On dispersion of compound DMCs

Yury Polyanskiy

***Abstract*—Code for a compound discrete memoryless channel (DMC) is required to have small probability of error regardless of which channel in the collection perturbs the codewords. Capacity of the compound DMC has been derived classically: it equals the maximum (over input distributions) of the minimal (over channels in the collection) mutual information. In this paper the expression for the channel dispersion of the compound DMC is derived under certain regularity assumptions on the channel. Interestingly, dispersion is found to depend on a subtle interaction between the channels encoded in the geometric arrangement of the gradients of their mutual informations. It is also shown that the third-order term need not be logarithmic (unlike single-state DMCs). By a natural equivalence with compound DMC, all results (dispersion and bounds) carry over verbatim to a common message broadcast channel.**

I. Introduction

An abstract compound channel is a triplet: measurable spaces of inputs A and outputs B and a collection of conditional probability measures $P_{Y_s|X} : \mathsf{A} \mapsto \mathsf{B}$ indexed by elements $s \in \mathcal{S}$ of a measurable space. Let M be a positive integer and $\epsilon \in [0, 1)$. An $(M, \epsilon)_{noCSI}$ *code*[1] is a pair of (possibly randomized) maps $\mathsf{f} : [M] \to \mathsf{A}$ (the encoder) and $\mathsf{g} : \mathsf{B} \to [M]$ (the decoder), satisfying

$$\mathbb{P}[\mathsf{g}(Y_s) \neq m | X = \mathsf{f}(m)] \leq \epsilon \qquad \forall s \in \mathcal{S}, \forall m \in [M] \quad (1)$$

An $(M, \epsilon)_{CSIR}$ *code* is a pair of (possibly randomized) maps $\mathsf{f} : [M] \to \mathsf{A}$ (the encoder) and $\mathsf{g} : \mathsf{B} \times \mathcal{S} \to [M]$ (the decoder with access to channel state), satisfying

$$\mathbb{P}[\mathsf{g}(Y_s, s) \neq m | X = \mathsf{f}(m)] \leq \epsilon \qquad \forall s \in \mathcal{S}, \forall m \in [M] \quad (2)$$

This paper focuses on the case where A and B are n-fold Cartesian products of finite alphabets $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{S}$ is a *finite* set and transformations $P_{Y_s|X}$ are n-fold i.i.d. products:

$$P_{Y_s|X}(y^n|x^n) \stackrel{\triangle}{=} \prod_{j=1}^{n} W^{(s)}(y_j|x_j)\,,$$

where $W^{(s)} : \mathcal{A} \to \mathcal{B}$ are stochastic matrices, n is the blocklength. An (M, ϵ) code for the n-fold product is denoted as (n, M, ϵ) code. Finally, finite-blocklength fundamental limits for both types of codes are defined to be

$$M^*_{noCSI}(n, \epsilon) \stackrel{\triangle}{=} \max\{M : \exists (n, M, \epsilon)_{noCSI}\text{-code}\} \quad (3)$$

$$M^*_{CSIR}(n, \epsilon) \stackrel{\triangle}{=} \max\{M : \exists (n, M, \epsilon)_{CSIR}\text{-code}\}\,. \quad (4)$$

The author is with the Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, 02139 USA. e-mail: yp@mit.edu. This material is based upon work supported by the National Science Foundation CAREER award under grant agreement CCF-12-53205 and by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

[1]Abbreviation "noCSI" stands for no channel state information, while "CSIR" stands for channel state information at the receiver.

A classical result of Blackwell, Breiman and Thomasian [1] states

$$\log M^*_{noCSI}(n, \epsilon) = nC + o(n)\,, \qquad n \to \infty \quad (5)$$

$$\log M^*_{CSIR}(n, \epsilon) = nC + o(n)\,, \qquad n \to \infty\,, \quad (6)$$

where

$$C = \max_{P_X} \min_{s \in \mathcal{S}} I(P_X, W^{(s)})\,. \quad (7)$$

Wolfowitz [2] established a refinement of (5) and (6) showing that

$$\log M^*(n, \epsilon) = nC + O(\sqrt{n}), \qquad n \to \infty$$

This note refines this expression further, by providing the exact coefficient in a $\sqrt{n}$-term. Results on $O(\sqrt{n})$ have been classically established for discrete memoryless channels (DMCs) by Dobrushin and Strassen [3], [4], and other channels more recently. Motivation for studying $\sqrt{n}$ terms comes from the problem of predicting finite blocklength behavior of fundamental limits [5]. See [6] for a survey.

