Channel Coherence in the Low SNR Regime *

Lizhong Zheng David N.C. Tse Muriel Médard

March 2, 2006

Abstract

Channel capacity in the limit of vanishing SNR per degree of freedom is known to be linear in SNR for fading and non-fading channels, regardless of channel state information at the receiver (CSIR). It is shown [1] that the significant engineering difference between the coherent and the non-coherent fading channels, including the requirement of peaky signaling and the resulting spectral efficiency, is determined by how the capacity limit is approached as SNR tends to zero, or in other words, the sub-linear term in the capacity expression. In this paper, we show that this sub-linear term is determined by the *channel coherence level*, which we define to quantify the relation between the SNR and the channel coherence time. This allows us to trace a continuum between the case with perfect CSIR and the case with no CSIR at all. Using this approach, we also evaluate the performance of suboptimal training schemes.

1 Introduction

The use of a large bandwidth to improve the power efficiency in wireless communications has been studied since the 1960s. Kennedy [2] showed that, for the Rayleigh fading channel at the infinite bandwidth limit, the amount of energy required to reliably transmit one information bit is $\frac{E_b}{N_0} = -1.59dB$, which is the same limit for the AWGN channels. Denote the signal-to-noise ratio per degree of freedom as SNR, and the corresponding capacity as $C(\mathsf{SNR})$(nats/s/Hz), this result is equivalent to

$$\lim_{\mathsf{SNR}\to 0} \frac{C_{fading}(\mathsf{SNR})}{\mathsf{SNR}} = \lim_{\mathsf{SNR}\to 0} \frac{C_{\mathrm{AWGN}}(\mathsf{SNR})}{\mathsf{SNR}} = 1 \tag{1}$$

This result is very robust. It holds regardless of whether the instantaneous channel state information (CSI) is available at the receiver or not (although it does assume

*Zheng and Médard (lizhong, medard@mit.edu) are with Massachusetts Institute of Technology, their research supported in part by an NSF grant on UWB communications; Tse (dtse@eecs.berkeley.edu) is with the University of California, Berkeley, his research is supported partially by the Army Research Office under a MURI program on Ultra Wideband communications, and by Qualcomm Inc. with matching support from the Cailfornia MICRO program.

the absence of CSI at the transmitter). It was later shown [3] that (1) also holds for general fading distributions.

On the other hand, it is also observed [4, 3, 5, 6] that there are some important differences between the coherent and the non-coherent fading channels. Without receiver CSI, the optimal signaling requires a very high peak-to-average ratio, and the resulting capacity approaches the infinite bandwidth limit much slower. Reference [4] provided numerical evidence to support this latter observation for the Rayleigh fading model.

The issue of rate of convergence is addressed by Verdú [1] quantitatively by looking at the second order Taylor expansion of the capacity expression:

$$C(\mathsf{SNR}) = C'(0)\mathsf{SNR} + C''(0)\mathsf{SNR}^2 + o(\mathsf{SNR}^2)$$

where $C'(0)$ and $C''(0)$ are the first and second derivatives of the function $\mathsf{SNR} \mapsto C(\mathsf{SNR})$ at $\mathsf{SNR} = 0$. It is shown that, for a general class of channels, the second derivative $C''(0)$ is finite in the coherent case, when CSI is available at the receiver, and is $-\infty$ in the non-coherent case. This means that although a near linear capacity can be achieved in both cases eventually at low enough SNR, this limit is approached much more slowly for the non-coherent case.

From a signal design point-of-view, the results in [1] can be understood as describing a fundamental tension in non-coherent communications. On the one hand, communicating over an unknown channel is subject to a penalty of the channel uncertainty (sometimes in the form of the training costs); on the other hand, reducing such penalty by sending over only a fraction of the available degrees of freedom (peaky signaling) results in a loss of spectral efficiency. The tradeoff between resolving the channel uncertainty and improving the spectral efficiency gives rise to optimal non-coherent signaling, which can be solved for some special asymptotic cases [1, 7].

The difference between the coherent and the non-coherent channels is even more significant when there is a constraint on the peak-to-average ratio of the input signaling. For the coherent case, the optimal input has a Gaussian distribution, i.i.d. over symbols. Thus, an additional constraint on the peakiness of the input usually does not reduce the throughput much. In contrast, the optimal input for the non-coherent channel is increasingly "flashy" at low SNR. Such input is ruled out when a peakiness constraint presents. It is shown [5, 6] that the capacity of the non-coherent channel at low SNR with the peakiness constraint scales as SNR^2, which is much less than the coherent capacity, which is linear in SNR. Thus the impact of the peakiness constraint is much more severe in the non-coherent case than in the coherent case.

The goal of the current paper is to develop a unified view of these results. To do that, we observe that the above notions of the "coherent" and the "non-coherent" channels can be viewed as two extreme cases. While the coherent model assumes perfect CSI, the non-coherent model not only assumes no CSI available but implicitly ignores the correlation between the fading coefficients over time, thus eliminating the possibility to estimate the channel. In practice, however, the channel coherence time might span hundreds or thousands of data symbols. In such cases, even though the fading coefficients are unknown to the receiver at the beginning, it is possible to estimate these coefficients and communicate coherently thereafter; hence a performance close to that of the coherent case can be achieved. Thus, depending on the coherence time, there is a *continuum* between the coherent and the non-coherent extremes.

Moreover, the SNR per degree of freedom is an important component in determining the effect of channel coherence on capacity. If one views the channel coherence as the reward for estimating the channel, since longer coherence time means taking more advantage of the channel estimate, then the SNR indicates the price of such estimation. When there is limited amount signal energy, any meaningful channel estimate would require a large fraction of the total energy resource, thus one tends to estimate and use the channel less often. This effect is quantitatively described in [5, 6, 8].

In short, the tradeoff between resolving the channel uncertainty and improving the spectral efficiency is affected by both the coherence time and the SNR. To quantify this effect, a natural question is therefore: how slowly the channel has to change over time, with respect to SNR, so that the capacity for a non-coherent channel begins to resemble that for the coherent channel. The main result of this paper is a clear characterization of the continuum between the non-coherent and the coherent extremes as the coherence time increases, in terms of capacity.

Specifically, we study the flat fading channel, where the fading coefficient is assumed to be Rayleigh distributed, and remains constant over a coherence time of l symbol periods, before changing into an independent realization. We compute the capacity of the block fading channels without receiver side CSI, as a function of the average signal-to-noise ratio SNR, the coherence time, and the peak power constraint. We focus on the asymptotic case that SNR tends to zero and the coherence time l is large. Let $\mathsf{E}_{\mathsf{max}}$ denote the maximum allowed signal power transmitted over a symbol period[1], normalized by the noise variance, and C denote the capacity in nats per symbol period. Our main results can be summarized as:

$$C \approx \begin{cases} \mathsf{SNR} - \sqrt{\frac{1}{l}} \cdot \mathsf{SNR} & \mathsf{E}_{\mathsf{max}} \geq \sqrt{\frac{1}{l}} \\ \mathsf{SNR} - \frac{\mathsf{SNR}}{l\mathsf{E}_{\mathsf{max}}} & \frac{1}{l} \leq \mathsf{E}_{\mathsf{max}} \leq \sqrt{\frac{1}{l}} \\ l \cdot \mathsf{E}_{\mathsf{max}} \cdot \mathsf{SNR} & \mathsf{E}_{\mathsf{max}} \leq \frac{1}{l} \end{cases} \tag{2}$$

where the meaning of $\approx$ will be made clear later in the paper. The optimal signaling for these three cases are as follows:

- When $\mathsf{E}_{\mathsf{max}} \geq \sqrt{\frac{1}{l}}$, the optimal input is the i.i.d Gaussian random code, transmitted over $\delta = \sqrt{l}\mathsf{SNR}$ fraction of the time. The peak power constraint is not active. In particular, if l is of the order larger than or equal to $O(\mathsf{SNR}^{-2})$, the capacity is close to that of the perfectly coherent channel; in contrast, when l grows very slowly with SNR, we get close to the results reported in [1].
- When $\frac{1}{l} \leq \mathsf{E}_{\mathsf{max}} \leq \sqrt{\frac{1}{l}}$, the optimal input is also the i.i.d. Gaussian codes, but transmitted over $\delta = \mathsf{SNR}/\mathsf{E}_{\mathsf{max}}$ fraction of the time. The peak power constraint is active.
- When $\mathsf{E}_{\mathsf{max}} \leq \frac{1}{l}$, the optimal input is on-off signaling. In particular, if the peak-to-average ratio $\mathsf{E}_{\mathsf{max}}/\mathsf{SNR}$ and the coherence time l are both fixed, the capacity scales as $O(\mathsf{SNR}^2)$ as SNR approaches 0, consistently with [5].

[1]The "peak power constraint" considered in this paper is slightly different from . Here, we are interested in how the peak signal power increases as $\mathsf{SNR} \to 0$. The constraint $\mathsf{E}_{\mathsf{max}}$ limits the use of flash signals, where the signal energy is more and more concentrated in a even smaller fraction of d.o.f., as SNR tends to 0. We do, however, consider a Gaussian random code with average power less than $\mathsf{E}_{\mathsf{max}}$ to be valid input: although the peak power of a Gaussian code is higher than the average power, the difference does not depend on the SNR of the channel. Detailed mathematical justification can be found section 5.2

These three cases form a continuum that glue together several known results for different extreme cases. The techniques developed here can also be used in analyzing the performance of specific signaling such as training based schemes. Using this approach, we quantify the performance loss of training schemes due to cost of energy that is allocated for training purpose. In contrast, the optimal signaling scheme can be understood as using all signal energy for both estimating the channel and carrying information, thus avoiding the extra energy cost. Our results can also be generalized to the cases with other models of the channel time-variation, such as the Gauss-Markov model. The interactions among the bandwidth, channel coherence, and the signal peakiness discussed in this paper is thus shown to be a general issue for communication over wideband fading channels.

The rest of this paper is organized as follows, after describing the channel model in section 2, we continue in section 3 to review some known results of the low-SNR capacity of the coherent and non-coherent fading channels, and motivate our approach and the specific asymptotic scaling we choose. In section 4, we study the performance of training schemes for block fading channel, providing contrast to the main capacity results, presented in section 5. We discuss the Gauss-Markov model in section 6 before concluding the paper in section 7.

2 Channel Model: from Wideband to Narrowband

We consider a communication system over a wideband fading channel. The wideband channel can be modelled as a set of N parallel narrowband channels, such that each channel is flat faded. The n^{th} such channel is

$$y_i^{(n)} = \sqrt{\mathsf{SNR}} \cdot h_i^{(n)} x_i^{(n)} + w_i^{(n)}$$

where $h_i^{(n)}, x_i^{(n)}$, and $w_i^{(n)}$ are the fading coefficient, transmitted symbol, and the additive noise, respectively at symbol time i for the n^{th} channel. We assume that $h_i^{(n)}, w_i^{(n)}$ are complex Gaussian distributed with zero mean and unit variance, denoted as $h_i^{(n)}, w_i^{(n)} \sim \mathcal{CN}(0,1)$. The parameter SNR denotes the signal-to-noise ratio per symbol time per narrowband channel.

In general, the fading coefficients, $\mathbf{h}_i^{(n)}$, are correlated over time and frequency. We are thus interested in computing the capacity for the narrowband fading channel with arbitrarily correlated fading coefficients. For simplicity, we drop the superscript (n), and start our discussion by focusing on the block fading model. That is, the fading coefficient is assumed to remain constant for a block of l symbol periods, before taking independent realizations. The channel over one of such block can be written as

$$\mathbf{y} = \sqrt{\mathsf{SNR}} \cdot h\mathbf{x} + \mathbf{w} \tag{3}$$

where $\mathbf{x}, \mathbf{w}, \mathbf{y} \in \mathcal{C}^l$; $h \sim \mathcal{CN}(0,1)$ is the scalar fading coefficient associated with the block. The block length l indicates the channel coherence over both time and frequency.

We assume an average power constraint: for any codeword transmitted over K coherence blocks,

$$\frac{1}{Kl} \sum_{i=1}^{Kl} {\mathbf{x}_i}^2 \leq 1$$

For the moment, we assume that there is no peak power constraint for the transmitted signals, and leave the discussion about peak power constraint in Section 5.2.

For wideband communications, as $N \to \infty$, the SNR per degree of freedom $\mathsf{SNR} \to 0$, and hence our interest is in the low SNR regime. The cases with/without receiver side CSI is referred to as the coherent/non-coherent cases, and we always assume that transmitter side CSI is not available.

3 The Coherent and Non-Coherent Extremes

3.1 Non-coherent Penalty

In general, the capacity of a channel, including Gaussian channels and fading channels under the coherent and non-coherent assumptions, increases with the signal-to-noise ratio per channel use in a sub-linear fashion. It is only at the low SNR limit, the channel capacity can increase with SNR linearly.

$$C(\mathsf{SNR}) = \mathsf{SNR} + o(\mathsf{SNR}) \quad \text{(nats/channel use)} \tag{4}$$

This makes low SNR channels desirable in energy efficient communications.

In many cases, $C(\mathsf{SNR})$ is close to linear at low SNR; thus the difference among these channels, including the peakiness of the signaling, and the spectral efficiency, etc., can only be explained by the difference in the sub-linear term $o(\mathsf{SNR})$ in (4). To emphasize the sub-linear term of the capacity, we use the following notation throughout this paper:

$$\Delta(\mathsf{SNR}) := \mathsf{SNR} - C(\mathsf{SNR})$$

Clearly, at the low SNR limit, if

$$\frac{\Delta(\mathsf{SNR})}{\mathsf{SNR}} \to 0 \tag{5}$$

the highest energy efficiency of $-1.59(dB)$ per information bit can be achieved. Following the definition in [1], (5) is equivalent to first order optimality. As the SNR increases, the gap $\Delta(\mathsf{SNR})$ increases, yielding a larger gap between the capacity $C(\mathsf{SNR})$ and its linear approximation SNR, and the energy efficiency decreases. In the following, we start with some examples to understand the operational meaning of this sub-linear term.

Example: AWGN Channel

For the AWGN channel, the capacity is

$$C_{\text{AWGN}}(\mathsf{SNR}) = \log(1 + \mathsf{SNR}) = \mathsf{SNR} - \frac{1}{2}\mathsf{SNR}^2 + o(\mathsf{SNR}^2) \tag{6}$$

the sub-linear term is

$$\Delta_{\text{AWGN}}(\mathsf{SNR}) = \frac{1}{2}\mathsf{SNR}^2 + o(\mathsf{SNR}^2).$$

$\circ$

Example: Coherent fading channel (perfect CSI)
For fading channels, when the CSI is available at the receiver side, the optimal input distribution is i.i.d. Gaussian. Assume that the fading coefficients h_i are generated from an ergodic process, with marginal distribution $h \sim \mathcal{CN}(0,1)$, then the ergodic capacity is

$$\begin{aligned} C_{coherent}(\mathsf{SNR}) &= E[\log(1+\mathsf{SNR}\,\|h\|^2)] \\ &= \mathsf{SNR} - \frac{1}{2}E[\|h\|^4](\mathsf{SNR})^2 + o(\mathsf{SNR}^2). \end{aligned}$$

The corresponding sub-linear term is thus

$$\Delta_{coherent}(\mathsf{SNR}) = \frac{1}{2}E[\|h\|^4](\mathsf{SNR})^2 + o(\mathsf{SNR}^2) \tag{7}$$

$\circ$

Example: Non-coherent i.i.d. fading channel
For fading channels, when the CSI is not available at the receiver side, the sub-linearity is larger than that for the AWGN channel. It is shown in [1] that for an i.i.d. Rayleigh fading channel, where h_i's are independently $\mathcal{CN}(0,1)$ distributed, when receiver CSI is not available,

$$\Delta_{iid}(\mathsf{SNR}) = \mathsf{SNR} - C_{iid}(\mathsf{SNR}) \gg O(\mathsf{SNR}^2)$$

where $\gg$ means the ratio between the two sides approaches ∞ as $\mathsf{SNR} \to 0$. This implies a much larger gap between the capacity $C_{iid}(\mathsf{SNR})$ and SNR, at a strictly positive signal-to-noise ratio. In other words, in order to obtain a high energy efficiency in the non-coherent fading channel, it requires the SNR to be very low, or equivalently the bandwidth to be much larger. $\circ$

The i.i.d. non-coherent model can be viewed as the opposite extreme to the perfectly coherent model, since the channel changes so fast that there is no hope to have any knowledge of the future channel state at all. In this paper, we are interested in the general non-coherent cases, where the channel state information is not available at the receiver, but can be obtained partially using channel coherence over time. To quantify the performance of such channels, we will focus on computing $\Delta(\mathsf{SNR})$ for such channels. Specializing to the block fading model, we will compute $\Delta_l(\mathsf{SNR}) \triangleq \mathsf{SNR} - C_l(\mathsf{SNR})$, where the subscript indicates the dependence on the channel coherence time l. Obviously, the i.i.d. fading model corresponds to $l = 1$. We are interested in finding out how does the sublinear term change from $\Delta_{iid}(\mathsf{SNR}) \gg O(\mathsf{SNR}^2)$ to $\Delta_{coherent}(\mathsf{SNR}) = O(\mathsf{SNR}^2)$ as l increases.

