-n\beta(\epsilon)}$*, where* $\beta(\epsilon)$ *does not depend on* K_{S^n} *and* K_{Z^n}*.*

Proof: Construct a $(e^{n(R_n - 5\epsilon)}, n)$ code as follows. Randomly generate $e^{I(U^n;Y^n)-4n\epsilon}$ i.i.d. Gaussian sequences u^n according to the marginal distribution $p(u^n)$ derived from the joint distribution $p(x^n|u^n, s^n)p(u^n|s^n)p(s^n)$. Put equal number of u^n's into e^{nR_n} bins at random, where $R_n = \frac{1}{n}(I(U^n; Y^n) - I(U^n; S^n)) - 5\epsilon$. Call this codebook $\mathcal{C}_u$. Reveal $\mathcal{C}_u$ to the receiver. Also, randomly generate $e^{I(S^n,U^n;X^n)+n\epsilon}$ i.i.d. Gaussian sequences x^n according to the marginal distribution $p(x^n)$ in a separate codebook. Call this codebook $\mathcal{C}_x$.

To encode a message $m \in \{1, \cdots, e^{nR_n}\}$ with the knowledge of s^n, the encoder looks for a u^n in bin m in $\mathcal{C}_u$ such that (u^n, s^n) is jointly ϵ'-typical, where ϵ' will be determined later[2]. If more than one pair can be found, choose one at random. Declare an error if no such u^n can be found. Next, for the given pair (u^n, s^n), find x^n in $\mathcal{C}_x$ such that (x^n, u^n, s^n) is jointly ϵ''-typical, where ϵ'' will also be determined later. Send x^n.

To decode the message based on the received sequence $y^n = x^n + s^n + z^n$, the decoder looks for a $\hat{u}^n$ in $\mathcal{C}_u$ such that $(\hat{u}^n, y^n)$ is jointly ϵ-typical. Declare an error if no or more than one such $\hat{u}^n$ exists. Decode $\hat{m}$ as the index of the bin containing $\hat{u}^n$.

We now compute the probability of error averaged over S^n and over all codeword functions in all possible codebooks. Without loss of generality, let the message index be $m = 1$. There are five sources of possible errors. Let E_1 be the event that given s^n and

[2]This code construction differs from the one in Gel'fand and Pinsker [6] in that it does not use strong typicality. Here, weak typicality with $\epsilon' \ll \epsilon$ is sufficient.

message m, there is no u^n in bin m that is jointly ϵ'-typical with s^n. Let E_2 be the event that for the selected u^n, there is no x^n that is jointly ϵ''-typical with (u^n, s^n). Let E_3 be the event that the average power of codewords in the codebook, averaged also over S^n, exceeds the power constraint. Index sequence u^n's as $\{u_1^n, \cdots u_{e^{I(U^n;Y^n)-4n\epsilon}}^n\}$. Let $E_{4,k}$ be the event that (u_k^n, y^n) is jointly ϵ-typical. Without loss of generality, assume u_1^n is the typical u^n chosen. The fourth source of error is the event $E_{4,1}^c$, i.e. when (u_1^n, y^n) is not jointly typical. The fifth source of error is the event $E_{4,k}$ for $k \geq 2$, i.e. when some other (u_k^n, y^n) is jointly typical. Define the notation $P_{e,j}^{(n)} = \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_j\}]$ for $j = 1, 2, 3$. Let $P_{e,4}^{(n)} = \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_{4,1}^c\}]$, and $P_{e,5}^{(n)} = \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_{4,k}\}]$, $k \neq 1$. By the union bound, the average probability of error is bounded by:

$$\begin{aligned}
P_e^{(n)} &= \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{\text{Error}\}] \\
&\leq \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_1 \cup E_2 \cup E_3 \cup E_{4,1}^c \cup E_{4,2} \cup \cdots \cup E_{4,e^{I(U^n;Y^n)-4n\epsilon}}\}] \\
&\leq \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_1\}] + \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_2\}] + \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_3\}] + \\
&\quad \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_{4,1}^c\}] + \sum_{k=2}^{e^{I(U^n;Y^n)-4n\epsilon}} \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_{4,k}\}] \\
&= P_{e,1}^{(n)} + P_{e,2}^{(n)} + P_{e,3}^{(n)} + P_{e,4}^{(n)} + (e^{I(U^n;Y^n)-4n\epsilon} - 1)P_{e,5}^{(n)} \qquad (18)
\end{aligned}$$

First, let's consider the first source of error E_1. The computation of $P_{e,1}^{(n)}$ mimics that in [5, pp.352-356]. The following two lemmas will be useful.

Lemma 3 *For all* $(u^n, s^n) \in A_\epsilon^n$, $p(u^n) \geq p(u^n|s^n)e^{-I(U^n;S^n)-3n\epsilon}$.

Lemma 4 *For* $0 \leq x, y \leq 1$, *and* $n \geq 0$, $(1 - xy)^n \leq 1 - x + e^{-yn}$.