We find that the channel dispersion [5] of a compound DMC is given by:

$$\sqrt{V} = \max_{v: \sum_{x \in \mathcal{A}} v_x = 0} \min_{s} dI_s(v) - \sqrt{V(P^*_X, W^{(s)})}\,, \quad (8)$$

where minimum is over states s with $I(P^*_X, W^{(s)}) = C$ and dI_s is the differential of the mutual information:

$$dI_s(v) \stackrel{\triangle}{=} \sum_{x \in \mathcal{A}} v_x D(W^{(s)}_x || P^*_{Y_s}) \quad (9)$$

$$W^{(s)}_x(y) \stackrel{\triangle}{=} W^{(s)}(y|x) \quad (10)$$

$$P^*_{Y_s}(y) \stackrel{\triangle}{=} \sum_{x \in \mathcal{A}} W^{(s)}(y|x) P^*_X(x) \quad (11)$$

(see Section III for more on notation). More precisely, we prove the following:

Theorem 1: Consider a finite-state compound DMC. Assume

1) The capacity achieving input distribution P^*_X (maximizer in (7)) is unique.
2) $P^*_X(x) > 0$ for all $x \in \mathcal{A}$.
3) $V(P^*_X, W^{(s)}) > 0$ for all $s \in \mathcal{S}$.

Then for any $\epsilon \in (0, \frac{1}{2})$ we have[2]

$$\log M^*(n, \epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + o(\sqrt{n}), n \to \infty \quad (12)$$

for both the noCSI and CSIR codes.

[2]$Q^{-1}(\epsilon)$ is the functional inverse of the Q-function: $Q(x) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt$

Remarks:

1) Somewhat counter-intuitively, the dispersion V is not the maximal (worst) dispersion among channels s attaining $I(P_X^*, W^{(s)}) = C$. Rather it depends on a subtle interaction between channels' mutual informations and dispersions.
2) For two-state channel the expression for V simplifies:

$$\sqrt{V} = \frac{a_2}{a_1+a_2}\sqrt{V_1} + \frac{a_2}{a_1+a_2}\sqrt{V_2}\,, \qquad (13)$$

where

$$a_s^2 = \sum_{x\in\mathcal{A}} D(W_x^{(s)}\|P_{Y_s}^*)^2 - \frac{1}{|\mathcal{A}|}\left(\sum_{x\in\mathcal{A}} D(W_x^{(s)}\|P_{Y_s}^*)\right)^2 \qquad (14)$$

and provided $a_1 \neq 0$ or $a_2 \neq 0$. If both are zero then

$$V = \max(V_1, V_2)\,. \qquad (15)$$

3) Unlike [5], in this paper we do not provide experimental validation for the tightness of approximation (12) at realistic blocklengths. Thus results here are purely asymptotic although we did attempt to provide bounds that (we expect) to be quite competitive at finite blocklengths.
4) Section IV constructs an example of the channel for which the $o(\sqrt{n})$ term is $\theta(n^{\frac{1}{4}})$ – this is in contrast to all the known examples of expansions (12) (such as DMCs, Gaussian channels, etc), for which the $o(\sqrt{n})$ term is known to be $O(\log n)$.
5) It should be noted that for *composite* channels one assumes a prior over states $\mathcal{S}$ and consequently defines probability of error as averaged over the state $s \in \mathcal{S}$ (as opposed to worst-case definitions (1) and (2)). For such channels, the capacity becomes a function of probability of error ϵ. For finite-state channels, the dispersion term is similar to (12) with argument of Q^{-1} modified, see [7]. However, for the continuum of states the dispersion term may disappear, a surprising effect arising for example in (single- or multiple-antenna) wireless channels, see [8].
6) Finally, we note that coding for a compound channel (with CSIR) is equivalent to a problem of *common message broadcast channel*. Thus, Theorem 1 and the rest of this note applies equally well to this question in multi-user information theory.

II. Abstract achievability bounds

In this section we present two general achievability bounds (noCSI and CSIR). Although the proof of Theorem 1 requires only one of these and only a very special particularization of it, we prefer to formulate general versions for two reasons:

1) The proof of Theorem 1 reduces noCSI case to CSIR case by training. This is not possible for infinite alphabet/state cases, and hence a direct noCSI bound is necessary.
2) For numerical evaluations, crude bounds sufficient to establish (12) will need to be replaced with exact computation of the theorems in this section.

Given a pair of distributions P and Q on common measurable space W, a randomized test between those two distributions is defined by a random transformation $P_{Z|W} : \mathsf{W} \mapsto \{0,1\}$ where 0 indicates that the test chooses Q. Performance of the best possible hypothesis test (HT) is given by

$$\beta_\alpha(P,Q) \triangleq \min \int P_{Z|W}(1|w)Q(dw)\,, \qquad (16)$$

where the minimum is over all probability distributions $P_{Z|W}$ satisfying

$$P_{Z|W} : \int P_{Z|W}(1|w)P(dw) \geq \alpha\,. \qquad (17)$$

The minimum in (16) is guaranteed to be achieved by the Neyman-Pearson lemma. An abbreviated version of this definition is:

$$\beta_\alpha(P,Q) \stackrel{\triangle}{=} \inf_{E:P[E]\geq\alpha} Q[E]\,.$$

With this convention we similarly define HT between collections of distributions as follows:

$$\beta_\alpha(\{P_s, s\in\mathcal{S}\}, \{Q_s, s\in\mathcal{S}'\}) \stackrel{\triangle}{=} \inf_{E:\min_{\mathcal{S}} P_s[E]\geq\alpha} \max_{\mathcal{S}'} Q_s[E]\,.$$

