Dispersion of the Discrete Arbitrarily-Varying Channel with Limited Shared Randomness

Oliver Kosut
School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ 85287
Email: okosut@asu.edu

Jörg Kliewer
Department of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102
Email: jkliewer@njit.edu

***Abstract*—The second-order behavior of the discrete memoryless arbitrarily-varying channel is considered in the fixed error regime when the encoder and decoder share randomness that is independent from the adversarial choice of state. The dispersion (coefficient of the second-order term) is exactly characterized for most channels of interest when infinite shared randomness is allowed, and it is shown that precisely the same dispersion is achievable with only $O(\log n)$ bits of shared randomness. We also show that the dispersion is identical to that of the non-adversarial channel induced by the adversary simply choosing an i.i.d. state sequence according to the correct distribution. Further, we present some remarks on the connection to the compound channel, as well as on cost constraints for input and state sequences.**

I. Introduction

Active adversaries are a significant threat against modern communication systems. This is particularly true of wireless systems, which make use of an inherently open medium that allows for a jammer to transmit into the same medium so as to disrupt the legitimate users. The arbitrarily-varying channel (AVC), one of the oldest problems in information theory, captures such an adversarial setup. The problem, with one legitimate transmitter, one legitimate receiver, and one adversarial transmitter, is modeled by a channel distribution $W(y|x,s)$, where x represents the legitimate input, s the adversarial input or "state", and y is the channel output. The adversary may choose a state sequence $(s_1,\dots,s_n)$ across the coding block in an arbitrary fashion, subject to certain knowledge constraints. In particular, in this paper we assume that the adversary knows the code used by the legitimate users and the message to be sent, but it does not know a random quantity shared by the legitimate users. Classical work on this problem in [1], [2] has found the AVC capacity with infinite shared randomness, and also shown that this same capacity can achieved with shared randomness of only $O(\log n)$ bits.

Second-order or dispersion results for information theory problems—in which the behavior of the maximum achievable rate as a function of blocklength is determined with fixed probability of error—date back to Strassen [3]. There has been significant renewed interest of late in such characterizations, particularly since the work of [4]. This work showed that the maximum number of messages M that can be sent across a memoryless channel with blocklength n can be written

$$\log M = nC - \sqrt{nV}Q^{-1}(\epsilon) + O(\log n)$$

where C is the capacity, V is the dispersion, and Q is the complementary Gaussian CDF. However, similar results for the AVC have not been discussed in the open literature to the best of our knowledge. In this work, we fill this void by exactly characterizing the dispersion with infinite shared randomness for most AVCs of interest—our bounds do not match only if both optimal input and state distributions are not unique. We find that the dispersion is identical to that of the non-adversarial channel that is induced if the adversary chooses an i.i.d. state sequence according to the correct distribution. Moreover, we show that, as with first-order behavior of the AVC, the dispersion does not change if the encoder and decoder share only $O(\log n)$ bits of randomness. In Sec. IV, we relate our results to earlier work on the dispersion of the compound channel from [5]. We further remark on input and state cost constraints in Sec. V.

This material is based upon work supported by the National Science Foundation under grants CNS-1526547 and CCF-1453718.

II. Preliminaries

A. Notation

Calligraphic letters such as $\mathcal{X}$ always refer to finite alphabets. Let $\mathscr{P}(\mathcal{X})$ be the simplex of distributions on $\mathcal{X}$. For some $P_X \in \mathscr{P}(\mathcal{X})$, we write $X \sim P_X$ to mean that X is a random variable drawn from distribution P_X. The probability measure is denoted by $\mathbb{P}$, and the expectation operator is denoted $\mathbb{E}$; the underlying distribution will be specified in context. Let $\mathscr{P}(\mathcal{Y}|\mathcal{X})$ be the set of conditional distributions $\tilde{W}(y|x)$ for $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. For any $P_X \in \mathscr{P}(\mathcal{X})$ and $\tilde{W} \in \mathscr{P}(\mathcal{Y}|\mathcal{X})$, and any blocklength n, we define the stationary-memoryless extensions of P_X and $\tilde{W}$ as $P_X^n \in \mathscr{P}(\mathcal{X}^n)$ and $\tilde{W}^n \in \mathscr{P}(\mathcal{Y}^n|\mathcal{X}^n)$ respectively, where

$$P_X^n(x^n) = \prod_{i=1}^n P_X(x_i), \qquad \tilde{W}^n(y^n|x^n) = \prod_{i=1}^n \tilde{W}(y_i|x_i).$$