In most cases of interests, $\Delta(\mathsf{SNR})$ is much larger than $O(\mathsf{SNR}^2)$. The capacity loss due to the lack of channel knowledge,

$$\begin{aligned} C_{coherent}(\mathsf{SNR}) - C(\mathsf{SNR}) &= (\mathsf{SNR} - C(\mathsf{SNR})) - (\mathsf{SNR} - C_{coherent}(\mathsf{SNR})) \\ &= \Delta(\mathsf{SNR}) + O(\mathsf{SNR}^2) \approx \Delta(\mathsf{SNR}). \end{aligned}$$

Therefore, we also refer to $\Delta(\mathsf{SNR})$ for the channels of interest as the *non-coherent penalty* of the channel.

3.2 The I.I.D. Rayleigh Fading Channel

In this section, we start by giving a more precise characterization of the capacity and sub-linearity for the non-coherent fading channel (3), with block length $l = 1$. The capacity of this channel has been partially characterized in [1]. It is shown that

$$\mathsf{SNR} - C_{iid}(\mathsf{SNR}) \gg O(\mathsf{SNR}^2),$$

or equivalently,

$$\frac{\Delta_{iid}(\mathsf{SNR})}{\mathsf{SNR}^2} \to \infty.$$

This result is sufficient to provide a clear distinction between the coherent and the non-coherent channels: $\Delta_{iid}(\mathsf{SNR}) \gg \Delta_{coherent}(\mathsf{SNR}) = O(\mathsf{SNR}^2)$. It is however of interest to find out more precisely what the "right" order of the sub-linear term is.

Proposition 1 *For the i.i.d Rayleigh fading without receiver side CSI, the channel capacity $C_{iid}(\mathsf{SNR})$ at low SNR is bounded by*

$$\mathsf{SNR} - \mathsf{SNR}\frac{\log\log\frac{1}{\mathsf{SNR}}}{\log\frac{1}{\mathsf{SNR}}} \leq C_{iid}(\mathsf{SNR}) \leq \mathsf{SNR} - \mathsf{SNR}\frac{\left(\log\log\frac{1}{\mathsf{SNR}}\right)^2 + 1}{\log\frac{1}{\mathsf{SNR}}}.$$

Proof:

We just give a sketch of the proof; a complete proof is given in Appendix A.

At low SNR, the optimal input distribution is i.i.d. on-off signaling, with

$$\mathsf{SNR} \cdot |x|^2 = \begin{cases} A & \text{with probability } \delta = \frac{\mathsf{SNR}}{A} \\ 0 & \text{with probability } 1 - \delta \end{cases}.$$

Under this input, the mutual information can be written as

$$\begin{aligned} I(x;y) &= I(x;y,h) - I(x;h|y) \\ &= \log(1+\mathsf{SNR}) - D(y||y_G) - \delta\log(1+A) \end{aligned}$$

where $D(\cdot||\cdot)$ is the Kullback Leibler divergence, and y_G is a Gaussian random variable with zero mean and variance $1 + \mathsf{SNR}$.

Intuitively, to maximize the mutual information, one would like to choose peaky inputs with larger A, so that the term $\delta\log(1 + A)$, which can be interpreted as the penalty of channel uncertainty, is minimized. On the other hand, peaky input makes the distribution of y more different from Gaussian, i.e., increases $D(y||y_G)$. The optimal choice of A that balances these two trends is given by

$$\frac{\log\frac{1}{\mathsf{SNR}}}{\log\log\frac{1}{\mathsf{SNR}}} \leq A^* \leq \log\frac{1}{\mathsf{SNR}}. \tag{8}$$

The detailed computation is given in Appendix A.

∘

Ignoring the log log term, the result in Proposition 1 gives an approximation

$$\Delta_{iid}(\mathsf{SNR}) \quad = \quad \mathsf{SNR} - C_{iid}(\mathsf{SNR}) \approx \frac{\mathsf{SNR}}{\log \frac{1}{\mathsf{SNR}}}.$$

This means that the sub-linear term is much larger than SNR^2 as $\mathsf{SNR} \to 0$, which is consistent with [1]. In fact, the statement in Proposition 1 is stronger. It says that the sub-linear term of the capacity is in fact much larger than $O(\mathsf{SNR}^{1+\alpha})$ for any $\alpha \in (0,1)$, or in other words, it is almost linear in SNR. Consequently, in the capacity expression $C_{iid}(\mathsf{SNR}) = \mathsf{SNR} - \Delta_{iid}(\mathsf{SNR})$, the slope of the leading linear SNR term is almost changed, implying that the -1.59 dB wideband limit is almost not achievable, it is approached very slowly.

To capture this intuition. we introduce a new notation for approximations. We write

$$f(\mathsf{SNR}) \doteq \mathsf{SNR}^{\alpha} \tag{9}$$

if

$$\lim_{\mathsf{SNR} \to 0} \frac{\log f(\mathsf{SNR})}{\log \mathsf{SNR}} = \alpha.$$

Suppose that $f(\mathsf{SNR}) \doteq \mathsf{SNR}^{\alpha}$ and $g(\mathsf{SNR}) \doteq \mathsf{SNR}^{\beta}$, we write $f(\mathsf{SNR}) \doteq g(\mathsf{SNR})$ if $\alpha = \beta$, and say that f and g are of the same order. Similarly, $\dot{\leq}$ and $\dot{\geq}$ are used to denote that f and g are of different order. We write $f(\mathsf{SNR}) \dot{\leq} g(\mathsf{SNR})$ if $\alpha > \beta$, and vice versa. This notation is simply a variation of exponential approximation used in information theory. In particular, it allows us to ignore the "log" term multiplied on a polynomial term. With this notation, we write

$$\Delta_{iid}(\mathsf{SNR}) \doteq \mathsf{SNR}$$

which is very different from that for the coherence channel $\Delta_{coherent}(\mathsf{SNR}) \doteq \mathsf{SNR}^2$.

3.3 Spectral Efficiency and Energy Efficiency

In the low SNR regime, the non-coherent penalty $\Delta(\mathsf{SNR})$ also has the following simple relation with the energy efficiency. Let E_n denote the transmitted energy per information nat, and N_0 the spectral level of the background noise. We have

$$\left(\frac{E_n}{N_0}\right)_{(J/nat)} \times C(\mathsf{SNR})_{(nat)} = \mathsf{SNR}_{(J)}$$

where the subscripts are the units of each term. Replacing $C(\mathsf{SNR})$ by $\mathsf{SNR} - \Delta(\mathsf{SNR})$, we have

$$\frac{E_n}{N_0} = \frac{1}{1 - \frac{\Delta(\mathsf{SNR})}{\mathsf{SNR}}}.$$

Taking natural log of both sides, we have

$$\log \frac{E_n}{N_0} \approx \frac{\Delta(\mathsf{SNR})}{\mathsf{SNR}} \tag{10}$$

Thus, the energy efficiency is directly connected to the non-coherent penalty, as well as the percentage loss of the capacity. In Figure 1, we plot the capacity loss $\Delta(\mathsf{SNR})/\mathsf{SNR}$ vs. SNR for the Gaussian channel and the coherent/non-coherent fading channels. Notice that in the low SNR regime, the channel capacity is approximately SNR, thus the resulting graph is a good approximation, as well as a new interpretation, of the "spectral efficiency vs. E_b/N_0(dB)" graph in Figure 4 of [1].

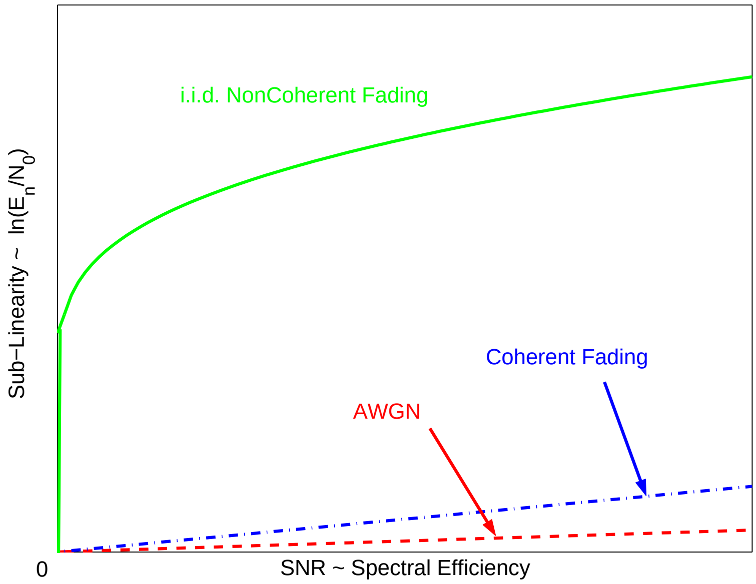

Figure 1: The Sub-linearity of Gaussian and Fading channels at Low SNR.

Figure 1 can also be interpreted as the tradeoff between the energy efficiency and the spectral efficiency of a wideband system. With a fixed total transmitted signal power, one can choose to lower the SNR per degree of freedom, to improve the energy efficiency, at the cost of requiring larger bandwidth, or vice versa. Thus in rest of this paper, we will focus on characterizing $\Delta(\mathsf{SNR})$, in order to capture this fundamental tradeoff.

Discussion:

At a low but strictly positive SNR, the effect of the sub-linearity on the channel capacity is rather marginal. Even in the i.i.d. non-coherent channel, the capacity $C_{iid}(\mathsf{SNR})$ is quite close to that of the AWGN channel at the same SNR level. For example, at $\mathsf{SNR} = -30$dB, the percentage loss of the capacity due to the lack of the channel knowledge is approximately

$$\frac{\Delta_{iid}(\mathsf{SNR})}{\mathsf{SNR}} = \frac{1}{\log \frac{1}{\mathsf{SNR}}} \approx 14.5\%$$

As $\mathsf{SNR} \longrightarrow 0$, this percentage loss approaches 0. In terms of the energy per information nat at a given SNR, the difference between the non-coherent fading channel and the Gaussian channel is also quite small. Again take $\mathsf{SNR} = -30$dB, by (10), we have

$$\frac{E_n}{N_0} \approx \exp\left(\frac{\Delta(\mathsf{SNR})}{\mathsf{SNR}}\right)$$

For the i.i.d. fading channel, $\Delta_{iid}/\mathsf{SNR} \approx \frac{1}{\log \frac{1}{\mathsf{SNR}}} \approx 0.145$, and $E_n/N_0 \approx 1.156$; for the AWGN channel, $\Delta_{\text{AWGN}}/\mathsf{SNR} = \frac{1}{2}\mathsf{SNR} = 5 \times 10^{-4}$, and $E_n/N_0 \approx 1.001$. The difference between the two is still only about 15%.

Such comparisons show that since the capacities of both the Gaussian channel and the fading channel are close to linear in SNR, the effect from $\Delta(\mathsf{SNR})$ is secondary, at least from a capacity point-of-view. However, we emphasize here that this small difference in the channel capacity is indeed the reason of many dramatic differences, signaling and the spectral efficiency at a finite bandwidth, between the coherent and the non-coherent channels.

For example, in order to achieve a certain level of energy efficiency, the required SNR and thus the bandwidth for the AWGN channel and the non-coherent fading channel can be dramatically different. In the above example, if we want to achieve an energy efficiency of $E_n/N_0 = 1.156$ in an AWGN channel, the signal to noise ratio should be $\mathsf{SNR} = 2 \cdot (\Delta/\mathsf{SNR}) = 2\log(E_n/N_0) = 0.29 \approx -5dB$, instead of $-30dB$ for the fading channel. This translates to a factor of more than 300 times in the required bandwidth. Thus the study of the sub-linear term of the channel capacity is more important when there is a hard constraint on the energy efficiency.Intuitively, since the energy efficiency $\log(E_n/N_0) \to 0$ as the bandwidth $W \to \infty$, the energy efficiency of the non-coherent channel at low SNR is only slightly worse than that of the coherent channel; however, it takes a much larger bandwidth to make up for this difference. ◦

4 Training schemes for the Block fading model

In this section and the next, we will focus on the block fading channel model. Before computing the channel capacity, we will first study a sub-optimal approach to use the block fading channel, by using training schemes, and establish the natural scaling to address the problem. Training schemes are widely used in communicating over fading channels when the fading coefficients are unknown to the receiver. At the beginning of each coherence block, a training sequence, known to the receiver, is transmitted to help the receiver to estimate the channel coefficients, and then these estimates are used to communicate during the rest of the coherence block. We follow the approach used in [7] and [9] to compute a lower bound of the achievable throughput, by optimizing the amount of energy used in training.

We start by describing the training scheme in details. We rewrite the block fading channel model, within one coherence block, as follows

$$y_i = \sqrt{\mathsf{SNR}} \cdot hx_i + w_i, \quad i = 1, \ldots, l$$

where the fading coefficient h is assumed to remain constant within a block of l symbols. For convenience, we refer to $\sqrt{\mathsf{SNR}} \cdot x_i$ as the transmitted signal, which has an average energy per symbol time of SNR.

We denote the total energy transmitted within one coherence block as

$$E_{total} = l \cdot \mathsf{SNR}.$$

At the beginning of a block, we use γ fraction of the total energy in training. For convenience, denote the energy used in training as

$$E_{tr} = \gamma E_{total} = \gamma l \cdot \mathsf{SNR}.$$

The receiver computes the minimum mean square estimate of the fading coefficient h. Since the quality of this estimate depends only on the energy, rather than the duration, of the training signal, we assume in the following that the training signal is transmitted within 1 symbol period, i.e., the signal transmitted in the first symbol of a block is

$$\sqrt{\mathsf{SNR}} \cdot x_1 = \sqrt{E_{tr}}$$

and the received signal is

$$y_1 = \sqrt{E_{tr}} h + w_1.$$

Using $\hat{h}$ and $\tilde{h}$ to denote the minimum mean square estimate of h and the estimation error, respectively, we have

$$\begin{aligned} E[|\hat{h}|^2] &= \frac{E_{tr}}{1 + E_{tr}} \\ E[|\tilde{h}|^2] &= \frac{1}{1 + E_{tr}}. \end{aligned}$$

For the remaining $(l-1)$ symbols within the same block, we communicate by using an i.i.d. Gaussian random code with average power of $(1-\gamma)\mathsf{SNR}$. The channel in this communication phase can be written as

$$\begin{aligned} y_i &= \sqrt{(1-\gamma)\mathsf{SNR}} \cdot \hat{h} x_i + \sqrt{(1-\gamma)\mathsf{SNR}} \cdot \tilde{h} x_i + w_i \\ &= \sqrt{(1-\gamma)\mathsf{SNR}} \cdot \hat{h} x_i + w_i' \end{aligned} \tag{11}$$

for $i = 2, \ldots, l$, where x_i is normalized to have $E[|x_i|^2] = 1$. Notice that the second term in (11), the extra noise due to the channel estimation error, is uncorrelated with the signal term, $\hat{h} x_i$. Thus, to obtain a lower bound of the mutual information, we can replace it with the additive Gaussian noise with the same power [10]. The overall noise w_i' is replaced by a Gaussian noise with variance

$$\begin{aligned} \sigma^2 &= 1 + (1-\gamma) E[|\tilde{h}^2] \mathsf{SNR} \\ &= 1 + (1-\gamma) \frac{1}{1 + E_{tr}} \mathsf{SNR}. \end{aligned}$$

The resulting mutual information per symbol time is lower bounded by

$$I_{tr}(\mathsf{SNR}) \geq \frac{l-1}{l} E\left[\log\left(1 + \frac{(1-\gamma)\mathsf{SNR}|\hat{h}|^2}{\sigma^2}\right)\right], \tag{12}$$

where the factor $(l-1)/l$ is due to the fact that 1 symbol time out of a block is used in training. Now we optimize over the power allocation γ, and compute the resulting throughput for this training scheme.

The question we try to answer here is that how long the coherence time l has to be in order that the throughput of a fading channel, without CSI at the receiver, to start look like that of a coherent channel. Our focus is on the sub-linear term in the throughput

$$\Delta_{tr}(\mathsf{SNR}) \stackrel{\Delta}{=} \mathsf{SNR} - I_{tr}(\mathsf{SNR}).$$

Recall that the two reference points from the extreme cases are: $\Delta_{coherent} \doteq \mathsf{SNR}^2$ and $\Delta_{iid} \doteq \mathsf{SNR}^1$. Thus a natural question to ask is: how long l has to be in order that $\Delta_{tr}(\mathsf{SNR}) \doteq \mathsf{SNR}^{1+\alpha}$ for a given constant $\alpha \in (0,1)$. The following lemma gives our first result for the training scheme.