Theorem 2 (noCSI codes): Fix a distribution Q_Y on B, $\tau \in (0,\epsilon)$ and a subset $\mathsf{F} \subseteq \mathsf{A}$. There exists an $(M,\epsilon)_{noCSI}$ code with encoder $\mathsf{f} : [M] \to \mathsf{F}$ and

$$M \geq \frac{\tilde{\kappa}_\tau}{\beta_{1-\epsilon+\tau}}\,,$$

where

$$\beta_\alpha = \sup_{x\in\mathsf{F}, s\in\mathcal{S}} \beta_\alpha(P_{Y_s|X=x}, Q_Y) \qquad (18)$$

$$\tilde{\kappa}_\tau = \inf_E Q_Y[E] \qquad (19)$$

and infimum in the definition of κ is over all sets E with the property

$$\forall x \in \mathsf{F}\ \exists s : P_{Y_s|X=x}[E] \geq \tau\,. \qquad (20)$$

Proof: The proof is a natural extension of the original $\kappa\beta$ bound [5, Theorem 25] and is omitted. ■

Theorem 3 (CSIR codes): Fix distributions $Q_{Y_s}, s\in\mathcal{S}$ on B, $\tau \in (0,\epsilon)$ and a subset $\mathsf{F} \subseteq \mathsf{A}$. There exists an $(M,\epsilon)_{CSIR}$ code with encoder $\mathsf{f} : [M] \to \mathsf{F}$ and

$$M \geq \frac{\kappa_\tau}{\beta_{1-\epsilon+\tau}}\,, \qquad (21)$$

where

$$\beta_\alpha = \sup_{x\in\mathsf{F}, s\in\mathcal{S}} \beta_\alpha(P_{Y_s|X=x}, Q_{Y_s}) \qquad (22)$$

$$\kappa_\tau = \inf_E \sup_{s\in\mathcal{S}} Q_{Y_s}[E] \qquad (23)$$

and infimum in the definition of κ is over all sets E with the property (20).

Proof: Again we assume familiarity with the (Feinstein-type) argument in the proof of [5, Theorem 25]. Suppose codewords $c_1, \dots, c_M$ have already been selected. To each codeword c there is a collection of sets $\{E_{c,s}, s \in \mathcal{S}\}$ satisfying for each $s \in \mathcal{S}$

$$P_{Y_s|X=c}[E_{c,s}] \geq 1 - \epsilon + \tau \,, \tag{24}$$

$$Q_{Y_s}[E_{c,s}] \leq \beta_{1-\epsilon+\tau} \,. \tag{25}$$

The decoder g inspects channel state s, the channel output Y_s and declares the message estimate as follows:

$$\mathsf{g}(s, y) \stackrel{\triangle}{=} \min\{j : y \in E_{c_j,s}\} \,.$$

Suppose that probability of error criterion (2) is satisfied with this decoder and codebook $\{c_1, \dots, c_M\}$, but that we can not grow the codebook without violating (2). This means

$$\forall x \exists s : P_{Y_s|X=x}\left[E_{x,c} \setminus \bigcup_{j=1}^{M} E_{c_j,s}\right] < 1 - \epsilon \,.$$

Applying the union bound and (24) with $c = x$ we find out

$$\forall x \exists s : P_{Y_s|X=x}\left[\bigcup_{j=1}^{M} E_{c_j,s}\right] \geq \tau \,.$$

Thus by the definition of κ_τ we must have

$$\sup_s Q_{Y_s}\left[\bigcup_{j=1}^{M} E_{c_j,s}\right] \geq \kappa_\tau \tag{26}$$

But from (25)

$$Q_{Y_s} \leq M\beta_{1-\epsilon+\tau} \,. \tag{27}$$

Clearly, (26) and (27) imply (21). ∎

In applications computation of κ_τ either requires certain symmetrization tricks, cf. [5, Appendix D], or the following method (applicable to finite-state channels only). Suppose that Q_{Y_s} in Theorem 3 have the following property:

$$Q_{Y_s}[E] = \int_{\mathsf{A}} P_{Y_s|X=x}[E] P_X(dx) \,, \tag{28}$$

for some distribution P_X. In words: Q_{Y_s} is the distribution induced by the channel $P_{Y_s|X}$ under input P_X. Then for any set satisfying (20) we have:

$$\sum_{s\in\mathcal{S}} P_{Y_s|X=x}[E] \geq \tau 1_{\mathsf{F}}(x)$$

Averaging this over P_X we obtain

$$\sum_{s\in\mathcal{S}} Q_{Y_s}[E] \geq \tau P_X[\mathsf{F}] \,,$$

thus implying that

$$\max_{s\in\mathcal{S}} Q_{Y_s}[E] \geq \frac{\tau P_X[\mathsf{F}]}{|\mathcal{S}|} \,.$$

Since the set E was arbitrary we have shown that under assumption (28) the κ_τ in Theorem 3 is lower-bounded as

$$\kappa_\tau \geq \frac{\tau P_X[\mathsf{F}]}{|\mathcal{S}|} \,. \tag{29}$$

Same argument shows that for $\tilde{\kappa}_\tau$ defined in (19), the lower bound (29) holds when $Q_Y = \frac{1}{|\mathcal{S}|}\sum_{s\in\mathcal{S}} Q_{Y_s}$ and distributions Q_{Y_s} defined in (28).