Given $P_X \in \mathscr{P}(\mathcal{X})$ and $\tilde{W} \in \mathscr{P}(\mathcal{Y}|\mathcal{X})$, we write $P_X\tilde{W} \in \mathscr{P}(\mathcal{Y})$ where

$$(P_X\tilde{W})(y) = \sum_{x\in\mathcal{X}} P_X(x)W(y|x).$$

978-1-5090-4096-4/17/$31.00 ©2017 IEEE

Similarly, given $P_S \in \mathscr{P}(\mathcal{S})$ and $W \in \mathscr{P}(\mathcal{Y}|\mathcal{X}\times\mathcal{S})$, we write $P_S W \in \mathscr{P}(\mathcal{Y}|\mathcal{X})$ where

$$(P_S W)(y|x) = \sum_{s\in\mathcal{S}} P_S(s) W(y|x,s).$$

Note that $P_X P_S W \in \mathscr{P}(\mathcal{Y})$ is now also well defined. Given two distributions $P_X, Q_X \in \mathscr{P}(\mathcal{X})$, their relative entropy is

$$D(P_X\|Q_X) = \sum_{x\in\mathcal{X}} P_X(x)\log\frac{P_X(x)}{Q_X(x)}.$$

Given a sequence $x^n \in \mathcal{X}^n$, its type is given by $P_{x^n}(x) = \frac{1}{n}|\{k : x_k = k\}$. Let $\mathscr{P}_n(\mathcal{X})$ be the set of all types for n-length sequences in $\mathcal{X}^n$. For $P \in \mathscr{P}_n(\mathcal{X})$, let T_P be the type class for P; i.e. the set of all sequences $x^n \in \mathcal{X}^n$ with $P_{x^n} = P$. For any integer M, we write $[M] = \{1,\dots,M\}$. Finally, $\log$ and $\exp$ are assumed to have base e.

B. Problem Description

Let $\mathcal{X},\mathcal{S},\mathcal{Y}$ be alphabets corresponding to the input, state, and output respectively. An AVC is described by a conditional distribution $W \in \mathscr{P}(\mathcal{Y}|\mathcal{X}\times\mathcal{S})$. For integers M (number of messages), n (blocklength), and K (size of shared randomness), an (M,n,K)-code is given by an encoding function

$$\phi : [M]\times[K] \to \mathcal{X}^n$$

and a decoding function

$$\psi : \mathcal{Y}^n \times [K] \to [M].$$

For any message $m \in [M]$ and state sequence $s^n \in \mathcal{S}^n$, the probability of error is given by

$$\lambda(m,s^n) = \mathbb{P}\{\psi(Y^n,U) = m\}$$

where U is a uniform random variable on $[K]$, $X^n = \phi(m,U)$, and $Y^n \sim W^n(\cdot|X^n,s^n)$. The maximal probability of error is given by

$$\lambda = \max_{m\in[M],s^n\in\mathcal{S}^n} \lambda(m,s^n).$$

Note the implicit assumption that the state sequence may depend on the code and the message, but must be independent of the shared random variable U. We define an (M,n,∞)-code as one where the uniform random variable on $[K]$ is replaced by a uniform random variable on the continuous interval $[0,1]$. Let $M^\star(W,n,\epsilon,K)$ be the largest integer M such that there exists an (M,n,K) code where $\lambda \le \epsilon$.

