Writing on Fading Paper and Causal Transmitter CSI

Shashi Borade Lizhong Zheng
Laboratory for Information and Decision Systems, MIT, Cambridge, MA 02139, USA.
Email: {spb,lizhong}@mit.edu

Abstract—[1] **A wideband fading channel is considered with causal channel state information (CSI) at the transmitter and no receiver CSI. A simple orthogonal code with energy detection rule at the receiver is shown to achieve the capacity of this channel in the limit of large bandwidth. This code transmits energy only when the channel gain is large enough. In this limit, this capacity without any receiver CSI is the same as the capacity with full receiver CSI–a phenomenon also true for dirty paper coding. For Rayleigh fading, this capacity (per unit time) is proportional to the logarithm of the bandwidth. Our coding scheme is motivated from the Gel'fand-Pinsker and dirty paper coding. Nonetheless, our scheme requires only causal transmitter CSI (CSIT) in contrast with Gel'fand-Pinsker and dirty paper coding, which require non-causal CSIT.**

A general discrete channel with i.i.d. states is considered later. Each input has an associated cost and a zero cost input "0" exists. The channel state is known at the transmitter in a causal manner. Capacity per unit cost is found for this channel and a simple orthogonal code is shown to achieve it. Later, a novel orthogonal coding scheme is proposed for the case of causal CSIT and a condition for equal capacity per unit cost with causal and non-causal CSIT is derived.

I. Introduction

We consider a wireless fading channel of a large bandwidth W. The input $\mathbf{x}_k[i]$ of band k at time $i \in \{1, 2 \cdots\}$ is related to the output $\mathbf{y}_k[i]$ as:

$$\mathbf{y}_k[i] = \mathbf{h}_k[i]\mathbf{x}_k[i] + \mathbf{n}_k[i] \quad \text{for} \quad k \in \{1, 2, \cdots W\} \qquad (1)$$

where $\mathbf{n}_k[i]$ is complex circularly symmetric white Gaussian noise of unit variance. Each $\mathbf{n}_k[i]$ is independent of all inputs, fading gains, and noise in other bands. The fading gains $\{\mathbf{h}_k[i]\}$ are complex Gaussian with variance 1 and are assumed i.i.d. over time and frequency. The transmitter has an average power constraint at any time i:

$$\sum_{k=1}^{W} \mathcal{E}\left[|\mathbf{x}_k[i]|^2\right] \leq P \qquad \forall\, i \in \{1, 2 \cdots\}$$

Note that the channel state at time i is completely described by the W channel gains $\{\mathbf{h}_k[i] : 1 \leq k \leq W\}$. We assume (for reasons discussed later) that at each time i, the transmitter knows this state, i.e. all W fading gains at that time and the receiver has no such knowledge. Thus we are assuming full transmitter CSI and no receiver CSI.

The case of causal transmitter CSI and no receiver CSI was studied by Shannon for discrete channels [1]. Later, [2] studied the following modification of this scenario. There the channel state for the entire codeword is known to the transmitter before beginning its transmission. Thus the CSI is available to the transmitter in a *non-causal* manner, whereas the receiver has no CSI at all. The optimal coding in this case has a large number of candidate codewords for each message. The candidate which is suitable to the entire state-sequence (spanning the code-length) is used for transmission. More precisely, a candidate which is jointly typical with the state-sequence is used for transmission. This motivates our coding scheme for this wideband fading channel, where the codeword candidate which *benefits* the most from the state sequence is used for transmission.

For the wideband fading channel above, the capacity without any receiver and transmitter CSI can be achieved by an orthogonal coding scheme like Pulse-Position Modulation or Frequency-Shift Keying [4]. In the limit of large bandwidth, this capacity without any CSI equals the capacity with full receiver CSI, which is $P \log_2 e$ bits per unit time.