III. Proof

A. Notation

We recall the notation and relevant results from [5]. Let W be a stochastic matrix, P distribution on $\mathcal{A}$.

- *conditional output distribution* $W_x(y) \stackrel{\triangle}{=} W(y|x)$
- *output distribution* PW $PW(y) = \sum_{x\in\mathcal{A}} P(x)W(y|x)$.
- *mutual information*

$$I(P, W) = \sum_{x\in\mathcal{A}}\sum_{y\in\mathcal{B}} P(x)W(y|x)\log\frac{W(y|x)}{PW(y)} \,. \tag{30}$$

- *divergence variance*

$$V(P\|Q) = \sum_{x\in\mathcal{A}} P(x)\left[\log\frac{P(x)}{Q(x)}\right]^2 - D(P||Q)^2 \,. \tag{31}$$

- *conditional information variance*

$$V(P, W) = \sum_{x\in\mathcal{A}} P(x)V(W_x\|PW) \tag{32}$$

- Asymptotic estimate of β_α: Let U be a subset of distributions on $\mathcal{A}$ with the property that $\inf_{P\in U} V(P, W) > 0$. Then there exists a constant K such that for every $x^n \in \mathcal{A}^n$ with type P in U we have

$$\log\beta_\alpha(\prod W_{x_j}, (PW)^n) = \\ - nI(P, W) - \sqrt{nV(P, W)}Q^{-1}(\alpha) + \frac{1}{2}\log n \pm K \,. \tag{33}$$

  (see [9, Lemma 14]). For all x^n we have

$$\log\beta_\alpha(\prod W_{x_j}, (PW)^n) \geq -nI(P, W) - \sqrt{\frac{nK}{\alpha}} + \log\alpha \tag{34}$$

  (see [9, Lemma 15]).
- Functions $P \mapsto I(P, W)$ and $P \mapsto V(P, W)$ are smooth on the interior of the simplex of distributions on $\mathcal{A}$[3]
- Differential of the mutual information at a point P in the interior:

$$d_P I(v) \stackrel{\triangle}{=} \sum_{x\in\mathcal{A}} v_x D(W_x||PW)$$

- Linear-quadratic property of mutual information: For each P and direction v the function

$$t \mapsto I(P + tv, W) \tag{35}$$

  is constant if and only if $d_P I(v) = 0$ and $v \in \ker W$. If $d_P I(v) \neq 0$ then function (35) is upper-bounded by

$$t \mapsto -t \cdot d_P I(v)$$

[3]Here and everywhere below, we consider the simplex $\{P : \sum_{x\in\mathcal{A}} P(x) = 1, P(x) \geq 0\}$ as a manifold with boundary. Consequently, when computing differentials and gradients we should remember that $P(x) = P_x$ are *not* independent coordinate functions because $\sum_x dP_x = 0$.

everywhere in the domain of the definition. If $d_P I(v) = 0$ but $v \notin \ker W$ then the function (35) is upper-bounded by

$$t \mapsto -t^2 \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} v_x W(y|x) \right)^2 \tag{36}$$

everywhere in the domain of the definition. In the latter case, the function (35) is also lower-bounded by

$$t \mapsto -Kt^2 \sum_{y\in\mathcal{B}} \left(\sum_{x\in\mathcal{A}} v_x W(y|x) \right)^2 \tag{37}$$

in some neighborhood of zero with $K = 2(\min\{PW(y) : PW(y) > 0\})^{-1}$. These statements follow from the formula for the Hessian of I, see [5, (504)].

B. Maximization lemma

Lemma 4: Let U be a compact convex neighborhood of zero in $\mathbb{R}^d$, with $\mathbb{R}$-valued functions f_s, g_s, $s \in \mathcal{S}$ defined on U. Assume

1) $\mathcal{S}$ is finite
2) f_s are concave and continuous on U, and differentiable at 0
3) g_s are continuous and bounded on U
4) function $f_{min}(x) \stackrel{\triangle}{=} \min_s f_s(x)$ possesses unique maximum at 0.

Then as $\delta \to 0$

$$\max_{x\in U} \min_{s\in\mathcal{S}} f_s(x) + \delta g_s(x) = f_{min}(0) + \delta G + o(\delta), \tag{38}$$

where G is a solution to a piecewise-linear program:

$$G = \max_{x\in\mathbb{R}^d} \min_s (x, \nabla f_s(0)) + g_s(0) \tag{39}$$

minimum taken over s satisfying $f_s(0) = f_{min}(0)$. Furthermore,

$$\min_s g_s(0) \le G \le \max_s g_s(0)\,. \tag{40}$$

Proof: Without loss of generality, assume $f_{min}(0) = 0$. Also by boundedness of g_s, for sufficiently small δ we may restrict the minimization over s in (38) to states s achieving $f_s(0) = 0$. Therefore, we may further assume that $f_s(0) = 0$ for all s.

