# Information Rates Over Multi-View Channels

V. Arvind Rameshwar and Nir Weinberger

**Abstract**

We investigate the fundamental limits of reliable communication over multi-view channels, in which the channel output is comprised of a large number of independent noisy views of a transmitted symbol. We consider first the setting of multi-view discrete memoryless channels and then extend our results to general multi-view channels (using multi-letter formulas). We argue that the channel capacity and dispersion of such multi-view channels converge exponentially fast in the number of views to the entropy and varentropy of the input distribution, respectively. We identify the exact rate of convergence as the smallest Chernoff information between two conditional distributions of the output, conditioned on unequal inputs. For the special case of the deletion channel, we compute upper bounds on this Chernoff information. Finally, we present a new channel model we term the Poisson approximation channel — of possible independent interest — whose capacity closely approximates the capacity of the multi-view binary symmetric channel for any fixed number of views.

## I. INTRODUCTION

Consider a communication setting, in which the decoder receives multiple, potentially noisy, views of a single transmitted sequence, which is sent via independent transmissions through a stochastic noise channel. Clearly, multiple noisy views (or multiple draws) of a transmitted sequence can only improve the error probabilities of decoding, as compared to decoding a single received sequence. Viewed differently, it is also clear that the Shannon capacity of the *multi-view* channel, as described above, is at least as large as the capacity of the standard single-view channel.

In this paper, we consider first the setting where the noisy channel is a discrete memoryless channel (DMC) and then extend our results to general multi-letter channels. Multi-view DMCs were first studied in the classic work by Levenshtein [1] on sequence reconstruction. Among several other results, [1] presented bounds on the probability of error under ML decoding and characterized the number of noisy views required to guarantee recovery of the transmitted sequence, with *decaying error probability*. These results are determined by the minimum Chernoff information between conditional distributions of outputs of the DMC, conditioned on unequal

inputs. Levenshtein's result on the number of noisy views required for reconstruction, however, depends crucially on the assumption that the error probability decays with the blocklength, and leaves open the question of characterizing the rates achievable over DMCs with fixed error probability, using finitely many channel uses. Furthermore, the proof strategy in [1] resists extension to channels, such as those that model synchronization errors, wherein the lengths of the input and output sequences could be different.

In this work, we characterize exactly the rate of convergence of the information rate and the channel dispersion of general multi-view channels, under arbitrary input distributions, to the input entropy and the varentropy, respectively, when the number of views is large. As our main result, we show that the rates of (exponential) convergence are again determined by the minimum Chernoff information between conditional distributions, as above. This result does not assume any additional structure on the channel such as symmetry, or binary alphabets (though require the input and output alphabet to be finite)

While our result on the convergence of information rate to the input entropy can be derived from the results in [1] (and [2, Thm. 5])[1], we present new tools for deriving such a result, which also helps us obtain the rate of convergence of the channel dispersion to the input varentropy. An important consequence of our results for DMCs (under some regularity conditions) is our characterization of the largest rates achievable over finitely many channel uses, with any *fixed error probability*, using results from finite-blocklength information theory, thereby resolving a natural question that stems from Levenshtein's work [1]. We then show that our results also hold for multi-letter channels with potentially different lengths of input and output sequences. Furthermore, for the special case of the (multi-letter) deletion channel, which is a popular synchronization error channel [3], we compute explicit upper bounds on the above Chernoff information, and discuss some interesting consequences.

Our interest in this setting is strongly motivated by the recently proposed model for DNA-based storage systems (see, e.g., [4]–[7] and [8], [9] for coding schemes). In a nutshell, in this channel, information is encoded into a large multiset of short sequences from a quarternary alphabet, called molecules (or strands). These sequences are then stored in a pool without order, and reading the information is performed by sampling molecules from the pool. After being inserted into the pool, the molecules are "amplified" by a process called Polymerase Chain Reaction (PCR), to produce several copies of each molecule. Therefore the sampling process during reading can be accurately modelled by sampling *with* replacement. Each sampled molecule is then sequenced to read the written symbols it consist of. This sequencing operation is noisy, and the noise is independent from one molecule read to the other. Furthermore, in some cases, this noisy operation is well modelled by a DMC, which tends to be *asymmetric* [10].

---

[1]We show, moreover, that for binary-input memoryless symmetric (BIMS) channels with a uniform input distribution, it is possible to derive the convergence of the information rate (which, in this case, is the capacity of the channel) to the input entropy, using a simple argument from binary hypothesis testing.

Evidently, in this DNA-storage model, each molecule is possibly read many times, and so the multi-view mutual information between the molecule sequence and its multiple output reads plays a pivotal role in determining the capacity of this storage system. Indeed, it was shown in [7, Thm. 5], that the capacity is composed of a weighted average of such mutual information terms, which accounts for the multiple input molecules, and the possibility that each of them is sampled a different number of times.