For the case of full CSI at both ends, the capacity is achieved by water-filling which transmits power only when the channel gain is large enough and this capacity was shown to be $\approx P \log_2 W$ bits per unit time for the Rayleigh fading case [5]. We use the notation $f(x) \approx g(x)$ to denote $\lim_{x\to\infty} \frac{f(x)}{g(x)} = 1$. For the intermediate case of only transmitter CSI, we combine these two ideas of orthogonal coding and water-filling. We show that one can combine these two ideas without loss of optimality, that is, a code combining these two ideas is shown to achieve the capacity of this channel. This capacity with only transmitter CSI turns out to be essentially the same as the capacity ($\approx P \log_2 W$ bits per unit time) with both transmitter and receiver having CSI. This is another example where receiver CSI (or lack of it) does not affect the wideband capacity. In fact, it turns out that this capacity can be achieved by the proposed code with only one bit of transmitter CSI for each channel gain without any receiver CSI.

After noting that transmitter CSI can significantly (by a factor of $\ln W$) increase the capacity of a wideband fading channel irrespective of receiver CSI, we address the assumption of having transmitter CSI without any receiver CSI. This may seem to be a peculiar assumption for a wireless system because the transmitter in a typical wireless system obtains its CSI through feedback from the receiver itself. Nonetheless, after feeding back CSI to the transmitter, the receiver may want to ignore the CSI for multiple reasons–especially since this does not hurt capacity. For example, ignoring CSI at the receiver might help the decoding–the orthogonal structure of the proposed code (for a receiver with no CSI) might simplify the decoder. Obtaining CSI at the receiver is intrinsically costly (e.g. in terms of energy spent in training for CSI). We see later (in section 2) that if receiver CSI is ignored, CSI needs to be obtained only for a small fraction of channels and for that fraction too, only one bit of CSI is enough. This reduces the overall cost (energy) of obtaining CSI and can yield significant

[1]This research is supported by NSF CAREER award #0347395. The writing on fading paper scheme was first mentioned briefly in [5] Allerton Conference, October 2004.

gains in this wideband system, where the energy available per degree of freedom is severely limited.

II. Capacity achieving scheme

Our coding scheme is a modification of a scheme like Frequency-Shift Keying scheme or Pulse-Position Modulation, that is, here the transmitter only transmits if the fading gain is large. The purpose here is to exploit the channel randomness instead of combating it. We will split the total bandwidth W into K pieces, each of bandwidth $w = W/K$ and these pieces would be used separately for communication. The available power P is also equally divided into these pieces. Next, we illustrate our coding scheme for one such piece and analyze its achievable rate r. The total achievable rate would be number of pieces K times the rate per piece r.

The code for such a piece of bandwidth w spans T symbols in time. This code uses each of the T time indices to denote a message from the set $\{1, 2 \cdots T\}$. Thus $\ln T$ information nats[2] are transmitted in time T and hence the code rate is $\ln T/T$ nats per unit time. Say a total of λ energy units are available for transmitting a message. For transmitting message j, these entire λ energy units are transmitted at time j. Moreover, these entire λ units of energy are transmitted on a single frequency band (see Figure 1). This is the first band where the channel gain for time j is larger than a threshold $\Phi = \ln w - \ln(2\ln w)$. Note that causal transmitter CSI is enough for this threshold based energy transmission. An encoding error happens if no such band exists at time j of the message.

The decoder calculates the average (over w bands) received energy E_i for each time index i

$$E_i = \frac{1}{w}\sum_{k=1}^{w} |\ \mathbf{y}_k[i]\ |^2 \quad i \in \{1, 2, \cdots T\}$$

The time index corresponding to the maximum E_i is declared as the decoded message. No receiver CSI is needed for this decoding method. A decoding error happens when the transmitted message is different from the decoded message.

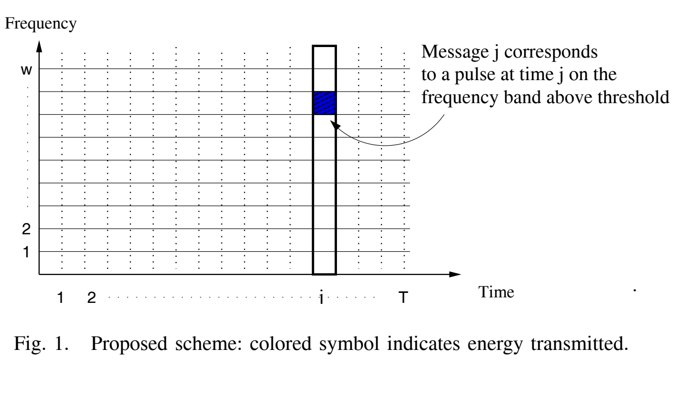

Fig. 1. Proposed scheme: colored symbol indicates energy transmitted.