The lemmas are proved in [5]. Let $\mathbf{1}_{\{\}}$ be the indicator function. Given a particular s^n, the probability that a randomly chosen u^n is not ϵ'-typical with s^n is:

$$\text{Prob}\{(U^n, s^n) \notin A_{\epsilon'}^n\} = 1 - \int_{u^n} p(u^n)\mathbf{1}_{(u^n,s^n)\in A_{\epsilon'}^n} du^n. \qquad (19)$$

By code construction, there are in total $e^{I(U^n;Y^n)-4n\epsilon}$ u^n's, and $e^{I(U^n;Y^n)-I(U^n;S^n)-5n\epsilon}$ bins. So, each bin contains $e^{I(U^n;S^n)+n\epsilon}$ u^n's. Thus, the probability that a bin does not contain a u^n jointly ϵ'-typical with s^n, averaged over S^n and over all codebooks $\mathcal{C}_u$, is:

$$P_{e,1}^{(n)} = \int_{s^n} p(s^n) \left(1 - \int_{u^n} p(u^n)\mathbf{1}_{(u^n,s^n)\in A_{\epsilon'}^n} du^n\right)^{e^{I(U^n;S^n)+n\epsilon}} ds^n \qquad (20)$$

By Lemma 3,

$$\int_{u^n} p(u^n)\mathbf{1}_{(u^n,s^n)\in A^n_{\epsilon'}} du^n \geq \int_{u^n} p(u^n|s^n) e^{-I(U^n;S^n)-3n\epsilon'} \mathbf{1}_{(u^n,s^n)\in A^n_{\epsilon'}} du^n. \tag{21}$$

Substitute this into the expression (20), and use Lemma 4:

$$\begin{aligned} P^{(n)}_{e,1} &\leq \int_{s^n} p(s^n)\left(1 - e^{-I(U^n;S^n)-3n\epsilon'} \int_{u^n} p(u^n|s^n)\mathbf{1}_{(u^n,s^n)\in A^n_{\epsilon'}} du^n\right)^{e^{I(U^n;S^n)+n\epsilon}} ds^n \\ &\leq \int_{s^n} p(s^n)\left(1 - \int_{u^n} p(u^n|s^n)\mathbf{1}_{(u^n,s^n)\in A^n_{\epsilon'}} du^n + e^{-e^{-I(U^n;S^n)-3n\epsilon'} e^{I(U^n;S^n)+n\epsilon}}\right) ds^n \\ &= 1 - \int_{s^n}\int_{u^n} p(s^n)p(u^n|s^n)\mathbf{1}_{(u^n,s^n)\in A^n_{\epsilon'}} du^n ds^n + e^{-e^{n(\epsilon-3\epsilon')}}, \end{aligned} \tag{22}$$

The last term in the probability of error bound goes to zero doubly exponentially as n goes to infinity if $\epsilon > 3\epsilon'$. The first two terms represent the probability that a randomly chosen pair (s^n, u^n) is not ϵ'-joint typical. This happens with probability $3e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}$ by Lemma 1 and the assumption $|K_{S^n}| > 0$ and $|K_{U^n}| > 0$. Thus, if $\epsilon' < \epsilon/3$,

$$P^{(n)}_{e,1} \leq 3e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))} + e^{-e^{n(\epsilon-3\epsilon')}}, \tag{23}$$

which goes to zero as n goes to infinity in a way that does not depend on the finite-dimensional covariance structures of S^n.

The second type of error E_2 occurs when for the given s^n and chosen u^n, there does not exist a x^n in $\mathcal{C}_x$ that is jointly ϵ''-typical with (u^n, s^n). Now, $\mathcal{C}_x$ contains $e^{I(X^n;U^n,S^n)+n\epsilon}$ independently generated x^n's. If a (u^n, s^n) were randomly generated with the joint Gaussian distribution $p(u^n, s^n)$, then the same technique used in the computation of $P^{(n)}_{e,1}$ can be immediately used here to bound the probability of not being able to find a jointly typical x^n. However, by the code construction, s^n and u^n are generated independently with their respective Gaussian marginal distributions, then an ϵ'-typical u^n is chosen from a codebook once s^n is revealed. So, (U^n, S^n) does not have a joint Gaussian distribution any more. Nevertheless, it turns out that a distribution-free error probability bound can still be derived. The key is to observe that since u^n and s^n are generated independently and a (u^n, s^n) pair is chosen only if (u^n, s^n) is jointly ϵ'-typical, the distribution of (U^n, S^n) is a Gaussian distribution $p(u^n)p(s^n)$ conditioned on the typical set $A^n_{\epsilon'}(U^n, S^n)$. We will show that by choosing ϵ' sufficiently small, the probability of error can be bounded using the Gaussian AEP for a jointly Gaussian (U^n, S^n). More precisely, let E_{u^n,s^n} be the event that for a given (u^n, s^n), a randomly chosen codebook does not contain at least one x^n