The total number of sampled molecules in the reading operation is a design parameter, say the number of moelcueles sampled may be a constant factor of the total number of molecules, say $d_0$. In this case, it is expected that each input molecule is also sampled about $d_0$ times, and the mutual information term between the input molecule sequences and its multiple output reads falls into the setting considered in this paper. Moreover, as said, each molecule on its own is rather short, and so finite-blocklength bounds appears to be relevant. Specifically, given a molecule length, and $d_0$, one may determine using our results when the maximal multi-view rate of this molecule is approached, for a given error probability. A detailed investigation of this setting is beyond the scope of this paper, but of interest for future work.

Previous works considered the mutual information for a fixed number of views. In this non-asymptotic setting, the capacity of a multi-view binary symmetric channel (BSC) was identified and an explicit expression was provided in [11]. More generally, bounds on the capacities of multi-view DMCs were derived in the literature on "information combining" (see [12], [13] and the references therein). Also related is the well-studied problem on *trace reconstruction* [14]–[16], wherein an input string is passed independently through a deletion channel and the goal is to find the number of resultant outputs required for reconstruction with high-probability. By contrast, in this work we focus on the scaling to mutual information and dispersion in the asymptotic regime of large number of views.

As an additional contribution, we also explore the non-asymptotic capacity of the so-called *Poisson approximation channel*, which we believe is of possible independent interest. We show that the capacity of this channel is a tight lower bound on the capacity of the multi-view BSC. Such a technique also allows us to obtain simple bounds on the non-asymptotic capacities of general binary-input memoryless symmetric (BIMS) channels.

The rest of the paper is organized as follows: Section II sets down the notation; Section III presents our main results for DMCs; Section IV provides a proof sketch of the rate of convergence of the information rate and channel dispersion of a DMC to their respective limits. Section V then extends our results to general multi-letter channels and presents upper bounds on Chernoff information discussed above for the deletion channel. Section VI contains a proof that the capacity of the Poisson approximation channel is a tight lower bound on the capacity of the multi-view BSC. Section VII then provides full proofs of our main convergence results. Section VIII concludes the paper and proposes problems for further research.

## II. NOTATION CONVENTIONS

All logarithms are assumed to be natural logarithms. Random variables are denoted by capital letters, e.g., $X, Y$, and small letters, e.g., $x, y$, denote their instantiations. Sets are denoted by calligraphic letters, e.g., $\mathcal{X}, \mathcal{Y}$. The set of positive natural numbers $\{1, 2, \ldots\}$ is denoted as $\mathbb{N}$. Notation such as $P(x), P(y|x)$ are used to denote the probabilities $P_X(x), P_{Y|X}(y|x)$, when it is clear which random variables are being referred to. For $\alpha \in [0, 1]$ and $n \geq 1$, $\mathrm{Ber}(\alpha)$ and $\mathrm{Bin}(n, \alpha)$, denote, respectively, the Bernoulli distribution, supported on $\{0, 1\}$, with $\Pr_{X \sim \mathrm{Ber}(\alpha)}[X = 1] = \alpha$, and the Binomial distribution, supported on $\{0, 1, \ldots, n\}$, with

$$\Pr_{X \sim \mathrm{Bin}(n, \alpha)}[X = k] = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k}, \tag{1}$$

for $k \in \{0, 1, \ldots, n\}$.

The entropy of $X \sim P_X$ is denoted by $\mathsf{H}(X) = \mathsf{H}(P_X) := \mathbb{E}[-\log P_X(X)]$ and its varentropy is denoted by

$$\mathsf{V}(X) := \mathbb{E}\left[(-\log P_X(X) - \mathsf{H}(X))^2\right]. \tag{2}$$

Both only depend only on the distribution of $X$. For $\gamma \in [0, 1]$, the binary entropy function is denoted by $h_b(\gamma) := -\gamma \log \gamma - (1 - \gamma) \log(1 - \gamma)$. The Chernoff information $\mathsf{C}(P, Q)$ between two distributions $P$ and $Q$ defined on the same alphabet $\mathcal{X}$ is given by (see [17, Ch. 11])

$$\mathsf{C}(P, Q) := -\min_{\lambda \in [0,1]} \log\left(\sum_{x \in \mathcal{X}} P(x)^{1-\lambda} Q(x)^{\lambda}\right). \tag{3}$$

Furthermore, the standard Kullback-Leibler (KL) divergence between two distributions $P, Q$ on the same alphabet $\mathcal{X}$ is given by

$$D(P||Q) := \mathbb{E}_{X \sim P}\left[\log \frac{P(X)}{Q(X)}\right], \tag{4}$$

where we take $0 \log \frac{0}{z} = 0$ and we set $D(P||Q) = \infty$ if there exists $x \in \mathcal{X}$ such that $Q(x) = 0$ but $P(x) > 0$.