Denote for convenience $L_s = \nabla f_s(0)$ and notice that by uniqueness of the maximum of f_{min} we have

$$\max_{x\in\mathbb{R}^d} \min_{s\in\mathcal{S}} (L_s, x) = 0\,.$$

Therefore, the value G defined by (39) is finite and satisfies (40). Next, we show that for sufficiently small δ maximum in (38) can be restricted to any compact ball $B \subset U$ surrounding 0. Indeed, by continuity of f_{min} and compactness of U we have

$$\sup_{x\in U\setminus B} f_{min}(x) < -\epsilon_1$$

for some $\epsilon_1 > 0$. Thus, if c is constant lower-bounding all g_s on U we have

$$\sup_{x\in U\setminus B} \min_s f_s(x) + \delta g_s(x) \le \sup_{x\in U\setminus B} f_{min}(x) + \delta c < -\frac{\epsilon_1}{2}\,,$$

for all sufficiently small δ. Therefore, in solving (38) any choice of $x \in U \setminus B$ is worse than $x = 0$ for all sufficiently small δ.

Fix arbitrary $\epsilon > 0$ and select compact ball $B \subset U$ so that it includes 0 and

$$g_s(x) \le g_s(0) + \epsilon \qquad \forall x \in B\,. \tag{41}$$

We have then the following chain of estimates:

$$\begin{aligned}
\max_{x\in U} \min_{s\in\mathcal{S}} f_s(x) + \delta g_s(x) &= \max_{x\in B} \min_{s\in\mathcal{S}} f_s(x) + \delta g_s(x) && (42)\\
&\le \max_{x\in B} \min_{s\in\mathcal{S}} f_s(x) + \delta g_s(0) + \delta\epsilon && (43)\\
&\le \max_{x\in B} \min_{s\in\mathcal{S}} (L_s, x) + \delta g_s(0) + \delta\epsilon && (44)\\
&\le \max_{x\in\mathbb{R}^d} \min_{s\in\mathcal{S}} (L_s, x) + \delta g_s(0) + \delta\epsilon && (45)\\
&= \delta G + \delta\epsilon && (46)
\end{aligned}$$

where (42) holds for sufficiently small δ by the previous argument, (43) is by (41), (44) is by concavity of f_s, (45) by extending the domain of maximization and (46) by noticing that solution of (39) scales linearly with scaling of $g_s(0)$ by δ. Finally, by arbitrariness of $\epsilon > 0$ we have shown

$$\max_{x\in U} \min_{s\in\mathcal{S}} f_s(x) + \delta g_s(x) \le \delta G + o(\delta)\,. \tag{47}$$

For the lower bound, let x^* be a solution in (38).

$$\begin{aligned}
&\liminf_{\delta\to 0} \frac{1}{\delta} \max_{x\in U} \min_s f_s(x) + \delta g_s(x)\\
&\ge \liminf_{\delta\to 0} \min_s \frac{1}{\delta} f_s(\delta x^*) + g_s(\delta x^*) && (48)\\
&= \min_s \liminf_{\delta\to 0} \left(\frac{1}{\delta} f_s(\delta x^*) + g_s(\delta x^*) \right) && (49)\\
&= \min_s (L_s, x^*) + g_s(0) && (50)\\
&= G\,, && (51)
\end{aligned}$$

where (48) follows since $\delta x^* \in U$ for sufficiently small δ, (49) is by continuity of the minimum of finitely many arguments, (50) is by differentiability of f_s and continuity of g_s at 0, and (51) is by the definition of x^*. ∎

C. Converse part

For the converse part of Theorem 1 we observe that any $(n, M, \epsilon)_{CSIR}$ code contains an $(n, M', \epsilon)_{CSIR}$-subcode of constant composition P and size

$$\log M' \ge \log M - O(\log n)\,.$$

Therefore, it is sufficient to show

$$\log M' \le nC - \sqrt{nV} Q^{-1}(\epsilon) + o(\sqrt{n})\,. \tag{52}$$

The subcode has maximal probability of error upper-bounded by ϵ on every constituent DMC $W^{(s)}$. By the meta-converse method, see [5, Theorem 30], we have

$$\log M' \le \inf_{Q_{Y^n}} \sup_{x^n \in T_P^n} -\log \beta_{1-\epsilon}(\prod W_{x_j}^{(s)}, Q_{Y^n}), \qquad \forall s \in \mathcal{S} \tag{53}$$

where T_P^n is the n-type of composition P. We will further relax the bound by selecting $Q_{Y^n} = (PW^{(s)})^n$.