Using standard techniques like Chernoff bound, [9] shows that overall (encoding or decoding) error probability of this orthogonal code goes to zero for any rate essentially smaller than $\Phi\lambda/T$ nats per unit time. Hence a essentially rate of $\Phi\lambda/T$ is achievable with this scheme[3].

This scheme is used only for δ fraction of the time where $\delta \ll 1$ is a suitably chosen parameter. No communication happens in the remaining fraction of time. Thus if p is the overall average power available for this piece of bandwidth, p/δ is the average power available when communication is being done. Thus the peakiness denoted by δ boosts the power level for actual communication by a factor of $1/\delta$. This boost in power level is necessary for the success of this orthogonal code. It ensures that the energy pulse transmitted (for the correct message) is strong enough to be identified from the ambient noise at the receiver.

Since the time-length of this code is T, total transmit energy λ for this code is Tp/δ. As communication happens for only δ fraction of time, the overall maximum achievable rate is

$$r = \delta\frac{\Phi\lambda}{T} = \delta\frac{\Phi T p/\delta}{T} = p\Phi \approx p\ln w$$

Since the total available power P is divided equally amongst K pieces of the total bandwidth, power p available per piece equals P/K. Hence the total rate over these K pieces equals $Kr = P\ln w$ nats per unit time. We choose $K = \ln w$ so that the total bandwidth of these K pieces ($W = wK$) satisfies $\ln W \approx \ln w$. Then the total rate is given by $R \approx P\ln W$ nats per unit time.

This rate is the same as the capacity of this fading channel when the receiver and transmitter both have full CSI [5]. This proves that the proposed coding scheme achieves the capacity for this channel with no receiver CSI. Thus the lack of receiver CSI does not reduce capacity.

Theorem 1: Capacity of the Rayleigh fading wideband channel with causal transmitted CSI and no receiver CSI is achieved by the proposed coding scheme. In the limit of large bandwidth, this capacity $C \approx P\ln W$ nats per unit time and is unchanged if even the receiver has full CSI.

Note that as mentioned in Section 1, full transmitter CSI is not needed for the proposed scheme; only one bit of CSI is enough for each channel gain $\mathbf{h}_i[j]$. This bit indicates whether or not the channel gain is above the threshold Φ. Also note that CSI is not needed at every time for this scheme. Since there is no activity for $(1-\delta)$ fraction of time and only δ fraction of the time is used for communication, only this δ fraction of time needs CSI and hence the cost of obtaining CSI is significantly reduced. Since the capacity of a wideband channel with full receiver and transmitter CSI (at all times) is essentially the same as the capacity of our channel with only transmitter CSI (for only a fraction δ of time), one may want to mimic no receiver CSI even when it is available!

[2] $\ln 2$ nats $= 1$ bit. Hence $\ln T$ nats equals $\log_2 T$ bits. Unless mentioned otherwise, units of information are nats.

[3] We have shown that the error probability of this orthogonal code goes to zero as w goes to infinity. However, a subtle point is that for showing a rate $\ln T/T$ is achievable, we have to show that arbitrarily small error probability can be achieved for a given (but large) w and T. This can be achieved by coding over many blocks of our orthogonal code by treating the orthogonal code as the inner code of this concatenated code. The orthogonal code provides an almost noiseless discrete memoryless channel (with input cardinality T) for the outer code. Thus a rate $\approx \ln T/T$ can be achieved.