The the length-$n$ vector $(a_1, \ldots, a_n)$ is denoted by $a^n$, and the Hamming weight, i.e., the number of ones in $a^n$, is denoted by $w(a^n)$. The indicator that $x$ equals $b$ is denoted by $\mathbb{1}\{x = b\}$; this indicator equals 1 if $x = b$ and 0, otherwise. For sequences $(a_n)_{n \geq 1}$ and $(b_n)_{n \geq 1}$ of positive reals, $a_n = o(b_n)$ if $\lim_{n \to \infty} \frac{a_n}{b_n} = 0$, $a_n = O(b_n)$ if for all $n$ large enough, $a_n \leq c \cdot b_n$, for some $c > 0$, and $a_n = \Theta(b_n)$ if for all $n$ large enough, we have that $m \cdot b_n \leq a_n \leq M \cdot b_n$, for some constants $m, M \in \mathbb{R}$.

## III. MAIN RESULTS FOR DMCS

We now formalize the problem for the case the noisy channel is a DMC, and present our main results. Later on, in Section V we will extend Theorem III.1 here to general multi-letter channels. Consider a DMC $W$ with finite input and output alphabets $\mathcal{X}$ and $\mathcal{Y}$, respectively, specified by the channel law $\{P_{Y|X}(y|x) = W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$. Let $X \sim P_X$ be the input to the channel. Our setting of interest is when $d \geq 1$ noisy outputs

are obtained by independently passing the input through the DMC. Let these outputs be denoted by $Y_1, \ldots, Y_d$; their conditional distribution obeys

$$P(Y^d = y^d \mid X = x) = \prod_{i=1}^{d} W(y_i|x). \tag{5}$$

We call a DMC with a channel law as above as the *d-view* DMC $W^{(d)}$. We assume further that $|\mathcal{X}|$ and $|\mathcal{Y}|$ do not depend on $d$.

Now, given any DMC $W$, we denote its capacity by [18]

$$C(W) = \max_{P_X} I(X;Y) = \max_{P_X} \mathbb{E}\left[\log \frac{W(Y|X)}{\sum_{x \in \mathcal{X}} P(x)W(Y|x)}\right]. \tag{6}$$

Hence, the capacity of $W^{(d)}$ is given by $C^{(d)} := \max_{P_X} I^{(d)}$, where $I^{(d)} := I(X;Y^d)$. Further, for any input distribution $P_X$, the channel dispersion of the *d-view* DMC $W$ is given by

$$V^{(d)} := \mathbb{E}\left[\left(\iota(X;Y^d) - I(X;Y^d)\right)^2\right], \tag{7}$$

where

$$\iota(x;y^d) := \log \frac{P(y^d|x)}{P(y^d)} \tag{8}$$

is the *information density*. We are interested in characterizing the behaviour of $C^{(d)}$ (or more generally, of $I^{(d)}$) and $V^{(d)}$ when $d$ is large. From operational considerations, we can argue that $I^{(d)}$ and $V^{(d)}$ are non-decreasing in $d$. More formally, for $I^{(d)}$, the data processing inequality [17, Thm. 2.8.1] guarantees that $I^{(d)}$ is non-decreasing in $d$. Furthermore, it is reasonable to expect that $I^{(d)}$ and $V^{(d)}$ converge to $\mathsf{H}(X)$ and $\mathsf{V}(X)$, respectively, since for very large $d$, the "uncertainty" introduced by the channel is expected to drop to 0. In this work, we confirm these convergence results by showing that the $I^{(d)}$ and $V^{(d)}$ in fact converge to their limiting values exponentially quickly in $d$, and explicitly identify the rate of convergence.

To build some intuition why the convergence is exponentially fast in $d$, at least for $I^{(d)}$, consider the simple example of the *d-view* $\mathrm{BSC}(p)$, which we denote by $\mathrm{BSC}^{(d)}(p)$ (here, $p < 1/2$ is the crossover probability of the BSC), with a uniform input distribution. At the decoder end, we construct a simple "data processor" that takes the (random) outputs $Y^d \in \{0,1\}^d$ and computes their majority $M \in \{0,1\}$. Intuitively, the more the number of noisy output sequences (or views) that are available to the decoder, the better the estimate that results from majority decoding. More precisely, we have

$$I^{(d)} = I(X;Y^d) \tag{9}$$

$$\geq I(X;M) = \log 2 - h_b(\Pr[X \neq M]), \tag{10}$$

where the inequality holds by the data processing inequality and the last equality holds by symmetry. Now, observe that since $p < 1/2$, we have by the Chernoff bound (see [19, Example 1.6.2]) that