C. Information Definitions

Given $P_X \in \mathscr{P}(\mathcal{X})$ and $\tilde{W} \in \mathscr{P}(\mathcal{Y}|\mathcal{X})$, we write the mutual information as

$$I(P_X,\tilde{W}) = \sum_{x\in\mathcal{X}} P_X(x) D(\tilde{W}(\cdot|x)\|P_X\tilde{W}).$$

The random coding capacity of an AVC is given by

$$C = \max_{P_X\in\mathscr{P}(\mathcal{X})}\ \min_{P_S\in\mathscr{P}(\mathcal{S})}\ I(P_X,P_S W) \tag{1}$$

It was shown in [2] that there exists a polynomial $p(n)$ such that,

$$\log M^\star(W,n,\epsilon,K) = nC + o(n)$$

as long as $K \ge p(n)$. Note that $I(P_X,P_S W)$ is concave in P_X and convex in P_S; thus the maximum/minimum in (1) is a saddlepoint, and the maximum and minimum can be interchanged without changing the value. Moreover, we may define sets of optimal input and state distributions as

$$\Pi_X = \{P_X : \min_{P_S} I(P_X,P_S W) = C\}$$
$$\Pi_S = \{P_S : \max_{P_X} I(P_X,P_S W) = C\}.$$

By the properties of convex-concave functions, Π_X and Π_S are convex sets. Moreover, $I(P_X,P_S W) = C$ for any $P_X \in \Pi_X, P_S \in \Pi_S$.

The information density is given by

$$\imath(x;y) = \log\frac{(P_S W)(y|x)}{(P_X P_S W)(y)}.$$

This quantity obviously depends on P_X and P_S; to reduce clutter, we write simply $\imath(x;y)$ and infer the underlying specific distributions from context. We now define several information variance quantities, which will be useful for stating our results. In each of the following, $(X,S,Y) \sim P_X\times P_S\times W$. The unconditional information variance is given by

$$U(P_X,P_S W) = \mathrm{Var}(\imath(X;Y)).$$

The X-conditional information variance is given by

$$V_X(P_X,P_S W) = \mathbb{E}\,\mathrm{Var}(\imath(X;Y)|X).$$

The S-conditional information variance is given by

$$V_S(P_X,P_S,W) = \mathbb{E}\,\mathrm{Var}(\imath(X;Y)|S).$$

Note that U and V_X depend only on the induced X to Y channel $P_S W$, but V_S depends on both P_S and W. It is known for a non-adversarial channel (cf. [4]) that $U = V_X$ for optimal input distributions. The following lemma, proved in Appendix A, asserts that all three of these information variances are equal for optimal input and state distributions.

Lemma 1: For any $P_X \in \Pi_X$ and $P_S \in \Pi_S$,

$$U(P_X,P_S W) = V_X(P_X,P_S W) = V_S(P_X,P_S,W).$$

Now let[1]

$$V_+ = \min_{P_X\in\Pi_X}\ \max_{P_S\in\Pi_S} V_S(P_X,P_S,W) \tag{2}$$
$$V_- = \max_{P_S\in\Pi_S}\ \min_{P_X\in\Pi_X} V_X(P_X,P_S W). \tag{3}$$

Note that if either Π_X or Π_S contain only a single element, then by Lemma 1, $V_+ = V_-$.

III. Main Results

All results are proved in Sec. VI. The following are our main converse and achievability results for infinite shared randomness. The following converse result is proved by assuming the state sequence is drawn from an i.i.d. distribution, and then appealing to the converse result from [4].

[1] In light of Lemma 1, these definitions may be equivalently written with U, V_X, or V_S. However, we specifically write V_+ in terms of V_S and V_- in terms of V_X to facilitate the discussion in Sec. V about cost constraints.

Theorem 2: For any $0 < \epsilon < 1/2$,

$$\log M^\star(W, n, \epsilon, \infty) \le C - \sqrt{nV_-}Q^{-1}(\epsilon) + O(\log n). \quad (4)$$

The following achievability result is proved using an i.i.d. codebook and a decoding rule in which the information density is maximized over all possible induced distributions by the state, followed by a dependence-testing bound.

Theorem 3: For any $0 < \epsilon < 1/2$,

$$\log M^\star(W, n, \epsilon, \infty) \ge C - \sqrt{nV_+}Q^{-1}(\epsilon) - O(\log n). \quad (5)$$

The following is a finite blocklength random code reduction result; i.e. it demonstrates that similar performance to infinite shared randomness can be achieved with limited randomness. The proof is along the lines of that of [6, Lemma 12.8]; the main difference is that a Chernoff bound is replaced by Hoeffding's inequality.