Lemma 2 *For the training scheme described above achieves a data rate $I_{tr}(\mathsf{SNR})$ with*

$$\Delta_{tr}(\mathsf{SNR}) = \mathsf{SNR} - I_{tr}(\mathsf{SNR}) \dot{\leq} \mathsf{SNR}^{1+\alpha}$$

for $\alpha \in [0,1]$, the sufficient and necessary condition is

$$l \dot{\geq} \mathsf{SNR}^{-(1+2\alpha)}.$$

Proof:
See Appendix C ◦

Remarks:
In general, the channel capacity is a function of both the coherence time l and the SNR. Although there is no physical connection between these two parameters, the channel capacity and the optimum signaling depend on the relation between the two, rather than on each of them in isolation. The simultaneous scaling of the two parameters in the Lemma 2 allows the characterization of the throughput in terms of how they are related.

A special case of interests in Lemma 2 is the case $\alpha = 1$. That is, if we wish to achieve a throughput whose sub-linear term, and hence the energy vs. spectrum efficiency tradeoff, is of the same order as that of the coherent channel,

$$I_{tr}(\mathsf{SNR}) = \mathsf{SNR} - O(\mathsf{SNR}^2)$$

it is sufficient and necessary that $l \doteq \mathsf{SNR}^{-3}$. We say in this case that a "near coherent" performance is achieved. ◦

The training based scheme proposed above transmits signals in every coherence block. Intuitively, this is optimal if one operates near the coherent extreme, with $\alpha = 1$. Therefore, the requirement of $l \doteq \mathsf{SNR}^{-3}$ to achieve $\Delta_{tr}(\mathsf{SNR}) \doteq \mathsf{SNR}^2$ is necessary. For intermediate values of $\alpha \in (0,1)$, however, the performance of the training schemes can be improved by allowing flashiness in the transmission. That is, we can train and communicate only in δ fraction of the available coherence blocks, thereby concentrate the energy and avoid estimating too many fading coefficients. By optimizing over the peakiness δ, one can reduce the gap between the training performance and the optimal. The result is summarized in the following lemma.

Lemma 3 *For a block fading channel with coherence time $l \doteq \mathsf{SNR}^{-3\alpha}$, using a flash training scheme with $\delta \doteq \mathsf{SNR}^{(1-\alpha)}$, one can achieve a data rate of the order $\mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})$*

Proof:
Write the throughput of the described flashy training scheme as

$$I_{tr}(\mathsf{SNR}, \delta) \quad = \quad \delta I_{tr}\left(\frac{\mathsf{SNR}}{\delta}\right)$$

where $I_{tr}(\mathsf{SNR})$ is the throughput for a non-flashy training scheme with average power per symbol time as

$$\mathsf{SNR}' = \frac{\mathsf{SNR}}{\delta}.$$

Now, for $\delta \doteq \mathsf{SNR}^{(1-\alpha)}$, we have

$$\mathsf{SNR}' \doteq \mathsf{SNR}^{\alpha}.$$

Lemma 2 says that, if the coherence time

$$l \doteq (\mathsf{SNR}')^{-3} = \mathsf{SNR}^{-3\alpha},$$

we have

$$I_{tr}(\mathsf{SNR}') \doteq \mathsf{SNR}' - O((\mathsf{SNR}')^2) = \mathsf{SNR}^{\alpha} - O(\mathsf{SNR}^{2\alpha})$$

and thus

$$\begin{aligned} I_{tr}(\mathsf{SNR}, \delta) &= \delta I_{tr}(\mathsf{SNR}') \\ &= \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha}) \end{aligned}$$

is achievable. ○

Comparing Lemma 2 and 3, we observe that the difference between the performance of the two schemes can be clearly quantified by the SNR exponent of the sub-linear term. It is however not clear, even at the given scaling, whether the performance of the training scheme in Lemma 3 is optimal or not. To gain more insights to the requirements of the coherence time l, we take a closer look at the special case with $\alpha = 1$.

Discussion: Near Coherent Performance
Intuitively, if we want a training scheme for a non-coherent channel to have a throughput close to the capacity of the perfect coherent channel, it is necessary that both the following conditions be satisfied:

- The energy used in training, E_{tr}, is large enough such that the channel estimation error is ignorable.
- The fraction of energy used in training, γ, is small enough such that its effect is ignorable.

In the scaling of interest, these two conditions can be quantitatively specified.

For convenience, we rewrite the RHS of (12) [2]

[2] Strictly speaking, $R_{tr}(\mathsf{SNR})$ is a lower bound of the achievable rate for the training schemes. However, in order to achieve a rate higher than $R_{tr}(\mathsf{SNR})$, a receiver that can take the advantage of the estimation error term, $\tilde{h}\mathbf{x}$, is required, which is usually difficult. Therefore, we take $R_{tr}(\mathsf{SNR})$ as a "practical" approximation of $I_{tr}(\mathsf{SNR})$.

$$R_{tr}(\mathsf{SNR}) = \frac{l-1}{l} E\left[\log\left(1+\frac{(1-\gamma)\mathsf{SNR}|\hat{h}|^2}{\sigma^2}\right)\right].$$

Notice that $\sigma^2 \geq 1$. By using Jensen's inequality, we obtain a simple upper bound for $R_{tr}(\mathsf{SNR})$:

$$\begin{aligned} R_{tr}(\mathsf{SNR}) &\leq \log(1+\mathsf{SNR}E[|\hat{h}|^2]) \\ &\leq \mathsf{SNR}\frac{E_{tr}}{1+E_{tr}} \\ &= \mathsf{SNR} - \mathsf{SNR}\frac{1}{1+E_{tr}}. \end{aligned}$$

For $\mathsf{SNR} - R_{tr}(\mathsf{SNR}) \doteq \mathsf{SNR}^2$, the training energy must be as large as:

$$E_{tr} \dot{\geq} \mathsf{SNR}^{-1}. \tag{13}$$

Noting that $E[|\hat{h}|^2] \leq 1$, we can find another upper bound to $R_{tr}(\mathsf{SNR})$,

$$\begin{aligned} R_{tr}(\mathsf{SNR}) &\leq \log(1+(1-\gamma)\mathsf{SNR}) && (14) \\ &\leq \mathsf{SNR} - \gamma\mathsf{SNR}. && (15) \end{aligned}$$

In order for $\mathsf{SNR} - R_{tr}(\mathsf{SNR}) \doteq \mathsf{SNR}^2$, the fraction of energy used in training must be as small as

$$\gamma \dot{\leq} \mathsf{SNR}. \tag{16}$$

Combining (13) and (16), and recalling that $E_{tr} = \gamma E_{total} = \gamma l \mathsf{SNR}$, we have

$$l \geq \mathsf{SNR}^{-3}$$

as a necessary condition that the near coherent throughput to be achieved.

From the above development, we observe that if one can *reuse* the training energy for communication purpose, i.e., eliminating the second condition above, the requirement to achieve the near coherent performance, $l \doteq \mathsf{SNR}^{-3}$, can be improved to $l \doteq \mathsf{SNR}^{-2}$. It turns out that this intuition indeed gives the optimal performance, which is shown in the next section.

5 Capacity of Block Fading Channels

5.1 Channel Capacity Without Peakiness Constraint

In this section, we compute the capacity of the block fading channel without any constraint on the peakiness of the transmitted signal. The following Theorem summarizes our main results on the interplay between coherence level of the block fading channel and the non-coherent penalty.

Theorem 4 *Consider a sequence of non-coherent block Rayleigh fading channels, indexed by the average signal to noise ratio* SNR*. Let the block length and the capacity of the channel corresponding to a particular value of* SNR *be* $l(\mathsf{SNR})$ *and* $C_l(\mathsf{SNR})$*, respectively. The capacity has a non-coherent penalty*

$$\Delta_l(\mathsf{SNR}) \dot{\leq} \mathsf{SNR}^{1+\alpha}$$

if and only if

$$l \dot{\geq} \mathsf{SNR}^{-2\alpha}$$

or equivalently,

$$\lim_{\mathsf{SNR}\to 0} \frac{\log(\mathsf{SNR} - C_l(\mathsf{SNR}))}{\log \mathsf{SNR}} \geq 1 + \alpha \qquad \Longleftrightarrow \qquad \lim_{\mathsf{SNR}\to 0} \frac{\log l}{\log \mathsf{SNR}} \leq -2\alpha.$$

Remarks:

Here, we focus on the cases that $\alpha \in [0, 1]$, i.e., the cases ranging from l being fixed and $l \doteq \mathsf{SNR}^{-2}$. For the case that the coherence time l is even larger than SNR^{-2}, the effect of l becomes less important. In [11], the authors studied the case with a fixed SNR and $l \to \infty$. Under this assumption, it is shown that the capacity

$$\mathcal{C}_l(\mathsf{SNR}) \quad \approx \quad C_{coherent}(\mathsf{SNR}) - O\left(\sqrt{\frac{\log l}{l}}\right) = \mathsf{SNR} - O(\mathsf{SNR}^2) - O\left(\sqrt{\frac{\log l}{l}}\right).$$

Note that this result is for general SNR values. Specializing in the low SNR case, however, it addresses a different regime than Theorem 4. With a fixed SNR and l tends to infinity, the sub-linear term, $\mathsf{SNR} - C_l(\mathsf{SNR})$, is dominated by $O(\mathsf{SNR}^2)$ from the coherent capacity. We say in this case that a "near coherent" throughput is already achieved, and the effect of changing l on the resulting energy efficiency is insignificant. $\circ$

Proof: Achievability

We first show that if $l \doteq \mathsf{SNR}^2$, a near coherent performance can be achieved. In particular, we show that if

$$l \geq \mathsf{SNR}^{-2} \left(\log \frac{1}{\mathsf{SNR}}\right)^2$$

then a throughput $R(\mathsf{SNR})$ with

$$\mathsf{SNR} - R(\mathsf{SNR}) = O(\mathsf{SNR}^2)$$

can be achieved.

Consider the communication over a coherence block. To obtain an achievable performance, we choose to use the i.i.d. Gaussian input distribution of $\mathbf{x} \in \mathcal{C}^l$. The resulting mutual information is lower bounded by

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= I(\mathbf{x};\mathbf{y}|h) + I(h;\mathbf{y}) - I(h;\mathbf{y}|\mathbf{x}) \\ &\geq I(\mathbf{x};\mathbf{y}|h) - I(h;\mathbf{y}|\mathbf{x}). \end{aligned}$$

The input is chosen to maximize $I(\mathbf{x};\mathbf{y}|h)$, thus the first term above coincides with the capacity of a coherent channel, and the second can be interpreted as the amount of information about h that one obtain from observing the output $\mathbf{y}$ given the input $\mathbf{x}$. We have that

$$\begin{aligned} I(h;\mathbf{y}|\mathbf{x}) &= E[\log(1+\|\mathbf{x}\|^2)] \\ &\leq \log[1+l\cdot\mathsf{SNR}]. \end{aligned}$$

Now, letting $l(\mathsf{SNR}) = \mathsf{SNR}^{-2}\left(\log\frac{1}{\mathsf{SNR}}\right)^2$, the mutual information per symbol time is:

$$\begin{aligned} \frac{1}{l}I(\mathbf{x};\mathbf{y}) &\geq \frac{1}{l}I(\mathbf{x};\mathbf{y}|h) - \frac{1}{l}I(h;\mathbf{y}|\mathbf{x}) \\ &\geq C_{coherent}(\mathsf{SNR}) - \mathsf{SNR}^2\cdot\left(\log\frac{1}{\mathsf{SNR}}\right)^{-2}\log\left[1+\mathsf{SNR}^{-1}\left(\log\frac{1}{\mathsf{SNR}}\right)^2\right] \\ &= C_{coherent}(\mathsf{SNR}) - o(\mathsf{SNR}^2) \\ &= \mathsf{SNR} - O(\mathsf{SNR}^2). \end{aligned} \tag{17}$$

To generalize this result, for any given $\alpha \in (0,1)$,we choose to concentrate the energy and transmit only in

$$\delta(\mathsf{SNR}) = \mathsf{SNR}^{1-\alpha} \tag{18}$$

fraction of the blocks, while leaving the other blocks in silence. In each block that we do transmit, we use an i.i.d. Gaussian random code with average power

$$\mathsf{SNR}' := \frac{\mathsf{SNR}}{\delta(\mathsf{SNR})} = \mathsf{SNR}^{\alpha}.$$

In such a block with coherence time $l \doteq \mathsf{SNR}^{-2\alpha} = \left(\mathsf{SNR}'\right)^{-2}$, thus by (17), a throughput per symbol time $\mathsf{SNR}' - O(\mathsf{SNR}'^2)$ can be achieved. Now the overall throughput

$$\begin{aligned} C_l(\mathsf{SNR}) &\geq \delta(\mathsf{SNR})\left[\mathsf{SNR}' - O(\mathsf{SNR}'^2)\right] \\ &= \mathsf{SNR}^{1-\alpha}\left[\mathsf{SNR}^{\alpha} - O(\mathsf{SNR}^{2\alpha})\right] \\ &= \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha}) \end{aligned}$$

is achievable. ◦

In the above proof, we choose the input to be the i.i.d. Gaussian random code over δ fraction of the available blocks. Here, $1/\delta$ can be interpreted as the peak-to-average ratio of the transmission; it can be determined from the coherence level α through the relation $\delta \doteq \mathsf{SNR}^{1-\alpha}$. For larger values of α, the system is "more coherent", and the input becomes less peaky. $\alpha = 1$ means that all available degrees of freedom are used.

In the following, we prove the converse part of the Theorem. We first recall [11] that the optimal input distribution for this block fading channel is

$$\mathbf{x} = \|\mathbf{x}\|\,\Theta_{\mathbf{x}},$$

where $\Theta_{\mathbf{x}}$ is isotropically distributed, and is independent of $\|\mathbf{x}\|$. Therefore, we only need to find the distribution of $\|\mathbf{x}\|$. It turns out that, for the low SNR channel, the key is to determine how peaky the optimal input is.

Proof: Converse
In the following, we prove that in order the capacity to satisfy

$$SNR - C_l(\mathsf{SNR}) \dot{\leq} \mathsf{SNR}^{1+\alpha} \tag{19}$$

or equivalently, if there exists a distribution of $\mathbf{x}$, such that

$$\mathsf{SNR} - \frac{1}{l} I(\mathbf{x};\mathbf{y}) \dot{\leq} \mathsf{SNR}^{1+\alpha} \tag{20}$$

then it is necessary that

$$l \dot{\geq} \mathsf{SNR}^{-2\alpha}. \tag{21}$$

Note that (20) implies

$$\mathsf{SNR} - \frac{1}{l} I(\mathbf{x};\mathbf{y}) \leq \mathsf{SNR}^{1+\alpha'} \tag{22}$$

where $\alpha' = \alpha - \epsilon$, for some $\epsilon > 0$. (22) is more convenient in the following development since it has $\leq$ instead of $\dot{\leq}$. Since ϵ can be made arbitrarily small, we do not distinguish α and α' in the sequel.

We first give an overview of the proof. Write

$$I(\mathbf{x};\mathbf{y}) \quad = \quad h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x}).$$

Let $\mathbf{y}_G$ be the Gaussian random vector with the same total power constraint as $\mathbf{y}$, but i.i.d. over symbols. It holds that

$$h(\mathbf{y}_G) = l \log\left[\pi e(1 + \mathsf{SNR})\right]$$

and

$$h(\mathbf{y}) \leq h(\mathbf{y}_G)$$

Let $D_y \stackrel{\Delta}{=} h(\mathbf{y}_G) - h(\mathbf{y}) \geq 0$, which gives a measure of how much the distribution of $\mathbf{y}$ is different from i.i.d. Gaussian.

On the other hand, since given $\mathbf{x}$, $\mathbf{y}$ is Gaussian distributed, with a covariance matrix $\mathsf{SNR}\mathbf{x}\mathbf{x}^\dagger + I$, we have

$$h(\mathbf{y}|\mathbf{x}) = l \cdot \log \pi e + E\left[\log(1 + \mathsf{SNR}\,\|\mathbf{x}\|^2)\right]. \tag{23}$$

Thus, the mutual information is

$$I(\mathbf{x};\mathbf{y}) = l\log(1 + \mathsf{SNR}) - D_y - E\left[\log(1 + \mathsf{SNR}\,\|\mathbf{x}\|^2)\right].$$

Now, to maximize the mutual information is equivalent to minimizing the sum $D_y + E[\log(1 + \mathsf{SNR}\,\|\mathbf{x}\|^2)]$. This minimization can be interpreted intuitively as the following.