$$\Pr[X \neq M] \leq \exp(-d \cdot D(\mathrm{Ber}(1/2)||\mathrm{Ber}(p))) \tag{11}$$

$$= \exp(-d \cdot Z(p)), \tag{12}$$

where $Z(p) = 2\sqrt{p(1-p)}$ is the Bhattacharya parameter for the BSC($p$). Therefore, for large enough $d$, we have that $\Pr[X \neq M]$ is sufficiently small so that

$$h_b(\Pr[X \neq M]) \leq -2\Pr[X \neq M] \cdot \log \Pr[X \neq M]. \tag{13}$$

Hence, for sufficiently large $d$, we see that $I^{(d)} \geq \log 2 - dZ(p) \cdot \exp(-d \cdot Z(p))$, thereby providing reason for the intuition that the convergence of $I^{(d)}$ to $\log 2$ is exponentially fast in $d$. In fact, in Theorem III.1, we argue that the rate $Z(p)$ is in fact tight for the BSC($p$).

Our main theorem is given below.

**Theorem III.1.** *We have that*

$$I^{(d)} = \mathsf{H}(X) - \exp\left(-d\rho + \Theta(\log d|\mathcal{X}|)\right), \tag{14}$$

*and*

$$\left|V^{(d)} - \mathsf{V}(X)\right| = \exp\left(-d\rho + \Theta(\log d|\mathcal{X}|)\right), \tag{15}$$

*where*

$$\rho = \min_{x,x':x\neq x'} \mathsf{C}(P_{Y|x}, P_{Y|x'}). \tag{16}$$

Note that the theorem above captures the dependence of the speed of convergence of mutual information and dispersion to their limits in terms of both the number of views $d$ and the size of the input alphabet (assumed to be independent of $d$). The dependence on $|\mathcal{X}|$ will assume significance when we transition to multi-letter channels in Section V.

We now proceed to interpret Theorem III.1. Suppose we define the rates

$$\rho(I) := \liminf_d -\frac{1}{d} \cdot \log\left(\mathsf{H}(X) - I^{(d)}\right) \tag{17}$$

$$= \liminf_d -\frac{1}{d} \cdot \log \mathsf{H}(X|Y^d), \tag{18}$$

and

$$\rho(V) := \liminf_d -\frac{1}{d} \cdot \log\left|\mathsf{V}(X) - V^{(d)}\right|, \tag{19}$$

we get by the above result that these rates of convergence obey $\rho(I) = \rho(V) = \rho$, which in turn is independent of the input distribution $P_X$, with possible dependence on its support. Further, the limits in the definitions of these rates exist. Moreover, note that for a binary-input memoryless symmetric (BIMS) channel $W$ (see [20, Ch. 4] for a definition), the function

$$f(\lambda) := \log\left(\sum_{y\in\mathcal{Y}} P(y|0)^{1-\lambda}P(y|1)^{\lambda}\right) \tag{20}$$

is symmetric about $\lambda = 1/2$. Hence, using the fact that $f$ is convex, we obtain that the minimum in (3) is attained at $\lambda = 1/2$, giving

$$\mathsf{C}(P_{Y|0}, P_{Y|1}) = -\log \sum_{z \in \mathcal{Z}} \sqrt{P_{Y|0}(z|0) P_{Y|1}(z|1)} \tag{21}$$

$$= -\log Z_b(W), \tag{22}$$

where $Z_b(W)$ is the Bhattacharya parameter of the BIMS channel $W$. Hence, following the results in [21], the rates of convergence for a BIMS channel $W$ are extremized by a BEC and BSC having the same capacity as $W$, respectively, with the BEC having the largest and BSC having the smallest Bhattacharya parameters, respectively. Here, for the $d$-view BSC($p$), we have $\rho(I) = \rho(V) = -\log 2\sqrt{p(1-p)}$, and for the $d$-view BEC($\epsilon$), we have $\rho(I) = \rho(V) = -\log \epsilon$. Furthermore, for the Z-channel with parameter $\delta$, $\rho(I) = \rho(V) = -\log \delta$. In fact, for most well-behaved DMCs $W$, we will have that $\rho(I) = \rho(V) > 0$, implying exponential convergence of mutual information and dispersion to $\mathsf{H}(X)$ and $\mathsf{V}(X)$, respectively.