Let U denote the compact neighborhood of P_X^* on the simplex of distributions on $\mathcal{A}$ such that $\inf_{P\in U, s\in\mathcal{S}} V(P, W^{(s)}) > 0$. If $P \notin U$ then by uniqueness assumption on P_X^* we have

$$\min_s I(P, W^{(s)}) < C - \epsilon_1$$

for some $\epsilon_1 > 0$ which only depends on U. Thus there exists some state s such that $I(P, W^{(s)}) < C - \epsilon_1$. Consequently, from (34) we get

$$\sup_{x^n \in T_P^n} -\log \beta_{1-\epsilon}(\prod W_{x_j}^{(s)}, (PW^{(s)})^n) \le nC - n\epsilon_1 + \sqrt{nK'} \tag{54}$$

for some $K' > 0$. Then (53) and (54) evidently imply (52).

If $P \in U$ then by (33) we have

$$-\log \beta_{1-\epsilon}(\prod W_{x_j}^{(s)}, (PW^{(s)})^n) \le$$
$$nI(P, W^{(s)}) - \sqrt{nV(P, W^{(s)})}Q^{-1}(\epsilon) + \frac{1}{2}\log n + K \tag{55}$$

From (53) and the above we get (by minimizing over s)

$$\log M' \le \frac{1}{2}\log n + K$$
$$+ \min_s nI(P, W^{(s)}) - \sqrt{nV(P, W^{(s)})}Q^{-1}(\epsilon). \tag{56}$$

Taking maximum over $P \in U$ of the second term and applying Lemma 4 with $\delta = \frac{Q^{-1}(\epsilon)}{\sqrt{n}}$ we get (52).

D. Achievability part

We aim to invoke Theorem 3. However, since the claim in Theorem 1 is made for noCSI and CSIR codes, we first notice that for some $c > 0$

$$M_{CSIR}^*(n, \epsilon) \le M_{noCSI}^*(n + c|\mathcal{A}|\log n, \epsilon + \frac{1}{\sqrt{n}}) \tag{57}$$

Indeed, as a first step the encoder for noCSI channel may send $c\log n$ repetitions of each symbol $x \in \mathcal{A}$. The corresponding first $c|\mathcal{A}|\log n$ channel outputs are used by the decoder to compute empirical estimate of the stochastic matrix $\hat{W^{(s)}}$. By Chernoff bound the probability that any row of this estimate deviates by more than $\delta > 0$ from the true $W^{(s)}$ is at most $e^{-O(\log n)}$. Hence by choosing c sufficiently large and δ sufficiently small we may ensure that the empirical estimate $\hat{W^{(s)}}$ is closer to the true $W^{(s)}$ than to any other one with probability at least $1 - \frac{1}{\sqrt{n}}$. The rest of the communication proceeds using the optimal $(n, M, \epsilon)_{CSIR}$ code, whose decoder is fed the estimate of state $\hat{s}$. (The possible mistake in determining state estimate contributes $\frac{1}{\sqrt{n}}$ to the right-hand side of (57).)

Thus, for the purpose of establishing a lower bound in (12) there is no difference between considering CSIR and noCSI scenarios. We proceed to lower-bounding $\log M_{CSIR}^*$ then.

Fix (large) blocklength n and a distribution P on $\mathcal{A}$ in a small neighborhood of P^*. Let P' be the closest n-type approximating P, then $\|P - P'\| \le O(\frac{1}{n})$, where $\|P - P'\|$ is Euclidean distance (induced by the canonical embedding of the simplex into $\mathbb{R}^{|\mathcal{A}|}$). Therefore replacing P with P' in expressions like

$$nI(P, W^{(s)}) - \sqrt{nV(P, W^{(s)})}Q^{-1}(\epsilon)$$

incurs an $O(1)$ difference. We therefore may simplify the reasoning below by pretending that P is an n-type, ignoring the need to replace P with P' in certain places.

We set parameters for Theorem 3 as follows:

- $\mathsf{A} = \mathcal{A}^n$, $\mathsf{B} = \mathcal{B}^n$, $P_{Y_s|X} = (W^{(s)})^n$
- $Q_{Y_s} = (PW^{(s)})^n$
- $\mathsf{F} = T_P^n$ – the collection of all strings $x^n \in \mathcal{A}^n$ of composition P.
- $\tau = \frac{1}{\sqrt{n}}$

Then by permutation symmetry and (33) we have simultaneously for all $x^n \in \mathsf{F}$ and all $s \in \mathcal{S}$:

$$\log \beta_\alpha(P_{Y_s|X=x^n}, Q_{Y_s}) =$$
$$-nI(P, W^{(s)}) - \sqrt{nV(P, W^{(s)})}Q^{-1}(\alpha) + O(\log n), \tag{58}$$

where $O(\log n)$ is uniform in P in a small neighborhood around P^*. Consequently, for the $\beta_{1-\epsilon+\tau}$ in (22) we have

$$\log \beta_{1-\epsilon+\tau} = -nR(n, P) + O(\log n) \tag{59}$$

where

$$R(n, P) = \min_s I(P, W^{(s)}) - \sqrt{\frac{V(P, W^{(s)})}{n}}Q^{-1}(\epsilon)$$

Since

$$P_X[\mathsf{F}] \ge (1 + n)^{1-|\mathcal{A}|}$$

the bound (29) implies

$$\log \kappa_\tau = O(\log n)$$

uniformly in P.