ϵ''-jointly typical with (u^n, s^n). Let $p(\cdot)$ denote Gaussian distributions. Then,

$$\begin{aligned}
P_{e,2}^{(n)} &= \mathbf{E}_{(S^n,\mathcal{C}_u)}[\text{Prob}\{E_{U^n,S^n}^c\}] \\
&= \frac{\displaystyle\int_{A_{\epsilon'}^n} \text{Prob}\{E_{u^n,s^n}^c\} p(u^n)p(s^n)du^n ds^n}{\displaystyle\int_{A_{\epsilon'}^n} p(u^n)p(s^n)du^n ds^n} \\
&\leq \frac{\displaystyle\int_{A_{\epsilon'}^n} \text{Prob}\{E_{u^n,s^n}^c\} e^{-h(U^n)+n\epsilon'} e^{-h(S^n)+n\epsilon'} du^n ds^n}{\displaystyle\int_{A_{\epsilon'}^n} e^{-h(U^n)-n\epsilon'} e^{-h(S^n)-n\epsilon'} du^n ds^n} \\
&= e^{4n\epsilon'} \int_{A_{\epsilon'}^n} \text{Prob}\{E_{u^n,s^n}^c\} \frac{1}{|A_{\epsilon'}^n|} du^n ds^n \\
&\leq \frac{e^{4n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \int_{A_{\epsilon'}^n} \text{Prob}\{E_{u^n,s^n}^c\} e^{-h(U^n,S^n)+n\epsilon'} du^n ds^n \\
&\leq \frac{e^{6n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \int_{A_{\epsilon'}^n} \text{Prob}\{E_{u^n,s^n}^c\} p(u^n,s^n) du^n ds^n \\
&\leq \frac{e^{6n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \int_{u^n,s^n} \text{Prob}\{E_{u^n,s^n}^c\} p(u^n,s^n) du^n ds^n, \qquad (24)
\end{aligned}$$

where the definition of typicality and Lemma 2 have been used repeatedly. Now, we can bound $P_{e,2}^{(n)}$ as if (u^n, s^n) were generated by a joint Gaussian process. Using the same technique as in (19)-(22), since there are $e^{I(U^n,S^n;X^n)+n\epsilon}$ x^n's in $\mathcal{C}_x$:

$$\begin{aligned}
\text{Prob}\{E_{u^n,s^n}^c\} &= \text{Prob}\{(X^n,u^n,s^n) \notin A_{\epsilon''}^n\}^{e^{I(U^n,S^n;X^n)+n\epsilon}} \\
&= \left(1 - \int_{x^n} p(x^n) \mathbf{1}_{(x^n,u^n,s^n)\in A_{\epsilon''}^n} dx^n\right)^{e^{I(U^n,S^n;X^n)+n\epsilon}} \\
&\leq \left(1 - e^{-I(X^n;U^n,S^n)-3n\epsilon''} \int_{x^n} p(x^n|u^n,s^n) \mathbf{1}_{(x^n,u^n,s^n)\in A_{\epsilon''}^n} dx^n\right)^{e^{I(X^n;U^n,S^n)+n\epsilon}} \\
&\leq 1 - \int_{x^n} p(x^n|u^n,s^n) \mathbf{1}_{(x^n,u^n,s^n)\in A_{\epsilon''}^n} dx^n + e^{-e^{-n(\epsilon-3\epsilon'')}} \qquad (25)
\end{aligned}$$

Continuing with (24) by substituting in the above,

$$\begin{aligned}
P_{e,2}^{(n)} &\leq \frac{e^{6n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \int_{u^n,s^n} \text{Prob}\{E_{u^n,s^n}^c\} p(u^n,s^n) du^n ds^n \\
&\leq \frac{e^{6n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \left(\int_{u^n,x^n,s^n} \mathbf{1}_{(x^n,u^n,s^n)\notin A_{\epsilon''}^n} p(x^n,u^n,s^n) dx^n du^n ds^n + e^{-e^{-n(\epsilon-3\epsilon'')}}\right)
\end{aligned}$$

$$\leq \frac{e^{6n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon''))}}\left(7e^{-n(\epsilon''-\frac{1}{2}\ln(1+2\epsilon''))}+e^{-e^{n(\epsilon-3\epsilon'')}}\right), \quad (26)$$

where the last inequality follows from Lemma 1. The conditions $|K_{U^n}|>0$, $|K_{X^n}|>0$ and $|K_{S^n}|>0$ are needed for Lemma 1 to hold. This probability of error bound is distribution-free, and it goes to zero as n goes to infinity if $3\epsilon''<\epsilon$ and $6\epsilon'<\epsilon''-\frac{1}{2}\ln(1+2\epsilon'')$.