Theorem III.1 along with the normal approximation for finite blocklengths [22] then allow us to characterize, up to exponential tightness, the largest finite-blocklength rates achievable over a $d$-view DMC $W^{(d)}$, under some regularity conditions. Our presentation follows that in [23] (see also [24, Thm. 22.2]). Suppose that the DMC $W^{(d)}$ is non-singular (see the definition in [23, Sec. 4.2.1]). Let $M^\star(n, \epsilon)$ be the largest integer $M$ such that there exists a blocklength-$n$ code with maximal probability of error $\epsilon > 0$ over $W^{(d)}$. Further, suppose that under the uniform input distribution, we have that $V^{(d)} > 0$. For $t \in \mathbb{R}$, let $\Phi(t)$ be the Gaussian cumulative distribution function (c.d.f.), i.e.,

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \mathrm{d}z, \tag{23}$$

with $\Phi^{-1}$ denoting its inverse. We then have the following theorem:

**Theorem III.2.** *For any non-singular DMC $W^{(d)}$, for all $n \geq 1$ and $\epsilon \in (0, 1)$, we have*

$$\frac{\log M^\star(n, \epsilon)}{n} \geq \log |\mathcal{X}| - e^{-d\rho + \Theta(\log d|\mathcal{X}|)} + \Phi^{-1}(\epsilon) \cdot \frac{e^{-d\rho/2 + \Theta(\log d|\mathcal{X}|)}}{\sqrt{n}} + \frac{\log n}{2n} + O(1), \tag{24}$$

*where $\rho = \rho(I) = \rho(V)$.*

*Proof.* The proof follows from [23, Thm. 4.1] by choosing the input distribution $P_X$ to be the uniform distribution. $\square$

Clearly, it follows from the above theorem that for any finite $n$, by choosing $d = \rho^{-1} \log n$, we can obtain rates

$$\frac{\log M^\star(n, \epsilon)}{n} \geq \log |\mathcal{X}| - O\left(\frac{\log n}{n}\right) \tag{25}$$

over any $d$-view DMC $W^{(d)}$.

Next, we consider the special case of the $d$-view BSC$(p)$, which we denote by BSC$^{(d)}(p)$; the capacity of this channel is known to be equal to the capacity of the so-called binomial channel [11, Sec. III.A], denoted by $\mathsf{Bin}_d(p)$. This capacity $C(\mathsf{Bin}_d(p))$ obeys

$$C(\mathsf{Bin}_d(p)) = \log 2 + \sum_{i=0}^{d} \binom{d}{i} \left( p^i (1-p)^{d-i} \log \frac{p^i (1-p)^{d-i}}{p^i (1-p)^{d-i} + p^{d-i}(1-p)^i} \right), \tag{26}$$

and the capacity is achieved by the input distribution $P_X = \text{Ber}(1/2)$.

While Theorem III.1 determines the rate of convergence of $C\left(\text{BSC}^{(d)}(p)\right) = C(\mathsf{Bin}_d(p))$ (via the rate of convergence of the mutual information $I(X; Y^d)$ under a uniform input distribution $P_X = \text{Ber}(1/2)$) to $\mathsf{H}(P_X) = 1$ in the asymptotic regime of large $d$, it will be useful if a more fine-grained, non-asymptotic understanding of the capacity of this basic multi-view channel is obtainable. To this end, we identify a closely-related DMC, which we call the *Poisson approximation channel* and denote by $\mathsf{Poi}_d(p)$, whose capacity turns out to be a good approximation to $C(\mathsf{Bin}_d(p))$. The DMC $\mathsf{Poi}_d(p)$ has input alphabet $\mathcal{X} = \{0, 1\}$ and output alphabet $\mathcal{R} = (\mathbb{N} \cup \{0\}) \times (\mathbb{N} \cup \{0\})$. We denote the input bit as $X$ and the output as the pair $(R_1, R_2)$. The channel law obeys

$$P_{R_1, R_2 | X}(r_1, r_2 | x) = P_{R_1 | X}(r_1 | x) P_{R_2 | x}(r_2 | x), \text{for all } x \in \{0, 1\} \text{ and } (r_1, r_2) \in (\mathbb{N} \cup \{0\}) \times (\mathbb{N} \cup \{0\}), \tag{27}$$

where

$$P_{R_1 | 0} = \text{Poi}(d(1-p)), \ \ P_{R_2 | 0} = \text{Poi}(dp), \ \text{and} \ P_{R_1 | 1} = \text{Poi}(dp), \ \ P_{R_2 | 1} = \text{Poi}(d(1-p)). \tag{28}$$

Here, the random variables $R_1$ and $R_2$ are approximations for the number of 0s and 1s, respectively, that are received as part of the $d$ outputs of the binomial channel. Further, $\text{Poi}(\gamma)$ denotes the Poisson distribution with parameter $\gamma$. Note that $\mathsf{Poi}_d(p)$ is a BIMS channel; hence, its capacity is achieved by $P_X = \text{Ber}(1/2)$ (see [20, Problem 4.8]). The following theorem then holds.