Thus from Theorem 3 we conclude: For every P in a neighborhood of P^* there exists an $(n, M, \epsilon)_{CSIR}$ code with

$$\log M \ge nR(n, P) + O(\log n)$$

with $O(\log n)$ uniform in P. Maximizing $R(n, P)$ over P and applying Lemma 4 (with $\delta = \frac{Q^{-1}(\epsilon)}{\sqrt{n}}$) we conclude

$$\log M_{CSIR}^*(n, \epsilon) \ge nC - \sqrt{nV}Q^{-1}(\epsilon) + o(\sqrt{n})$$

IV. On the $o(\sqrt{n})$ term

For DMCs it is known that when $\epsilon < 1/2$

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

see [4], [5]. For many channels, it has also been established that the $O(\log n)$ term is in fact equal to $\frac{1}{2}\log n + O(1)$, see [3], [5], [9]–[11]. It is natural to ask therefore, whether the estimate on the remainder term in Theorem 1 can be improved to $O(\log n)$. The answer is negative:

Proposition 5: Let W^1 and W^2 be a pair of stochastic matrices defining a compound DMC satisfying all assumptions of Theorem 1 and also:

1) $I(P^*, W^1) = I(P^*, W^2) = C$
2) P^* achieves global maximum of $I(P, W^1)$.
3) There exists $v \in \mathbb{R}^{|\mathcal{A}|}$ such that $t \mapsto I(P + tv, W^1)$ is constant.
4) $\sum_{x\in\mathcal{A}} v_x V(W^1_x \| P^* W^1) < 0$
5) $\sum_{x\in\mathcal{A}} v_x D(W^2_x \| P^* W^2) = 0$ (i.e. $v \perp \nabla_P I(P, W^2)$)
6) $\sum_{x\in\mathcal{A}} v_x W^2_x(y) \neq 0$ for at least one $y \in \mathcal{B}$ (i.e. $v \notin \ker W^2$))
7) $V_1 > V_2$ where $V_s \stackrel{\triangle}{=} V(P^*, W^{(s)})$ for $s = 1, 2$.

Then for any $\epsilon \in (0, \frac{1}{2})$ there exists $K > 0$ s.t.

$$\log M^*(n,\epsilon) \geq nC - \sqrt{nV}Q^{-1}(\epsilon) + Kn^{\frac{1}{4}} + o(n^{\frac{1}{4}}) \quad (60)$$

Proof: It is instructive to understand what the assumptions imply. First, channel 1's dispersion V_1 determines the dispersion of the compound channel (see (13) and assumption 2). However, P^*, although optimal from the W^1-capacity point of view, is not optimal from the W^1-dispersion point of view. Thus by deviating very slightly from P^* we may improve slightly the dispersion of the W^1 channel, while not affecting too significantly mutual information $I(P, W^2)$.

We proceed to formal proof. By assumption 2 gradient of $I(P, W^1)$ is zero at P^* and we get from either (13) or (15) that

$$V = V_1 \,.$$

Next, choose a sequence of distributions

$$P_n = P^* + \frac{c}{n^{\frac{1}{4}}} v$$

with $c > 0$ to be specified shortly. For the first channel mutual information $I(P_n, W^1) = C$ and hence we get:

$$\begin{aligned} &nI(P_n, W^1) - \sqrt{nV(P_n, W^1)}Q^{-1}(\epsilon) \\ &= nC - \sqrt{nV}Q^{-1}(\epsilon) + K_1 c n^{\frac{1}{4}} + o(n^{\frac{1}{4}}) \end{aligned} \quad (61)$$

with $K_1 > 0$ due to assumption 4. For the second channel, due to assumptions 5-6 and (37) for all sufficiently large n we must have

$$I(P_n, W^2) \geq C - \frac{K_2 c^2}{\sqrt{n}}$$

for some $K_2 > 0$. Therefore, we get

$$\begin{aligned} &nI(P_n, W^2) - \sqrt{nV(P_n, W^2)}Q^{-1}(\epsilon) \\ &\geq nC - K_2 c^2 \sqrt{n} - \sqrt{nV_2}Q^{-1}(\epsilon) - K_3 c n^{\frac{1}{4}} + o(n^{\frac{1}{4}}) \,, \end{aligned} \quad (62)$$

for some $K_3 > 0$. Then since $V_2 < V_1$ we can always select c small enough so that the minimum of (61) and (62) exceeds

$$nC - \sqrt{nV}Q^{-1}(\epsilon) + Kn^{\frac{1}{4}} + o(n^{\frac{1}{4}}) \,,$$

for some $K > 0$. The rest of the proof proceeds by applying the $\kappa\beta$ bound exactly as in Section III-D. ∎