The third type of error $E_3(\mathcal{C}_x)$ is the event that the average power of the codewords in the codebook $\mathcal{C}_x$ exceeds the power constraint. Let $(X^n(m,S^n))_i$ denote the ith sample in mth codeword function in the codebook. Let $P_{X^n(m)}=\frac{1}{n}\sum_i(X^n(m,S^n))_i^2$ be the power of the codeword m. We wish to compute $P_{e,3}^{(n)}=\mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_3\}]=\mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}\left[\text{Prob}\left\{\frac{1}{M_n}\sum_{m=1}^{M_n}P_{X^n(m)}>P\right\}\right]$. If X^n were Gaussian distributed, the derivation in equations (12) and (13) could have been used here again to derive this probability of error. However, x^n's in $\mathcal{C}_x$ are generated differently. First, s^n is chosen from a Gaussian distribution. Second, a u^n is chosen so that (u^n,s^n) is jointly ϵ'-typical. Third, a x^n is chosen so that (x^n,u^n,s^n) is jointly ϵ''-typical. More precisely, (x^n,u^n,s^n) is generated independently from their respective Gaussian marginal distributions, and the triple is chosen only if it is in the following set $A_1\cap A_{\epsilon''}^n$, where $A_1=\{(x^n,u^n,s^n):(u^n,s^n)\in A_{\epsilon'}^n(U^n,S^n)\}$ and $A_{\epsilon''}^n$ is the ϵ''-typical set on the joint distribution (X^n,U^n,S^n). In other words, the joint distribution of (U^n,X^n,S^n) is the Gaussian distribution $p(u^n)p(x^n)p(s^n)$ conditioned on $A_1\cap A_{\epsilon''}^n$. The idea is to choose ϵ' and ϵ'' sufficiently small so that (u^n,x^n,s^n) looks Gaussian. In the following, set $\epsilon'<\epsilon''$, so that $A_{\epsilon'}^n(X^n,U^n,S^n)\subset(A_1\cap A_{\epsilon''}^n)$. Then, using a similar chain of inequalities as (24), $\text{Prob}\{E_3\}$ averaged over S^n and over all codebooks can be computed as follows:

$$\begin{aligned}
P_{e,3}^{(n)} &= \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{E_3\}] \\
&= \frac{\int_{A_1\cap A_{\epsilon''}^n}\text{Prob}\{E_3\}p(u^n)p(x^n)p(s^n)du^ndx^nds^n}{\int_{A_1\cap A_{\epsilon''}^n}p(u^n)p(x^n)p(s^n)du^ndx^nds^n} \\
&\leq \frac{\int_{A_1\cap A_{\epsilon''}^n}\text{Prob}\{E_3\}e^{-h(U^n)+n\epsilon''}e^{-h(X^n)+n\epsilon''}e^{-h(S^n)+n\epsilon''}du^ndx^nds^n}{\int_{A_1\cap A_{\epsilon''}^n}e^{-h(U^n)-n\epsilon''}e^{-h(X^n)-n\epsilon''}e^{-h(S^n)-n\epsilon''}du^ndx^nds^n} \\
&= e^{6n\epsilon''}\int_{A_1\cap A_{\epsilon''}^n}\text{Prob}\{E_3\}\frac{1}{|A_1\cap A_{\epsilon''}^n|}du^ndx^nds^n
\end{aligned}$$

$$
\begin{aligned}
&\leq e^{6n\epsilon''} \int_{A_1 \cap A^n_{\epsilon''}} \text{Prob}\{E_3\} \frac{1}{|A^n_{\epsilon'}(X^n, U^n, S^n)|} du^n dx^n ds^n \\
&\leq \frac{e^{6n\epsilon''}}{1 - e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}} \int_{A_1 \cap A^n_{\epsilon''}} \text{Prob}\{E_3\} e^{-h(U^n, X^n, S^n) + n\epsilon'} du^n dx^n ds^n \\
&\leq \frac{e^{6n\epsilon''} e^{2n\epsilon'}}{1 - e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}} \int_{A_1 \cap A^n_{\epsilon''}} \text{Prob}\{E_3\} p(u^n, x^n, s^n) du^n dx^n ds^n \\
&\leq \frac{e^{6n\epsilon''} e^{2n\epsilon'}}{1 - e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}} \int_{u^n, x^n, s^n} \text{Prob}\{E_3\} p(u^n, x^n, s^n) du^n dx^n ds^n \qquad (27)
\end{aligned}
$$

Now, we can use the Chebyshev inequality to bound $P^{(n)}_{e,3}$ as if X^n is Gaussian as conditioning has now been removed. Let $M_n = e^{n(R_n - 5\epsilon)}$ be the size of the codebook. By the assumption $R_n > 6\epsilon$, $M_n > e^{n\epsilon}$. Use the same technique as in (12)-(13), the probability that the average power of the codebook, averaged over M_n messages and over S^n, exceeds the power constraint, is:

$$
\begin{aligned}
P^{(n)}_{e,3} &\leq \frac{e^{6n\epsilon''} e^{2n\epsilon'}}{1 - e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}} \int_{x^n} \text{Prob}\left\{ \frac{1}{M_n} \sum_{m=1}^{M_n} P_{x^n(m)} > P \right\} p(x^n) dx^n \\
&\leq \frac{e^{6n\epsilon''} e^{2n\epsilon'} e^{-n\epsilon} \epsilon^{-1} 4n^2 P}{1 - e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}}, \qquad (28)
\end{aligned}
$$

where the last inequality follows from (13). Thus, if $6\epsilon'' + 2\epsilon' < \epsilon$, $P^{(n)}_{e,3}$ goes to zero as n goes to infinity in a way that does not depend on K_{X^n}.