**Theorem III.3.** *We have that*

$$C(\mathsf{Poi}_d(p)) \le C(\mathsf{Bin}_d(p)) \le C(\mathsf{Poi}_d(p)) + \exp(-d(1 - Z(p))) - Z(p)^{2d}, \tag{29}$$

*where $Z(p) = 2\sqrt{p(1-p)}$ is the Bhattacharya parameter of the BSC$(p)$.*

Figure 1 plots the capacities of the $\mathsf{Bin}(p)$ and $\mathsf{Poi}(p)$ channels for varying values of $d$. It is clear from the plots and from Theorem III.3 that $C(\mathsf{Poi}_d(p))$ is a tight lower bound on $C(\mathsf{Bin}_d(p))$; furthermore, the tightness of the lower bound improves as $d$ increases.

The result of Theorem III.3 can be easily generalized to $d$-view BIMS channels $W^{(d)}$ with a finite output alphabet. It is well-known (see, e.g., Section II in [25]) that any BIMS channel $W$ with a finite output alphabet can be decomposed into finitely many BSC subchannels, i.e.,

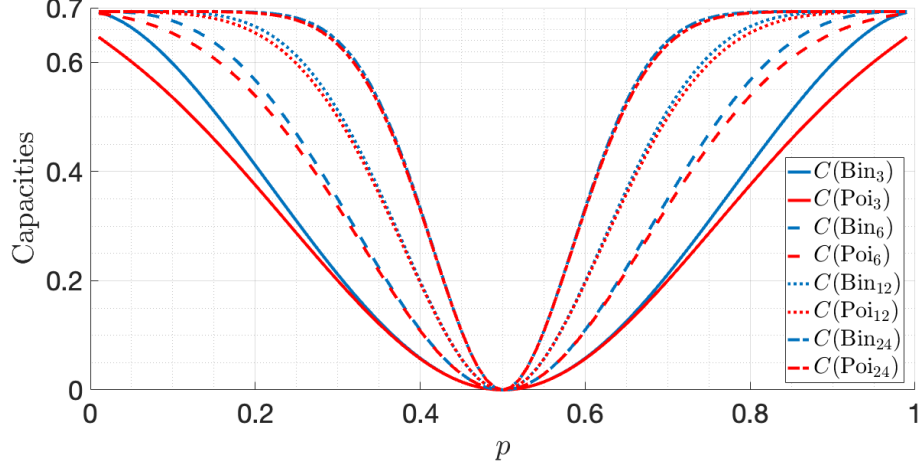$$W_{Y|X=x} = \sum_{i=1}^{K} \epsilon_i \cdot W_{Y|X=x}^{(i)}, \tag{30}$$

Fig. 1: Plots comparing $C(\mathsf{Bin}(p))$ and $C(\mathsf{Poi}(p))$ for $p \in [0, 1]$, and $d \in \{3, 6, 12, 24\}$.

for some $K < \infty$, for any $x \in \{0, 1\}$, with $W_{Y|X=x}^{(i)}$ being the channel law of a BSC with crossover probability $p_i$. Here, $\epsilon_i$ is the probability of choosing subchannel $i$, with $\sum_i \epsilon_i = 1$. Let $P_s$ be the distribution on $\{1, \ldots, K\}$ with mass $\epsilon_i$ at point $i$. We then have that

$$C^{(d)}(W) = \mathbb{E}\left[C(J_1, \ldots, J_d)\right], \tag{31}$$

where the random variables $J_i$, $1 \le i \le K$, are drawn i.i.d. according to $P_s$, and the notation $C(j_1, \ldots, j_d)$ refers to the capacity of that channel with the channel law

$$P(Y^d = y^d | X = x) = \prod_{\ell=1}^{d} W^{(j_\ell)}(y_\ell | x), \tag{32}$$

for $x \in \{0, 1\}$ and $y^d \in \mathcal{Y}^d$. Now, suppose that $0 < p_1 \le p_2 \le \ldots \le p_K < 1/2$. We then have from Theorem III.3 and (31) that

$$C(\mathsf{Poi}_d(p_K)) \le C^{(d)}(W) \le C(\mathsf{Poi}_d(p_1)) + \exp(-d(1 - Z(p_1))) - Z(p_1)^{2d}. \tag{33}$$

## IV. Proof Sketch of Main Results

In this section, we provide a sketch of our strategy for proving Theorem III.1 for a DMC $W$ with an aribitrary distribution $P_X$. To this end, we shall first consider the special case when $W$ is a BIMS channel, with a uniform input distribution, and provide a simple proof for the convergence of the mutual information $I^{(d)}$ to the input entropy. We then provide a brief sketch of the proof of Theorem III.1 for general DMCs, which makes use of well-known results from large-deviations theory. The full proof, owing to its length, is presented in Section VII.