Here is an example ensuring assumptions of the Proposition are satisfiable. Let

$$W^1 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{4} & \frac{1}{4} \\ e & \frac{1}{2}-e & g & \frac{1}{2}-g \\ \frac{1}{2}-e & e & \frac{1}{2}-g & g \end{pmatrix}$$

and let P_1 be the first row, P_4 the last row of W^1 and P_Y^* – the uniform distribution on $\{1,2,3,4\}$. Then, select $e, g \in (0, \frac{1}{2})$ so that

$$H(P_1) = \frac{3}{2} \text{ bit} \quad (63)$$

$$V(P_1 \| P_Y^*) < V(P_4 \| P_Y^*) \,, \quad (64)$$

where $H(\cdot)$ is the entropy. Existence of such assignment is easily verified numerically. For the second channel let

$$W^2 = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix}$$

It is easy to see that W^2 is an additive-noise channel (addition over $\mathbb{Z}/4\mathbb{Z}$) with capacity $1/2$ bit. The uniform input distribution also attains the capacity of W^1: indeed all conditional entropies for W^1 are equal (this is due to (63)) and thus maximizing $I(P, W^1)$ is equivalent to maximizing the output entropy $H(PW^1)$. The latter maximum is evidently attained at $H(P_Y^*) = 2$ bit. Therefore the compound capacity is

$$C = \frac{1}{2} \text{ bit}$$

achieved at P^* – uniform. Assumptions 1, 2 are verified then. Assumption 5 holds for every v since P^* is a global maximum of $I(P, W^2)$ and thus the gradient at P^* is zero. Assumption 6 holds because $\ker W^2 = \{0\}$ (e.g. compute the determinant). Assumption 7 holds due to (64) and

$$V_1 = \frac{1}{2}V(P_1 \| P_Y^*) + \frac{1}{2}V(P_4 \| P_Y^*) \quad (65)$$

$$V_2 = V(P_1 \| P_Y^*) \quad (66)$$

For the assumption 3 take

$$v = \begin{pmatrix} 1 & 1 & -1 & -1 \end{pmatrix}$$

and note that $vW^1 = 0$. For the assumption 4 simply recall (64).

Finally, it is not hard to show that the estimate of $n^{\frac{1}{4}}$ in (60) is order-optimal. Indeed, from (36) the mutual information $I(P, W^2)$ satisfies:

$$I(P, W^2) \leq C - K_1 \|P - P^*\|^2$$

in a neighborhood of P. At the same time $V(P, W^{(s)})$ is Lipschitz:

$$V(P, W^{(s)}) \le V_s + K_2 \|P - P^*\|, \qquad s = 1, 2. \tag{67}$$

Thus, by inspecting (56) we can see that in order to not violate the $\sqrt{n}$-term estimate of Theorem 1 an optimizing P must satisfy

$$\|P - P^*\|^2 \le \frac{K_3}{\sqrt{n}}$$

Implying that $\|P - P^*\| \lesssim n^{-\frac{1}{4}}$. Applying (67) and Taylor expansion to (56) we conclude that the $o(\sqrt{n})$ term is upper-bounded by $K_4 n^{\frac{1}{4}} + o(n^{\frac{1}{4}})$ for some $K_4 > 0$.

ACKNOWLEDGMENT

This paper has benefited from discussions with Tsung-Yi Chen (UCLA) of the earlier draft.

REFERENCES

[1] D. Blackwell, L. Breiman, and A. Thomasian, "The capacity of a class of channels," *Ann. Math. Stat.*, pp. 1229–1241, 1959.

[2] J. Wolfowitz, *Coding Theorems of Information Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1962.

[3] R. L. Dobrushin, "Mathematical problems in the Shannon theory of optimal coding of information," in *Proc. 4th Berkeley Symp. Mathematics, Statistics, and Probability*, vol. 1, Berkeley, CA, USA, 1961, pp. 211–252.

[4] V. Strassen, "Asymptotische Abschätzungen in Shannon's Informationstheorie," in *Trans. 3d Prague Conf. Inf. Theory*, Prague, 1962, pp. 689–723.

[5] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[6] Y. Polyanskiy and S. Verdú, "Finite blocklength methods in information theory (tutorial)," in *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013. [Online]. Available: http://people.lids.mit.edu/yp/homepage/data/ISIT13_tutorial.pdf

[7] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Dispersion of the Gilbert-Elliott channel," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 1829–1848, Apr. 2011.

[8] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static SIMO fading channels at finite blocklength," in *Proc. 2013 IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013.

[9] Y. Polyanskiy, "Channel coding: non-asymptotic fundamental limits," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 2010, available: http://www.mit.edu/~ypol.

[10] M. Tomamichel and V. Tan, "A tight upper bound for the third-order asymptotics for most discrete memoryless channels," *IEEE Trans. Inf. Theory*, 2013, to appear.

[11] P. Moulin, "The log-volume of optimal codes for memoryless channels, within a few nats," in *2012 Inf. Theory and Appl. Workshop (ITA)*, San Diego, CA, Feb. 2012.