We can extend above analysis for the case of noisy transmitter CSI, where the channel gain $\mathbf{h}_i[j]$ equals the sum of two independent complex Gaussian components, $\mathbf{g}_i[j]$ and $\mathbf{f}_i[j]$, which are i.i.d. over frequency and time. Transmitter only knows $\{\mathbf{g}_i[j]\}$ and the unknown component $\mathbf{f}_i[j]$ is independent of $\mathbf{g}_i[j]$. The variance of the known component is $\beta \in (0, 1]$. A code similar to the perfect CSI case can be employed. For example, if message 1 is to be transmitted, the transmitter transmits energy at time 1 only in the frequency band where the known component's strength $|\mathbf{g}_i[1]|^2$ is larger than $\beta\Phi$. Thus the threshold for the perfect transmitter CSI case is reduced by a factor of β. This scheme can be shown to achieve a rate of $\beta P \ln W$ nats per unit time. This again equals the capacity when receiver also has full CSI [5]. Thus again receiver CSI is irrelevant for capacity in wideband limit.

Remark 1: Similar results can be proved when distribution of the fading gain $|\mathbf{h}_i[j]|^2$ is not exactly exponential but has an exponential tail. If the tail behaves similar to an exponential with mean m, the capacity can be shown to be $mP \ln W$ nats per unit time. Similar analysis can be performed if the tail of the fading gain distribution is a polynomial, that is, $P(|\mathbf{h}_i[j]|^2 \geq x) \approx x^{-n}$ for some $n > 0$. In that case, the proposed code essentially achieves a rate $PW^{\frac{1}{n+1}}$ nats per unit time. This again turns out to be the same as the capacity when the receiver also has full CSI.

III. Capacity per cost with causal transmitter CSI

We saw in the previous section how the proposed orthogonal code achieved the capacity of the wide-band fading channel with causal CSIT and no receiver CSI. This section analyzes the case of causal CSIT for a more general channel.

The random variables at time $i \in \{1, 2, 3 \cdots\}$ corresponding to the channel input X_i, output Y_i and channel state S_i take values from the sets $\mathcal{X}, \mathcal{Y}$ and $\mathcal{S}$ respectively. State S defines a channel transition matrix denoted by $P_{Y|XS}$. Unless stated otherwise, capital letters denote random variables and small letters denote their values. Notation X_1^l is used as a shorthand for the sequence $X_1 X_2 \cdots X_l$. The states are assumed to change i.i.d. over time, that is, if $P_S(\cdot)$ denotes the distribution of S_i then the probability of a state-sequence s_1^l equals $\prod_{i=1}^{l} P_S(s_i)$. Conditioned on the state sequence, the channel is assumed memoryless, that is, $P(Y_1^l | X_1^l, S_1^l)$ equals $\prod_{i=1}^{l} P_{Y|XS}(y_i | x_i, s_i)$.

Each input $x \in \mathcal{X}$ incurs a cost $b(x) \in [0, \infty)$. A zero cost input is assumed to exist and denoted by "0". In a code of length l, the codeword for message j is denoted by the sequence $x_1^l(j)$. A length l code having $M \in \{1, 2 \cdots\}$ messages is denoted by a (l, M, ν, ϵ) code if the average probability of error is at most ϵ and codeword for every message j satisfies the total cost constraint $\sum_{i=1}^{l} b(x_i(j)) \leq \nu$. The capacity per unit cost for this channel is defined as [8].

Definition 2: For a given $0 \leq \epsilon < 1$, rate (in nats) per unit cost R is said to be ϵ-achievable if for all every $\gamma > 0$, there exists a ν_0 such that for all $\nu \geq \nu_0$, a (l, M, ν, ϵ) code can be found with $\ln M \geq \nu(R - \gamma)$. Rate per unit cost of R is said to be achievable if R is ϵ-achievable for every $\epsilon > 0$. Capacity per unit cost is the maximum achievable rate per unit cost.

We assume no receiver CSI and causal CSIT, which means that the transmitter gets to know S_i at time i before transmitting X_i. Let $U : \mathcal{S} \rightarrow \mathcal{X}$ denote a mapping from states to inputs. This mapping U is equivalent to a vector in $\mathcal{X}^{|\mathcal{S}|}$, where its each entry denotes the input mapped from the corresponding state. Let $P_{Y|U=u}(y)$ denote the output distribution induced when mapping $U = u$ is chosen.

$$P_{Y|U=u}(y) = \sum_{s \in \mathcal{S}} P_S(s) P_{Y|XS}(y | u(s), s) \tag{2}$$

where $u(s)$ denotes mapping of state s under u. We next prove the following theorem.