*A. Convergence of Mutual Information for BIMS Channels*

To build intuition, we present a simple proof for the case when the DMC $W$ is a BIMS channel, with $P_X = \text{Ber}(1/2)$. The choice of the input distribution being uniform assumes relevance since the capacity of a BIMS channel is achieved by uniform inputs (see [20, Problem 4.8]).

The following bounds on the conditional entropy of the input to a BIMS channel given the output, which are drawn from [26, Prop. 2.8], will prove useful. Recall the definition of the Bhattacharya parameter of a BIMS channel $W$ with output alphabet $\mathcal{Y}$ given by $Z_b(W) := \sum_{y \in \mathcal{Y}} \sqrt{P_{Y|0}(y)P_{Y|1}(y)}$.

**Lemma IV.1.** *Given a BIMS channel $\overline{W}$ with input $\overline{X} \sim P_X$, where $P_X = \text{Ber}(1/2)$, and output $\overline{Y}$, we have that*

$$Z_b(\overline{W})^2 \leq \mathsf{H}(\overline{X} \mid \overline{Y}) \leq \log(1 + Z_b(\overline{W})). \tag{34}$$

We thus obtain the following simple corollary.

**Corollary IV.1.** *Given a BIMS channel $W$ with input $X \sim P_X$, where $P_X = \text{Ber}(1/2)$, we have that for the $d$-view DMC $W^{(d)}$,*

$$Z_b(W)^{2d} \leq \mathsf{H}(X \mid Y^d) \leq Z_b(W)^d. \tag{35}$$

*Proof.* We first observe that when $W$ is a BIMS channel, then so is the $d$-view DMC $W^{(d)}$. To see why this is true, note that corresponding to a BIMS channel $W$, there exists a permutation $\pi : \mathcal{Y} \to \mathcal{Y}$ such that

$$W(y \mid 0) = W(\pi(y) \mid 1), \tag{36}$$

with $\pi$ being an involution, i.e., $\pi^{-1} = \pi$. Now, for the $d$-view DMC $W^{(d)}$, we define the permutation $\pi^{(d)} : \mathcal{Y}^d \to \mathcal{Y}^d$ such that

$$\pi^{(d)}(y_1, y_2, \ldots, y_d) = (\pi(y_1), \pi(y_2), \ldots, \pi(y_d)), \tag{37}$$

for $y_1, \ldots, y_d \in \mathcal{Y}$. Clearly, we have that $\pi^{(d)}$ is an involution, with

$$W^{(d)}((y_1, \ldots, y_d) \mid 0) = W(\pi^{(d)}(y_1, \ldots, y_d) \mid 1), \tag{38}$$

from (5). This hence implies that $W^{(d)}$ is also a BIMS channels.

Now, the Bhattacharya parameter

$$Z_b(W^{(d)}) = \sum_{y^d \in \mathcal{Y}^d} \sqrt{P_{Y|0}(y^d|0)P_{Y|1}(y^d|1)} \tag{39}$$

$$= \sum_{y^d \in \mathcal{Y}^d} \prod_{i=1}^{d} \sqrt{P_{Y|0}(y_i|0)P_{Y|1}(y_i|1)} \tag{40}$$

$$= \prod_{i=1}^{d} \sum_{y_i \in \mathcal{Y}} \sqrt{P_{Y|0}(y_i|0)P_{Y|1}(y_i|1)} \tag{41}$$

$$= Z_b(W)^d. \tag{42}$$

The proof is then completed by appealing to Lemma IV.1 and noting that $\log(1 + z) \leq z$, for all $z > -1$. $\qquad\square$

Moreover, we have the following bound on $\rho(I)$ for general binary-input DMCs and for arbitrary input distributions $P_X$.

**Lemma IV.2.** *Given a binary-input DMC W with input $X \sim P_X$, where $P_X$ is arbitrary, we have that*

$$\mathsf{H}(X|Y^d) \geq e^{-d \cdot \mathsf{C}(P_{Y|X=0}, P_{Y|X=1}) + \Theta(\log d)}. \tag{43}$$

*Proof.* Observe that

$$\mathsf{H}(X|Y^d) = \mathbb{E}\left[\log \frac{1}{P_{X|Y^d}(X \mid Y^d)}\right]. \tag{44}$$

Further, recall that in a binary Bayesian hypothesis testing setting (see [17, Ch. 11]) with prior $P_X$ and hypotheses $\{P_{Y|0}, P_{Y|1}\}$, we have that the probability of correct decision is given by

$$\mathbb{E}\left[P(X \mid Y^d)\right] = 1 - e^{-d \cdot \mathsf{C}(P_{Y|X=0}, P_{Y|X=1}) + \Theta(\log d)}. \tag{45}$$

Hence, we see that for large enough $d$,

$$\mathsf{H}(X|Y^d) = \mathbb{E}\left[\log \frac{1}{P(X \mid Y^d)}\right] \tag{46}$$

$$\geq -\log\left(1 - e^{-d \cdot \mathsf{C}(P_{Y|X=0}, P_{Y|X=1}) + \Theta(\log d)}\right) \tag{47}$$

$$\geq e^{-d \cdot \mathsf{C}(P_{Y|X=0}, P_{Y|X=1}) + \Theta(\log d)}, \tag{48}$$

where the first inequality follows from Jensen's inequality and the second from the fact that $\log(1 + z) \leq z$, for $z > -1$. Also, the $\Theta(\log d)$ term in the exponent in (47) is due either to the Laplace method (see [27, Sec. 4.2]) or to the results of [28, Thm. 2.1 and Corollary 1].

