Capacity of Sparse Wideband Channels with Partial Channel Feedback

Gautham Hariharan†, Vasanthan Raghavan†*, Akbar M. Sayeed

Abstract

We study the non-coherent wideband capacity of multipath channels with limited feedback. This work builds on recent results that have established the very significant capacity gains achievable in the wideband/low-SNR regime, when there is perfect channel state information (CSI) at the transmitter. Furthermore, this benchmark gain can be obtained with just 1-bit of CSI at the transmitter about each channel coefficient. However, the capacity achieving signals are peaky; they have large instantaneous transmit power, especially in the non-coherent scenario. Signal peakiness is related to channel coherence and, in contrast to the prevalent assumption of rich multipath, we investigate the ergodic capacity of sparse multipath channels. Sparsity naturally leads to coherence in time and frequency. With perfect receiver CSI, it is shown that limited feedback, even with an instantaneous power constraint, is sufficient to achieve the benchmark capacity under an average power constraint. Our analysis reveals the benefits of channel sparsity in the non-coherent scenario, where we employ a training-based communication scheme. With an average power constraint, it is shown that the benchmark is achievable, provided the channel coherence scales at a sufficiently fast rate with signal space dimension. Furthermore, even when we impose an instantaneous power constraint, we show that the benchmark is still attainable for a sparse channel. We present rules of thumb on choosing the signaling parameters as a function of the channel parameters so that the full benefits of sparsity can be realized.

I. INTRODUCTION

Emerging applications in ultra-wideband (UWB) communication systems and sensor networks have renewed the focus on investigating the fundamental performance limits in the wideband/low

This work was supported in part by the National Science Foundation through grant #CCF-0431088. The first and third authors are with the University of Wisconsin-Madison, Madison WI 53706, USA. The second author was with the University of Wisconsin-Madison, Madison WI 53706, USA while a part of this work was done. He is currently with the University of Illinois, Urbana-Champaign, Urbana IL 61801, USA. E-mail: gauthamh@cae.wisc.edu, vasanthan_raghavan@ieee.org, akbar@engr.wisc.edu. †Equal contribution by first and second authors. *Corresponding author.
SNR regime. In the wideband realm, the impact of channel state information (CSI), particularly the non-coherent regime, when there is no CSI at the receiver a priori merits special attention. The key contributions in this regard include the seminal work of [1], [2], [3], where, from a capacity perspective, the necessity of peaky signaling schemes in achieving wideband optimality is illustrated. However these results were all derived based on an implicit assumption of rich multipath, where channel variability or uncertainty is fixed as bandwidth scales. Recent work by Zheng et al [4] has emphasized the crucial role of channel coherence in the low SNR regime and the importance of channel learning schemes that can bridge the gap between the coherent and non-coherent extremes.

Motivated by these works as well as by recent measurement campaigns for UWB [5], [6], we recently introduced the notion of multipath sparsity as a physical source of channel coherence and proposed a channel modeling framework in [7] that captures the effect of sparsity in delay and Doppler. A key implication of sparsity is that the degrees of freedom (DoF) in the wideband channel scale sub-linearly with the signal space dimension (time-bandwidth product). Based on this model, we investigated the ergodic capacity of training-based communication schemes. The analysis in [7] reveals the impact of channel sparsity on channel coherence scaling and the role played by sparsity in reducing/eliminating peaky signaling to achieve wideband capacity.

Building on the results in [7], the focus of this paper is on the impact of feedback on the ergodic capacity of sparse wideband channels. Although the effect of transmitter CSI on the capacity of fading channels has been explored in great detail (see for example [8], [9], [10] and references therein), it is only recently [11], [12] that the impact of feedback in the wideband, non-coherent regime has received attention. In particular, it is shown in [11] that at low SNR, the capacity gain with perfect transmitter and receiver CSI (over the case when there is only receiver CSI) equals $\log \left( \frac{1}{\text{SNR}} \right)$ and is obtained with the well known water-filling solution [8]. More interestingly, it is shown that this gain can actually be achieved with limited feedback; when there is just 1 bit of CSI about each channel coefficient at the transmitter and on-off signaling is employed. However, for both water-filling and on-off signaling, the capacity achieving input tends to be peaky (or) bursty in time, leading to a high peak-to-average (PAR) power ratio, and difficulties from an implementation standpoint. The peakiness aspect is much more relevant in the no receiver CSI (non-coherent) scenario, for which [11] proposes peaky training and communication. The necessity of peakiness is tied in with the need to achieve perfect channel
learning at the receiver.

In this paper, we analyze the capacity of sparse wideband channels with limited feedback as in [11], [12]. We start with a brief description in Section II of the sparse channel model [7]. Our main focus is on the case when there is no receiver CSI \textit{a priori} and training-based communication is employed. As a benchmark, we first analyze in Section III, the perfect receiver CSI scenario. The analysis is performed with both an average or long-term power constraint as well as with an instantaneous or short-term constraint. We restrict our attention to \textit{causal} signaling schemes that are realizable in practice. With an average power constraint, we show that an optimal threshold given by $h_t = \lambda \log \left( \frac{1}{\text{SNR}} \right)$ for any $\lambda \in (0, 1)$ directly provides a measure of capacity which behaves as $(1 + h_t) \text{SNR}$ in the wideband limit. Thus with $\lambda \to 1$, we achieve the perfect transmitter CSI capacity, which is the benchmark for all limited feedback schemes. We also show that the same capacity can be achieved even when there is an instantaneous power constraint. We quantify the necessary conditions to attain this capacity when the channel is rich and when it is sparse.

In Section IV, we discuss the capacity of training-based communication schemes. With an average power constraint, it is shown that as long as the channel coherence dimension $N_c$ scales with SNR as $N_c = \frac{1}{\text{SNR}}$ for some $\mu > 1$, the capacity of the training-based scheme converges to the coherent capacity in the wideband limit. This condition is achievable only when the channel is sparse and we provide guidelines on choosing the signal space parameters (signaling duration, bandwidth and transmit power) so that $\mu > 1$ is satisfied. The critical role of channel sparsity is further revealed when we impose an instantaneous power constraint. We present two necessary conditions: (i) the scaling requirement on $N_c$ and (ii) the scaling of the DoF, $D$, with SNR so that the capacity of the training-based communication scheme with an instantaneous power constraint converges to the coherent capacity with perfect transmitter and receiver CSI. We conclude with a discussion on the achievability of these two conditions under different assumptions of rich and sparse multipath.

\section{System Model}

In this section, we briefly summarize the model developed in [7] for sparse multipath channels. Our results are based on an orthogonal short-time Fourier (STF) signaling framework [13], [14] that naturally relates multipath sparsity in delay-Doppler to coherence in time and frequency.
A. Sparse Multipath Channel Modeling

A physical discrete multipath channel can be modeled as

\[ h(\tau, \nu) = \sum_n \beta_n \delta(\tau - \tau_n) \delta(\nu - \nu_n) \quad y(t) = \sum_n \beta_n y(t - \tau_n) e^{j2\pi \nu_n t} + w(t) \]

where \( h(\tau, \nu) \) is the delay-Doppler spreading function of the channel, \( \beta_n, \tau_n \in [0, T_m] \) and \( \nu_n \in [-W_d/2, W_d/2] \) denote the complex path gain, delay and Doppler shift associated with the \( n \)-th path. \( T_m \) and \( W_d \) denote the delay and Doppler spreads, respectively. The quantities \( x(t), y(t) \) and \( w(t) \) denote the transmitted, received and additive white Gaussian noise waveforms, respectively. Throughout this paper, we assume an underspread channel: \( T_m W_d \ll 1 \).

We use a virtual representation [15], [16] of the physical model in (1) that captures the channel characteristics in terms of resolvable paths and greatly facilitates system analysis from a communication-theoretic perspective. The virtual representation uniformly samples the multipath in delay and Doppler at a resolution commensurate with signaling bandwidth \( W \) and signaling duration \( T \), respectively [15], [16]:

\[ y(t) = \sum_{\ell=0}^{[T_m W]} \sum_{m=-[T_W/2]}^{[TW_d/2]} h_{\ell,m} x(t - \ell/W) e^{j2\pi mt/T} + w(t) \]

\[ h_{\ell,m} \approx \sum_{n \in S_{\ell,\ell} \cap S_{\nu,m}} \beta_n. \]

The sampled representation (2) is linear and is characterized by the virtual delay-Doppler channel coefficients \( \{h_{\ell,m}\} \). Each \( h_{r,m} \) consists of the sum of gains of all paths whose delays and Doppler shifts lie within the \((\ell, m)\)-th delay-Doppler resolution bin \( S_{\ell,\ell} \cap S_{\nu,m} \) of size \( \Delta \tau \times \Delta \nu \), \( \Delta \tau = \frac{1}{WT} \), \( \Delta \nu = \frac{1}{T} \) as shown in Fig. 1(a). Distinct \( h_{\ell,m} \)'s correspond to approximately disjoint subsets of paths and are hence approximately statistically independent. In this work, we assume that the channel coefficients \( \{h_{\ell,m}\} \) are perfectly independent. We also assume Rayleigh fading in which \( \{h_{\ell,m}\} \) are zero-mean Gaussian random variables.