Theorem 4 (Random code reduction): For any ϵ and K, if

$$\epsilon' > \epsilon + \sqrt{\frac{\log M^\star(W, n, \epsilon, \infty) + n \log |\mathcal{S}|}{2K}}. \quad (6)$$

then $M^\star(W, n, \epsilon', K) \ge M^\star(W, n, \epsilon, \infty)$.

Theorem 4 leads to dispersion bounds for shared randomness consisting of a logarithmic number of bits (in particular, a uniform random variable on $[n^2]$) that are identical to those for infinite shared randomness. This stated as follows.

Corollary 5: The bounds in (4)–(5) hold for $M^\star(W, n, \epsilon, n^2)$.

IV. Relationship to the Compound Channel

It was shown by Ahlswede [7] that there is a close relationship between an AVC and the compound channel wherein an adversary selects a single memoryless channel from

$$\{(P_S W)^n : P_S \in \mathscr{P}(\mathcal{S})\}. \quad (7)$$

Indeed, it is easy to see that the capacity of this compound channel is exactly the random coding capacity C of the AVC. Moreover, [7] shows, using the so-called *robustification technique*, that any code for this compound channel can be used to construct a code for the AVC with only a polynomial increase in the probability of error.[2]

Due to the close relationship between these channels, it is natural to ask whether their dispersions match as well. In [5], dispersion results were derived for discrete memoryless compound channels under certain regularity conditions that include that described by (7), provided Π_X contains a single element. In fact, in this case the dispersion found in [5] exactly matches $V_+ = V_-$. However, it is unclear whether a version of Ahlswede's robustification technique can be used to derive the AVC dispersion directly from the compound channel dispersion, due to the polynomial increase in probability of error. Moreover, our results give tight dispersion bounds under slightly more general conditions than [5], such as when Π_X does not contain a single element but Π_S does. Additionally, we prove a $O(\log n)$ error term, which is not the case for all compound channels as shown in [5].[3]

V. Remarks on Input and State Constraints

A natural extension of the AVC is to consider cost constraints on the channel input and state. This has been considered classically in, for example, [8], [9]. In particular, if we impose a constraint on the input sequence of the form $\frac{1}{n}\sum_{i=1}^n g(x_i) \le \Gamma$ and a constraint on the state sequence of the form $\frac{1}{n}\sum_{i=1}^n \ell(s_i) \le \Lambda$, then the AVC capacity is again given by (1), where we restrict to P_X satisfying $\mathbb{E}g(X) \le \Gamma$ and P_S satisfying $\mathbb{E}\ell(S) \le \Lambda$.

As for the dispersion, the proofs in Sec. VI can indeed yield bounds on the dispersion of the AVC with cost constraints in some cases. This is formalized in the following corollary.

Corollary 6: Consider an AVC with input cost constraint Γ but no state cost constraint. Let Π_X (under a cost constraint) and Π_S be the optimal sets of distributions in (2)–(3). Then, (4) provides an upper bound for $\log M^\star(W, n, \epsilon, \infty)$. Likewise, consider an AVC with a state constraint Λ but no input cost constraint. Let Π_S (under a cost constraint) and Π_X be the optimal sets of distributions in (2)–(3). Then, (5) provides a lower bound for $\log M^\star(W, n, \epsilon, \infty)$.

Even though these bounds do not precisely characterize the dispersion, they do suggest that the dispersion under cost constraints differs from that of the associated compound channel in (7). For example, consider a state-constrained AVC given by $Y = X \oplus S$, where $X, S, Y \in \{0, 1\}$ and $\oplus$ denotes modulo-2 addition; the state cost function is $\ell(s) = s$ for $s \in \{0, 1\}$. For $\Lambda \le 1/2$, the capacity is $C = \log 2 - H(\Lambda)$, where $H(\cdot)$ is the binary entropy function. It is easy to see that Π_X contains only $\mathrm{Ber}(1/2)$ and Π_S contains only $\mathrm{Ber}(\Lambda)$, so $V_- = 0$. Thus, by Corollary 6, the dispersion is zero. On the other hand, the associated compound channel, which consists of all BSCs with crossover probabilities no more than Λ, has positive dispersion.

Finding the dispersion for cost constraints in general appears to require characterizing the asymptotic behavior of $\sum_{i=1}^n \imath(X_i; Y_i)$ where X^n and S^n are independent and uniform over type classes T_{P_X} and T_{P_S} respectively, and $Y^n \sim W^n(\cdot|X^n, S^n)$. A Berry-Esseen-type bound in this context is the subject of ongoing work.