The term $E[\log(1 + \mathsf{SNR}\,\|\mathbf{x}\|^2)]$ from (23) can be viewed as the penalty due to the channel uncertainty, since it is the difference between $h(\mathbf{y}|\mathbf{x})$ and its counterpart for the coherent channel. To minimize this penalty, one would set the input $\|\mathbf{x}\|$ to be peaky. On the other hand, using peaky input signals causes the distribution of the output $\mathbf{y}$ to appear increasingly different than Gaussian, thus increasing D_y, which consequently has the interpretation that channel uncertainty leads to a wasting of degrees of freedom. The minimization of the sum of these two types of penalty therefore captures the tension, in selecting the peakiness of the signal, between using the degrees of freedom and being affected by channel uncertainty. In the rest of the proof, we will quantify this effect, and thereby solve the optimization problem. Throughout the proof, we occasionally remark on the intuition that guides the proof, although the mathematical derivation is completely rigorous.

First we observe that both of the penalty terms are non-negative. Now suppose that (22) is satisfied with some input distribution, it follows that

$$\frac{1}{l}\left[D_y + E[\log(1+\mathsf{SNR}\,\|x\|^2)]\right] \leq \mathsf{SNR}^{1+\alpha}$$
$$\Rightarrow \frac{1}{l}E[\log(1+\mathsf{SNR}\,\|\mathbf{x}\|^2)] \leq \mathsf{SNR}^{1+\alpha} \tag{24}$$

Remark:
(24) is a constraint that the input $\|\mathbf{x}\|$ must be peaky. To see that, we use Jensen's inequality on the LHS of (24) and get

$$\frac{1}{l}E[\log(1+\mathsf{SNR}\,\|\mathbf{x}\|^2)] \leq \frac{1}{l}\log(1+\mathsf{SNR}E[\|\mathbf{x}\|^2]) \approx \mathsf{SNR}.$$

Equality holds if the input $\|\mathbf{x}\|^2$ is a constant. Expression (24) says that, in order to achieve the desired performance (22), the distribution of $\|\mathbf{x}\|^2$ must be such that the expectation of the log function is much less than its upper bound given by Jensen's bound. ◦

Now we derive another necessary condition so that the performance (24) can be achieved. We write

$$I(\mathbf{x};\mathbf{y}) \quad = \quad I(\|\mathbf{x}\|\,;\mathbf{y}) + I(\mathbf{x};\mathbf{y}|\,\|\mathbf{x}\|).$$

In the following, we first show that the first term, the information conveyed by the norm $\|\mathbf{x}\|$ is small:

$$\frac{1}{l}I(\|\mathbf{x}\|\,;\mathbf{y}) \dot{\leq} \mathsf{SNR}^{1+\alpha}. \tag{25}$$

Since the optimal input distribution is isotropically distributed, we have

$$I(\|\mathbf{x}\|\,;\mathbf{y}) = I(\|\mathbf{x}\|\,;\|\mathbf{y}\|).$$

To obtain an upper bound of this mutual information, we give the receiver the side information of the direction of the transmitted signal vector, $\mathbf{x}/\|\mathbf{x}\|$. With this side information, the optimal receiver filters out the additive noise that is perpendicular to $\mathbf{x}$, and get a new sub-channel

$$\|\mathbf{x}\| \rightarrow \mathbf{y}' = \sqrt{\mathsf{SNR}}\cdot h\,\|\mathbf{x}\| + \mathbf{w}'$$

where $\mathbf{w}' \sim \mathcal{CN}(0,1)$. This becomes a scalar fading channel with signal-to-noise ratio of $\mathsf{SNR}' = l \cdot \mathsf{SNR}$. Clearly, as l increases, $\frac{1}{l}I(\|\mathbf{x}\|\,; \mathbf{y}')$ decreases. Thus, it suffices to consider the case where $l \cdot \mathsf{SNR} \ll 1$.

We choose the input distribution to maximize $I(\|\mathbf{x}\|\,; \|\mathbf{y}\|)$, subject to the power constraint, $E[\|\mathbf{x}\|^2] \leq l$, and the peakiness constraint (24). At low SNR, with only the average power constraint, the mutual information is maximized with the on-off input distribution,

$$\mathsf{SNR}\,\|\mathbf{x}\|^2 = \begin{cases} 0 & 1-\delta \\ A & \delta \end{cases},$$

where $A > 1$. The same input also minimizes $E[\log(1 + \mathsf{SNR}\,\|\mathbf{x}\|^2)]$ in (24), thus this extra peakiness constraint does not change the optimal input, and it suffices to consider only the on-off signaling.

From (24), we have

$$\begin{aligned} & \frac{1}{l}\delta \log(1+A) \leq \mathsf{SNR}^{1+\alpha} \\ \Rightarrow\quad & \frac{1}{l}\delta \leq \mathsf{SNR}^{1+\alpha} \\ \Rightarrow\quad & \frac{1}{l}\left(-\delta \log \delta\right) \dot{\leq} \mathsf{SNR}^{1+\alpha}. \end{aligned}$$

Now with the on-off input distribution of $\|\mathbf{x}\|$,

$$\frac{1}{l}I(\|\mathbf{x}\|\,; \mathbf{y}') \leq \frac{1}{l}H(\|\mathbf{x}\|) \approx \frac{1}{l}(-\delta \log \delta) \dot{\leq} \mathsf{SNR}^{1+\alpha}$$

and (25) follows.

Expression (25) can be viewed as a direct consequence of the peakiness constraint given in (24). Consider the on-off distribution of $\|\mathbf{x}\|^2$, and for the time being, assume $l \doteq \mathsf{SNR}^{-2\alpha}$, even though it is not shown as necessary yet. Expression (24) implies that $\delta \leq \mathsf{SNR}^{1+\alpha} \cdot l$, which means the input norm has the non-zero point mass at

$$\mathsf{SNR}\,\|\mathbf{x}\|^2 = A = \frac{\mathsf{SNR} \cdot l}{\delta} \doteq \mathsf{SNR}^{-\alpha}$$

and the per symbol period signal energy is

$$\frac{1}{l}\mathsf{SNR}\,\|\mathbf{x}\|^2 = \frac{\mathsf{SNR}^{-\alpha}}{l} \doteq \mathsf{SNR}^{\alpha}.$$

Let us compare the above expression with the optimal input for an i.i.d. non-coherent channel. We observe that the energy per symbol time decrease and the input signal is less peaky, the energy per block increases to $\mathsf{SNR}^{-\alpha} \gg 1$. Hence, using the transmitted energy over the entire block to convey information becomes very inefficient. ◦

We have that

$$\frac{1}{l}I(\mathbf{x};\mathbf{y}) \quad \leq \quad \frac{1}{l}I(\|\mathbf{x}\|\,;\mathbf{y}) + \frac{1}{l}I(\mathbf{x};\mathbf{y}|\,\|\mathbf{x}\|),$$

where the first term on the right side is smaller than $O(\mathsf{SNR}^{1+\alpha})$. Thus, for (22)to be satisfied, we need

$$\frac{1}{l}I(\mathbf{x};\mathbf{y}|\,\|\mathbf{x}\|) \geq \mathsf{SNR} - \mathsf{SNR}^{1+\alpha}.$$

Note that, given the norm of $\mathbf{x}$, this mutual information is upper bounded by the capacity of AWGN channels with the same input power. Thus, we have

$$E\left[\log\left(1+\frac{\mathsf{SNR}\,\|\mathbf{x}\|^2}{l}\right)\right] \geq \mathsf{SNR} - \mathsf{SNR}^{1+\alpha}. \tag{26}$$

Remark:
Expression (26) says that, in order to achieve a certain data rate, the input signal cannot be too peaky; otherwise even if the receiver has CSI, the desired throughput cannot be achieved. This means the signal power has to be spread out over the available degrees of freedom.

We have therefore a tension between the effect of the lack of channel coherence, which leads to (24), and the inefficient use of degrees of freedom yielded by peaky signaling, which leads to (26). In order to investigate this tension further, we define the random variable

$$B \triangleq \frac{\|\mathbf{x}\|^2}{l}.$$

If the desired performance given by (22) is achieved, B satisfies

$$E[B] = 1 \tag{27}$$

$$E[\log(1+\mathsf{SNR}\cdot B)] \geq \mathsf{SNR} - \mathsf{SNR}^{1+\alpha} \tag{28}$$

$$\frac{1}{l}E[\log(1+\mathsf{SNR}\cdot l\cdot B)] \leq \mathsf{SNR}^{1+\alpha}. \tag{29}$$

These constraints imply that l must be large enough, in way that is quantified in the following Lemma, which concludes the proof of the Theorem.

Lemma 5 *If there exists a random variable B that satisfies (27)-(29), for any $\alpha \in (0,1)$, then it is necessary that*

$$l \dot{\geq} \mathsf{SNR}^{-2\alpha}.$$

Proof: See Appendix B. ◦

Discussion: Coherence level and Peakiness

The result in Theorem 4 is surprisingly simple: the system behavior can be determined by a single parameter α, the coherence level. To summarize, at a coherence time $l \doteq \mathsf{SNR}^{-2\alpha}$, the optimal input is the Gaussian code in $\delta \doteq \mathsf{SNR}^{1-\alpha}$ fraction of the blocks; the resulting capacity is $C_l(\mathsf{SNR}) = \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})$, and the resulting energy efficiency is $\log(E_n/N_0) \approx \mathsf{SNR}^{\alpha}$. While some potential improvements may be ignored by the coarse scaling we chose, the main relation between the channel coherence level and the peakiness of the optimal input is captured in our development.

One important engineering observation to be made is that, with the optimal Gaussian input, the information transmitted through the channel is dominated by that conveyed by $\theta_{\mathbf{x}}$, the direction of the transmitted signal vector over the block, instead of the total energy. Decomposing the mutual information as

$$I(\mathbf{x};\mathbf{y}) = I(\|\mathbf{x}\|\,;\mathbf{y}) + I(\mathbf{x};\mathbf{y}|\,\|\mathbf{x}\|)$$

it is shown in (25) that the first term, the information conveyed by the power of the transmitted signal, is negligible for any $\alpha > 0$. This is surprising since, traditionally, schemes such as PPM and FSK, which only use the position of impulses to convey information, are considered efficient in low SNR applications. This extreme non-coherent view turns out to be valid only for the system has a very short coherence time. Theorem 4 says that, in other cases, when the channel coherence time l is not negligible, signaling by pulse positions quickly becomes inefficient, when compared to coherent communication using Gaussian block codes.

To illustrate further the choice of the peakiness of the input, we compute the total energy transmitted over a coherence block, where signal is transmitted, as

$$\mathsf{SNR}\,\|\mathbf{x}\|^2 \;=\; \frac{l\cdot\mathsf{SNR}}{\delta} = \mathsf{SNR}^{-(1-\alpha)}\mathsf{SNR}^{-2\alpha}\mathsf{SNR} = \mathsf{SNR}^{-\alpha} \gg 1.$$

Note that if $\alpha \le \frac{1}{2}$, the total average energy available per coherence block is $l\mathsf{SNR} = \mathsf{SNR}^{1-2\alpha} \ll 1$. Under such coditions, it seems impossible to have any meaningful channel estimate and, hence, coherent communication. The peakiness of the input signaling, however, allows the signal energy to be concentrated to create locally large enough energy to communicate coherently over a small fraction of the blocks. In contrast, if $\alpha \ge \frac{1}{2}$, the average energy per block is much larger than 1, therefore we have more than enough energy to estimate all the fading coefficients. However, we may still not wish to communicate over all blocks. Indeed, peakiness is still desirable to keep the penalty of the channel uncertainty low. The signal power per symbol period, on the other hand, is

$$\frac{1}{l}\mathsf{SNR}\,\|\mathbf{x}\|^2 = \mathsf{SNR}^{\alpha} \ll 1$$

remains small, to maintain a high energy efficiency.

The choice of the optimal peakiness therefore balances the requirements of low channel uncertainty penalty and high energy efficiency. This can be illustrated with the following simple derivation. Consider using the signaling that transmits i.i.d. Gaussian random codes in a coherence block with probability δ. We focus on the data rate achievable in each of these blocks, where the equivalent signal to noise ratio per symbol time is $\mathsf{SNR}' = \mathsf{SNR}/\delta$. The achievable data rate per block is

$$\begin{aligned} I(\mathbf{x};\mathbf{y}) &= I(\mathbf{x};\mathbf{y}|h) + I(h;\mathbf{y}) - I(h;\mathbf{y}|\mathbf{x}) \\ &\ge I(\mathbf{x};\mathbf{y}|h) - I(h;\mathbf{y}|\mathbf{x}). \end{aligned}$$

Since i.i.d. Gaussian random code is used, $I(\mathbf{x};\mathbf{y}|h)$ is indeed the coherent capacity, thus

$$\begin{aligned} \frac{1}{l}I(\mathbf{x};\mathbf{y}) &\ge C_{coherent}(\mathsf{SNR}') - E_{\mathbf{x}}\left[\log(1+\mathsf{SNR}'\cdot\|\mathbf{x}\|^2)\right] \\ &\ge l\cdot C_{coherent}(\mathsf{SNR}') - \log(1+\mathsf{SNR}'\cdot l). \end{aligned}$$

The average data rate per symbol period is thus

$$\begin{aligned}\delta \cdot \frac{1}{l} I(\mathbf{x};\mathbf{y}) &\geq \delta \cdot C_{coherent}(\mathsf{SNR}') - \delta \frac{1}{l} \cdot \log(1 + \mathsf{SNR}' \cdot l) \\ &= \mathsf{SNR} - k \cdot \frac{\mathsf{SNR}^2}{\delta} - \frac{\delta}{l} \log\left(1 + \frac{l \cdot \mathsf{SNR}}{\delta}\right)\end{aligned} \tag{30}$$

for some constant k. Let $l \doteq \mathsf{SNR}^{-2\alpha}, \delta \doteq \mathsf{SNR}^{\beta}$, and assume $l \cdot \mathsf{SNR}/\delta \gg 1$,

$$\delta \cdot \frac{1}{l} I(\mathbf{x};\mathbf{y}) \quad \dot{\geq} \quad \mathsf{SNR} - O(\mathsf{SNR}^{2-\beta}) - O(\mathsf{SNR}^{2\alpha+\beta}). \tag{31}$$

The two negative terms in the RHS can be interpreted as the sublinearity due to log(.) function, and to the cost of the channel estimation, respectively. To find the optimal peakiness, one needs to choose β to balance these two terms, i.e., $2-\beta = 2\alpha+\beta$, which is equivalent as

$$\delta^* \doteq \mathsf{SNR}^{1-\alpha} \tag{32}$$

While (30) is only an achievable data rate, it is a tight lower bound and can be used as an approximation of capacity. An important conclusion to be drawn here is that the optimal peakiness of transmission is a function of the channel coherence level. Thus when there is a peakiness constraint in the system, as long as the above optimal peakiness level is allowed, such constraint is no active. Thus in the designs of wideband wireless systems, it is not always "the more peaky, the better." The case that the peakiness constraint indeed rules out the optimal value of δ^* will be further discussed in the next section.

Theorem 4 can be viewed as a result on the rate that the capacity of the non-coherent block fading channel converges to that of the coherent channel. Eliminating the parameter α, we can write

$$\Delta_l(\mathsf{SNR}) = \mathsf{SNR} - C_l(\mathsf{SNR}) \doteq \mathsf{SNR}^{1+\alpha} \doteq \sqrt{\frac{1}{l}} \cdot \mathsf{SNR} \tag{33}$$

with the optimal peakiness as

$$\delta(\mathsf{SNR}) \doteq \mathsf{SNR}^{1-\alpha} = \mathsf{SNR} \cdot \sqrt{l} \tag{34}$$

and the resulting energy efficiency, using (10), is

$$\log\left(\frac{E_n}{N_0}\right) \approx \frac{\Delta_l(\mathsf{SNR})}{\mathsf{SNR}} \doteq \sqrt{\frac{1}{l}} \tag{35}$$

It is important to read (33)-(35) with the corresponding level of approximations. The notion of the coherence level α, although not explicitly present in these equations, indicates the precision level of these statements. In general, because of the use of approximation $\doteq$, these results are more useful in comparing different systems than in making absolute statements.

A rule of thumb in designing the input signals for practical system can thus be obtained from these results. If one wishes to apply a given system to a new environment where mobile users moves 4 times faster, i.e., the coherence time decreases by a factor of 1/4, then the optimal signaling should be twice as peaky as the original system. The resulting energy efficiency, in terms of the energy to transmit a nat $\log(E_n/N_o) \approx \Delta_l/\mathsf{SNR} \doteq \sqrt{\frac{1}{l}}$ is increased by $3dB$.