Theorem 3: Capacity per unit cost with no receiver CSI and causal CSIT is given by

$$\sup_u \frac{D(P_{Y|U=u} || P_{Y|U=0})}{\mathcal{E}\left[b(X)|U = u\right]}$$

where $D(P_{Y|U=u} || P_{Y|U=0})$ denotes the Kullback-Leibler divergence (in nats) between the output distributions induced when mapping u is chosen and when identically zero mapping is chosen. $\mathcal{E}\left[b(X)|U = u\right] = \sum_{s \in \mathcal{S}} P_S(s) b(u(s))$ denotes the average cost incurred when mapping u is chosen.

Achievability: We first show achievability with an orthogonal coding scheme. We use the shorthand $f(n) \doteq g(n)$ to denote $\lim_{n \to \infty} \frac{\ln f(n)}{\ln g(n)} = 1$. Similarly, $f(n) \dot{\leq} g(n)$ and $f(n) \dot{<} g(n)$ are defined.

Choose a mapping $u : \mathcal{S} \rightarrow \mathcal{X}$. Our code of M messages spans Mn symbols. Each message corresponds to a non-overlapping interval of length n, that is, message $j \in \{0, 1, \cdots, M - 1\}$ corresponds to the integer interval $[jn + 1, jn + n]$. If message j is to be transmitted, "0" is transmitted at all times except interval $[jn + 1, jn + n]$. During each time $i \in [jn + 1, jn + n]$, input $u(S_i)$ is transmitted. This requires only causal CSI at the encoder.

Assuming message j was transmitted, the output distribution at each time in interval $[jn + 1, jn + n]$ is given by $P_{Y|U=u}$. Outputs in all other intervals are distributed as $P_{Y|U=0}$. For each of the M intervals of length n, the decoder finds the empirical output distribution of that interval. Let P_Y^k denote this empirical distribution for interval $[kn + 1, kn + n]$. The interval k for which $D(P_Y^k || P_{Y|U=0})$ is the largest is declared as the decoded message. An error happens either if the decoded message is different from the transmitted message. Using Sanov's theorem and other standard tools, [9] shows that error probability of this scheme vanishes for any rate per unit cost smaller than $\frac{D(P_{Y|U=u} || P_{Y|U=0})}{\mathcal{E}[b(X)|U=u]}$.

Remark 2: Note the similarity of this scheme with the coding scheme in previous section for the wideband fading channel. In particular, the decoder chooses the message which scores the best in terms $D(P_Y^i || P_{Y|U=0})$ or the average received energy E_i. Moreover, distribution of $D(P_Y^i || P_{Y|U=0})$ and E_i are both of exponential nature. Hence one can interpret the divergence $D(P_Y^i || P_{Y|U=0})$ for interval i as the discrete channel analogue of the average received energy E_i for the wideband fading channel.

Converse: We first note the following upper bound in [8] on capacity per unit cost of a discrete memoryless channel with input V and output Z

$$\sup_{v} \frac{D(P_{Z|V=v}||P_{Z|V=0})}{c(v)} \tag{3}$$

where $P_{Z|V=v}$ denotes the output transition probability for input v, $c(v)$ denotes the cost of input v and $V = 0$ denotes the zero cost input.

Now recall Shannon's idea [1] that this channel with causal CSIT and i.i.d. states is equivalent to a discrete memoryless channel (DMC) with the same output alphabet but a larger input alphabet. The input alphabet U of that equivalent DMC corresponds to a mapping from $\mathcal{S}$ to $\mathcal{X}$. If the symbol u_i was transmitted at time i on the DMC, the transmitter with causal CSI transmits input $u_i(S_i)$ at time i after observing state S_i.

Finally, note that $\mathcal{E}[b(X)|U = u]$ denotes the (average) cost incurred due to choosing the DMC input u. The converse follows by applying (3) after replacing $P_{Z|V=v}$ by $P_{Y|U=u}$, $P_{Z|V=0}$ by $P_{Y|U=0}$ and $c(v)$ by $\mathcal{E}[b(X)|U = u]$. See [9] for a detailed proof using Fano's inequality.