VI. Proofs

A. Converse (Theorem 2)

Choose state distribution $P_S \in \Pi_S$ such that

$$\min_{P_X \in \Pi_X} V_X(P_X, P_S W) = V_-.$$

Any code that achieves probability of error ϵ for all state sequences s^n also achieves probability of error ϵ if the state sequence is drawn randomly from P_S^n. That is, any AVC code with probability of error ϵ also functions as a code for the ordinary memoryless channel with distribution $P_S W$, again with probability of error at most ϵ. The upper bound follows from [4].

[2] These observations from [7] are focused on the scenario in which the encoder knows the state sequence, but they apply equally well to the shared randomness setting, including when the encoder/decoder share $O(\log n)$ bits of randomness.

[3] On the other hand, it is likely that the particular compound channel in (7) would lead to a $O(\log n)$ error term.

B. Achievability (Theorem 3)

Choose input distribution $P_X \in \Pi_X$ such that

$$\max_{P_S \in \Pi_S} V_S(P_X, P_S, W) = V_+.$$

In the infinite randomness case, the encoder and decoder may share arbitrary random quantities that are independent from the state sequence. In particular, the encoder and decoder share a random codebook composed of M message drawn from the i.i.d. distribution P_X^n. Let $X^n(m)$ be the mth message in this codebook. To send message m, the encoder transmits $X^n(m)$. Upon receiving y^n, the the decoder chooses its message estimate as the message m that maximizes the following quantity, breaking ties arbitrarily:

$$\max_{P_S \in \mathscr{P}_n(\mathcal{S})} \log \frac{(P_S W)^n(y^n|X^n(m))}{(P_X P_S W)^n(y^n)}. \tag{8}$$

Suppose the state sequence is s^n. For any γ, we decode correctly if, for the correct codeword and $P_S = P_{s^n}$ the quantity in (8) is at least γ, and for all incorrect codewords and all $P_S \in \mathscr{P}_n(\mathcal{S})$ the quantity in (8) is less than γ. Thus we may upper bound the probability of error by

$$\begin{aligned}\lambda(m, s^n) \le{}& \mathbb{P}\left\{\log \frac{(P_{s^n} W)^n(Y^n|X^n)}{(P_X P_{s^n} W)^n(Y^n)} < \gamma\right\} \\ &+ M\,\mathbb{P}\left\{\max_{P_S \in \mathscr{P}_n(\mathcal{S})} \log \frac{(P_S W)^n(Y^n|\bar{X}^n)}{(P_X P_S W)^n(Y^n)} \ge \gamma\right\}, \end{aligned} \tag{9}$$

where $(X^n, Y^n, \bar{X}^n) \sim P_X^n(x^n) W^n(y^n|x^n, s^n) P_X^n(\bar{x}^n)$.

To bound the first term in (9), note that the random quantity is a sum of independent variables with mean

$$\begin{aligned}&\sum_{i=1}^n \mathbb{E} \log \frac{(P_{s^n} W)(Y_i|X_i)}{(P_X P_{s^n} W)(Y_i)} \\ &= n \sum_{x,s,y} P_X(x) P_{s^n}(s) W(y|x,s) \log \frac{(P_{s^n} W)(y|x)}{(P_X P_{s^n} W)(y)} \\ &= n I(P_X, P_{s^n} W).\end{aligned}$$

Moreover, the variance is given by

$$\sum_{i=1}^n \mathrm{Var} \log \frac{(P_{s^n} W)(Y_i|X_i)}{(P_X P_{s^n} W)(Y_i)} = n V_S(P_X, P_{s^n}, W).$$