Discussion: Comparing to training schemes
Comparing the capacity result in Theorem 4 and the throughput of training schemes in Lemma 3, we observe that the optimal training scheme uses the same duty cycle as the capacity achieving scheme, $\delta \doteq \mathsf{SNR}^{1-\alpha}$, for any value of $\alpha \in (0,1)$. The optimal signaling scheme differs from the training based ones in that there is no energy separately used in training. It is crucial that the optimal receiver performs joint channel estimation and detection of the transmitted messages. Effectively, this is as if that the energy used for training is not "wasted" in the optimal communication scheme. ◦

Discussion: Wideband Slope
There is more than one scaling that can be used to describe the coherence level of a channel. For example, in [1], the wideband slope is defined as a measure of the sub-linear term in the capacity expression. If we write the channel capacity for small SNR as

$$C(\mathsf{SNR}) = C'(0)\mathsf{SNR} + \frac{1}{2}C''(0)\mathsf{SNR}^2 + o(\mathsf{SNR}^2) \text{ (nats/channel use)}$$

where $C'(0), C''(0)$ are the first and second derivative of the capacity at $\mathsf{SNR} = 0$, then the wideband slope

$$S_0 = \frac{2(C'(0))^2}{-C''(0)}$$

is the slope of the $C(\mathsf{SNR}) \sim \log(E_b/N_0)$ curve at $E_b/N_0 = -1.6$(dB).

The non-coherent penalty defined in this paper is related the wideband slope as following, assume that the first order optimality is achieve, that is, $C'(0) = 1$, then

$$\lim_{\mathsf{SNR}\to 0} \frac{\Delta(\mathsf{SNR})}{\mathsf{SNR}^2} = \lim_{\mathsf{SNR}\to 0} \frac{\mathsf{SNR} - C(\mathsf{SNR})}{\mathsf{SNR}^2} = \frac{-C''(0)}{2} = \frac{1}{S_0}. \tag{36}$$

Verdú shows that the wideband slope is strictly positive for the perfectly coherent channel and is 0 at the non-coherent extreme. To describe a continuum between these two extreme cases, one can ask the question “how long must the coherence time be for S_0 to attain a given value”. From (36), it is clear that describing the non-coherent penalty by the wideband slope is equivalent to linear curve fitting

$$\log \frac{E_n}{N_0} \approx \frac{\Delta}{\mathsf{SNR}} \longleftrightarrow \frac{1}{S_0}\mathsf{SNR}.$$

Theorem 4 suggests that any $S_0 > 0$ can be achieved if l is of the order SNR^{-2}. However, the precise value of l is hard to compute. Moreover, since the difference between the two extremes lies in the SNR exponent of $\Delta(\mathsf{SNR})$, it seems more natural to approximate the $\log E_n/N_0$ curve by powers of SNR, as used in Theorem 4. Figure 2 illustrates the continuum in performance with coherence levels.

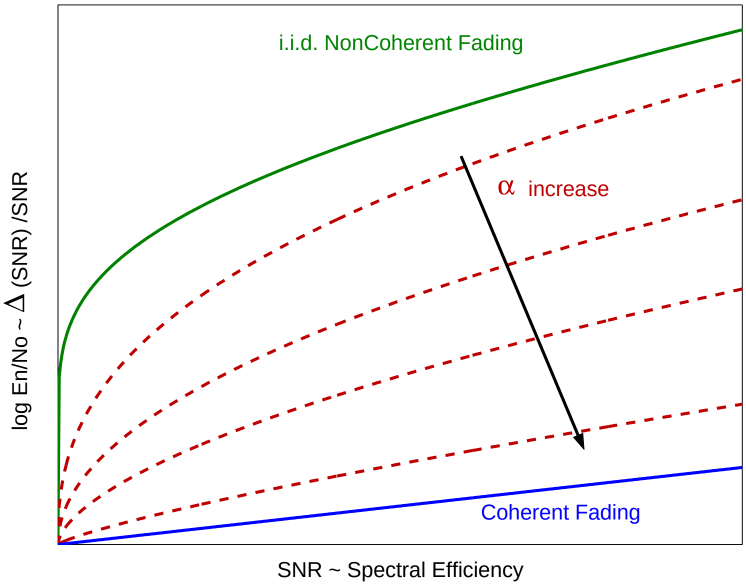

Figure 2: Energy efficiency vs. spectral efficiency at different coherence levels.

5.2 Channel Capacity with Peakiness Constraint

The optimal input for the non-coherent fading channel at low SNR has high peak-to-average ratio, and is thus usually difficult to implement in practice. It is therefore of interest to investigate capacity with an extra constraint of the peakiness of the input signals. In this section, we continue to study the block fading channel model, with an additional constraint on the peakiness of the transmitted signals. From the previous section, we have observed that the optimal input has a peakiness $\delta^* \doteq \mathsf{SNR}^{1-\alpha}$. In this section, we will find out the channel capacity when such small duty cycles are not allowed by the peakiness constraint. A natural form of such constraint is

$$\delta \dot{\geq} \mathsf{SNR}^{\beta} \tag{37}$$

Clearly, this constraint is only active if $\beta < 1 - \alpha$. Thus, the impact of the peakiness constraint depends on the coherence level of the channel α.

One can write the equivalent constraint on the maximum power transmitted in one symbol period,

$$\mathsf{SNR}\, \|x_i\|^2 \dot{\leq} \mathsf{E_{max}} = \mathsf{SNR}^{1-\beta} \tag{38}$$

Strictly speaking, $\mathsf{E_{max}}$ is somewhat different from the commonly used notion of peak signal power. Here, we consider the use of a Gaussian random code with duty cycle δ satisfying (37) as a valid input. Although the fluctuation in the Gaussian ensemble might cause the signal energy in a particular symbol period to be larger than its average value, the difference does not depend on SNR, and is thus ignored in the following order-of-magnitude analysis. For convenience, we will refer to $\mathsf{E_{max}}$ as the "maximum peak power", with the operational meaning that the signaling scheme we consider valid in this section will satisfy the peak power constraint of $K \cdot \mathsf{E_{max}}$, with a probability arbitrarily close to 1, for a large enough constant K.

Consider the case that the transmitted energy is concentrated in $\delta \doteq \mathsf{SNR}^{\beta}$ fraction of the coherence blocks. Within each of such blocks, the signal to noise ratio is increased to

$$\mathsf{SNR}' = \frac{1}{\delta} \cdot \mathsf{SNR} \doteq \mathsf{SNR}^{1-\beta}$$

the achievable rate is, from (30) (although it is a lower bound of the data rate, we use it as an approximation, and the converse can be carried in a manner similar to the proof of Theorem 4)

$$\begin{aligned} R(\mathsf{SNR}) &\approx \delta \left[\mathsf{SNR}' - k(\mathsf{SNR}')^2\right] - \delta \frac{1}{l} \log\left(1 + l \cdot \mathsf{SNR}'\right) \\ &= \mathsf{SNR} - k\mathsf{SNR}^{2-\beta} - \mathsf{SNR}^{2\alpha+\beta} \log\left(1 + \mathsf{SNR}^{1-\beta-2\alpha}\right) \end{aligned}$$

If we assume that $1 - \beta - 2\alpha < 0$, the last term is of the order $\mathsf{SNR}^{2\alpha+\beta}$. In the case that $\alpha + \beta < 1$, i.e., the peakiness constraint is active, this dominates the second term $\mathsf{SNR}^{2-\beta}$. These conditions can be written in a compact form as

$$\left.\begin{array}{lll} 1 - \beta - 2\alpha < 0 & \Leftrightarrow & l \cdot \mathsf{E}_{\mathsf{max}} \dot{\geq} 1 \\ \alpha + \beta < 1 & \Leftrightarrow & \mathsf{E}_{\mathsf{max}} \dot{\leq} \sqrt{\frac{1}{l}} \end{array}\right\} \Longrightarrow \Delta_l(\mathsf{SNR}) \doteq \mathsf{SNR}^{2\alpha+\beta} \doteq \frac{1}{l} \cdot \frac{\mathsf{SNR}}{\mathsf{E}_{\mathsf{max}}}. \quad (39)$$

Now the only case left is when $l \cdot \mathsf{E}_{\mathsf{max}} < 1$. For this case, we can use the following lemma from Médard and Gallager [5], which is also reported as a special case in the recent paper by Subramanian and Hajek [6].

Lemma 6 *For the non-coherent block fading channel with coherence time l, assume a forth moment constraint γ such that*

$$E[\|x_i\|^4] \leq \gamma$$

the capacity per symbol time satisfies

$$\frac{1}{l} I(\mathbf{x}; \mathbf{y}) \leq \frac{1}{2} \cdot l \cdot \gamma \cdot \mathsf{SNR}^2.$$

Notice that, under the average and peak power constraints, this lemma implies

$$\begin{aligned} \frac{1}{l} I(\mathbf{x}; \mathbf{y}) &\leq \frac{1}{2} \cdot l \cdot \left(\frac{\mathsf{E}_{\mathsf{max}}}{\mathsf{SNR}}\right) \cdot \mathsf{SNR}^2 \\ &= \frac{1}{2} \cdot l \cdot \mathsf{E}_{\mathsf{max}} \cdot \mathsf{SNR} \end{aligned} \quad (40)$$

where $\mathsf{E}_{\mathsf{max}}/\mathsf{SNR}$ is the maximum allowed peak-to-average ratio.

Expression (40) applies to the case when both l and $\mathsf{E}_{\mathsf{max}}$ go to their limits as $\mathsf{SNR}^{-2\alpha}$ and $\mathsf{SNR}^{1-\beta}$, respectively, when SNR approaches 0. In such case, the capacity per symbol period is upper bounded by $C_l(\mathsf{SNR}) \dot{\leq} \mathsf{SNR}^{2-2\alpha-\beta}$. This upper bound is only meaningful when $1 - 2\alpha - \beta > 0$, since otherwise one can use the trivial bound $C(\mathsf{SNR}) \leq \mathsf{SNR}$.

While Lemma 6 is only an upper bound of the capacity, under the scaling of interest, a throughput

$$\frac{1}{l} I(\mathbf{x}; \mathbf{y}) \doteq l \cdot \mathsf{E}_{\mathsf{max}} \cdot \mathsf{SNR}$$

is in fact achievable, provided that $l \cdot \mathsf{E}_{\mathsf{max}} \dot{\leq} \mathsf{SNR}^0$, i.e., the maximum total energy in a coherence block is much less than order 1. This result is given in a more precise form in [5, 6], where the order of the leading term of the capacity, as well as its coefficient, are precisely computed as a function of l and $\mathsf{E}_{\mathsf{max}}$. However, for the purpose of this paper, a crude scaling suffices.

The following theorem summarizes all these results.

Theorem 7 *For block fading channel with a coherence time l and a peak transmitted energy per symbol period constraint $\mathsf{E}_{\mathsf{max}}$, the capacity and the optimal input are given in the following three cases:*

(I) If $\mathsf{E}_{\mathsf{max}} > \sqrt{\frac{1}{l}}$, the capacity has $\Delta_l(\mathsf{SNR}) \doteq \sqrt{\frac{1}{l}} \cdot \mathsf{SNR}$, and the optimal input is i.i.d. Gaussian code transmitted in $\delta \doteq \sqrt{l} \cdot \mathsf{SNR}$ fraction of the available degrees of freedom.

(II) If $\mathsf{E}_{\mathsf{max}} < \sqrt{\frac{1}{l}}$, and $l \cdot \mathsf{E}_{\mathsf{max}} > 1$, the capacity has $\Delta_l(\mathsf{SNR}) \doteq \mathsf{SNR}/(l \cdot \mathsf{E}_{\mathsf{max}})$, and the optimal input is i.i.d. Gaussian code transmitted in $\delta \doteq \mathsf{SNR}/\mathsf{E}_{\mathsf{max}}$ fraction of the available degrees of freedom.

(III) If $l \cdot \mathsf{E}_{\mathsf{max}} < 1$, the capacity is $C_l(\mathsf{SNR}) \doteq l \cdot \mathsf{E}_{\mathsf{max}} \cdot \mathsf{SNR}$, the optimal input is on-off signaling over coherence blocks with a probability of being on as $\mathsf{SNR}/\mathsf{E}_{\mathsf{max}}$.

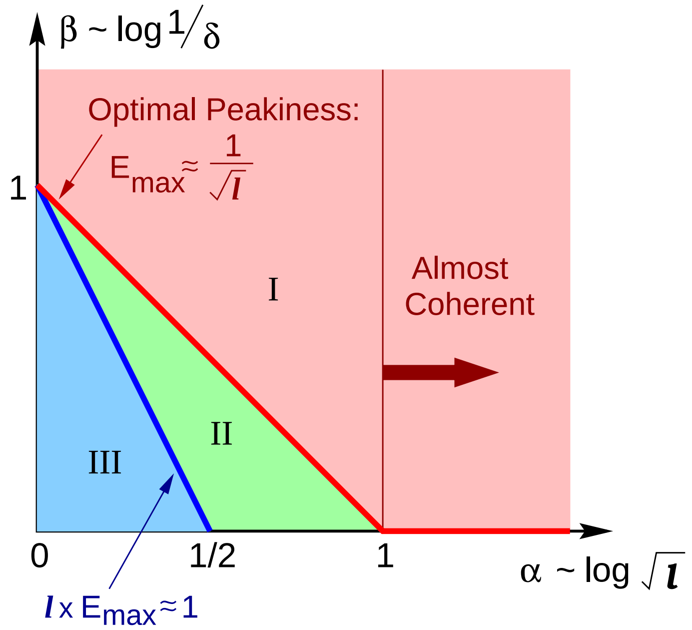

Figure 3: Peakiness constraint and Channel coherence

Remark:

The results in Theorem 7 are depicted in Figure 3. The parameters α and β can be viewed as scaled versions of the $\log \sqrt{l}$ and $\log \delta$ (scaled by $\log \frac{1}{\mathsf{SNR}}$ as $\mathsf{SNR} \to 0$). Case I corresponds to the case when $\alpha + \beta > 1$, when the peakiness constraint is not active; case III corresponds to the region $2\alpha + \beta < 1$, where the maximum signal energy in a coherence block is simply not enough to estimate the fading coefficient, and Shannon's wideband limit is not achievable; case II corresponds to the range in the middle. Thus, a complete picture of the effect of the peakiness constraint and the channel coherence is given.

The coherent and the non-coherent extreme cases correspond to $\alpha = 1$ and $\alpha = 0$, respectively. As long as $\alpha \geq 1$, near coherent performance can be achieved. Decomposing the mutual information into the part carried by the norm of the input vector, $I(\|\mathbf{x}\|; \mathbf{y})$, and the part carried by the direction, $I(\theta_{\mathbf{x}}; \mathbf{y}| \|\mathbf{x}\|)$, we observe that the information carriers are different in these three cases. In case I and II, the information is mostly carried by $\theta_{\mathbf{x}}$. It is thus nearly optimal to transmit in a prescribed fraction (positions) of coherence blocks, and encode the information bits only in $\theta_{\mathbf{x}}$. In case III, however, the information is mostly carried by $\|\mathbf{x}\|^2$. Thus the blocks where signals are transmitted should not be predetermined, but rather decided according to the information bits, however, $\theta_{\mathbf{x}}$, can be fixed, without significantly decreasing the throughput. The boundary between these two cases is the line $l \cdot \mathsf{E}_{\max} \doteq 1$. Thus, we conclude that the techniques based on energy detection, such as PPM, are nearly optimal only in range 3, when the transmitted energy, even if being focused as much as possible, is not enough to estimate any fading coefficient. ◦

6 Gauss-Markov Model

The study of the block fading channel model in the previous sections established the sufficient and necessary conditions on the coherence time (block length) in order that the channel has a coherence level α. In this section, we study another model of the time-variation of the fading process, namely, the Gauss-Markov model. Our goal is to understand the impact of the channel modelling assumptions on the performance, and to build a connection among different models.

We shall focus on a first order Gauss-Markov model of the fading process:

$$h_{t+1} = \sqrt{1-\epsilon} \cdot h_t + \sqrt{\epsilon} \cdot z_t$$

where $\mathbf{z}_t$'s are i.i.d. $\mathcal{CN}(0,1)$ random variables. The parameter ϵ is referred to as the innovation rate, which indicates the speed that the channel changes over time. The channel at time t is

$$y_t = \sqrt{\mathsf{SNR}} \cdot h_t x_t + w_t. \tag{41}$$

It is shown in [1] that, for any fixed innovation rate ϵ, as SNR approaches 0, the wide band slope is 0. This means that the sub-linearity of the channel is much larger than that of the coherent channel

$$\frac{\mathsf{SNR} - C_\epsilon(\mathsf{SNR})}{\mathsf{SNR}} \gg \mathsf{SNR}$$

or equivalently,

$$\mathsf{SNR} - C_\epsilon(\mathsf{SNR}) \gg \mathsf{SNR}^2.$$

As for the block fading channel, in order that the Gauss-Markov channel to have an intermediate level of coherence, or even have a capacity approaches that of the perfectly coherent channel, we need to let the innovation rate ϵ approach 0 as SNR decreases, while maintaining a certain relation $\epsilon(\mathsf{SNR})$ between the two. The following Theorem summarizes our result.