Let \( D \) denote the number of dominant\(^1\) non-zero channel coefficients; it reflects the (dominant) statistically independent degrees of freedom (DoF) in the channel and also signifies the delay-Doppler diversity afforded by the channel [15]. We decompose \( D \) as \( D = D_T D_W \) where \( D_T \)

\(^1\)For which \( \Psi(\ell, m) > \gamma \) for some prescribed threshold \( \gamma > 0 \). The choice of \( \gamma \) is beyond the scope of this paper.
denotes the Doppler/time diversity and $D_W$ the frequency/delay diversity. The channel DoF or delay-Doppler diversity is bounded as:

$$D = D_T D_W \leq D_{\text{max}} = D_{T,\text{max}} D_{W,\text{max}}, \quad D_{T,\text{max}} = \lfloor TW_d \rfloor, \quad D_{W,\text{max}} = \lfloor T_m W \rfloor$$

(4)

where $D_{T,\text{max}}$ denotes the maximum Doppler diversity and $D_{W,\text{max}}$ denotes maximum delay diversity. Note that $D_{T,\text{max}}$ and $D_{W,\text{max}}$ increase linearly with $T$ and $W$, respectively, and represent a rich multipath environment in which each resolution bin in Fig. 1(a) corresponds to a dominant channel coefficient.

However, there is growing experimental evidence [5], [17], [6] that the dominant channel coefficients get sparser in delay as the bandwidth increases. Furthermore, we are also interested in modeling scenarios with Doppler effects, due to motion. In such cases, as we consider large bandwidths and/or long signaling durations, the resolution of paths in both delay and Doppler domains gets finer, leading to the scenario in Fig. 1(a) where the delay-Doppler resolution bins are sparsely populated with paths, i.e. $D < D_{\text{max}}$.

We model multipath sparsity with a sub-linear scaling in $D_T$ and $D_W$ with $T$ and $W$:

$$D_W \sim g_1(W), \quad D_T \sim g_2(T)$$

(5)

where $g_1$ and $g_2$ are arbitrary sub-linear functions. A concrete example that we will use later in the analysis is a power-law scaling of the form:

$$D_T = \frac{T^{\delta_1}}{W_d^{\delta_1}}, \quad D_W = \frac{W^{\delta_2}}{T_m^{\delta_2}}$$

(6)

for $\delta_1, \delta_2 \in (0, 1)$. Note that (4) and (5) imply that the total number of delay-Doppler DoF, $D = D_T D_W$, scales sub-linearly with the signal space dimension $N = TW$ in sparse multipath.

Remark 1: With perfect CSI at the receiver, the parameter $D$ denotes the delay-Doppler diversity afforded by the channel, whereas with no CSI, it reflects the level of channel uncertainty; the number of channel parameters that need to be estimated at the receiver for coherent processing.

B. Orthogonal Short-Time Fourier Signaling

We consider signaling using an orthonormal short-time Fourier (STF) basis [13], [14] that is a natural generalization of orthogonal frequency-division multiplexing (OFDM) for time-

\[ ^2 \text{STF signaling can be considered as OFDM signaling over a block of OFDM symbol periods and with an appropriately chosen OFDM symbol duration.} \]
varying channels. An orthogonal STF basis \( \{ \phi_{\ell m}(t) \} \) for the signal space is generated from a fixed prototype waveform \( g(t) \) via time and frequency shifts: \( \phi_{\ell m}(t) = g(t - \ell T_o)e^{j2\pi W_o t} \), where \( T_o W_o = 1, \quad \ell = 0, \cdots, N_T - 1, \quad m = 0, \cdots, N_W - 1 \) and \( N = N_T N_W = TW \) with \( N_T = T/T_o, N_W = W/W_o \). The transmitted signal can be represented as

\[
x(t) = \sum_{\ell=0}^{N_T-1} \sum_{m=0}^{N_W-1} x_{\ell m} \phi_{\ell m}(t) \quad 0 \leq t \leq T
\]  

(7)

where \( \{ x_{\ell m} \} \) represent the \( N \) transmitted symbols that are modulated onto the STF basis waveforms. The received signal is projected onto the STF basis waveforms to yield

\[
y_{\ell m} = \langle y, \phi_{\ell m} \rangle = \sum_{\ell', m'} h_{\ell m, \ell' m'} x_{\ell' m'} + w_{\ell m}.
\]  

(8)

We can represent the system using an \( N \)-dimensional matrix equation

\[
y = Hx + w
\]  

(9)

where \( w \) represents the additive noise vector whose entries are i.i.d. \( CN(0, 1) \). The \( N \times N \) matrix \( H \) consists of the channel coefficients \( \{ h_{\ell m, \ell' m'} \} \) in (8). We assume that the input symbols that form the transmit codeword \( x \) satisfy an average power constraint

\[
\frac{1}{T} \cdot E[\|x\|^2] \leq P
\]  

(10)

Since there are \( N = TW \) symbols per codeword, we define the parameter SNR (transmit energy per modulated symbol) for a given average transmit power \( P \) as \( \text{SNR} = \frac{TP}{TW} = \frac{P}{W} \). In this work, the focus is on the wideband regime where \( \text{SNR} \to 0 \) as \( W \to \infty \) for a fixed \( P \).

For sufficiently underspread channels, the parameters \( T_o \) and \( W_o \) can be matched to \( T_m \) and \( W_d \) so that the STF basis waveforms serve as approximate eigenfunctions of the channel [14], [13]; that is, (8) simplifies to \( y_{\ell m} \approx h_{\ell m} x_{\ell m} + w_{\ell m} \). Thus the channel matrix \( H \) is approximately diagonal. In this work, we assume that \( H \) is exactly diagonal; that is,

\[
H = \text{diag}\left[ h_{11}, \cdots h_{1N_c}, \ h_{21}, \cdots h_{2N_c}, \ \cdots, \ h_{D1}, \cdots h_{DN_c} \right].
\]  

(11)

The diagonal entries of \( H \) in (11) admit an intuitive block fading interpretation in terms of time-frequency coherence subspaces [13] illustrated in Fig. 1(b). The signal space is partitioned

\[\text{Subspace 1} \quad h_{11} \cdots h_{1N_c} \quad \text{Subspace 2} \quad h_{21} \cdots h_{2N_c} \quad \cdots \quad \text{Subspace D} \quad h_{D1} \cdots h_{DN_c}.\]
as \( N = TW = N_c D \) where \( D \) represents the number of statistically independent time-frequency coherence subspaces, reflecting the DoF in the channel, and \( N_c \) represents the dimension of each coherence subspace, which we refer to as the **coherence dimension**. In the block fading model in (11), the channel coefficients over the \( i \)-th coherence subspace \( h_{i1}, \ldots, h_{iN_c} \) are assumed to be identical (denoted by \( h_i \)), whereas the coefficients across different coherence subspaces are independent and identically distributed. Thus, the channel is characterized by the \( D \) distinct STF channel coefficients, \( \{ h_i \} \), that are i.i.d. zero-mean Gaussian random variables (Rayleigh fading) with (normalized) variance equal to \( \mathbb{E}[|h_i|^2] = \sum_n \mathbb{E}[|\beta_n|^2] = 1 \) [13].

Using the DoF scaling for sparse channels in (5), the scaling behavior for the coherence dimension can be computed as

\[
W_{coh} = \frac{W}{D_W} \sim f_1(W), \quad T_{coh} = \frac{T}{D_T} \sim f_2(T), \quad N_c = W_{coh} T_{coh} \sim f_1(W) f_2(T) \tag{12}
\]

where \( T_{coh} \) is the **coherence time** and \( W_{coh} \) is the **coherence bandwidth** of the channel, as illustrated in Fig. 1(b). As a consequence of the sub-linearity of \( g_1 \) and \( g_2 \) in (5), \( f_1 \) and \( f_2 \) are also sub-linear. In particular, corresponding to the power-law scaling in (6), we obtain

\[
T_{coh} = \frac{T^{1-\delta_1}}{W_d^{\delta_2}}, \quad W_{coh} = \frac{W^{1-\delta_2}}{T_m^{\delta_1}} \tag{13}
\]

**Remark 2:** Note that when the channel is sparse, both \( N_c \) and \( D \) increase sub-linearly with \( N \), whereas when the channel is rich, \( D \) scales linearly with \( N \), while \( N_c \) is fixed.

In this work, our focus is on computing non-coherent channel capacity with feedback and as we will see later in Sections III and IV, capacity turns out to be a function only of the parameters \( N_c \) and SNR. Thus, in order to analyze the low SNR asymptotics, the following relation between \( N_c \) and SNR \((= P/W)\) plays a key role:

\[
N_c = \frac{1}{\text{SNR}^\mu}, \quad \mu > 0, \tag{14}
\]

where the parameter \( \mu \) reflects the level of channel coherence. We will revisit (14) and discuss its achievability and implications in Section IV.

### III. Capacity with Perfect Receiver CSI and Limited Feedback

The ergodic capacity of the wideband fading channel in (11) can be calculated under different assumptions on receiver/transmitter CSI. We assume throughout this paper that both the transmitter and the receiver have statistical CSI - knowledge of \( T_m, W_d, g_1, g_2, f_1 \) and \( f_2 \) so that the
scaling in $D$ and $N_c$ is known. On the one extreme, with perfect receiver CSI and no transmitter CSI (no feedback), the coherent capacity per dimension (in b/s/Hz) equals 

$$C_{coh,0}(\text{SNR}) = \frac{\text{sup}_{Q: \text{Tr}(Q) \leq TP} E \left[ \log_2 \det \left( I_{N_c,D} + HQH^H \right) \right]}{N_cD} \tag{15}$$

The optimization is over the set of $N_cD$-dimensional positive definite input covariance matrices $Q = E[xx^H]$ satisfying the average power constraint in (10). Due to the diagonal nature of $H$ in (11), the optimal $Q$ is also diagonal. Furthermore, with no transmitter CSI, the uniform power allocation $Q = \frac{TP}{N_cD} I_{N_c,D} = \text{SNR} I_{N_c,D}$ achieves this optimum. The corresponding capacity in the limit of low SNR equals [7]

$$C_{coh,0}(\text{SNR}) \approx \log_2(e) \cdot [\text{SNR} - \text{SNR}^2] \tag{16}$$

On the other extreme is the case of perfect transmitter CSI, where the receiver feeds back the channel coefficients, $\{h_i\}_{i=1}^D$, exactly to the transmitter, corresponding to the $D$ independent coherence subspaces. The optimum transmitter power allocation policy in this case is water-filling [8] over the different coherence subspaces. In the low SNR extreme, it is shown in [11] that the water-filling threshold is given by $h_w \sim \log \left( \frac{1}{\text{SNR}} \right)$ and the capacity with perfect transmitter CSI scales as $\log \left( \frac{1}{\text{SNR}} \right) \cdot \text{SNR}$. That is, the capacity gain over the receiver only CSI case is directly proportional to the water-filling threshold, $\log \left( \frac{1}{\text{SNR}} \right)$ and this gain serves as a benchmark for all limited feedback schemes. More interestingly, it is shown in [11] that this maximum capacity gain can be achieved with just one bit of feedback per channel coefficient.