Thus, if $V_S(P_X, P_{s^n}, W) > 0$, then the non-identically-distributed version of the Berry-Esseen theorem allows us to upper bound the first term in (9) by

$$Q\left(-\frac{\gamma - n I(P_X, P_{s^n} W)}{\sqrt{n V_S(P_X, P_{s^n}, W)}}\right) + \frac{B}{\sqrt{n}}$$

where B is a constant proportional to the third absolute moment of $\imath(X;Y)$. The second term in (9) is at most

$$\begin{aligned}&M \sum_{P_S \in \mathscr{P}_n(\mathcal{S})} \mathbb{P}\left\{\log \frac{(P_S W)^n(Y^n|\bar{X}^n)}{(P_X P_S W)^n(Y^n)} \ge \gamma\right\} &(10)\\ &= M \exp\{-\gamma\}\, |\mathscr{P}_n(\mathcal{S})| &(11)\end{aligned}$$

where (11) follows by first applying Markov's inequality and because, for any y^n, $\mathbb{E}(P_S W)^n(y^n|\bar{X}^n) = (P_X P_S W)^n(y^n)$. Thus $\lambda(m, s^n)$ is upper bounded by

$$Q\left(-\frac{\gamma - n I(P_X, P_{s^n} W)}{\sqrt{n V_S(P_X, P_{s^n}, W)}}\right) + \frac{B}{\sqrt{n}} + M \exp\{-\gamma\} |\mathscr{P}_n(\mathcal{S})|.$$

If we choose

$$\begin{aligned}\gamma = {}& \min_{P_S \in \mathscr{P}(\mathcal{S})} n I(P_X, P_S W) \\ &- \sqrt{n V_S(P_X, P_S, W)} Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n}}\right), &(12)\end{aligned}$$

$$\log M = \gamma - \log |\mathscr{P}_n(\mathcal{S})| - \frac{1}{2} \log n \tag{13}$$

then $\lambda(m, s^n) \le \epsilon$ for any state sequence s^n. Note that this holds even if $V_S = 0$, since in this case the choice of γ in (12) ensures that the first term in (9) is zero.

We now need to show that the right-hand side of (13) can be written as the right-hand side of (5). Recall that $\min_{P_S} I(P_X, P_S W) = C$ and $\max_{P_S \in \Pi_S} V_S(P_X, P_S, W) = V_+$. For any δ, let $\Pi_S(\delta) = \{P_S \in \mathscr{P}(\mathcal{S}) : d(P_S, \Pi_S) \le \delta\}$ where $d(P_S, \Pi_S)$ is the Euclidean distance. By [4, Lemma 64][4], if there exists constants $\delta > 0, f_1 > 0, f_2$ such that, for all $P_S \in \Pi_S(\delta)$,

$$I(P_X, P_S W) \ge C + f_1 d(P_S, \Pi_S)^2 \tag{14}$$

$$\left|\sqrt{V_S(P_X, P_S, W)} - \sqrt{V_+}\right| \le f_2 d(P_S, \Pi_S) \tag{15}$$

then (13) is lower bounded by

$$n C - \sqrt{n V_+} Q^{-1}\left(\epsilon - \frac{B+1}{\sqrt{n}}\right) - \log |\mathscr{P}_n(\mathcal{S})| - \frac{1}{2} \log n - O(1)$$

which proves the theorem since Q^{-1} has bounded derivative around ϵ and $\log |\mathscr{P}_n(\mathcal{S})| = O(\log n)$.

It remains to prove (14)–(15). Condition (15) follows from the facts that V_S is infinitely differentiable and $\Pi_S(\delta)$ is a compact set. Consider any $P_S' \notin \Pi_S$, and let $P_S^\star$ be the projection of P_S' onto Π_S. Let v be an $|\mathcal{X}| \cdot |\mathcal{Y}|$-dimensional vector given by the difference between distributions $P_S' W - P_S^\star W$. Let ∇I and $\mathcal{H}$ be the gradient vector and Hessian matrix respectively of $I(P_X, \tilde{W})$, as a function of $\tilde{W}$, evaluated at $\tilde{W} = P_S^\star W$. By Taylor expansion and the fact that $I(P_X, P_S^\star W) = C$,

$$I(P_X, P_S' W) - C = v^T \nabla I + \frac{1}{2} v^T \mathcal{H} v + o(\|v\|^2).$$

We have

$$\mathcal{H}_{xy,x'y'} = \delta_{y=y'} \left[\frac{P_X(x)}{(P_S^\star W)(y|x)} \delta_{x=x'} - \frac{P_X(x) P_X(x')}{(P_X P_S^\star W)(y)}\right].$$