Theorem 8 *The Gauss-Markov channel has a coherence level of α, i.e.,*

$$C_\epsilon(\mathsf{SNR}) = \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})$$

for $\alpha \in [0,1]$, if and only if

$$\epsilon(\mathsf{SNR}) \doteq \mathsf{SNR}^{2\alpha}$$

Sketch of the Proof:
We first consider the case with $\alpha = 1$, corresponding to the perfectly coherent case. In this case, it is necessary to transmit signals in all available degrees of freedom. That is, the input signal in not peaky. The average signal-to-noise ratio for each symbol time is thus SNR.

To establish the necessary condition on $\epsilon(\mathsf{SNR})$, we assume that at each symbol time, the transmitted signal $\mathbf{x}_t$ can be perfectly decoded, after which it can be reused to update the channel estimate. Denote $\hat{h}_t$ and $\tilde{h}_t$ as the channel estimate and estimation error, respectively, at the end of the t^{th} symbol period. By the orthogonality principle, we have

$$E[|\hat{h}_t|^2] + E[|\tilde{h}_t|^2] = E[|h_t|^2] = 1.$$

We observe that, in order to achieve a capacity

$$C_\epsilon(\mathsf{SNR}) = \mathsf{SNR} - O(\mathsf{SNR}^2), \tag{42}$$

it is necessary that

$$E[|\tilde{h}_t|^2] = 1 - E[|\hat{h}_t|^2] \dot{\leq} \mathsf{SNR} \tag{43}$$

Otherwise even if we ignore the channel estimation error, the upper bound of the throughput

$$\begin{aligned} C_\epsilon(\mathsf{SNR}) &\leq E\left[\log(1 + \mathsf{SNR} \cdot |\hat{h}_t|^2)\right] \\ &\leq \log(1 + \mathsf{SNR} \cdot E[|\hat{h}_t|^2]) \\ &\leq \mathsf{SNR} \cdot E[|\hat{h}_t|^2] \\ &= \mathsf{SNR} - \mathsf{SNR} \cdot E[|\tilde{h}_t|^2] \end{aligned}$$

is smaller than the desired throughput (42).

Now, in the next symbol period, the channel innovation increases the variance of the unknown part of the fading coefficient to

$$(1-\epsilon)E[|\tilde{h}_t|^2] + \epsilon.$$

After decoding the transmitted symbol, we estimate this unknown coefficient, and the new estimation error has a variance

$$E[|\tilde{h}_{t+1}|^2] = \frac{1}{1+\mathsf{SNR}}\left[(1-\epsilon)E[|\tilde{h}_t|^2] + \epsilon\right]. \tag{44}$$

In order that the desired throughput be achieved in all symbol periods, we need $E[|\tilde{h}_{t+1}|^2] = E[|\tilde{h}_t|^2]$. That is, the channel innovation rate ϵ needs to be small enough that, with a transmitted signal power of SNR per symbol time, we can maintain the variance of the channel estimation error at the desired level.

Remarks:

(44) is indeed a key observation of using the Markov channel. As we are interested in the upper bound of the performance in general, we should not be limited to the schemes based on channel estimation. The optimal non-coherent communication scheme might not require any channel estimation, and even if it does, it is not always true that one would like to maintain a constant channel estimation error variance through the time.

In this section, however, we are interested in the coherence level of the channel. If no channel estimation can be made, then the time correlation of the fading coefficients becomes irrelevant, and only the performance of the non-coherent extreme can be achieved. On the other hand, the Markov structure of the channel implies that it is always more desirable to estimate the channel base on the recent history of the fading process, rather than having a "fresh start" after the channel changes completely. Thus, it is intuitively clear that the optimal way of using the Markov channel is to use it continuously in a long block while maintaining the channel estimates. ∘

Reorganizing (44), we have

$$(\mathsf{SNR} + \epsilon) E[|\tilde{h}_t|^2] = \epsilon.$$

Combining the latter expression with (43), we have that

$$\epsilon \doteq \mathsf{SNR}^2$$

as a necessary condition to achieve the performance in (42).

We may in fact attain this upper bound of the performance, up to the scaling of interest. To do so, we need to use a communication scheme that allows joint channel estimation and communication; we also need to show that the effect of channel estimation error is negligible.

To allow joint channel estimation and communications, we adopt the interleaved decision-oriented training scheme proposed in [12]. This scheme is depicted in Figure 4.

In Figure 4, each box denotes a transmitted symbol, and the boxes with the same pattern form a codeword. Training signals are transmitted periodically once in T symbol periods. We deliberately distinguish T from the coherence time l used in the previous sections, since the value of T can be chosen regardless of the actually speed that channel changes over time. The use of training signals is only to set the symbols from the same codeword at the same footing. At the receiver, the first symbol after the training signal in each block is first received, and the corresponding codeword is decoded. With an arbitrarily large number of such blocks, the codeword can be decoded reliably. After decoding a codeword, the channel estimates are updated, before receiving the next symbol. As shown in (44), with $\epsilon \doteq \mathsf{SNR}^2$, the channel estimation error can be maintained at

$$E[|\tilde{h}_t|^2] \doteq \mathsf{SNR}.$$

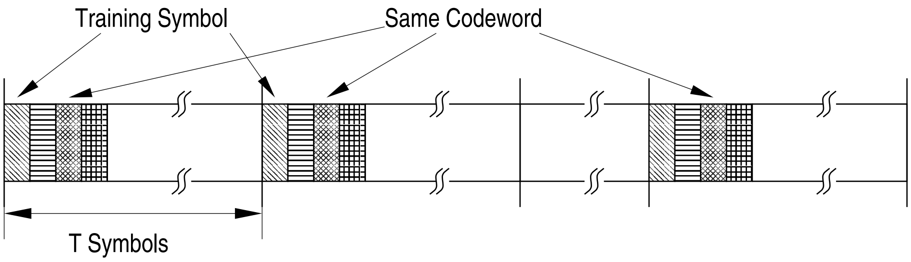

Figure 4: Interleaved Training Scheme for the Gauss-Markov Channel.

Thus we can choose T to be arbitrarily large, and the cost of training, in time and energy, can be ignored.

Now, in each symbol period, the transmitted signal is Gaussian with variance SNR. By treating the effect of the channel estimation error as Gaussian noise, we obtain an achievable throughput,

$$R_1(\mathsf{SNR}) \geq E\left[\log\left(1+\frac{|\hat{h}_t|^2\mathsf{SNR}}{1+E[|\tilde{h}_t|^2]\mathsf{SNR}}\right)\right].$$

Denoting

$$f(\mathsf{SNR}) \triangleq \frac{|\hat{h}_t|^2\mathsf{SNR}}{1+E[|\tilde{h}_t|^2]\mathsf{SNR}}$$

as the effective SNR at each symbol, and using the fact that

$$E[|\tilde{h}_t|^2] = 1 - E[|\hat{h}_t|^2] \doteq \mathsf{SNR},$$

we have

$$f(\mathsf{SNR}) = \mathsf{SNR}\cdot|\hat{h}_t|^2 - O(\mathsf{SNR}^2).$$

We observe here the channel estimation error causes an increase in the noise variance, $E[|\tilde{h}_t|^2]\mathsf{SNR}$, which is no larger than $O(\mathsf{SNR}^2)$, and can thus be ignored.

$$\begin{aligned} R_1(\mathsf{SNR}) &\geq E\left[f(\mathsf{SNR}) - \frac{1}{2}f(\mathsf{SNR})^2\right] \\ &\doteq \mathsf{SNR} - O(\mathsf{SNR}^2) \end{aligned}$$

thus the desired near coherent throughput is achieved.

Finally, in order to achieve a data rate of the order

$$\mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})$$

for $\alpha \in [0,1]$, we can choose to use flashy signaling, that is, we transmit only in $\delta \doteq \mathsf{SNR}^{1-\alpha}$ fraction of the available time slots. The resulting throughput is

$$R_\alpha(\mathsf{SNR}) = \delta\cdot R_1\left(\frac{\mathsf{SNR}}{\delta}\right).$$

Let $\mathsf{SNR}' = \frac{\mathsf{SNR}}{\delta} \doteq \mathsf{SNR}^\alpha$. We know from our previous discussion that, if $\epsilon \doteq (\mathsf{SNR}')^2 \doteq \mathsf{SNR}^{2\alpha}$, a throughput

$$\begin{aligned} R_1(\mathsf{SNR}') &\doteq \mathsf{SNR}' - O((\mathsf{SNR}')^2) \\ &\doteq \mathsf{SNR}^\alpha - O(\mathsf{SNR}^{2\alpha}) \end{aligned}$$

can be achieved. Thus, the resulting throughput is

$$R_\alpha(\mathsf{SNR}) = \delta \cdot R_1\left(\frac{\mathsf{SNR}}{\delta}\right) \doteq \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})$$

and the throughput corresponding to a coherence level α is achieved.

The choice of $\alpha \doteq \mathsf{SNR}^{1-\alpha}$ is indeed the optimal choice of the duty cycle. The proof of this fact is similar to that in the previous sections, and is therefore omitted. $\circ$

Comparing our above results for Gauss-Markov channels to the capacity results for the block fading model in section 5.1, we observe that, for the channel to have a coherence level of α, for a block fading channel, the block length has to be $l \dot{\geq} \mathsf{SNR}^{-2\alpha}$; for the Gauss-Markov model, the innovation rate has to be $\epsilon \dot{\leq} \mathsf{SNR}^{2\alpha}$. Thus, through the notion of the coherence level, we can build a connection between the block fading model and the Gauss-Markov model. A block fading model with a block length l and a Gauss-Markov model with an innovation rate ϵ are equally coherent if $l \approx \epsilon^{-1}$, provided that optimal signaling is used. This means the peakiness δ in the optimal input signals is of the same order for both models, and the resulting non-coherent penalty of the capacity is of the same order for both cases. Although the optimal signaling presented in 5.1 for the block fading model is rather different from the interleaved scheme in this section, it can be verified that this interleaved scheme, with the key component of joint channel estimation and communication, is also optimal in a block fading channel, in the sense that the requirement on the block length l to achieve a certain coherence level is the minimum.

Furthermore, if one insist to have separated training and communication phases, the achievable sub-optimal performance for the Gauss-Markov model is also similar to that of the block fading channels, with the correspondence $l \leftrightarrow \epsilon^{-1}$. That is, in order to achieve the throughput of $\mathsf{SNR} - \mathsf{SNR}^{1+\alpha}$ for the Gauss-Markov model, the sufficient and necessary condition is $\epsilon \doteq \mathsf{SNR}^{3\alpha}$. The proof of this statement is similar to Lemma 3, and is therefore omitted. The only differences in the design of the training signals for the Gauss-Markov channel are: 1) with the channel memory, each training signal is used to update the channel state information from the previous estimates, instead of estimating the channel from scratch; 2) training signals have to be sent frequent enough before the quality of the channel estimates degrades too much: in particular, the interval between two training symbols should be much smaller than $1/\epsilon$, which is different from the block fading case, where one could use 1 training symbol per block of length l.

The above comparison shows much similarity between the Gauss-Markov model and the block fading model, in terms of the capacity limits, the optimal signaling schemes, as well as the penalty of sub-optimal operations. We thus conjecture that the low SNR fading channel is in fact insensitive to the details of the time variation of the channel. The performance is mainly determined by the average changing rate of the channel. While the precise mathematical statement of this intuition is beyond the scope of this paper, we will use the following discussion to further illustrate the idea.

Discussion:
We consider a family of models, which lies between the block fading model and the Gauss-Markov model. Suppose that the fading coefficient remains constant for a block of T symbol periods, and then evolves as a Markov process. That is, for any integer n,

$$\begin{aligned} h_{nT} &= h_{nT+1} = \ldots = h_{nT+T-1} \\ h_{(n+1)T} &= \sqrt{1-\mu} \cdot h_{nT} + \sqrt{\mu} \cdot z_n \end{aligned}$$

where z_n's are i.i.d. $\mathcal{CN}(0,1)$ random variables. The parameters T and μ specify a particular model in the family. To make comparisons among these models, we let $\mu = T\epsilon$ for a fixed ϵ. Clearly the model with $T = 1$ is the Gauss-Markov model consider in this section, and $T = \frac{1}{\epsilon} = l$ corresponds to the block fading model with a block length l.

For simplicity, we focus on the case $\epsilon \doteq \mathsf{SNR}^2$, which can be easily generalized using the analysis of peaky signaling in the previous sections. We use a communication scheme as follows. At the beginning of each block of length T, a channel estimate is formed. Communication symbols with average energy SNR are sent throughout the block and received using the channel estimate. Assuming these symbols are decoded perfectly, which can be assured using an interleaving scheme similar to Figure 4, at the end of the block, we can use the communication signals in the block, with total energy of $T\mathsf{SNR}$, as training signals to update the channel estimate. The estimation error variance at the end of the n^{th} block is

$$E[|\tilde{h}_{nT+T-1}|^2] = \frac{1}{1+T\mathsf{SNR}} \left[(1-\mu)E[|\tilde{h}_{nT-1}|^2] + \mu\right]$$

In steady state, we have $E[|\tilde{h}_{nT+T-1}|^2] = E[|\tilde{h}_{nT-1}|^2]$, and

$$(T\mathsf{SNR} + \mu)E[|\tilde{h}_{nT+T-1}|^2] = \mu,$$

which implies $E[|\tilde{h}_{nT+T-1}|^2] \approx \mathsf{SNR}$. At the beginning of the next block, the channel estimation error becomes $E[|\tilde{h}_{(n+1)T}|^2] \approx \mathsf{SNR} + \mu$. Now as long as $T \dot{\leq} \mathsf{SNR}^{-1}$, we have $E[|\tilde{h}_{(n+1)T}|^2] \doteq \mathsf{SNR}$ and thus the near coherent performance is achieved.

The message of this discussion is that the family of models gives essentially the same performance limit, for a large range of values of T. Thus the capacity of the channel is determined by the parameter ϵ, which can be viewed as a long term average changing rate of the fading process. No matter the channel variation occurs gradually over time, or in bursts once every T symbols, the resulting capacity and coherence level are the same, hence confirming with our intuition of the insensitivity of the channel models at low SNR. ◦

7 Conclusion

We have characterized in this paper how the rate of change of channel, signal peakiness and SNR interplay to determine channel capacity. Our results lead to compact characterization of different regions, of varying operational coherence, that span the range between the coherent and non-coherent extremes. The regions may be interpreted as the ranges in which certain effects dominate capacity behavior. While we have considered only a block fading model and a Gauss-Markov model for the channel fading, the agreement in the results we obtain for these two types of channels indicate that our results are robust to reasonable modelling assumptions and necessary simplifications. Our results may be readily extended to multiple-user systems, since for low SNR systems, the channel noise will dominate the interference.

The results we have presented naturally lead to questions regarding the effect of channel coherence in the multiple-input multiple-output (MIMO) case with low SNR. In such systems, traditional approaches, geared towards high SNR systems, have relied heavily on channel coherence to mitigate interference. We have extended the results in this paper to MIMO systems and shown that the concept of coherence level can be readily extended [13].

A Proof of Proposition 1

For convenience, we rewrite the channel here,

$$\mathbf{y} = \sqrt{\mathsf{SNR}} \cdot h\mathbf{x} + \mathbf{w}$$

where $\mathbf{x}, \mathbf{w}, \mathbf{y}$, and $h \in \mathcal{C}$, and the power constraint is normalized such that $E[\|\mathbf{x}\|^2] = 1$. It is shown in [14] that, at low SNR, the optimal input is on-off signaling with

$$\sqrt{\mathsf{SNR}} \cdot \mathbf{x} = \begin{cases} \sqrt{A}, & \text{with probability } \delta \\ 0, & \text{with probability } 1-\delta \end{cases}$$

where $\delta A = \mathsf{SNR}$.

With this input, we directly compute the mutual information

$$I(\mathbf{x}; \mathbf{y}) = h(\mathbf{y}) - h(\mathbf{y}|\mathbf{x})$$

We have that, given $\mathbf{x}$, $\mathbf{y}$ is Gaussian distributed,

$$\begin{aligned} h(\mathbf{y}|\mathbf{x}) &= E[\log \pi e(1 + \mathsf{SNR} \cdot \|\mathbf{x}\|^2)] \\ &= \log(\pi e) + \frac{\mathsf{SNR}}{A} \log(1 + A). \end{aligned}$$

On the other hand,

$$h(\mathbf{y}) = h(\mathbf{y}_G) - D(\mathbf{y}||\mathbf{y}_G),$$

where $\mathbf{y}_G \sim \mathcal{CN}(0, 1 + \mathsf{SNR})$ is Gaussian distributed with the same variance as $\mathbf{y}$, $h(\mathbf{y}_G) = \log \pi e(1 + \mathsf{SNR})$; and $D(\cdot||\cdot)$ denotes the Kullback-Leibler divergence.