The achievability result also could have been proved using this method of conversion to a DMC. However, the earlier detailed proof based on threshold based decoding is expected to be more insightful in view of writing on fading paper.

IV. Causal vs. non-causal transmitter CSI

For the wireless fading channel, we noted that in the wideband limit, the capacity with causal CSIT was the same as that with non-causal CSIT. Equivalently, the capacity per unit cost was the same with causal or non-causal CSIT. Similar phenomenon is shown in [7] for the dirty paper channel (AWGN channel with additive Gaussian interference known at the transmitter). It shows that for writing on dirty paper at low SNR, interference does not reduce capacity even with causal CSIT. Thus causal CSIT is as good as non-causal CSIT in this case. This result is proved by modifying the scheme in [6] to a threshold-based scheme similar to section II.

If with causal or non-causal CSIT, the capacities (at any given cost constraint) are the same for a channel, then the capacities per unit cost would naturally be the same. This is because for a channel with a 0 cost alphabet, capacity per unit cost is given by the slope of the capacity-cost curve at 0. More interesting problem is to characterize the class of channels for which the capacity-cost curves are not the same for causal and non-causal CSIT, but the capacity per unit cost is the same with causal or non-causal CSIT. Above mentioned wideband AWGN channel with additive interference and the wideband fading channel are two such channels.

A. Review of the non-causal transmitter CSI case

We briefly summarize the coding scheme that achieves the capacity per unit cost with non-causal CSIT [6]. This code of M messages spans Mqn symbols. Each message in this orthogonal code corresponds to a separate interval of length qn. For transmitting a message j, non-zero symbols can be only transmitted in the j'th interval of length qn. This message interval of length qn can be thought as the set of q subintervals, each of length n.

A distribution of states $\hat{P}_S(\cdot)$ is chosen beforehand. Out of these q subintervals in the interval for message j, the subinterval whose empirical distribution is like $\hat{P}_S(\cdot)$ is chosen. Since the actual distribution of states is P_S, the probability of a subinterval having empirical distribution like $\hat{P}_S(\cdot)$ essentially (in $\doteq$ sense) equals $\exp(-nD(\hat{P}_S||P_S))$. We can find such a subinterval with high probability if the number of subintervals per message interval is $q \doteq \exp(nD(\hat{P}_S||P_S))$. Non-zero symbols are only transmitted in this subinterval. A mapping $u : \mathcal{S} \to \mathcal{X}$ is also chosen beforehand. Similar to previous section, input $u(s)$ is transmitted for state s in this subinterval. Note that non-causal CSIT is necessary to determine the subinterval of empirical state distribution $\hat{P}_S(\cdot)$. The output distribution in this subinterval is

$$\hat{P}_Y(y) = \sum_{s\in\mathcal{S}} \hat{P}_S(s) P_{Y|XS}(y|u(s), s) \tag{4}$$

$$\text{and} \quad P_{Y|U=0}(y) = \sum_{s\in\mathcal{S}} P_S(s) P_{Y|XS}(y|0, s) \tag{5}$$

is the output distribution in all other subintervals (where only input 0 is transmitted).

Note that non-zero symbols are transmitted in a small fraction $(1/q)$ of the interval corresponding to message j. This fraction decays exponentially to 0 with increasing n. At the receiver, empirical output distribution is found for all the q subintervals for each of the M message intervals. If one of these Mq subintervals has distribution like $\hat{P}_Y$ in (4), the message interval containing that subinterval is declared as the transmitted message. An error is declared otherwise. Applying Stein's lemma and union bound on these subintervals, error probability vanishes for large n if $M \doteq \exp\Big(n(D(\hat{P}_Y||P_{Y|U=0}) - D(\hat{P}_S||P_S))\Big)$. By law of large numbers, the total cost incurred for transmission is essentially $n\mathcal{E}_{\hat{P}_S}[b(u(S))]$ where $\mathcal{E}_{\hat{P}_S}[b(u(S))] = \sum_{s\in\mathcal{S}} \hat{P}_S(s)b(u(s))$. So rate per unit cost achieved is

$$\frac{\ln M}{n\mathcal{E}_{\hat{P}_S}[b(u(S))]} \doteq \frac{D(\hat{P}_Y||P_{Y|U=0}) - D(\hat{P}_S||P_S)}{\mathcal{E}_{\hat{P}_S}[b(u(S))]} \tag{6}$$

Optimizing above expression over the choice of $\hat{P}_S$ and $u(\cdot)$ can be shown to yield the capacity per unit cost for this channel with non-causal CSIT. We denote an optimum choice by $\hat{P}_S^*$ and $u^*(\cdot)$, respectively.