In this case of limited feedback, both the transmitter and the receiver have a priori knowledge of a common threshold $h_t$. The receiver compares the channel strength ($|h_i|^2$, $i = 1, 2, \ldots, D$) in each coherence subspace with $h_t$, and feeds back

$$b_i = \begin{cases} 1 & \text{if } |h_i|^2 \geq h_t \\ 0 & \text{if } |h_i|^2 < h_t. \end{cases} \tag{17}$$

At the transmitter, power allocation is uniform across the coherence subspaces for which $b_i = 1$ and no power is allocated to those subspaces for which $b_i = 0$. Conditioned on the $\{b_i\}_{i=1}^D$, the input power allocation, which we still denote by $Q$ with a little abuse of notation, takes the form

$$Q = \text{diag} \left( |x_1|^2, |x_2|^2, \ldots, |x_N|^2 \right) = \text{diag} \left( q_1, \ldots, q_1, q_2, \ldots, q_2, \ldots, q_D, \ldots, q_D \right)$$

where $q_i = P_0 \cdot \chi(|h_i|^2 \geq h_t). \tag{18}$
The choice of $P_0$ depends on the nature of transmit power constraint and also on the nature of feedback. Let $D_{\text{eff}}$ denote the number of active subspaces, those which exceed the threshold $h_t$. We have $D_{\text{eff}} = \sum_{i=1}^{D} \chi(|h_i|^2 \geq h_t)$ and hence

$$\mathbb{E}[D_{\text{eff}}] \overset{(a)}{=} D \mathbb{E}\left[\chi(|h|^2 \geq h_t)\right] \overset{(b)}{=} De^{-h_t}$$

(19)

where (a) is due to the fact that $\{h_i\}_{i=1}^{D}$ are i.i.d. and (b) is because for a standard Gaussian $h_i$, $\mathbb{E}\left[\chi(|h_i|^2 \geq h_t)\right] = \text{Pr}(|h_i|^2 \geq h_t) = e^{-h_t}$.

If we assume knowledge of all $\{b_i\}_{i=1}^{D}$ at the start of each codeword, then we can uniformly divide power among the active subspaces. That is

$$P_{0,nc} = \frac{TP}{N_c D_{\text{eff}}}.$$  

(20)

The capacity with this power allocation, using (15) and (18) equals

$$C_{\text{coh},1,LT}(\text{SNR}) = \max_{h_t} \frac{1}{D} \sum_{i=1}^{D} \mathbb{E}\left[\log_2 \left(1 + \frac{TP}{N_c D_{\text{eff}} h_t} \cdot |h_i|^2\right) \chi(|h_i|^2 \geq h_t)\right].$$  

(21)

The above power allocation satisfies the power constraint instantaneously as well as on average. To see this, note that

$$P_{\text{inst},nc} = \frac{1}{T} \|x\|^2 = N_c \sum_{i=1}^{D} q_i = \frac{N_c}{T} \sum_{i=1}^{D} \frac{TP}{N_c D_{\text{eff}}} \chi(|h_i|^2 \geq h_t) = P$$

(22)

and clearly $\mathbb{E}[P_{\text{inst},nc}] \leq P$ as well. However, the above scheme is not realizable in practice since it is not causal. This is especially relevant in the more practical scenario when the receiver estimates the channel and feeds back $\{b_i\}_{i=1}^{D}$ based on the estimated channel coefficients. This motivates us to instead consider a causal power allocation scheme. From (18), we have

$$\mathbb{E} [\|x\|^2] = N_c \sum_{i=1}^{D} \mathbb{E}[g_i] = N_c \sum_{i=1}^{D} P_0 \cdot \mathbb{E}\left[\chi(|h_i|^2 \geq h_t)\right] \overset{(a)}{=} N_c P_0 \mathbb{E}[D_{\text{eff}}]$$

(23)

where (a) follows from (19). Thus to satisfy $\mathbb{E} [\|x\|^2] \leq TP$, the power allocation for this causal scheme is given by

$$P_{0,c} = \frac{TP}{N_c \mathbb{E}[D_{\text{eff}}]} = \frac{TP}{N_c De^{-h_t}}$$

(24)

and the capacity in this case is given by

$$\hat{C}_{\text{coh},1,LT}(\text{SNR}) = \max_{h_t} \frac{1}{D} \sum_{i=1}^{D} \mathbb{E}\left[\log_2 \left(1 + \frac{TP}{N_c De^{-h_t}} |h_i|^2\right) \chi(|h_i|^2 \geq h_t)\right].$$

(25)
While both schemes satisfy the average power constraint, the causal scheme can have a large instantaneous power since
\[
P_{\text{inst},c} = \frac{1}{T} \|x\|^2 = \frac{N_c}{T} \sum_{i=1}^{D} \frac{TP}{N_c D e^{-h_t}} \chi(|h_i|^2 \geq h_t) = \left( \frac{D_{\text{eff}}}{D} \right) P e^{h_t}
\] (26)
and hence \(P_{\text{inst},c} \in [0, \infty)\). We will address this important issue later, but first, we solve the capacity problem in (25), considering only the causal average power constraint in (24).

A. Capacity with Average Power Constraint

The following theorem establishes that a threshold of the form \(h_t \sim \lambda \log \left( \frac{1}{\text{SNR}} \right)\) for some \(\lambda \in (0, 1)\) provides the solution to (25). Further, we also show that the rate achievable with this causal scheme is asymptotically (in low SNR) the same as the non-causal capacity in (21).

**Theorem 1:** Given any \(\lambda \in (0, 1)\), a causal signaling scheme satisfying the average power constraint achieves \(\hat{C}_{LB} \leq \hat{C}_{\text{coh,1,LT}(\text{SNR})} \leq \hat{C}_{UB}\) where
\[
\hat{C}_{UB} = \text{SNR}^\lambda \cdot \left[ \log_2 \left( 1 + \lambda \text{SNR}^{1-\lambda} \log \left( \frac{1}{\text{SNR}} \right) \right) + \log_2 \left( 1 + \frac{\text{SNR}^{1-\lambda}}{1 + \lambda \text{SNR}^{1-\lambda} \log \left( \frac{1}{\text{SNR}} \right)} \right) \right]
\] (27)
\[
\hat{C}_{LB} = \text{SNR}^\lambda \cdot \left[ \log_2 \left( 1 + \lambda \text{SNR}^{1-\lambda} \log \left( \frac{1}{\text{SNR}} \right) \right) + \frac{1}{2} \log_2 \left( 1 + \frac{2^{\text{SNR}^{1-\lambda}}}{1 + 2 \lambda \text{SNR}^{1-\lambda} \log \left( \frac{1}{\text{SNR}} \right)} \right) \right]
\] (28)

The above rate is achieved by using an optimal threshold satisfying
\[
\lim_{\text{SNR} \to 0} \frac{h_t}{\lambda \log \left( \frac{1}{\text{SNR}} \right)} = 1.
\] (29)

Further, \(\hat{C}_{\text{coh,1,LT}(\text{SNR})}\) is a tight bound to \(C_{\text{coh,1,LT}(\text{SNR})}\) and for all \(\lambda \in (0, 1)\), we have
\[
\lim_{\text{SNR} \to 0} \frac{|C_{\text{coh,1,LT}(\text{SNR})} - \hat{C}_{\text{coh,1,LT}(\text{SNR})}|}{C_{\text{coh,1,LT}(\text{SNR})}} = 0.
\] (30)

**Proof:** We start with (25)
\[
\hat{C}_{\text{coh,1,LT}(\text{SNR})} = \max_{h_t} \frac{1}{D} \sum_{i=1}^{D} \mathbb{E} \left[ \log_2 \left( 1 + \frac{TP}{N_c D e^{-h_t}} |h_i|^2 \right) \chi(|h_i|^2 \geq h_t) \right]
\]
\[= (a) \mathbb{E} \left[ \log_2 \left( 1 + \text{SNR} e^{h_t} |h|^2 \right) \chi(|h|^2 \geq h_t) \right]
\] (31)
where (a) follows from the fact that \(\{h_i\}\) are i.i.d. \(CN(0, 1)\). In Appendix A, we show that \(\hat{C}_{\text{coh,1,LT}(\text{SNR})}\) is a tight approximation to \(C_{\text{coh,1,LT}(\text{SNR})}\) and satisfies (30) for any choice of \(h_t\).
Computing the expectation in (31) using [18, 4.337(1), p. 574] and defining \( \alpha = \frac{1 + \text{SNR} h_t e^{ht}}{\text{SNR} e^{ht}} \)
for convenience, we have

\[
\hat{C}_{\text{coh,1,LT}}(\text{SNR}) = e^{-ht} \cdot \left[ \log_2 \left( 1 + \text{SNR} h_t e^{ht} \right) + \log_2(e) \cdot \exp\left( \alpha \int_{\alpha}^{\infty} \frac{e^{-t}}{t} \, dt \right) \right] = e^{-ht} \cdot \left[ \log_2 \left( 1 + \text{SNR} h_t e^{ht} \right) + \log_2(e) \cdot \nu_\alpha \right]
\]