The eigenvalues of $\mathcal{H}$ are those λ such that, for some y,

$$(P_X P_S^\star W)(y) = \sum_x \frac{P_X(x)^2}{\frac{P_X(x)}{(P_S^\star W)(y|x)} - \lambda}.$$

Note that for each y, $\lambda = 0$ satisfies this equation. Moreover, the right-hand side is strictly increasing for $\lambda < \min_x \frac{P_X(x)}{(P_S^\star W)(y|x)}$, so the next smallest value of λ satisfying

[4] In [4], this lemma was used to *upper* bound a maximization over P_X in the converse proof; here we use it to lower bound a minimization over P_S in the achievability proof.

the above at a value greater than $\min_x \frac{P_X(x)}{(P_S^\star W)(y|x)}$. Thus, the smallest non-zero eigenvalue may be bounded by

$$\lambda_{\min+}(\mathcal{H}) > \min_{x,y} \frac{P_X(x)}{(P_S^\star W)(y|x)} \geq \min_x P_X(x) = P_{X,\min}.$$

Note that the null space of $\mathcal{H}$ is composed of vectors $v_{xy} = c_y(P_S^\star W)(y|x)$ for some constants c_y. Let $v = v^0 + v^1$, where v^0 is in the null space of $\mathcal{H}$ and v^1 is orthogonal to it; i.e. $\sum_x v^1_{xy}(P_S^\star W)(y|x) = 0$ for all y. Note that

$$\begin{aligned} &I(P_X, P_S^\star W + v^0) \\ &= \sum_{x,y}(1+c_y)P_X(x)(P_S^\star W)(y|x)\log\frac{(P_S^\star W)(y|x)}{(P_X P_S^\star W)(y)} \\ &= C + \sum_{x,y} v^0_{xy} P_X(x)\log\frac{(P_S^\star W)(y|x)}{(P_X P_S^\star W)(y)}. \end{aligned}$$

Moreover, since $P^\star + \lambda P' \notin \Pi_S$ for any $\lambda > 0$, it must be that $I(P_X, P_S^\star W + \lambda v^0)$ is linear and strictly decreasing, so there exists a constant $\Gamma_1 > 0$ such that

$$I(P_X, P_S^\star W + v^0) - C \geq \Gamma_1\|v^0\| \geq \Gamma_1\|v^0\|^2$$

where in the latter inequality we have assumed $\delta \leq 1$. On the other hand,

$$v^T\mathcal{H}v = (v^1)^T\mathcal{H}v^1 \geq \lambda_{\min+}(\mathcal{H})\|v^1\|^2 \geq P_{X,\min}\|v^1\|^2.$$

Thus we have

$$I(P_X, P_S' W) - C \geq \Gamma_1\|v^0\|^2 + \frac{1}{2}P_{X,\min}\|v^1\|^2 + o(\|v\|^2).$$

This proves (14).

C. Random Code Reduction (Theorem 4)

Consider the code (ϕ, ψ) achieving $M^\star(W, n, \epsilon, \infty)$. Recall that both ϕ and ψ are functions of a random variable U uniformly drawn from the unit interval. For each $u \in [0,1]$, let $\lambda(m, s^n, u)$ be the probability of error for message m and state sequence s^n, conditioned on $U = u$. In particular, $\epsilon \geq \lambda(m, s^n) = \mathbb{E}\lambda(m, s^n, U)$. Let $U_1, \ldots, U_K$ be independent random variables, each uniform on the unit interval. Fix ϵ' satisfying (6). For fixed message m and state sequence s^n, we have, by Hoeffding's inequality and the fact that $\lambda \in [0,1]$,

$$\mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^K \lambda(m, s^n, U_k) > \epsilon'\right\} \leq e^{-2K(\epsilon'-\epsilon)^2}.$$

Thus, by the union bound

$$\begin{aligned} \mathbb{P}\left\{\frac{1}{K}\sum_{k=1}^K \lambda(m, s^n, U_k) > \epsilon' \text{ for any } m \in [M], s^n \in \mathcal{S}^n\right\} \\ \leq M|\mathcal{S}|^n e^{-2K(\epsilon'-\epsilon)^2} < 1 \end{aligned}$$

where the strict inequality follows from (6). Thus, there exists a set of deterministic $u_1, \ldots, u_K \in [0,1]$ where $\frac{1}{K}\sum_{k=1}^K \lambda(m, s^n, u_k) \leq \epsilon'$ for all m, s^n. Therefore, the (M, n, K)-code (ϕ', ψ') given by $\phi'(m,k) = \phi(m, u_k)$, $\psi'(y^n, k) = \psi(y^n, u_k)$ achieves maximal probability of error at most ϵ'.