B. Adapting to the causal transmitter CSI case

With causal CSIT, transmitter does not a priori know the subinterval having empirical state distribution $\hat{P}_S$. To overcome this issue, let there be only one subinterval per interval i.e. let $q = 1$. Thus each message corresponds to an interval of length n. Now a fraction θ is chosen by the transmitter. The transmitter can only transmit energy (non-zero symbols) in a fraction θ of message interval. For each state $s \in \mathcal{S}$, the transmitter will transmit input $u(s)$ for the first $n\theta\hat{P}_S(s)$ occurrences of state s. Thus the states where energy is transmitted will have an empirical distribution $\hat{P}_S$. Since the actual state distribution is P_S, (by law of large numbers) an interval of length n will have $n\theta\hat{P}_S(s)$ occurrences of state s only if

$$\theta\hat{P}_S(s) \leq P_S(s) \quad \forall s \in \mathcal{S} \quad \Rightarrow \theta \leq \inf_{s \in \mathcal{S}} \{P_S(s)/\hat{P}_S(s)\} \quad (7)$$

Note that the $n\theta$ symbols where energy is transmitted need not be in a contiguous block. By law of large numbers, the total cost incurred here is (in $\approx$ sense) essentially $n\theta\mathcal{E}_{\hat{P}_S}[b(u(S))]$.

At the decoder, for each message interval of length n, the empirical output distribution is found for all $\binom{n}{n\theta}$ subsequences of length $n\theta$. (The term "subinterval" is used for a contiguous block of symbols, whereas a "subsequence" need not be contiguous.) Out of these $M\binom{n}{n\theta}$ subsequences, if all the subsequences having distribution like $\hat{P}_Y$ in (4) belong to a single message interval, the message corresponding to that message interval is declared as the transmitted message. An error is declared if more than one or none of the message intervals have such subsequences. By law of large numbers, the correct subsequence of length $n\theta$ where energy is transmitted will have an empirical output distribution like $\hat{P}_Y$ with high probability for large n. A length $n\theta$ subsequence in an incorrect message interval will have an empirical output distribution like $\hat{P}_Y$ with probability $p_1 \doteq \exp(-\theta n D(\hat{P}_Y||P_{Y|U=0}))$. Applying union bound, the probability of a subsequence of an incorrect message interval having empirical output distribution like $\hat{P}_Y$ is bounded by $M\binom{n}{n\theta}p_1$. Hence vanishing error probability can be achieved if

$$M \doteq \frac{1}{\binom{n}{n\theta}p_1} \doteq \frac{\exp(\theta n D(\hat{P}_Y||P_{Y|U=0}))}{\exp(nH_b(\theta))}$$

by Sterling's approx. Thus rate per unit cost achieved here is

$$\begin{aligned} \frac{\ln M}{n\theta\mathcal{E}_{\hat{P}_S}[b(u(S))]} &= \frac{\theta D(\hat{P}_Y||P_{Y|U=0}) - H_b(\theta)}{\theta\mathcal{E}_{\hat{P}_S}[b(u(S))]} \\ &= \frac{D(\hat{P}_Y||P_{Y|U=0}) - H_b(\theta)/\theta}{\mathcal{E}_{\hat{P}_S}[b(u(S))]} \end{aligned}$$

For this to equal the capacity per unit cost with non-causal CSIT in (6), an optimum $\hat{P}_S^*$ maximizing (6) should satisfy

$$\begin{aligned} H_b(\theta)/\theta &= D(\hat{P}_S^*||P_S) \quad \text{Now apply (7) to get} \\ D(\hat{P}_S^*||P_S) &\leq \sum_{s \in \mathcal{S}} \hat{P}_S^*(s) \ln\frac{1}{\theta} = \ln(1/\theta) \leq H_b(\theta)/\theta \end{aligned}$$

Last step is met with equality either when $\theta = 1$ or when θ tends to zero. Equality in second step needs $\hat{P}_S^*/P_S = 1/\theta$ for all states having $\hat{P}_S^*(s) > 0$.