In the following, we derive upper and lower bounds for $D(\mathbf{y}||\mathbf{y}_G)$. The end results are summarized in (56).

By the change of variable $t = \|\mathbf{y}\|^2$, we have

$$D(\mathbf{y}||\mathbf{y}_G) = \int_0^\infty p(y)(t) \log \frac{p_y(t)}{p_G(t)} dt$$

where

$$\begin{aligned}
p_y(t) &= (1-\delta)p_0(t) + \delta p_1(t) \\
p_0(t) &= \exp(-t) \\
p_1(t) &= \frac{1}{1+A}\exp\left(\frac{-t}{1+A}\right) \\
p_G(t) &= \frac{1}{1+\mathsf{SNR}}\exp\left(\frac{-t}{1+\mathsf{SNR}}\right).
\end{aligned}$$

Now write the divergence as

$$\begin{aligned}
D(y||y_G) &= \int p_y(t)\log\left[(1-\delta)\frac{p_0(t)}{p_G(t)} + \delta\frac{p_1(t)}{p_G(t)}\right]dt \\
&= \int p_y(t)\log\left[(1-\delta)\frac{p_0(t)}{p_G(t)}\right]dt \qquad (45)\\
&\quad + \int p_y(t)\log\left[1 + \frac{\delta}{1-\delta}\frac{p_1(t)}{p_0(t)}\right]dt. \qquad (46)
\end{aligned}$$

Expression (45) can be easily computed and satisfies

$$(45) = -\frac{\mathsf{SNR}}{A} + O(\mathsf{SNR}^2)$$

$$\begin{aligned}
(46) &= \int (1-\delta)p_0(t)\log\left[1 + \frac{\delta}{1-\delta}\frac{p_1(t)}{p_0(t)}\right]dt \qquad (47)\\
&\quad + \delta\int p_1(t)\log\left[1 + \frac{\delta}{1-\delta}\frac{p_1(t)}{p_0(t)}\right]dt. \qquad (48)
\end{aligned}$$

Let us first consider (48),

$$\begin{aligned}
(48) &= \delta\int p_1(t)\log\left[1 + \frac{\delta}{1-\delta}\frac{p_1(t)}{p_0(t)}\right]dt \\
&= \frac{\mathsf{SNR}}{A}\int \frac{1}{1+A}e^{-\frac{t}{1+A}}\log\left[1 + \frac{\mathsf{SNR}}{(A-\mathsf{SNR})(1+A)}\exp\left(\frac{At}{1+A}\right)\right]dt \\
&= \frac{\mathsf{SNR}}{A}\int \frac{1}{1+A}e^{-\frac{t}{1+A}}\log\left[1 + \frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right]dt + O(\mathsf{SNR}^2).
\end{aligned}$$

For convenience, we denote

$$G(\mathsf{SNR}, A) \triangleq \frac{\mathsf{SNR}}{A}\int \frac{1}{1+A}e^{-\frac{t}{1+A}}\log\left[1 + \frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right]dt.$$

Now, let t^* satisfy

$$\exp\left(\frac{-At^*}{1+A}\right) = \frac{\mathsf{SNR}}{A(1+A)}. \qquad (49)$$

We may write

$$\begin{aligned}
G(\mathsf{SNR}, A) &= \int_0^{t^*} \frac{\mathsf{SNR}}{A(1+A)}e^{-\frac{t}{1+A}}\log\left[1 + \frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right]dt \ldots\ldots G_1(\mathsf{SNR}, A)\\
&\quad + \int_{t^*}^{\infty} \frac{\mathsf{SNR}}{A(1+A)}e^{-\frac{t}{1+A}}\log\left[1 + \frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right]dt \ldots\ldots G_2(\mathsf{SNR}, A)
\end{aligned}$$

where

$$\begin{aligned}
G_1(\mathsf{SNR},A) &= \int_0^{t^*} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \log\left[1+\frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right] dt \\
&\leq \int_0^{t^*} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right) dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^2 \int_0^{t^*} e^{\frac{A-1}{A+1}t} dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^2 \frac{A+1}{A-1}\left[e^{\frac{A-1}{A+1}t^*}-1\right] \\
&= \frac{A+1}{A-1}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} + O(\mathsf{SNR}^2)
\end{aligned}$$

and

$$\begin{aligned}
G_2(\mathsf{SNR},A) &= \int_{t^*}^{\infty} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \log\left[1+\frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right] dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \int_{t^*}^{\infty} e^{-\frac{t-t^*}{1+A}} \log\left[1+\exp\left(\frac{A(t-t^*)}{1+A}\right)\right] dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \int_0^{\infty} e^{-\frac{t}{1+A}} \log\left[1+\exp\left(\frac{At}{1+A}\right)\right] dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \int_0^{\infty} e^{-\frac{t}{1+A}} \left[\frac{At}{1+A}+\log\left[1+\exp\left(\frac{-At}{1+A}\right)\right]\right] dt \\
&\leq \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \int_0^{\infty} e^{-\frac{t}{1+A}} \left[\frac{At}{1+A}+\exp\left(\frac{-At}{1+A}\right)\right] dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \left[A(1+A)+1\right].
\end{aligned}$$

Denote (47) as:

$$(47) \quad \stackrel{\Delta}{=} \quad (1-\delta)F(\mathsf{SNR},A)$$

where

$$\begin{aligned}
F(\mathsf{SNR},A) &= \int_0^{\infty} \exp(-t)\log\left[1+\frac{\mathsf{SNR}}{A(1+A)}\exp\left(\frac{At}{1+A}\right)\right] dt \\
&= e^{-t^*}\int_0^{\infty} e^{-(t-t^*)} \log\left[1+\exp\left(\frac{A(t-t^*)}{1+A}\right)\right] dt \\
&= e^{-t^*}\int_{-t^*}^{0} e^{-s}\log\left[1+\exp\left(\frac{As}{1+A}\right)\right] ds \qquad (50)\\
&\quad +e^{-t^*}\int_0^{\infty} e^{-s}\log\left[1+\exp\left(\frac{As}{1+A}\right)\right] ds. \qquad (51)
\end{aligned}$$

We may use a similar approach to obtain upper bounds:

$$\begin{aligned}
(50) \triangleq F_1(\mathsf{SNR}, A) &= e^{-t^*} \int_0^{t^*} e^s \log\left[1+\exp\left(\frac{-As}{1+A}\right)\right] ds \\
&\leq e^{-t^*} \int_0^{t^*} e^s \exp\left(\frac{-As}{1+A}\right) ds \\
&= (1+A)e^{-t^*}\left[\exp\left(\frac{t^*}{1+A}\right)-1\right] \\
&= \frac{\mathsf{SNR}}{A} - (1+A)\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \qquad (52)
\end{aligned}$$

and

$$\begin{aligned}
(51) \triangleq F_2(\mathsf{SNR}, A) &= e^{-t^*} \int_0^{\infty} e^{-s} \log\left[1+\exp\left(\frac{As}{1+A}\right)\right] ds \\
&= e^{-t^*} \int_0^{\infty} e^{-s} \left[\left(\frac{As}{1+A}\right) + \log\left[1+\exp\left(\frac{-As}{1+A}\right)\right]\right] ds \\
&\leq e^{-t^*} \int_0^{\infty} e^{-s} \left[\left(\frac{As}{1+A}\right) + \exp\left(\frac{-As}{1+A}\right)\right] ds \\
&= e^{-t^*}\left[\frac{A}{1+A}+\frac{1+A}{1+2A}\right] \\
&= \left[\frac{A}{1+A}+\frac{1+A}{1+2A}\right]\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A}. \qquad (53)
\end{aligned}$$

Now, we have (47) $= (1-\delta)F(\mathsf{SNR}, A) = F(\mathsf{SNR}, A) - O(\mathsf{SNR}^2)$. By combining all of the above we have

$$\begin{aligned}
D(y||y_G) \leq & \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \\
& \left[A(1+A)+1+\frac{A+1}{A-1}+\frac{A}{1+A}+\frac{1+A}{1+2A}-(1+A)\right] + O(\mathsf{SNR}^2) \qquad (54)
\end{aligned}$$

At low SNR, $A \to \infty$, this becomes

$$D(y||y_G) = \mathsf{SNR}^{1+1/A} A^{-2/A}(1+o(1)).$$

Similarly, a lower bound for every term can be found, and

$$\begin{aligned}
D(y||y_G) \geq & \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \left[A(1+A)+\frac{A+1}{A-1}-\frac{1+A}{4A-2}-(1+A)+\frac{A}{1+A}-\frac{A+1}{2(A-1)}\right] \\
& + O(\mathsf{SNR}^2). \qquad (55)
\end{aligned}$$

Although the above bounds are derived assuming A is fixed, as $\mathsf{SNR} \to 0$, it is optimal to choose $A \to \infty$, while the $O(\mathsf{SNR}^2)$ term remain at the same order of magnitude.

For completeness, we give the proof of the lower bound as follows:

$$
\begin{aligned}
G_1(\mathsf{SNR}, A) &= \int_0^{t^*} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \log\left[1 + \frac{\mathsf{SNR}}{A(1+A)} \exp\left(\frac{At}{1+A}\right)\right] dt \\
&\geq \int_0^{t^*} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \frac{\mathsf{SNR}}{A(1+A)} \exp\left(\frac{At}{1+A}\right) dt \\
&\quad - \frac{1}{2}\int_0^{t^*} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \left[\frac{\mathsf{SNR}}{A(1+A)} \exp\left(\frac{At}{1+A}\right)\right]^2 dt \\
&= \frac{A+1}{A-1}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} + O(\mathsf{SNR}^2) \\
&\quad - \frac{1}{2}\int_0^{t^*} \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^3 \exp\left(\frac{2A-1}{1+A}t\right) dt \\
&= \frac{A+1}{A-1}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} + O(\mathsf{SNR}^2) \\
&\quad - \frac{1}{2}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^3 \frac{1+A}{2A-1}\left[\exp\left(\frac{2A-1}{1+A}t^*\right) - 1\right] \\
&= \frac{A+1}{A-1}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} + O(\mathsf{SNR}^2) \\
&\quad - \frac{1}{2}\frac{1+A}{2A-1}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} + O(\mathsf{SNR}^3) \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A}\left[\frac{A+1}{A-1} - \frac{1}{2}\frac{1+A}{2A-1}\right] + O(\mathsf{SNR}^2)
\end{aligned}
$$

and

$$
\begin{aligned}
G_2(\mathsf{SNR}, A) &= \int_{t^*}^{\infty} \frac{\mathsf{SNR}}{A(1+A)} e^{-\frac{t}{1+A}} \log\left[1 + \frac{\mathsf{SNR}}{A(1+A)} \exp\left(\frac{At}{1+A}\right)\right] dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \int_0^{\infty} e^{-\frac{t}{1+A}} \left[\frac{At}{1+A} + \log\left[1 + \exp\left(\frac{-At}{1+A}\right)\right]\right] dt \\
&\geq \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \int_0^{\infty} e^{-\frac{t}{1+A}} \left[\frac{At}{1+A}\right] dt \\
&= \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \left[A(1+A)\right].
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
F_1(\mathsf{SNR}, A) &= e^{-t^*}\int_0^{t^*} e^s \log\left[1+\exp\left(\frac{-As}{1+A}\right)\right] ds \\
&\geq e^{-t^*}\int_0^{t^*} e^s \left[\exp\left(\frac{-As}{1+A}\right) - \frac{1}{2}\exp\left(\frac{-2As}{1+A}\right)\right] ds \\
&= \frac{\mathsf{SNR}}{A} - (1+A)\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \\
&\quad -\frac{1}{2}e^{-t^*}\int_0^{t^*} e^s \exp\left(\frac{-2As}{1+A}\right) ds \\
&= \frac{\mathsf{SNR}}{A} - (1+A)\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \\
&\quad -\frac{1}{2}e^{-t^*}\frac{1+A}{1-A}\left[\exp\left(\frac{1-A}{1+A}t^*\right) - 1\right] \\
&= \frac{\mathsf{SNR}}{A} - (1+A)\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} \\
&\quad -\frac{1}{2}\frac{1+A}{A-1}\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A} + \frac{1}{2}\frac{1+A}{A-1}\exp\left(-\frac{2A}{1+A}t^*\right) \\
&= \frac{\mathsf{SNR}}{A} - \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A}\left[(1+A) + \frac{1}{2}\frac{1+A}{A-1}\right] + O(\mathsf{SNR}^2).
\end{aligned}
$$

Note that $e^{-t^*} = \left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A}$. Finally,

$$
\begin{aligned}
F_2(\mathsf{SNR}, A) &= e^{-t^*}\int_0^{\infty} e^{-s} \log\left[1+\exp\left(\frac{As}{1+A}\right)\right] ds \\
&= e^{-t^*}\int_0^{\infty} e^{-s} \left[\left(\frac{As}{1+A}\right) + \log\left[1+\exp\left(\frac{-As}{1+A}\right)\right]\right] ds \\
&\geq e^{-t^*}\int_0^{\infty} e^{-s} \left[\left(\frac{As}{1+A}\right)\right] ds \\
&= e^{-t^*}\left[\frac{A}{1+A}\right] \\
&= \left[\frac{A}{1+A}\right]\left[\frac{\mathsf{SNR}}{A(1+A)}\right]^{1+1/A}
\end{aligned}
$$

Combining the terms for G_1, G_2, F_1 and F_2, we obtain the lower bound (55).

To summarize, $D(\mathbf{y}||\mathbf{y}_G)$ is upper and lower bounded as

$$
D(\mathbf{y}||\mathbf{y}_G) = \mathsf{SNR}^{1+1/A}A^{-2/A}(1+o(1)) + O(\mathsf{SNR}^2) \tag{56}
$$

and the mutual information is

$$
I(\mathbf{x};\mathbf{y}) = \log(1+\mathsf{SNR}) - D(\mathbf{y}||\mathbf{y}_G) - \frac{\mathsf{SNR}}{A}\log(1+A)
$$

To maximize the mutual information, we need choose A to minimize

$$
\min_A \frac{\mathsf{SNR}}{A}\log(1+A) + \mathsf{SNR}^{1+1/A}A^{-2/A}
$$

or approximately, for large A,

$$\min_A M(A, \mathsf{SNR})$$

where

$$M(A, \mathsf{SNR}) = \frac{\log(A)}{A} + \left[\frac{\mathsf{SNR}}{A^2}\right]^{1/A}.$$

Let $M^*(\mathsf{SNR})$ be the minimum value. We first show that

$$M^*(\mathsf{SNR}) \geq M_L(\mathsf{SNR}) \triangleq \frac{\log\log\frac{1}{\mathsf{SNR}}}{\log\frac{1}{\mathsf{SNR}}}. \tag{57}$$

To see why (57) holds, suppose that, for some A, $M(A, \mathsf{SNR}) < M_L(\mathsf{SNR})$. Then, both terms in $M(A, \mathsf{SNR})$ must be less than $M_L(\mathsf{SNR})$, since both are non-negative. Now $\log(A)/A < M_L(\mathsf{SNR})$ implies that

$$A > \log\frac{1}{\mathsf{SNR}}.$$

On the other hand, for any A satisfying this,

$$\begin{aligned}
\left[\frac{\mathsf{SNR}}{A^2}\right]^{1/A} &= \frac{\mathsf{SNR}^{\frac{1}{A}}}{\exp\left[2\frac{\log(A)}{A}\right]} \\
&\geq \mathsf{SNR}^{-\frac{1}{\log\mathsf{SNR}}}\exp[-M_L(\mathsf{SNR})] \\
&= e^{-1}\exp\left[-\frac{\log\log\frac{1}{\mathsf{SNR}}}{\log\frac{1}{\mathsf{SNR}}}\right].
\end{aligned}$$

As $\mathsf{SNR} \to 0$, the above right-hand side is much larger than $M_L(\mathsf{SNR})$. Thus, (57) is proved by contradiction.