The fourth type of error $E^c_{4,1}$ occurs when (u^n, y^n) is not jointly typical. Again, had (U^n, Y^n) been jointly Gaussian, $P^{(n)}_{e,4}$ could have been bounded by Gaussian AEP directly. But, again (u^n, s^n, x^n) is generated independently with their respective Gaussian marginal distributions conditioned on $A_1 \cap A^n_{\epsilon''}$ by the code construction process. So, the same argument as that in (27) has to be used again to remove conditioning and to bound the probability of error as if X^n is Gaussian. This will make $y^n = x^n + s^n + z^n$ also look Gaussian, and allows the distribution-free Gaussian AEP to be used to compute $P^{(n)}_{e,4}$. More specifically, $P^{(n)}_{e,4}$ averaged over S^n and over all codebooks is computed as follows:

$$
\begin{aligned}
P^{(n)}_{e,4} &= \mathbf{E}_{(S^n, \mathcal{C}_u, \mathcal{C}_x)}[\text{Prob}\{(U^n, Y^n) \notin A^n_\epsilon\}] \\
&= \frac{\displaystyle\int_{A_1 \cap A^n_{\epsilon''}} \text{Prob}\{(u^n, Y^n) \notin A^n_\epsilon\} p(u^n) p(x^n) p(s^n) du^n dx^n ds^n}{\displaystyle\int_{A_1 \cap A^n_{\epsilon''}} p(u^n) p(x^n) p(s^n) du^n dx^n ds^n} \\
&\leq \frac{e^{6n\epsilon''} e^{2n\epsilon'}}{1 - e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))}} \int_{u^n, x^n, s^n} \text{Prob}\{(u^n, Y^n) \notin A^n_\epsilon\} p(u^n, x^n, s^n) du^n dx^n ds^n
\end{aligned}
$$

$$\leq \frac{3e^{6n\epsilon''}e^{2n\epsilon'}e^{-n(\epsilon-\frac{1}{2}\ln(1+2\epsilon))}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}}, \tag{29}$$

where the last inequality follows from Lemma 1 since the probability of error is now averaged over Gaussian (u^n, s^n, x^n), so $Y^n = x^n + s^n + Z^n$ is also Gaussian. Lemma 1 requires $|K_{U^n}| > 0$, and also $|K_{Y^n}| > 0$, which is provided by the assumption that $|K_{X^n}| > 0$. This probability of error bound is distribution-free, and it goes to zero as n goes to infinity if $6\epsilon'' + 2\epsilon' < \epsilon - \frac{1}{2}\ln(1+2\epsilon)$.

The fifth type of error $E_{5,i}$ occurs when some other u^n is jointly typical with y^n. Let $\tilde{u}^n$ denote some other u^n in the codebook. $\tilde{U}^n$ is Gaussian distributed and it is independent of Y^n. Y^n is, however, not Gaussian distributed. So, in order to use the Gaussian AEP, the previous argument has to be used again. The probability of error, averaged over all codebooks and over S^n, is computed in the same way as in (27):

$$\begin{aligned}
P_{e,5}^{(n)} &= \mathbf{E}_{(S^n,\mathcal{C}_u,\mathcal{C}_x)}[\text{Prob}\{(\tilde{U}^n, Y^n) \in A_\epsilon^n\}] \\
&= \frac{\displaystyle\int_{A_1 \cap A_{\epsilon''}^n} \text{Prob}\{(\tilde{U}^n, Y^n) \in A_\epsilon^n\} p(u^n)p(x^n)p(s^n)du^n dx^n ds^n}{\displaystyle\int_{A_1 \cap A_{\epsilon''}^n} p(u^n)p(x^n)p(s^n)du^n dx^n ds^n} \\
&\leq \frac{e^{6n\epsilon''}e^{2n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \int_{u^n,x^n,s^n} \text{Prob}\{(\tilde{U}^n, Y^n) \in A_\epsilon^n\} p(u^n, x^n, s^n)du^n dx^n ds^n \\
&\leq \frac{e^{-I(U^n;Y^n)+3n\epsilon+6n\epsilon''+2n\epsilon'}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}},
\end{aligned} \tag{30}$$

where the last inequality follows from Lemma 2 because the probability of error is now averaged over Gaussian (u^n, s^n, x^n), so $Y^n = x^n + s^n + Z^n$ is also Gaussian, and also $\tilde{U}^n$ is Gaussian and independent of Y^n. Now, there are $e^{I(U^n;Y^n)-4n\epsilon} - 1$ other $\tilde{u}^n$'s. So, the probability of the fourth type of error is bounded by:

$$\left(e^{I(U^n;Y^n)-4n\epsilon} - 1\right) P_{e,5}^{(n)} \leq \frac{e^{-n(\epsilon-6\epsilon''-2\epsilon')}}{1-e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}}. \tag{31}$$

This probability of error bound is distribution-free and it goes down to zero as n goes to infinity if $6\epsilon'' + 2\epsilon' < \epsilon$.