(32)

where we define

\[
\nu_\alpha = \exp\left( \alpha \int_{\alpha}^{\infty} \frac{e^{-t}}{t} \, dt \right).
\]

(33)

Furthermore, in the limit of \( \alpha \to \infty \), we have the following bounds to \( \nu_\alpha \) [19, 5.1.20, p. 229]:

\[
\frac{1}{2} \log_2 \left( 1 + \frac{2}{\alpha} \right) \leq \nu_\alpha \leq \log_2 \left( 1 + \frac{1}{\alpha} \right).
\]

(34)

The choice of \( h_t \) that maximizes (32) is obtained by setting its derivative to zero and satisfies

\[
\Delta \equiv 1 - \log_2 \left( 1 + \text{SNR} h_t e^{ht} \right) - \frac{1}{\text{SNR} e^{ht}} \cdot \nu_\alpha = 0.
\]

(35)

Now if \( h_t \) is such that \( \lim_{\text{SNR} \to 0} \frac{h_t}{\lambda \log\left( \frac{1}{\text{SNR}} \right)} = 1 \) for some \( \lambda \in (0, 1) \), then as \( \text{SNR} \to 0 \), we have \( \text{SNR} h_t e^{ht} \to 0 \) and \( \alpha \to \infty \). Thus using (34), \( \nu_\alpha \approx \frac{1}{\alpha} \). Using this in (35), we have

\[
\frac{1}{\text{SNR} e^{ht}} \cdot \nu_\alpha \approx \frac{1}{1 + \text{SNR} h_t e^{ht}} \to 1.
\]

Therefore, with the choice of \( h_t \) as in (29), it follows that as \( \text{SNR} \to 0 \), \( \Delta \to 0 \). Substituting this choice of \( h_t \) in (32) and using the upper and lower bounds on \( \nu_\alpha \) in (34), we obtain the bounds in (27) and (28).

We have the following corollary that is an implication of Theorem 1.

**Corollary 1:** The capacity gain for the \( D \)-bit feedback, causal power allocation scheme over the receiver CSI only capacity in (16) satisfies

\[
\lim_{\text{SNR} \to 0} \frac{\hat{C}_{\text{coh,1,LT}}(\text{SNR})}{C_{\text{coh,0}}(\text{SNR})} = (1 + h_t) = \left( 1 + \lambda \log\left( \frac{1}{\text{SNR}} \right) \right).
\]

(36)

**Proof:** By performing a Taylor series expansion of the upper and lower bounds in (27) and (28), we note that they are equal up to a first-order and obtain, \( \hat{C}_{\text{coh,1,LT}}(\text{SNR}) = [1 + \lambda \log\left( \frac{1}{\text{SNR}} \right)] \text{SNR} = (1 + h_t) \text{SNR} \). On the other hand, with only receiver CSI, we have from (16), \( C_{\text{coh,0}}(\text{SNR}) = \text{SNR} \). Thus the desired result follows.

**Remark 3:** The capacity gain due to feedback is directly proportional to \( h_t \) and depends on the choice of \( \lambda \in (0, 1) \), which is a free parameter to control. Thus the highest gain is obtained by choosing \( \lambda \to 1 \), and equals the perfect CSI benchmark.

We now revert our attention back to the instantaneous transmit power described in (26). Note that as \( D \to \infty \), \( P_{\text{inst,c}} \to P \) as a consequence of the law of large numbers. However, for any
large but finite $D$, $P_{\text{inst,c}}$ may be much larger than $P$. This is a serious issue in practical systems that typically operate with peak power limitations. Thus it is important to analyze the impact of constraints on the instantaneous power in (26), as discussed next.

**B. Capacity with Instantaneous Power Constraint**

In addition to the average power constraint, we impose an additional constraint on the instantaneous transmit power of the form

$$P_{\text{inst,c}} \leq AP$$  \hspace{1cm} (37)

where $A > 1$ and finite. With this constraint, we would like to calculate the capacity, $\hat{C}_{\text{coh,1,ST}}(\text{SNR})$, of the causal signaling scheme. We are particularly interested in exploring conditions under which we do not take a hit in capacity and $\hat{C}_{\text{coh,1,ST}}(\text{SNR}) \approx \hat{C}_{\text{coh,1,LT}}(\text{SNR})$. To this end, we employ the following power allocation scheme

$$Q = \text{diag} \left( |x_1|^2, |x_2|^2, \ldots, |x_N|^2 \right) = \text{diag} \left( q_1, \ldots, q_1, q_2, \ldots, q_2, \ldots, q_D, \ldots, q_D \right)$$

where $q_i = P_{a,c} \cdot \chi(|h_i|^2 \geq h_t) \cdot \chi \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) \leq ADe^{-h_t} \right)$. (38)

The second indicator function in (38) checks for the constraint on the instantaneous power in (37) causally, during each time-frequency coherence slot, and allocates power only if this constraint is satisfied. The capacity obtained with such a scheme equals

$$\hat{C}_{\text{coh,1,ST}}(\text{SNR})$$

$$= \frac{1}{D} \mathbb{E} \left[ \sum_{i=1}^{D} \log_2 \left( 1 + \frac{TP}{N_c} |h_i|^2 \chi(|h_i|^2 \geq h_t) \frac{\chi \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) \leq ADe^{-h_t} \right)}{D} \right) \right]$$

$$= \frac{1}{D} \sum_{i=1}^{D} \mathbb{E} \left[ \log_2 \left( 1 + \text{SNR} \cdot e^{h_t} \cdot |h_i|^2 \chi(|h_i|^2 \geq h_t) \right) \chi \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) \leq ADe^{-h_t} \right) \right]$$

$$= \frac{1}{D} \sum_{i=1}^{D} \mathbb{P} \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) \leq ADe^{-h_t} \right) \cdot \mathbb{E} \left[ \log_2 \left( 1 + \text{SNR} \cdot e^{h_t} \cdot |h_i|^2 \chi(|h_i|^2 \geq h_t) \right) \right]$$

$$\overset{(a)}{=} \mathbb{E} \left[ \log_2 \left( 1 + \text{SNR} \cdot e^{h_t} \cdot |h_i|^2 \chi(|h_i|^2 \geq h_t) \right) \right] \cdot \frac{\sum_{i=1}^{D} \mathbb{P} \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) \leq ADe^{-h_t} \right)}{D}$$

$$= \hat{C}_{\text{coh,1,LT}}(\text{SNR}) \cdot \sum_{i=1}^{D} p_i / D$$
where \( \hat{C}_{\text{coh,1,LT}}(\text{SNR}) \) is the coherent capacity of the causal signaling scheme in (25), with only an average power constraint, and (a) follows from the fact that \( \{h_i\} \) are i.i.d. and \( p_i \triangleq \Pr\left( \sum_{j=1}^{J} \chi(|h_j|^2 \geq h_t) \leq AD e^{-h_t} \right) \). Thus, characterizing \( \hat{C}_{\text{coh,1,ST}}(\text{SNR}) \) is equivalent to characterizing \( p_i \). In particular, under what condition does \( \frac{\sum p_i}{D} \to 1? \) This is discussed in the following proposition.

**Proposition 1:** With \( h_t \sim \lambda \log \left( \frac{1}{\text{SNR}} \right) \) as in (29), \( \frac{\sum p_i}{D} \geq L \), where \( L \) satisfies

\[
L \approx \begin{cases} 
1 - \frac{4}{\text{SNR}^{\lambda}(1+\text{SNR}^{\lambda}/4)^{4/3}} - \frac{D(1-A/2)}{(1+\text{SNR}^{\lambda}/4)^{D(A-1)/4}}, & 1 < A < 2 \\
1 - \frac{4}{\text{SNR}^{\lambda}(1+\text{SNR}^{\lambda}/4)^{D(A-1)/4}}, & A > 2
\end{cases}
\tag{40}
\]

In particular, if

\[
\lim_{\text{SNR} \to 0} E[D_{\text{eff}}] = D\text{SNR}^{\lambda} = \infty
\tag{41}
\]

then \( L \to 1 \) for all \( A > 1 \) and \( \hat{C}_{\text{coh,1,ST}}(\text{SNR}) \to \hat{C}_{\text{coh,1,LT}}(\text{SNR}) \).

**Proof:** See Appendix B. \( \blacksquare \)

**C. Discussion**

The first thing to note is that unlike the average power constrained case (Theorem 1), where we could choose any \( \lambda \in (0, 1) \) (and hence \( \lambda \to 1 \) for maximum gain), the condition in (41) puts a constraint on the maximum value of \( \lambda \) so that \( \hat{C}_{\text{coh,1,ST}}(\text{SNR}) \approx \hat{C}_{\text{coh,1,LT}}(\text{SNR}) \). Since this constraint is related to the way the DoF, \( D \), scales as a function of SNR, we analyze this aspect further. We contrast the cases when the channel is rich and when it is sparse.

**A1) Rich multipath:** For a rich channel, we note from (4) that \( D \) scales linearly with \( T \) and \( W \). Therefore, for a fixed \( T \), \( D \sim \text{SNR}^{-1} \) (since by definition \( \text{SNR} = \frac{P}{W} \)). Thus, in (41) \( D\text{SNR}^{\lambda} = \text{SNR}^{\lambda-1} \to \infty \) as long as \( \lambda < 1 \). We conclude that when the channel is rich, there is no penalty for imposing the constraint in (37). We can choose any \( \lambda \in (0, 1) \), in particular \( \lambda \to 1 \).