Case of $\theta = 1$ corresponds to $\hat{P}_S^* = P_S$. By law of large numbers, empirical distribution for each interval would be P_S with high probability. Thus the non-causal nature of transmitter CSI is rendered useless in this case because only 1 subinterval (i.e. $q = 1$) suffices per message interval.

Thus this coding scheme gives the following sufficient condition for the capacity per unit cost with causal or non-causal CSIT to be the same.

Theorem 4: Let μ denote $\inf_{s\in\mathcal{S}}\{P_S(s)/\hat{P}_S^*(s)\}$ for an optimum $\hat{P}_S^*$ that achieves the capacity per unit cost for non-causal CSIT in (6). Capacity per unit cost with causal and non-causal CSIT is equal if μ is either equal to 1 or arbitrarily small, and for all states in the support of $\hat{P}_S^*$ (i.e. states s having $\hat{P}_S^*(s) > 0$) achieve the infimum $\mu = P_S(s)/\hat{P}_S^*(s)$.

If μ tends to zero, the divergence $D(\hat{P}_S^*||P_S)$ should tend to infinity to satisfy the above condition. In other words, its arbitrarily rare to observe the source distribution where energy is transmitted. This is because (by Stein's lemma) the larger $D(\hat{P}_S^*||P_S)$ is, the rarer it is to have empirical distribution like $\hat{P}_S^*$ when actual state distribution is P_S.

Note that the wideband fading channel and the wideband writing on dirty paper [6] satisfy the above Theorem, which guarantees that capacity per unit cost is the same with causal or non-causal CSIT. The fraction of states θ where energy was transmitted was arbitrarily small in both these examples. Thus Theorem 4 explains some reasons for the equivalence of the capacity per unit cost with causal and non-causal CSI for those channels.

Theorem 4 also provides a new interpretation of the optimum scheme in [6] for capacity per unit cost with non-causal CSIT. Think of a n-symbol subinterval in Section IV-A as a super-symbol. Energy is transmitted in arbitrarily rare states of this super-channel. Their probability vanishes as $\exp(-nD(\hat{P}_S||P_S))$. Hence by Theorem 4, the capacity per unit cost of this super-channel is the same with causal or non-causal CSIT. Thus even if the original channel cannot satisfy Theorem 4, the idea of subintervals achieves the non-causal capacity per unit cost by converting the original channel to a super-channel for which Theorem 4 is satisfied.

ACKNOWLEDGEMENTS

Thanks to Robert Gallager for suggesting a simple On-Off fading channel, which prompted the writing on fading paper scheme. Shashi Borade also acknowledges numerous insightful comments and suggestions by Ashish Khisti.

REFERENCES

[1] C. E. Shannon, "Channels With Side Information at the Transmitter," *IBM Journal Research and Development*, Vol. 2, pp. 289–293, 1958.

[2] S. Gel'fand and M. Pinsker, "Coding for Channel with Random Parameters," *Problems of Control and Information Theory*, Vol. 9, No. 1, pp. 19—31, 1980.

[3] M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. 29, no. 3, pp. 439-441, May 1983.

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

[5] S. Borade and L. Zheng, "Wideband fading channels with feedback," *Proc. Allerton Conference*, Sept. 2004.

[6] T. Liu and P. Vishwanath, "Opportunistic orthogonal writing on dirty paper," submitted to *IEEE Trans. on Information Theory*.

[7] S. Borade, "Wideband dirty paper writing with causal transmitter CSI," preprint.

[8] S. Verdu, "On channel capacity per unit cost," *IEEE Transactions on Information Theory*, Vol. 36, No. 5, pp. 1019–1030, September 1990.

[9] S. Borade and L. Zheng, "Writing on Fading Paper and Causal Transmitter CSI," preprint at http://arxiv.org/PS_cache/cs/pdf/0511/0511081.pdf