Putting everything together, set

$$\epsilon'' = \min\left\{\frac{\epsilon}{8}, \frac{\epsilon - \frac{1}{2}\ln(1+2\epsilon)}{8}\right\}. \tag{32}$$

$$\epsilon' = \min\left\{\epsilon'', \frac{\epsilon' - \frac{1}{2}\ln(1+2\epsilon')}{7}\right\}. \tag{33}$$

A $(e^{n(R_n-5\epsilon)}, n)$ code constructed using this (ϵ', ϵ'') will satisfy the power constraint P and will have an average probability of error:

$$\begin{aligned} P_e^{(n)} &= P_{e,1}^{(n)} + P_{e,2}^{(n)} + P_{e,3}^{(n)} + P_{e,4}^{(n)} + (e^{I(U^n;Y^n)-4n\epsilon} - 1)P_{e,5}^{(n)} \\ &\leq 3e^{-n(\epsilon' - \frac{1}{2}\ln(1+2\epsilon'))} + e^{-e^{n(\epsilon-3\epsilon')}} + \frac{7e^{-n(\epsilon''-\frac{1}{2}\ln(1+2\epsilon'')-6\epsilon')} + e^{e^{-n(\epsilon-3\epsilon'')}+6n\epsilon'}}{1 - e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \\ &\quad + \frac{e^{-n(\epsilon-6\epsilon''-2\epsilon')}\epsilon^{-1}4n^2P}{1 - e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} + \frac{3e^{-n(\epsilon-\frac{1}{2}\ln(1+2\epsilon)-6\epsilon''-2\epsilon')}}{1 - e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} + \frac{e^{-n(\epsilon-6\epsilon''-2\epsilon')}}{1 - e^{-n(\epsilon'-\frac{1}{2}\ln(1+2\epsilon'))}} \end{aligned} \tag{34}$$

This is the probability of error averaged over S^n and over all possible codebooks. Thus, there must exist at least one codebook whose probability of error averaged over S^n is at most this. This error probability bound is independent of the joint distribution $p(x^n|u^n, s^n)p(u^n|s^n)p(s^n)p(y^n|x^n, s^n)$. □

3.3 Capacity

The coding theorem assumes that both X^n and U^n are Gaussian random variables. In this section, we will further restrict the auxiliary variable to take the form $U^n = X^n + FS^n$, where X^n and S^n are independent. In this case, the encoding process reduces to finding a u^n jointly typical with s^n in the appropriate bin, then setting $x^n = u^n - Fs^n$. This is analogous to the i.i.d. case where $U = X + \alpha S$, and $\alpha = P/(P+N)$. The main result here is to show that the capacity of the "colored-paper" channel is achieved with an optimal choice of F, hence the above choice of (U^n, X^n) is without loss of generality. Just as in the i.i.d. case, it turns out that neither optimal F nor capacity depends on the distribution of S^n. Curiously, the optimal F takes the form of an optimal non-causal Wiener filter. In the i.i.d. case, $\alpha = P/(P+N)$ can be interpreted as the optimal filter to estimate X from $X + Z$. In the vector case, the optimal takes the form $F = K_{X^n}(K_{X^n} + K_{Z^n})^{-1}$, and it can be interpreted as the optimal filter to estimate X^n from $X^n + Z^n$.

Theorem 3 *Let $Y^n = X^n + S^n + Z^n$, where X^n, S^n and Z^n are Gaussian sequences with covariance matrices K_{X^n}, K_{S^n} and K_{Z^n} respectively. Let X^n, S^n and Z^n be independent, and let $U^n = X^n + FS^n$, where F is an $n \times n$ matrix. Then an optimal matrix F which maximizes $I(U^n; Y^n) - I(U^n; S^n)$ is $F = K_{X^n}(K_{X^n} + K_{Z^n})^{-1}$. Further, the maximum value of $I(U^n; Y^n) - I(U^n; S^n)$ is $I(X^n; Y^n|S^n)$.*

Proof: For the ease of notation, we drop the superscript n. It is understood that the result holds for each individual n. Compute:

$$\begin{aligned} I(U;Y) &= H(U) + H(Y) - H(U;Y) \\ &= \frac{1}{2}\log \frac{|K_X + FK_SF^T| \cdot |K_X + K_S + K_Z|}{\left| \begin{array}{cc} K_X + FK_SF^T & K_X + FK_S \\ K_X + K_SF^T & K_X + K_S + K_Z \end{array} \right|}, \end{aligned} \tag{35}$$

where $H(U)$, $H(Y)$ and $H(U;Y)$ are computed using the relation $U = X + FS$. We also need $I(U;S)$. Since $U = X + FS$, we can view S as the input and U as the output of a Gaussian channel, with X as noise. So,

$$I(U;S) = \frac{1}{2}\log \frac{|K_X + FK_SF^T|}{|K_X|} \tag{36}$$

Define the function:

$$\begin{aligned} R(F) &= I(U;Y) - I(U;S) \\ &= \frac{1}{2}\log \frac{|K_X| \cdot |K_X + K_S + K_Z|}{\left| \begin{array}{cc} K_X + FK_SF^T & K_X + FK_S \\ K_X + K_SF^T & K_X + K_S + K_Z \end{array} \right|}. \end{aligned} \tag{37}$$