D. Dispersion for Limited Randomness (Corollary 5)

The upper bound is obvious. To prove achievability, we may apply Theorem 4 with $K = n^2$. Fix ϵ' and $\delta > 0$, and let

$$\epsilon = \epsilon' - \sqrt{\frac{C + \delta + \log|\mathcal{S}|}{n}}.$$

Note that by Theorem 2, for sufficiently large n, $\log M^\star(W, n, \epsilon, \infty) \leq n(C+\delta)$. Thus this choice of ϵ satisfies (6). Therefore

$$\begin{aligned} &\log M^\star(W, n, \epsilon', n^2) \geq \log M^\star(W, n, \epsilon, \infty) \\ &\geq nC - \sqrt{nV_+}Q^{-1}\left(\epsilon' - \sqrt{\frac{C+\delta+\log|\mathcal{S}|}{n}}\right) - O(\log n) \\ &= nC - \sqrt{nV_+}Q^{-1}(\epsilon') - O(\log n). \end{aligned}$$

Appendix A
Proof of Lemma 1

Differentiating the mutual information with respect to P_X and P_S gives

$$\begin{aligned} \frac{\partial dI(P_X, P_S W)}{\partial dP_X(x)} &= D((P_S W)(\cdot|x)\|P_X P_S W) - 1 \\ \frac{\partial I(P_X, P_S W)}{\partial P_S(s)} &= \sum_{x,y} P_X(x)W(y|x,s)\log\frac{(P_S W)(y|x)}{(P_X P_S W)(y)}. \end{aligned}$$

Thus, by the optimality conditions for the maximum/minimum in (1), for all $P_X \in \Pi_X, P_S \in \Pi_S$ and x, s where $P_X(x) > 0$, $P_S(s) > 0$, we have

$$\begin{aligned} C &= D((P_S W)(\cdot|x)\|P_X P_S W) \\ C &= \sum_{x,y} P_X(x)W(y|x,s)\log\frac{(P_S W)(y|x)}{(P_X P_S W)(y)}. \end{aligned}$$

In particular, $C = \mathbb{E}[\imath(X;Y)|X] = \mathbb{E}[\imath(X;Y)|S]$ where $(X, S, Y) \sim P_X \times P_S \times W$. Therefore $U = V_X = V_S$.

References

[1] D. Blackwell, L. Breiman, and A. J. Thomasian, "The capacities of certain channel classes under random coding," *The Annals of Mathematical Statistics*, vol. 31, no. 3, pp. 558–567, 1960.

[2] R. Ahlswede, "Elemination of correlation in random codes for arbitrarily varing channels," *Probability Theory and Related Fields*, vol. 44, no. 2, pp. 159–175, 1978.

[3] V. Strassen, "Asymptotic approximations in Shannon's information theory," http://www.math.cornell.edu/ pmlut/strassen.pdf, Aug. 2009, english translation of original Russian article in Trans. Third Prague Conf. on Inform. Th., Statistics, Decision Functions, Random Processes (Liblice, 1962), Prague, 1964.

[4] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inform. Theory*, vol. 56, pp. 2307–2359, 2010.

[5] Y. Polyanskiy, "On dispersion of compound DMCs," in *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Oct 2013, pp. 26–32.

[6] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akademiai Kiado, 1981.

[7] R. Ahlswede, "Arbitrarily varying channels with states sequence known to the sender," *IEEE Transactions on Information Theory*, vol. IT-32, no. 5, pp. 621–629, Sep. 1986.

[8] I. Csiszár and P. Narayan, "Arbitrarily varying channels with constrained inputs and states," *IEEE Transactions on Information Theory*, vol. 34, no. 1, pp. 27–34, Jan 1988.

[9] ——, "The capacity of the arbitrarily varying channel revisited: positivity, constraints," *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 181–193, 1988.