**A2) Delay sparsity only:** In this case while \( D \) scales linearly with \( T \) through \( D_T \ (g_2(x) = x) \), the scaling of \( D \) with \( W \) is sub-linear (see (5)), through \( D_W \). Therefore, for a fixed \( T \), satisfying (41) critically depends on the type of sub-linear scaling. For example, for the power-law scaling in (6)

\[
D \sim W^{\delta_2} \sim \text{SNR}^{-\delta_2} \quad \text{and} \quad D\text{SNR}^{\lambda} = \text{SNR}^{\lambda-\delta_2}
\tag{42}
\]
and (41) now reduces to the condition $\lambda < \delta_2$ and consequently limits the maximum $\lambda$. On the other hand, if in (5), $g_1(W) = \log(W)$, then (41) can never be satisfied for any $\lambda \in (0, 1)$.

This issue is resolved by maintaining a canonical scaling relationship between $T$ and $W$, which provides an additional scaling of $D$ with SNR. For example, if $T \sim f_3(W)$, then by choosing

$$f_3(x) = \frac{x^\beta}{g_1(x)}$$

(43)

we get $D \sim g_1(W)g_2(f_3(W)) = g_1(W)f_3(W) = W^\beta \sim \text{SNR}^{-\beta}$ and consequently (41) becomes $D\text{SNR}^\lambda = \text{SNR}^{\lambda-\beta}$. Thus with $\beta \geq 1$ in (43), $\lambda \rightarrow 1$ is feasible and we obtain the capacity in (25). For the power-law scaling in (6), we get the following $T$ vs. $W$ scaling requirement from (43):

$$T \sim W^{\beta-\delta_2}$$

(44)

in which by choosing $\beta \geq 1$, we obtain $\tilde{C}_{\text{coh,1,ST}}(\text{SNR}) \rightarrow \tilde{C}_{\text{coh,1,LT}}(\text{SNR}) \rightarrow C_{\text{coh,1,LT}}(\text{SNR})$.

A3) Delay and Doppler sparsity: Now the scaling in $D$ is even slower, sub-linear with both $T$ and $W$. Again, we need to maintain a scaling relationship between $T$ and $W$ in order to satisfy (41). For example, if $T \sim f_3(W)$, then with

$$f_3(x) = g_2^{-1}\left(\frac{x^\beta}{g_1(x)}\right)$$

(45)

we get $D \sim g_1(W)g_2(f_3(W)) = W^\beta \sim \text{SNR}^{-\beta}$ and hence $D\text{SNR}^\lambda = \text{SNR}^{\lambda-\beta}$. Thus with $\beta \geq 1$ in (45), we once again ensure that $\lambda \rightarrow 1$ can be attained. For the power-law scaling in (6), the scaling in (45) simplifies to

$$T \sim W^{\frac{\beta-\delta_2}{\delta_1}}.$$

(46)

A4) Doppler sparsity only: Here we have $D$ scaling linearly with $W$ and hence, even for a fixed $T$, $D \sim \text{SNR}^{-1}$. Thus analogous to A1, (41) is trivially satisfied.

We conclude that with perfect receiver CSI, independent of whether the channel is rich or sparse, we obtain the same capacity with both average as well as instantaneous power constraint. However, it is easier to achieve this capacity in a rich channel, whereas certain conditions need to be satisfied in a sparse channel. Things change dramatically when there is no receiver CSI a priori and the channel has to be learnt at the receiver.
IV. FEEDBACK CAPACITY WITH CHANNEL ESTIMATION AT THE RECEIVER

The focus of this section is on the more realistic scenario when there is no CSI at the receiver \textit{a priori}. We show that the first-order term of coherent capacity with feedback can be achieved if the channel is sparse and the channel coherence dimension $N_c$ scales with SNR at an appropriate rate, allowing the receiver to learn the channel reliably. We also show that this is infeasible when the channel is rich, due to poor channel estimation. Within the non-coherent regime, we focus on training-based communication schemes.

A. Training-Based Communication Using STF Signaling

We consider a communication scheme where the transmitted signals include training symbols to enable channel estimation and coherent detection. The restriction to training schemes is motivated by their practical feasibility. The total energy available for training and communication is $PT$, of which a fraction $\eta$ is used for training and the remaining fraction $(1 - \eta)$ is used for communication. Due to the block fading model, our scheme uses one signal space dimension in each coherence subspace for training and the remaining $(N_c - 1)$ for communication, as illustrated in Fig. 1(c). We consider minimum mean squared error (MMSE) channel estimation and the two metrics that capture channel estimation performance in this setting are: (i) $\eta$, the fraction of energy used for estimation, and (ii) MSE, the mean squared error in estimating each channel coefficient. The reader is referred to [7, Sec. II(c)] for a more detailed description of the training scheme.

B. Capacity of the Training-based Communication Scheme

Let $\hat{C}_{\text{train},1,\text{LT}}(\text{SNR})$ denote the average mutual information achievable (per-dimension) with the causal training and communication scheme satisfying the average power constraint. To characterize $\hat{C}_{\text{train},1,\text{LT}}(\text{SNR})$, we proceed on the same lines as the case with no feedback [7, Lemma 1]. Let $H$ be the actual channel, $\hat{H}$ be the estimated channel and $\Delta = H - \hat{H}$. We begin with the following well-known lower-bound [20] to $\hat{C}_{\text{train},1,\text{LT}}(\text{SNR})$:

$$
\hat{C}_{\text{train},1,\text{LT}}(\text{SNR}) \geq \sup_{Q: \text{Tr}(Q) \leq (1-\eta)TP} \mathbb{E} \left[ \log_2 \det \left( I_{(N_c-1)D} + \hat{H}Q\hat{H}^H (I + \Sigma_{\Delta})^{-1} \right) \right]. 
$$

(47)
As before, the optimal $Q$ is diagonal and analogous to (18) and (24), equals

$$Q = \text{diag} \left( |x_1|^2, |x_2|^2, \ldots, |x_{N-D}|^2 \right) = \text{diag} \left( q_{N_c-1}, q_{N_c}, q_{N_c-1}, q_{N_c} \right)$$

where $q_i = \frac{(1 - \eta)TP}{(N_c - 1)D} \cdot \frac{\chi \left( |\hat{h}_i|^2 \geq h_{\text{train}}^t \right)}{E \left( |\hat{h}|^2 \geq h_{\text{train}}^t \right)}$.  

(48)

The following theorem describes the conditions under which the capacity of the training-based communication scheme converges to the coherent capacity.

**Theorem 2:** If $N_c = \frac{1}{\text{SNR}^\mu}$ for some $\mu > 1$, then

$$\lim_{\text{SNR} \to 0} \frac{\hat{C}_{\text{train,1,LT}}(\text{SNR})}{\hat{C}_{\text{coh,1,LT}}(\text{SNR})} = 1.$$  

(49)

**Proof:** We provide the key steps here and the finer details are relegated to Appendix C. Using the choice of $Q$ from (48) in (47) and proceeding on the lines of (39), we obtain the following simplification:

$$\hat{C}_{\text{train,1,LT}}(h_{\text{train}}^t, \eta, N_c, \text{SNR}) = \kappa_1 \cdot \left[ \log_2 \left( 1 + \frac{(1 - \eta)(1 + \eta N_c \text{SNR})h_{\text{train}}^t \text{SNR}}{(1 - \eta)\text{SNR} + \kappa_1 \kappa_2} \right) \right] + \nu \left( \frac{(1 - \eta)(1 + \eta N_c \text{SNR})h_{\text{train}}^t \text{SNR} + (1 - \eta)\text{SNR} + \kappa_1 \kappa_2}{\eta N_c \text{SNR}^2} \right),$$  

(50)

$$\kappa_1 = e^{-\frac{h_{\text{train}}^t (1 + \eta N_c \text{SNR})}{\eta N_c \text{SNR}^2}}, \quad \kappa_2 = \eta (N_c - 1) \text{SNR} + \left( 1 - \frac{1}{N_c} \right).$$  

(51)

where $\nu$ is as defined in (33). The tightest lower bound to (50) is obtained by maximizing $\hat{C}_{\text{train,1,LT}}(h_{\text{train}}^t, \eta, N_c, \text{SNR})$ over the fraction of energy spent on training, $\eta$, and over $h_{\text{train}}^t$:

$$C_{\text{train,1,LT}}^*(\text{SNR}) = \max_{h_{\text{train}}^t} \max_{\eta} \hat{C}_{\text{train,1}}(h_{\text{train}}^t, \eta, N_c, \text{SNR}).$$  

(52)

As such, performing the double optimization as formulated above seems difficult. However, motivated by our study in Section III, we will assume a specific form for the threshold, given by $h_{\text{train}}^t = \epsilon \log \left( \frac{1}{\text{SNR}} \right)$. As shown in Appendix C, with this choice of the threshold, the optimal choice of $\eta$ and $N_c$ can be obtained via simple manipulations and the desired result established.

**C. Discussion**

The above result is closely related to our earlier work [7], where we analyzed the capacity of training-based schemes without transmitter CSI and showed that when $N_c = \frac{1}{\text{SNR}^\mu}$ with
the channel is *asymptotically coherent*; channel estimation performance is near-perfect at a vanishing estimation cost. Here we have shown, analogous to [7], that under the assumption of an error-free $D$-bit feedback link, the capacity of the training-based scheme converges to the coherent capacity (in the first-order sense). Furthermore, the cost of feedback, measured in terms of the number of feedback bits per dimension ($\frac{D}{N}$) goes to zero asymptotically when the channel is sparse (since $D$ scales sub-linearly with $N$; see (5)). On the other hand, the cost of feedback is constant for a rich channel since $D$ scales linearly with $N$.