The task is to maximize $R(F)$ over F. By expanding the denominator using Schur's complement formula for matrix determinant $\left| \begin{array}{cc} A & B \\ C & D \end{array} \right| = |D| \cdot |A - BD^{-1}C|$, we obtain,

$$R(F) = \frac{1}{2}\log \frac{|K_X|}{|K_X + FK_SF^T - (K_X + FK_S)(K_X + K_S + K_Z)^{-1}(K_X + K_SF^T)|}. \tag{38}$$

So to maximize $R(F)$ over F, we only need to minimize the denominator in the above. The denominator is a quadratic function of F, so it can be minimized with the standard "complete-the-square" technique. Setting the denominator as $(Fa-b)(Fa-b)^T + c$, where a, b and c are all $n \times n$ matrices, and comparing the coefficients:

$$aa^T = K_S - K_S(K_X + K_S + K_Z)^{-1}K_S, \tag{39}$$

$$ba^T = K_X(K_X + K_S + K_Z)^{-1}K_S, \tag{40}$$

$$bb^T + c = K_X - K_X(K_X + K_S + K_Z)^{-1}K_X. \tag{41}$$

Then, the minimizing F is,

$$F = ba^{-1}$$

$$
\begin{aligned}
&= ba^T(aa^T)^{-1} \\
&= K_X(K_X+K_S+K_Z)^{-1}K_S(K_S-K_S(K_X+K_S+K_Z)^{-1}K_S)^{-1} \\
&= K_X(K_X+K_Z)^{-1}. \qquad (42)
\end{aligned}
$$

This is analogous to the scalar case where $\alpha = P/(P+N)$. Now, set $F = ba^{-1}$. So, the minimum value of the denominator is c. To find c, we need to solve for bb^T,

$$
\begin{aligned}
bb^T &= Faa^TF^T \\
&= K_X(K_X+K_Z)^{-1}(K_S-K_S(K_X+K_S+K_Z)^{-1}K_S)(K_X+K_Z)^{-1}K_X \\
&= K_X(K_X+K_Z)^{-1}(K_S^{-1}+(K_X+K_Z)^{-1})^{-1}(K_X+K_Z)^{-1}K_X \\
&= K_X(K_S^{-1}(K_X+K_Z)+I)^{-1}(K_X+K_Z)^{-1}K_X \\
&= K_X(K_X+K_Z+K_S)^{-1}K_S(K_X+K_Z)^{-1}K_X, \qquad (43)
\end{aligned}
$$

where the identity $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$ is used. The minimum value of the denominator in (38) can then be found:

$$
\begin{aligned}
c &= K_X - K_X(K_X+K_S+K_Z)^{-1}K_X - bb^T \\
&= K_X - K_X(K_X+K_Z+K_S)^{-1}(I+K_S(K_X+K_Z)^{-1})K_X \\
&= K_X - K_X(K_X+K_Z)^{-1}K_X. \qquad (44)
\end{aligned}
$$

Thus,

$$
\max_F R(F) = \frac{1}{2}\log\frac{|K_X|}{|K_X - K_X(K_X+K_Z)^{-1}K_X|}. \qquad (45)
$$

It remains to evaluate the above. Using the determinant formula for Schur's complement again, $\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D|\cdot|A-BD^{-1}C| = |A|\cdot|D-CA^{-1}B|$, we have:

$$
|K_X+K_Z|\cdot|K_X - K_X(K_X+K_Z)^{-1}K_X| = |K_X|\cdot|K_Z|. \qquad (46)
$$

Thus,

$$
\max_F R(F) = \frac{1}{2}\log\frac{|K_X+K_Z|}{|K_Z|}. \qquad (47)
$$

This is the mutual information formula for the vector Gaussian channel without the interfering S. Thus $\max_F I(U;Y) - I(U;S) = I(X;Y|S)$. The optimal $F = K_X(K_X+K_Z)^{-1}$. The capacity achieving K_U is equal to $K_X + K_X(K_X+K_Z)^{-1}K_S(K_X+K_Z)^{-1}K_X$. □

Theorem 3 by itself would have been sufficient if we only consider a memoryless Gaussian vector channels $Y^n = X^n + S^n + Z^n$, where each use of the channel involves a

vector input and a vector output, and coding is done over many uses of the channel. In this case, the general capacity result for memoryless channels with side information by Gel'fand and Pinsker [6] and Heegard and El Gamal [7] applies, and the evaluation of the capacity reduces to the maximization problem above. In fact, as Lapidoth pointed out to us [8], a memoryless vector channel may be transformed into n parallel sub-channels by a diagonalization of the noise covariance. This gives an alternative proof of Theorem 3. Let $K_Z = Q^T \Lambda_Z Q$ be an eigenvalue decomposition. Define $\tilde{Y} = QY$, $\tilde{X} = QX$, $\tilde{S} = QS$, and $\tilde{Z} = QZ$. Then, the vector channel reduces to $\tilde{Y} = \tilde{X} + \tilde{S} + \tilde{Z}$, where $\tilde{Z}$ is uncorrelated, but $\tilde{S}$ is correlated across the sub-channels. Such correlation is of no consequence to channel capacity because each sub-channel can be coded separately and the capacity of the scalar channel with known interference does not depend on the distribution of S. Thus, the capacity of the memoryless vector channel is the same as if S^n is not present, and $\max_{p(u,x|s)} I(U;Y) - I(U;S) = I(X;Y|S)$.