Now we simply choose a value of A to get an upper bound. Let

$$A = \frac{\log\frac{1}{\mathsf{SNR}}}{\log\log\frac{1}{\mathsf{SNR}}}.$$

The two terms are

$$\begin{aligned}
\frac{\log(A)}{A} &= \frac{(\log\log\frac{1}{\mathsf{SNR}} - \log\log\log\frac{1}{\mathsf{SNR}})\log\log\frac{1}{\mathsf{SNR}}}{\log\frac{1}{\mathsf{SNR}}} \\
&\leq \frac{\left(\log\log\frac{1}{\mathsf{SNR}}\right)^2}{\log\frac{1}{\mathsf{SNR}}}
\end{aligned}$$

and

$$\left[\frac{\mathsf{SNR}}{A^2}\right]^{\frac{1}{A}} = \mathsf{SNR}^{\frac{1}{A}}A^{-\frac{2}{A}}$$

Notice that

$$\begin{aligned}\mathsf{SNR}^{\frac{1}{A}} &= \mathsf{SNR}^{\frac{\log\log\frac{1}{\mathsf{SNR}}}{\log\frac{1}{\mathsf{SNR}}}}\\ &= \frac{1}{\log\frac{1}{\mathsf{SNR}}}.\end{aligned}$$

It can be verified that $A > 1$ and $A^{-\frac{2}{A}} \leq 1$. Combining everything, we have that

$$M^*(\mathsf{SNR}) \leq M(A, \mathsf{SNR}) \leq \frac{\left(\log\log\frac{1}{\mathsf{SNR}}\right)^2 + 1}{\log\frac{1}{\mathsf{SNR}}}.$$

In conclusion,

$$\mathsf{SNR} - \mathsf{SNR}\frac{\log\log\frac{1}{\mathsf{SNR}}}{\log\frac{1}{\mathsf{SNR}}} \leq C(\mathsf{SNR}) \quad \leq \quad \mathsf{SNR} - \mathsf{SNR}\frac{\left(\log\log\frac{1}{\mathsf{SNR}}\right)^2 + 1}{\log\frac{1}{\mathsf{SNR}}}.$$

B Proof of Lemma 5

We need to show that, if there exists a non-negative random variable $B \geq 0$ with

$$\begin{aligned}&E[B] = 1\\ &E[\log(1+\mathsf{SNR}B)] \geq \mathsf{SNR} - \mathsf{SNR}^{1+\alpha} \qquad (58)\\ &\frac{1}{l}E[\log(1+l\cdot\mathsf{SNR}B)] \leq \mathsf{SNR}^{1+\alpha}, \qquad (59)\end{aligned}$$

then $l \dot{\geq} \mathsf{SNR}^{-2\alpha}$.

The intuition of this lemma is quite clear. (58) says that the distribution of B is quite concentrated around its mean, so that the concavity of log(.) does not have much effect on the expectation; while (59) says that the distribution random variable $l \cdot B$ is, in contrast, widely spread out, pushing the expectation to be very low. The two conditions combine to the fact that l has to be very large.

To prove this lemma, we first use (58) to show that there is non-vanishing probability that $B \leq \mathsf{SNR}^{-(1-\alpha)+2\epsilon}$, e.g.,

$$\Pr[B \leq \mathsf{SNR}^{-(1-\alpha+2\epsilon)}] \geq \frac{1}{2} \qquad (60)$$

and that the conditional mean of $E[B|B \leq \mathsf{SNR}^{-(1-\alpha+\epsilon)}]$ is also of order 1.

We combine these two results with (59) to have

$$\begin{aligned}\mathsf{SNR}^{1+\alpha} &\geq \frac{1}{l}E[\log(1+l\mathsf{SNR}B)]\\ &\geq \Pr[B \leq \mathsf{SNR}^{-(1-\alpha+2\epsilon)}]\cdot\frac{1}{l}E[\log(1+l\mathsf{SNR}B)|B \leq \mathsf{SNR}^{-(1-\alpha+2\epsilon)}]\\ &\geq \frac{1}{2}E[\log(1+l\mathsf{SNR}\tilde{B})]\\ &\doteq E[\log(1+l\mathsf{SNR}\tilde{B})]\end{aligned}$$

where $\tilde{B}$ is the random variable that minimizes $E[\log(1+l\mathsf{SNR}B)]$, under the constraint that B has a support on $[0, \mathsf{SNR}^{-(1-\alpha+2\epsilon)}]$, and has an expectation of order 1. It is clear that $\tilde{B}$ must have a distribution as

$$\tilde{B} = \begin{cases} \mathsf{SNR}^{-(1-\alpha+2\epsilon)} & \text{with probability} \quad \eta \\ 0 & \text{with probability} \quad 1-\eta \end{cases}$$

with $\eta \doteq \mathsf{SNR}^{1-\alpha+2\epsilon}$, and that

$$\begin{aligned} \mathsf{SNR}^{1+\alpha} &\dot{\geq} \frac{1}{l} E[\log(1 + l\mathsf{SNR}\tilde{B})] \\ &= \frac{1}{l}\eta \log(1 + l\mathsf{SNR}\mathsf{SNR}^{-(1-\alpha+2\epsilon)}) \\ &= \frac{1}{l}\mathsf{SNR}^{1-\alpha+2\epsilon} \end{aligned}$$

thus $l \geq \mathsf{SNR}^{-2\alpha+2\epsilon}, \forall \epsilon > 0$, which proves the lemma.

Now it remains to prove that

$$\Pr[B \leq \mathsf{SNR}^{-(1-\alpha+2\epsilon)}] \geq \frac{1}{2}$$

and that

$$E[B|B \leq \mathsf{SNR}^{-(1-\alpha+2\epsilon)}] \doteq 1 \tag{61}$$

The first inequality can be directly obtained by using Markov inequality, with

$$\Pr[B \geq \mathsf{SNR}^{-(1-\alpha+2\epsilon)}] \leq \mathsf{SNR}^{(1-\alpha+2\epsilon)} \ll 1$$

(61) is more tricky. We prove that for $\beta = (1-\alpha+2\epsilon)$,

$$E[B|B \geq \mathsf{SNR}^{-\beta}] \Pr(B \geq \mathsf{SNR}^{-\beta}] < \frac{1}{2}$$

thus

$$E[B|B \leq \mathsf{SNR}^{-\beta}] \Pr(B \leq \mathsf{SNR}^{-\beta}] \geq \frac{1}{2}$$

since $E[B] = 1$.

To do that, we compute the expectation for different ranges of B. We first use (58)

$$\begin{aligned} \mathsf{SNR}^{1+\alpha} &\geq \mathsf{SNR} - E[\log(1+\mathsf{SNR}B)] \\ &= E[\mathsf{SNR}B - \log(1+\mathsf{SNR}B)] \\ &\geq \Pr(B \geq \mathsf{SNR}^{-\gamma}) E[\mathsf{SNR}B - \log(1+\mathsf{SNR}B)|B \geq \mathsf{SNR}^{-\gamma}] \end{aligned}$$

When $\gamma > 1$ and $B \geq \mathsf{SNR}^{-\gamma}$, $\mathsf{SNR}B \gg \log(1+\mathsf{SNR}B)$, thus we have

$$\Pr(B \geq \mathsf{SNR}^{-\gamma}) E[\mathsf{SNR}B|B \geq \mathsf{SNR}^{-\gamma}] \dot{\leq} \mathsf{SNR}^{1+\alpha}$$

and thus

$$\Pr(B \geq \mathsf{SNR}^{-\gamma}) E[B|B \geq \mathsf{SNR}^{-\gamma}] \dot{\leq} \mathsf{SNR}^{\alpha}$$

One can easily generalize this to the case $\gamma \geq 1$, using the fact that the probability that B lies in a neighborhood of SNR^{-1} is very small, thus we have

$$\Pr(B \geq \mathsf{SNR}^{-1})E[\mathsf{SNR}B|B \geq \mathsf{SNR}^{-1}] \ll 1 \tag{62}$$

Now consider for any $\gamma \leq 1$, again use (58),

$$\begin{aligned}
\mathsf{SNR}^{1+\alpha} &\geq \Pr[B \geq \mathsf{SNR}^{-\gamma}]E[\mathsf{SNR}B - \log(1+\mathsf{SNR}B)|B \geq \mathsf{SNR}^{-\gamma}] \\
&\geq \Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}]E[\mathsf{SNR}B - \log(1+\mathsf{SNR}B)|\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}] \\
&\geq \Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}]E\left[\frac{1}{2}(\mathsf{SNR}B)^2 - \frac{1}{3}(\mathsf{SNR}B)^3 \middle| \mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}\right] \\
&\dot{\geq} \Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}]E[(\mathsf{SNR}B)^2|\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}] \\
&\geq \Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}]\mathsf{SNR}^{2-2\gamma}.
\end{aligned}$$

Thus

$$\Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\gamma}] \leq \mathsf{SNR}^{2\gamma-(1-\alpha)} \tag{63}$$

Now for $\beta = (1-\alpha) + 2\epsilon < 1$, to compute

$$\Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\beta}]E[B|\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\beta}]$$

we break the range of B into many segments, $[\mathsf{SNR}^{-\beta}, \mathsf{SNR}^{-(\beta+\epsilon)})$, $[\mathsf{SNR}^{-(\beta+\epsilon)}, \mathsf{SNR}^{-(\beta+2\epsilon)})$, ... For each of the interval, we use (63), with $\gamma = \beta + k\epsilon$ for $k \geq 0$, to compute

$$\begin{aligned}
& E[B|B \in [\mathsf{SNR}^{-(\beta+k\epsilon)}, \mathsf{SNR}^{-(\beta+k\epsilon+\epsilon)})] \Pr[B \in [\mathsf{SNR}^{-(\beta+k\epsilon)}, \mathsf{SNR}^{-(\beta+k\epsilon+\epsilon)})] \\
\leq\ & \mathsf{SNR}^{-(\beta+k\epsilon+\epsilon)}\mathsf{SNR}^{2(\beta+k\epsilon)-(1-\alpha)} \\
=\ & \mathsf{SNR}^{(k-1)\epsilon}\mathsf{SNR}^{\beta-(1-\alpha)} \\
=\ & \mathsf{SNR}^{(k+1)\epsilon}
\end{aligned}$$

Thus,

$$\Pr[\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\beta}]E[B|\mathsf{SNR}^{-1} \geq B \geq \mathsf{SNR}^{-\beta}] \ll 1 \tag{64}$$

Combine (62) and (64), we proved (61), and completed the proof.

C Proof of Lemma 2

For convenience, we copy (12)

$$I_{tr}(\mathsf{SNR}) \geq \frac{l-1}{l}E\left[\log\left(1 + \frac{(1-\gamma)\mathsf{SNR}|\hat{h}|^2}{\sigma^2}\right)\right].$$

Notice that the term inside the expectation is the mutual information per symbol time in the training scheme, and is upper bounded by the SNR, thus dropping the factor $(l-1)/l$ yields an error less than $\mathsf{SNR}/l = o(\mathsf{SNR}^2)$ for $l \doteq \mathsf{SNR}^{-(1+2\alpha)}$.

$$I_{tr}(\mathsf{SNR}) \geq E\left[\log\left(1 + \frac{(1-\gamma)\mathsf{SNR}|\hat{h}|^2}{\sigma^2}\right)\right] - o(\mathsf{SNR}^2).$$

Using the inequality $\log(1+x) \geq x - \frac{1}{2}x^2$, we have

$$I_{tr}(\mathsf{SNR}) \geq f(\mathsf{SNR},\gamma) - \frac{1}{2}f(\mathsf{SNR},\gamma)^2 - o(\mathsf{SNR}^2),$$

where $f(\mathsf{SNR},\gamma)$ can be viewed as the effective signal to noise ratio for the training scheme.

$$\begin{aligned}
f(\mathsf{SNR},\gamma) &= \frac{(1-\gamma)\mathsf{SNR}E[|\hat{h}|^2]}{\sigma^2} \\
&= \frac{(1-\gamma)\mathsf{SNR}\cdot E[|\hat{h}|^2]}{1+(1-\gamma)\mathsf{SNR}\cdot E[|\tilde{h}|^2]} \\
&= \frac{(1-\gamma)\mathsf{SNR}\frac{E_{tr}}{1+E_{tr}}}{1+(1-\gamma)\mathsf{SNR}\frac{1}{1+E_{tr}}} \\
&= \frac{(1-\gamma)\mathsf{SNR}\gamma l\mathsf{SNR}}{1+\gamma l\mathsf{SNR}+(1-\gamma)\mathsf{SNR}}.
\end{aligned}$$

Maximizing $f(\mathsf{SNR},\gamma)$ we have that, for

$$\gamma^* = \frac{\sqrt{1+\frac{(l-1)\mathsf{SNR}}{1+\mathsf{SNR}}}-1}{\frac{(l-1)\mathsf{SNR}}{1+\mathsf{SNR}}},$$

$$f(\mathsf{SNR},\gamma^*) = \frac{l\mathsf{SNR}^2}{1+\mathsf{SNR}}\gamma^*.$$

It can be easily verified that replacing $(1+\mathsf{SNR})$ by 1 and $l-1$ by l causes an error of order no larger than $O(\mathsf{SNR}^2)$, thus

$$\begin{aligned}
f(\mathsf{SNR},\gamma^*) &= l\mathsf{SNR}^2\left(\frac{\sqrt{1+l\mathsf{SNR}}-1}{l\mathsf{SNR}}\right)^2 + O(\mathsf{SNR}^2) \\
&= l\mathsf{SNR}^2\frac{(1+l\mathsf{SNR}-2\sqrt{1+l\mathsf{SNR}}+1)}{(l\mathsf{SNR})^2} + O(\mathsf{SNR}^2) \\
&= \mathsf{SNR} - \frac{2\sqrt{1+l\mathsf{SNR}}}{l} + \frac{2}{l} + O(\mathsf{SNR}^2) \\
&= \mathsf{SNR} - \frac{2\sqrt{1+l\mathsf{SNR}}}{l} + O(\mathsf{SNR}^{1+2\alpha}) + O(\mathsf{SNR}^2)
\end{aligned}$$

for $l \doteq \mathsf{SNR}^{-(1+2\alpha)}$. For the same l, we have that

$$\frac{\sqrt{1+l\mathsf{SNR}}}{l} \doteq \mathsf{SNR}^{1+\alpha}.$$

Notice that neither $O(\mathsf{SNR}^{1+2\alpha})$ nor $O(\mathsf{SNR}^2)$ is larger than $O(\mathsf{SNR}^{1+\alpha})$. Thus, for $\alpha \in [0,1]$, we have

$$f(\mathsf{SNR},\gamma^*) \doteq \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})$$

and

$$\begin{aligned}
I_{tr}(\mathsf{SNR}) &\dot{\geq} f(\mathsf{SNR},\gamma^*) - \frac{1}{2}f(\mathsf{SNR},\gamma^*) + O(\mathsf{SNR}^2) \\
&\doteq \mathsf{SNR} - O(\mathsf{SNR}^{1+\alpha})
\end{aligned}$$

The desired data rate is thus achievable. ◦

References

[1] S. Verdu, "Spectral efficiency in the wideband regime," *IEEE Transactions on Information Theory*, vol. 48, pp. 1319 –1343, June 2002.

[2] R. S. Kennedy, *Fading Dispersive Communication Channels*. New York: Wiley-Interscience, 1969.

[3] E. Telatar and D. Tse, "Capacity and mutual information of wideband multipath fading channels," *IEEE Transactions on Information Theory*, vol. 46, pp. 1384 –1400, July 2000.

[4] E. Telatar, "Coding and multiaccess for the energy limited rayleigh fading channel," Master's thesis, Massachusetts Institute of Technology, 1988.

[5] M. Medard and R. Gallager, "Bandwidth scaling for fading multipath channels," *IEEE Transactions on Information Theory*, vol. 48, pp. 840–52, April 2002.

[6] V. G. Subramanian and B. Hajek, "Broad-band fading channels: signal burstiness and capacity," *IEEE Transactions on Information Theory*, vol. 48, pp. 908–27, April 2002.

[7] L. Zheng and D. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE Transactions on Information Theory*, vol. 48, pp. 359–83, February 2002.

[8] V. Sethuraman and B. Hajek, "Capacity per unit energy of fading channels with a peak constraint," *IEEE Transactions on Information Theory*, vol. 51, September 2005.

[9] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?," *IEEE Transactions on Information Theory*, vol. 49, pp. 951–63, April 2003.

[10] M. Médard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Transactions on Information Theory*, vol. 46, May 2000.

[11] T.L.Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in rayleigh flat fading," *IEEE Transactions on Information Theorey*, vol. 45, pp. 139–57, January 1999.

[12] R. Etkin and D. Tse, "Degrees of freedom in underspread mimo fading channels," *Submitted to IEEE Transactions on Information Theory*, 2003.

[13] S. Ray, M. Médard, L. Zheng, and J. Abounadi, "On the sublinear behavior of mimo channel capacity at low snr," in *International Symposium on Information Theory and its Applications (ISITA)*, 2004.

[14] I. Abou-Fayçal, M. Trott, and S. Shamai, "The capacity of discrete-time memoryless rayleigh-fading channels," *IEEE Transactions on Information Theory*, vol. 47, pp. 1290–301, May 2001.