From a system design perspective, it is of interest to investigate the scenarios under which we achieve $\mu > 1$ in the relation $N_c = \frac{1}{\text{SNR}^\mu}$. What are the conditions on the channel parameters ($T_m$, $W_d$, $f_1$, and $f_2$) and how do they interact with the signal space parameters ($T$, $W$ and $P$) so that $\mu > 1$ is feasible?. As we discuss next, by leveraging delay and Doppler sparsity and using peaky signaling (when necessary), $\mu > 1$ can be achieved in a variety of scenarios.

**B1) Delay and Doppler sparsity:** When the channel is sparse in both delay and Doppler we have, using (12), $W_{coh} \sim f_1(W)$ and $T_{coh} \sim f_2(T)$. Therefore, if we scale $T$ with $W$ according to

$$T \sim f_3(W) \quad \text{with} \quad f_3(x) = f_2^{-1}\left(\frac{x^\mu}{f_1(x)}\right)$$

so that $N_c = W_{coh}T_{coh} \sim f_1(W)f_2(f_3(W)) = f_1(W)f_2\left(f_2^{-1}\left(\frac{W^\mu}{f_1(W)}\right)\right) \sim \frac{1}{\text{SNR}^\mu}$. Thus with $\mu > 1$ in (53), we attain the desired scaling of $N_c$ with SNR. For example, for the power-law scaling in (13), the desired scaling in $N_c$ can be obtained by choosing $T$, $W$ and $P$ satisfying the following canonical relationship that is obtained using (13) in (53)

$$T = \left(\frac{T_m W_d^{\beta_2}}{P}\right)^{\frac{1}{1-\beta_1}}W^{\frac{\mu + \beta_2}{1-\beta_1}}\frac{W^{\mu + \beta_2}}{P^{\frac{\mu}{1-\beta_1}}}.$$  

**B2) Delay sparsity only:** In this case, $T_{coh} = \frac{1}{W_d}$ and $N_c = W_{coh}T_{coh}$ scales with SNR only through $W_{coh} \sim f_1\left(\frac{1}{\text{SNR}}\right)$. Therefore, for a sub-linear $f_1$, we cannot satisfy the required $\mu > 1$. A solution to this is to use peaky signaling where training and communication is performed only on a subset of the $D$ coherence subspaces. We model peakiness similar to [4], [7] and define $\zeta = \text{SNR}^\gamma$, $\gamma > 0$ as the fraction of $D$ over which signaling is performed. It can be shown that [7, Lemma 3] with peaky training schemes, the condition for asymptotic coherence gets relaxed to $N_c = \frac{1}{\text{SNR}^\mu_{\text{peaky}}}$ from the original $N_c = \frac{1}{\text{SNR}^\mu}$ where $\mu_{\text{peaky}} = \mu + \gamma$. Thus now we
require $\mu_{\text{peaky}} > 1$, that is $\mu > 1 - \gamma$. For example, with the power-law scaling in (13), we have $N_c \sim f_1(W) = W^{1-\delta_2} \sim \frac{1}{\text{SNR}^{1-\delta_2}}$. Thus with $\gamma > \delta_2$, we satisfy the desired condition.

**B3) Doppler sparsity only**: When the channel is sparse only in Doppler, $W_{\text{coh}} = \frac{1}{T_m}$ is fixed and the scaling in $N_c$ is only through $T_{\text{coh}} \sim f_2(T)$ (see (12)). Therefore, by scaling $T$ with $W$ according to $T \sim f_2^{-1}(W^\mu)$ and choosing $\mu > 1$, we have $N_c \sim T_{\text{coh}} \sim f_2(f_2^{-1}(W^\mu)) \sim \frac{1}{\text{SNR}^\mu}$. For the power-law scaling in (13), we obtain

$$T \sim W^{\frac{\mu}{1-\delta_1}}$$

**(B4) Rich multipath**: When the channel is rich in both delay and Doppler ($f_1$ and $f_2$ are both constant functions), $N_c = \frac{1}{T_m W_d}$ is fixed and does not scale with SNR. Thus we can never maintain the scaling relationship in $N_c$ as in Theorem 2.

**Remark 4**: We contrast the results obtained here with recently made observations in [11], [12], where the authors have independently shown that with training and 1-bit feedback, capacity scales in the limit of vanishing SNR as $\log(T_{\text{coh}})\text{SNR}$, where $T_{\text{coh}}$ is the coherence time. The focus in [11] is on the regime when $T_{\text{coh}}$ increases as SNR decreases. In particular, the authors show that capacity scales as $\log(T_{\text{coh}})\text{SNR}$ if $\log(T_{\text{coh}}) \leq \log\left(\frac{1}{\text{SNR}}\right)$ and equals the coherent capacity, $\log\left(\frac{1}{\text{SNR}}\right)\text{SNR}$ when $\log(T_{\text{coh}}) \geq \log\left(\frac{1}{\text{SNR}}\right)$. The result of Theorem 2 implies that as long as $\mu > 1$, we can achieve rates arbitrarily close to $\log\left(\frac{1}{\text{SNR}}\right)\text{SNR}$ by choosing $\lambda \rightarrow 1$. All that is required here is channel sparsity and even when the channel is not sparse in either delay or Doppler, the rate is achievable, as analyzed in B1-B4 above.

**Remark 5**: The type of signaling mechanisms discussed in this paper are also significantly different from what has been reported earlier. For example, in [11], peaky training schemes are shown to be necessary to achieve perfect training performance. Although such schemes satisfy the average power constraint, they would almost surely violate any finite instantaneous power constraint, for example as assumed in (37) (finite value of $A$). On the other hand, our findings here reveal that channel sparsity is a new degree of freedom that can be exploited in obtaining near-coherent performance with training-based communication schemes. Since it helps offset/eliminate peaky input signals, it should also provide a means to achieve capacity even under an instantaneous power constraint. This is the subject of discussion in the next section.
D. Capacity of Training-based Scheme with Instantaneous Power Constraint

With a finite constraint on the instantaneous transmit power as in (37) for the communication phase of the channel learning scheme, we explore an achievable lower bound for the capacity. We consider exactly the same power allocation scheme as in Section III-B and proceed on the same lines to compute the capacity. Thus, analogous to (39), we obtain

\[ C_{\text{train},1,\text{ST}}(\text{SNR}) = \left(1 - \frac{1}{N_c}\right) \frac{1}{D} \sum_{i=1}^{D} \mathbb{E} \left[ \log_2 \left( 1 + \frac{\widehat{h}^2 q_i (1 + E_{tr})}{1 + q_i + E_{tr}} \right) \right] \]

where

\[ p_{i}^{\text{train}} = \Pr \left( \sum_{j=1}^{i} \chi(\widehat{h}_j^2 \geq h_t) \leq \frac{ADe^{-h_{\text{train}} (1 + N_c \text{SNR})}}{\eta N_c \text{SNR}} \right) \]

Thus, once again the problem reduces to analyzing the sum of the probabilities \( \{p_{i}^{\text{train}}\}_{i=1}^{D} \) and we desire \( \sum_{i=1}^{D} p_{i}^{\text{train}} \rightarrow 1 \). Taking recourse to the analysis in Proposition 1 and by using a threshold of the form \( h_{\text{train}} = \frac{\eta}{1 + \eta N_c \text{SNR}} h_t \) for the training scheme with \( h_t \sim \lambda \log \left( \frac{1}{\text{SNR}} \right) \) as in (29), it can be shown that the lower bound \( L \) in (40) is also a lower bound to \( \sum_{i=1}^{D} p_{i}^{\text{train}} \). Thus we obtain same constraint on \( \lambda \) as in (41).

We conclude that the capacity of the training-based communication scheme with an instantaneous power constraint converges to the coherent capacity with only an average power constraint, which is the ultimate performance benchmark for the \( D \)-bit feedback scheme that we started with in Section III-A. In order to achieve this, we need to satisfy the following two conditions:

**C1**) The channel coherence dimension, \( N_c \), scales with SNR according to \( N_c \sim \frac{1}{\text{SNR}^\mu} \) with \( \mu > 1 \), and

**C2**) The independent degrees of freedom (DoF), \( D \), in the channel scales with SNR such that \( \mathbb{E}[D_{\text{eff}}] = D \text{SNR}^\lambda \rightarrow \infty \) as \( \text{SNR} \rightarrow 0 \).

While the first condition ensures the fidelity of the training performance, the second condition leads to a negligible capacity loss due to the instantaneous power constraint. For training and communication schemes, both factors come into play, in contrast with the perfect receiver CSI scenario (Section III-B), where only C2 is relevant.
Finally, we discuss the feasibility of achieving C1 and C2. For this purpose, we draw upon the insights gained from the discussions in Sections III-C and IV-C, where we analyzed the achievability of C2 and C1, respectively. Now we need to satisfy them simultaneously. It is clear that the two conditions are somewhat conflicting in nature since for a richer channel, it is easier to increase $D$ but a lot tougher to increase $N_c$, while for a sparser channel, it is vice versa. For ease of understanding, we will focus on the power-law scaling in (6) and (13).

Case 1 (Rich multipath): It follows from the discussion in B4 that we cannot satisfy C1. Thus, though C2 can always be satisfied, we can never satisfy them simultaneously.