The above approach works if the vector channel can be repeated so that the capacity on each sub-channel is achieved over many blocks. In the context considered in this paper where S^n and Z^n may be potentially non-stationary or non-ergodic, the coding theorem developed here which works on a single block of n transmissions is indeed needed.

Theorem 4 *Consider one block of n transmissions in a Gaussian channel $Y^n = X^n + S^n + Z^n$, where S^n and Z^n are independent Gaussian sequences with S^n known non-causally at the transmitter. Suppose $|K_{S^n}| > 0$. The capacity of the channel under a power constraint P is:*

$$C_n = \max_{K_{X^n}} \frac{1}{2n} \log \frac{|K_{X^n} + K_{Z^n}|}{|K_{Z^n}|}, \tag{48}$$

provided that the maximization is over covariance matrices K_{X^n} such that $\frac{1}{n}\mathrm{tr}(K_{X^n}) \leq P$, and the maximizing K_{X^n} is such that $|K_{X^n}| > 0$.

Proof: The achievability is a direct consequence of Theorem 2 and 3. Fix $\epsilon > 0$. By Theorem 2, $R_n = \frac{1}{n}(I(U^n;Y^n) - I(U^n;S^n)) - \epsilon$ is achievable, provided that $|K_{X^n}| > 0$ and $\mathrm{trace}(K_{X^n}) \leq P(1-\epsilon)$. Now, let $U^n = X^n + FS^n$, where X^n and S^n are independent, and $F = K_{X^n}(K_{X^n} + K_{Z^n})^{-1}$. By Theorem 3, this choice of U^n gives $I(U^n;Y^n) - I(U^n;S^n) = I(X^n;Y^n|S^n) = \frac{1}{2}\log\frac{|K_{X^n} + K_{Z^n}|}{|K_{Z^n}|}$. So, $R_n = \frac{1}{n}I(X^n;Y^n|S^n) - \epsilon$ is achievable. Now, $I(X^n;Y^n|S^n)$ depends only on $p(x)$, and not directly on $p(u|x,s)$, so it can be further optimized over all $p(x)$ satisfying the power constraint. Thus, $R_n = \max_{p(x)} \frac{1}{n} I(X^n;Y^n|S^n) - \epsilon = \max_{K_{X^n}} \frac{1}{2n}\log\frac{|K_{X^n} + K_{Z^n}|}{|K_{Z^n}|} - \epsilon$ is achievable. But,

$\max_{K_{X^n}} \frac{1}{2n} \log \frac{|K_{X^n} + K_{Z^n}|}{|K_{Z^n}|}$ is the capacity of the Gaussian vector channel without interference. Since ϵ is arbitrary, and the capacity of the channel with interference cannot exceed the capacity of the channel without interference, C_n must be the capacity of the "colored-paper" channel. □

4 Conclusion

We have shown that Costa's surprising results can be extended to the Gaussian channel with colored interference and noise. This result is proved by constructing a code over a single block of finite length which has a probability of error that does not depend on the noise and interference covariance matrices. This coding theorem does not require blocks to repeat, and is therefore well-suited to deal with arbitrary interference and noise that may have memory or may be non-stationary or non-ergodic. The Gaussian nature of the noise and the interference suffices to provide the necessary information stability and consistency. Thus, when the interference is known non-causally at the transmitter, a channel with arbitrary Gaussian interference and arbitrary Gaussian noise has the same capacity as if interference does not exist.

References

[1] M. Costa, "Writing on dirty paper," *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[2] A. Cohen and A. Lapidoth, "The Gaussian watermarking game: Part I," *submitted to IEEE Trans. Inform. Theory*, 2001.

[3] U. Erez, S. Shamai, and R. Zamir, "Capacity and lattice strategies for cancelling known interference," in *Int. Symp. Inform. Theory and Its Appl.*, Nov. 2000.

[4] T. M. Cover and S. Pombra, "Gaussian feedback capacity," *IEEE Trans. Inform. Theory*, vol. 35, no. 1, pp. 37–43, Jan. 1989.

[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, 1991.

[6] S. I. Gel'fand and M.S. Pinsker, "Coding for channel with random parameters," *Problems of Control and Information Theory*, vol. 9, no. 1, pp. 19–31, 1980.

[7] C. Heegard and A. El Gamal, "On the capacity of computer memories with defects," *IEEE Trans. Inform. Theory*, vol. 29, pp. 731–739, Sep. 1983.

[8] A. Lapidoth, "private communications," Feb. 2001.