Case 2 (Sparse multipath): We consider all the different possible cases of sparsity. With only delay sparsity, we have from B2 that peaky signaling is necessary to satisfy C1, but this violates any finite instantaneous power constraint and hence we can never satisfy both C1 and C2 simultaneously. With only Doppler sparsity, peaky signaling is not necessary to satisfy C1, and it follows from (55) that we require, $T \sim W^{\frac{\mu}{1+\delta_1}}$ with $\mu > 1$. On the other hand, from A4, we know that C2 is always satisfied, even for a fixed $T$. Thus by scaling $T$ with $W$ according to (55), both conditions are jointly satisfied. With sparsity in both delay and Doppler, it follows from (54) that to satisfy C1, we require $T \sim W^{\frac{\mu-1+\delta_2}{1+\delta_1}}$ with $\mu > 1$, whereas to satisfy C2, the necessary requirement is obtained from (46) as $T \sim W^{\frac{\beta-\delta_2}{\delta_1}}$ with $\beta \geq 1$. Thus to satisfy both, we require

$$T \sim W^\rho, \quad \rho = \max\left( \frac{\mu - 1 + \delta_2}{1 - \delta_1}, \frac{\beta - \delta_2}{\delta_1} \right)$$

Choosing the minimum possible $\mu = \beta \approx 1$, we observe that the least stringent scaling in $T$ with $W$ is obtained when both the exponents in (57) are equal. This happens when $\frac{\delta_2}{1-\delta_1} = \frac{1-\delta_2}{\delta_1}$. That is $\delta_1 + \delta_2 = 1$. This provides a channel sparsity condition when the two requirements are balanced. When $\delta_1 + \delta_2 < 1$, the channel is more sparse and hence it is easier to satisfy C1, but harder to attain C2. When $\delta_1 + \delta_2 > 1$, the channel is less sparse and as a consequence, it is that much easier to increase $D$ and satisfy C2, but tougher to scale $N_c$ and satisfy C1.
APPENDIX

A. Tightness of $\hat{C}_{\text{coh,1,LT}}(\text{SNR})$ to $C_{\text{coh,1,LT}}(\text{SNR})$ as SNR $\to 0$

Let $\chi_i$ denote the random variable $\chi(|h_i|^2 \geq h_t)$. Defining $\gamma \triangleq \frac{|C_{\text{coh,1,LT}}(\text{SNR}) - \hat{C}_{\text{coh,1,LT}}(\text{SNR})|}{C_{\text{coh,1,LT}}(\text{SNR})}$, we have

$$\gamma = \frac{1}{D} \left| \sum_{i=1}^{D} \mathbb{E} \left[ \log_2 \left( 1 + \frac{TP|h_i|^2 \chi_i (D e^{-h_t} - \sum_i \chi_i)}{\sum_i \chi_i N_c D e^{-h_t}} \right) \right] \right|$$

$$(a) \quad \frac{1}{D} \sum_{i=1}^{D} \mathbb{E} \left[ \frac{TP|h_i|^2 \chi_i (D e^{-h_t} - \sum_i \chi_i)}{\sum_i \chi_i N_c D e^{-h_t}} \right]$$

$$= \frac{TP}{N_c D^2 e^{-h_t}} \sum_{i=1}^{D} \mathbb{E} \left[ \frac{|h_i|^2 \chi_i D e^{-h_t} - \sum_i \chi_i}{\sum_i \chi_i (1 + \frac{TP|h_i|^2 \chi_i}{N_c D e^{-h_t}})} \right]$$

$$(b) \quad \frac{TP}{N_c D e^{-h_t}} \mathbb{E} \left[ \frac{|h_1|^2 \chi_1 |D e^{-h_t} - \sum_i \chi_i|}{\sum_i \chi_i (1 + \frac{TP|h_1|^2 \chi_1}{N_c D e^{-h_t}})} \right] \triangleq \gamma_0$$

where (a) follows from the log-inequality and (b) from the fact that $\{h_i\}$ are i.i.d. Conditioning on $\chi_1$, we now have

$$\gamma_0 = \frac{TP}{N_c D e^{-h_t}} \mathbb{E}[\chi_1] \mathbb{E}_{h_1,\{\chi_j, j > 1\}} \left[ \frac{|h_1|^2 |D e^{-h_t} - (1 + \sum_{j>1} \chi_j)|}{(1 + \sum_{j>1} \chi_j) (1 + \frac{TP|h_1|^2}{N_c D e^{-h_t}})} \right]$$

$$\triangleq \frac{\text{SNR} \cdot \mathbb{E}_{h_1} \left[ \frac{|h_1|^2}{1 + \frac{TP|h_1|^2}{N_c D e^{-h_t}}} \right] \cdot \mathbb{E}_{\{\chi_j, j > 1\}} \left[ \frac{D e^{-h_t} - (1 + \sum_{j>1} \chi_j)}{1 + \sum_{j>1} \chi_j} \right]}{(1 + \sum_{j>1} \chi_j)} \triangleq \gamma_1$$

where (a) follows from the fact that $h_1$ and $\{\chi_j, j > 1\}$ are independent.

To show the closeness of $\hat{C}_{\text{coh,1,LT}}(\text{SNR})$ to $C_{\text{coh,1,LT}}(\text{SNR})$, we now produce an upper bound for $\gamma_1$ that tends to 0 as SNR $\to 0$. Our goal is to show that given any choice of $D$, $\frac{\gamma_1}{\text{SNR}}$ is
bounded. Consider
\[
E_{\{x_j, j > 1\}} \left[ \frac{D e^{-h_t} - (1 + \sum_{j > 1} x_j)}{(1 + \sum_{j > 1} x_j)} \right] = E_{\{x_j, j > 1\}} \left[ \frac{D e^{-h_t}}{(1 + \sum_{j > 1} x_j)} - 1 \right]
\]
\[
(a) \sqrt{E_{x_j} \left[ \left( \frac{D e^{-h_t}}{(1 + \sum_{j > 1} x_j)} \right)^2 + 1 - 2 \frac{D e^{-h_t}}{(1 + \sum_{j > 1} x_j)} \right]} \overset{\Delta}{=} \gamma_2
\]
where (a) is a consequence of Cauchy-Schwarz inequality. Let \( E \) denote \( e^{-h_t} \). We then have
\[
\gamma_2 \overset{(b)}{\leq} \sqrt{1 + D^2 E^2 \cdot E_{X_j} \left[ \frac{1}{1 + \sum_{j > 1} x_j} \right] - \frac{2DE}{1 + (D - 1)E}} \tag{60}
\]
where in (b) we have used the fact that \( E \left[ \frac{1}{X} \right] \geq \frac{1}{E[X]} \) for a positive random variable \( X \). We now estimate \( \alpha \overset{\Delta}{=} E_{x_j} \left[ \frac{1}{1 + \sum_{j > 1} x_j} \right] \). It is easy to check that
\[
\alpha = \sum_{i=0}^{D-1} \binom{D-1}{i} \frac{E^i(1 - E)^{D-1-i}}{(i + 1)^2}. \tag{61}
\]
Noting that
\[
(1 + y)^{D-1} = \sum_{i=0}^{D-1} \binom{D-1}{i} y^i
\]
and integrating twice both sides of (62) with respect to \( y \), we have
\[
\frac{(1 + y)^{D+1}}{D(D + 1)} = \sum_{i=0}^{D-1} \binom{D-1}{i} \frac{y^{i+2}}{(i + 1)(i + 2)}. \tag{63}
\]
Using \( y = \frac{E}{1 - E} \) in (63), we have
\[
\frac{1}{D(D + 1)E^2} = \sum_{i=0}^{D-1} \binom{D-1}{i} \frac{E^i(1 - E)^{D-1-i}}{(i + 1)(i + 2)}. \tag{64}
\]
Observe that \( \frac{1}{(i+1)^2} \leq \frac{2}{(i+1)(i+2)} \) for all \( i \geq 0 \) and an upper bound for \( \gamma_2 \) is
\[
\gamma_2 \leq \sqrt{1 + \frac{2D^2 E^2}{D(D + 1)E^2} - \frac{2DE}{1 + (D - 1)E}} = \sqrt{\frac{D^2E - 4DE + 3D - E + 1}{(D + 1)(DE - E + 1)}} \tag{65}
\]
which is bounded for any choice of \( D \). (In fact, the upper bound converges to 1 as \( D \to \infty \)). Note that the bound in (65) is very loose and one might expect that \( \frac{\gamma_1}{SNR} \to 0 \) as \( D \to \infty \) as a consequence of the law of large numbers. However, for our purpose, the proposed loose upper bound in (65) is sufficient.
B. Proof of Proposition 1

To compute \( p_i \triangleq \Pr \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) \leq AD e^{-h_t} \right) \), we need the following result [21, Theorem 2.8, p. 57] on the tail probability of a sum of independent random variables.

**Lemma 1:** Let \( \mathbf{X}_i, i = 1, \ldots , n \) be independent random variables with \( \mathbb{E}[\mathbf{X}_i] = 0 \) and \( \mathbb{E}[\mathbf{X}_i^2] = \sigma_i^2 \). Define \( B_n = \sum_{i=1}^{n} \sigma_i^2 \). If there exists a positive constant \( H \) such that

\[
\mathbb{E}[\mathbf{X}_i^m] \leq \frac{1}{2} m! \sigma_i^2 H^{m-2}
\]

for all \( i \) and \( x \geq \frac{B_n}{H} \), then we have \( \Pr \left( \sum_{i=1}^{n} \mathbf{X}_i > x \right) \leq \exp \left( - \frac{x^2}{4H} \right) \). If \( x \leq \frac{B_n}{H} \), then we have \( \Pr \left( \sum_{i=1}^{n} \mathbf{X}_i > x \right) \leq \exp \left( - \frac{x^2}{4H} \right) \).

To apply Lemma 1, we set \( n = i \) and \( \mathbf{X}_j = \chi(|h_j|^2 \geq h_t) - \mathbb{E} \left[ \chi(|h_j|^2 \geq h_t) \right] = \chi(|h_j|^2 \geq h_t) - e^{-h_t} = \chi_j - \mathbb{E} \) for \( j = 1, \ldots , i \). Then, a simple computation of the higher moments of \( \mathbf{X}_j \) implies that \( \mathbb{E}[\mathbf{X}_j^2] = \sigma_j^2 = \mathbb{E}(1-E), B_i = i\mathbb{E}(1-E), \mathbb{E}[\mathbf{X}_j^m] = \mathbb{E}(1-E) \cdot ((1-E)^{m-1} + (-1)^m\mathbb{E}^{m-1}) \).

It can be checked that \( H = (1-E) \) is sufficient to satisfy the conditions of Lemma 1. With this setting, we have

\[
\Pr \left( \sum_{j=1}^{i} \chi(|h_j|^2 \geq h_t) - iE > (AD - i)E \right) \leq \begin{cases} 
\exp \left( - \frac{(AD-i)E}{4(1-E)} \right) & \text{if } i \leq \lfloor \frac{AD}{2} \rfloor, \\
\exp \left( - \frac{(AD-i)^2E}{4(1-E)} \right) & \text{if } i \geq \lfloor \frac{AD}{2} \rfloor + 1.
\end{cases}
\]

If \( 1 < A < 2 \), with \( \kappa = \frac{E}{4(1-E)} \) using (67), the following lower bound, \( L \), holds for \( \sum_{i=1}^{D} p_i \):

\[
L = 1 - \left[ \exp(-AD\kappa) \sum_{i \leq \lfloor \frac{AD}{2} \rfloor} e^{i\kappa} + \sum_{i \geq \lfloor \frac{AD}{2} \rfloor + 1} e^{-\frac{(AD-i)^2\kappa}{i}} \right]
\]

\[
\overset{(a)}{=} 1 - \left[ \frac{e^{-\kappa(AD-1)} \cdot (e^{\kappa(\frac{AD}{2})} - 1)}{e^{\kappa} - 1} + \left( D - \left\lfloor \frac{AD}{2} \right\rfloor \right) e^{-\frac{(A-1)^2D\kappa}{2}} \right]
\]

\[
\geq 1 - \left[ \frac{1}{e^{\kappa} - 1} \cdot e^{-\kappa(\frac{AD}{2}-1)} + (1 + D(1-A/2)) e^{-\frac{(A-1)^2D\kappa}{2}} \right]
\]

where (a) follows from first using \( \frac{(AD-i)^2}{i} \geq (A-1)^2D \) for all \( 1 \leq i \leq D \) and then upon further simplification using the sum of a geometric series.

If \( A \geq 2 \), we have the following lower bound to \( \sum_{i=1}^{D} p_i \):

\[
L = 1 - \exp(-AD\kappa) \sum_{1 \leq i \leq D} e^{i\kappa} \approx 1 - e^{-\kappa(D(A-1)-1)} \cdot \frac{1}{e^{\kappa} - 1}.
\]

With \( h_t = \lambda \log \left( \frac{1}{\text{SNR}} \right) \) as in (29), the dominant term of \( E \) is \( \text{SNR}^\lambda \) and hence in \( \kappa \) is \( \frac{\text{SNR}^\lambda}{4} \).

With this choice of \( h_t \) in (69) and (70) and simplifying, we obtain the desired bounds as in (40).
It is also straightforward to see that when \( D \) satisfies \( DSNR^{-\lambda} \to \infty \) as \( SNR \to 0, L \to 1 \) in both the cases.

\[ \]

C. Completing the Proof of Theorem 2

The choice of \( h_t \) we study is \( h_t = \epsilon \log \left( \frac{1}{SNR} \right) \) for some \( \epsilon > 0 \). First, with this fixed choice of \( h_t \), note that maximizing \( \tilde{C}_{train,1,LT}(\eta, N_c, SNR) \) is equivalent to setting its derivative (with respect to \( \eta \)) to zero. Then, it is straightforward to check that the derivative is

\[
\frac{\nu_\beta h_t}{\eta} + \frac{h_t}{\eta} \log e \left( 1 + \frac{(1 - \eta)(1 + \eta N_c SNR) h_t SNR}{(1 - \eta) SNR + \kappa_1 \kappa_2} \right) \\
+ \left( \frac{\nu_\beta - \frac{1}{\beta}}{SNR \eta} \right) \left[ \kappa_1 \left( 1 - \frac{1}{N_c} \right) \frac{N_c \eta^2 SNR + 2\eta - 1}{(1 - \eta)^2} + \frac{h_t (1 + \eta N_c SNR)}{\eta N_c SNR (1 - \eta)} - SNR (h_t + 1) \right] \\
+ \frac{h_t SNR^2 N_c \eta}{(1 - \eta) SNR + \kappa_1 \kappa_2} \cdot \frac{N_c \eta^2 SNR^2 (1 - \eta)^2 - \kappa_1 \kappa_2 (1 + \eta SNR N_c) \left( 1 + \frac{h_t (1 - \eta)}{N_c \eta^2 SNR} \right)}{(1 - \eta) SNR + \kappa_1 \kappa_2 + (1 - \eta) (1 + \eta N_c SNR) h_t SNR} .
\]

(71)

For simplicity, we will denote the four terms in (71) by I, II, III and IV. We will further assume that \( \eta = SNR^x, x \geq 0 \) and \( N_c = \frac{1}{SNR^y}, y > 0 \). For a given choice of \( \epsilon \), our goal is to determine the relationship between \( x \) and \( y \) such that the derivative in (71) can be zero. We consider three cases: i) \( y > 1 + x \), ii) \( y < 1 + x \) and iii) \( y = 1 + x \).

**Case i:** First, note that \( \eta N_c SNR = SNR^{-z} \) for some \( z > 0 \). The dominant terms of \( \beta \) can be seen to be \( \frac{1}{SNR^{1+x}} + \epsilon \log \left( \frac{1}{SNR} \right) \) and thus, up to first order \( \beta = \frac{1}{SNR^{1+x}} \). Similarly, \( (1 - \eta) SNR + \kappa_1 \kappa_2 \) up to first order equals \( SNR^{c-x} \). Note from \([19, 5.1.20, p. 229]\) that \( \nu_\beta = O \left( \frac{1}{\beta} \right) \) if \( \beta \to \infty \) and hence I is \( \epsilon \log \left( \frac{1}{SNR} \right) ^2 \frac{1}{SNR^{1+x+y}} \). It can also be checked that II is \( \epsilon \log \left( \frac{1}{SNR} \right) ^2 \frac{1}{SNR^{1+x+y}} \) as long as \( y < 1 + 2x \). Under the same assumption, \( y < 1 + 2x \), IV is \(-\epsilon \log \left( \frac{1}{SNR} \right) ^2 \frac{1}{SNR^{1+x+y}} \). Thus, by playing with constants the derivative can be set to zero in this case. If \( y \geq 1 + 2x \), I and II remain unchanged, but III is \( SNR^{2x+y-\epsilon} \) and IV is \(-\epsilon \log \left( \frac{1}{SNR} \right) SNR^{2x+y-\epsilon} \). By comparing the coefficients, we see that the only way the derivative can be zero is if \( y = 1 + 2x \).

**Case ii:** In this case, the first order terms show the following behavior. With \( w = 1 + x - y > 0 \), I is \( SNR^{w-x} \), II is \( \epsilon \log \left( \frac{1}{SNR} \right) \log \log \left( \frac{1}{SNR} \right) SNR^{1+x+y} \), III is \(-SNR^{2w-x} \frac{1}{\epsilon \log \left( \frac{1}{SNR} \right)} \), and IV is \( SNR^{2-2y+x} \). It can be seen that the derivative can never be zero and hence this case is ruled out.
Case iii: In this case, based on a similar analysis, we see that the derivative can again be set to zero.

Therefore, if $\epsilon \in (0, 1)$, $x \geq 0$ and $1 + x < y \leq 1 + 2x$, we have

$$\hat{C}_{\text{train},1,\text{LT}}(\text{SNR}) \geq \text{SNR}^x \log \left( 1 + \frac{\epsilon \log \left( \frac{1}{\text{SNR}} \right) \text{SNR}^{1-\epsilon}(1-\text{SNR}^x)}{1-\text{SNR}^y} \right) + \text{SNR}. \quad (72)$$

Thus, $\hat{C}_{\text{train},1,\text{LT}}(\text{SNR})$ is up to first order the same as $\hat{C}_{\text{coh},1,\text{LT}}(\text{SNR})$ and $C_{\text{coh},1,\text{LT}}(\text{SNR})$. If $y = 1 + x$ and $\eta N_c \text{SNR} = a$ for some choice of $a$ (positive, finite and independent of SNR), we need $a > \frac{\epsilon}{1-\epsilon}$ and we have

$$\hat{C}_{\text{train},1,\text{LT}}(\text{SNR}) \geq \text{SNR}^{(1+a)/a} \log \left( 1 + \epsilon \text{SNR}^{1-\frac{(1+a)}{a}} \log \left( \frac{1}{\text{SNR}} \right) \right) + \frac{a}{1+a} \cdot \text{SNR}. \quad (73)$$

If $y < 1 + x$, the training scheme is strictly sub-optimal (in the limit of SNR) from an ergodic capacity point-of-view. Putting things together, we obtain the desired condition, $\mu > 1$. ■

REFERENCES


Fig. 1. (a) Delay-doppler sampling commensurate with signaling duration and bandwidth. (b) Time-frequency coherence subspaces in STF signaling. (c) Illustration of the training-based communication scheme in the STF domain. One dimension in each coherence subspace (dark squares) represents the training dimension and the remaining dimensions are used for communication.