$\qquad\square$

The above ingredients suffice to afford a proof of Theorem III.1 for the special case of BIMS channels with a uniform input distribution.

*Proof of Thm. III.1 for BIMS channels with uniform inputs.* For the case when $W$ is a BIMS channel, we have that $\mathsf{C}(P_{Y|0}, P_{Y|1}) = -\log Z_b(W)$ (see the remark following the statement of Theorem III.1). Hence, when the input distribution is uniform, we have from Corollary IV.1, we have that $\mathsf{H}(X|Y^d) \leq e^{-d \cdot \mathsf{C}(P_{Y|0}, P_{Y|1})}$. The proof concludes by appealing then to Lemma IV.2, which shows that

$$\mathsf{H}(X|Y^d) \geq e^{-d \cdot \mathsf{C}(P_{Y|0}, P_{Y|1}) + \Theta(\log d)}. \tag{49}$$

Hence, overall, we have that

$$\mathsf{H}(X|Y^d) = e^{-d \cdot \mathsf{C}(P_{Y|0}, P_{Y|1}) + \Theta(\log d)}. \tag{50}$$

$\qquad\square$

It is possible, with some effort, to prove Theorem III.1 for $I^{(d)}$ for general DMCs with arbitrary input distributions with the aid of [1, Lemma 9] (see also [2, Thm. 5]) and an extension of Lemma IV.2. However, the analysis of dispersion requires more involved techniques. Owing to the length of the proofs, we present a detailed proof for $I^{(d)}$ that introduces our techniques, in Section VII-A, and then extend them to $V^{(d)}$, in Section VII-B.

We now sketch a broad outline of the proof strategy for the convergence of mutual information to the input entropy. We first express the conditional entropy $\mathsf{H}(X|Y^d)$ as

$$\mathsf{H}(X|Y^d) = \mathbb{E}\left[\log\frac{1}{P_{X|Y^d}(X \mid Y^d)}\right] \tag{51}$$

$$= \sum_{x \in \mathcal{X}} P_X(x)\mathbb{E}\left[\log\frac{1}{P_{X|Y^d}(x \mid Y^d)}\bigg| X = x\right]. \tag{52}$$

We then fix an $x \in \mathcal{X}$ and focus on the inner term in the summation above. Then,

$$\mathbb{E}\left[\log\frac{1}{P_{X|Y^d}(x \mid Y^d)}\bigg| X = x\right] = \int_0^\infty \mathbb{P}\left[\log\frac{1}{P_{X|Y^d}(x \mid Y^d)} \geq t\bigg| X = x\right]\mathrm{d}t. \tag{53}$$

Our approach is to obtain (exponentially tight) bounds on the probability $\mathbb{P}\left[-\log P_{X|Y^d}(x \mid Y^d) \geq t| X = x\right]$, using techniques from large-deviations theory—Sanov's theorem, in particular. We show that

$$\mathbb{E}\left[\log\frac{1}{P_{X|Y^d}(x \mid Y^d)}\bigg| X = x\right] = \exp\left(-d \cdot \min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{54}$$

This then allows us to use Laplace's method (e.g., see [27, Sec. 4.2]) and the independence of the input distribution $P_X$ on $d$ to conclude that

$$\mathsf{H}(X|Y^d) = \max_{x,\tilde{x}:\tilde{x} \neq x} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right), \tag{55}$$

and hence that

$$I^{(d)} = \mathsf{H}(X) - \exp\left(-d\rho + \Theta(\log d|\mathcal{X}|)\right). \tag{56}$$

Interestingly, our proof strategy allows us to obtain insights into the "higher-order" terms in the exponent of $\mathsf{H}(X|Y^d)$ and their dependence on both $d$ and the size of the input alphabet $|\mathcal{X}|$. The proof for the convergence of the channel dispersion to the input varentropy is similar, and relies on the results obtained during the derivation of the convergence of the mutual information to the input entropy.

In the next section, we extend Theorem III.1 to general *multi-letter* channels, which includes as a special case, synchronization channels such as the deletion channel. We then derive explicit, analytical bounds on the rates of convergence of the mutual information and dispersion of the deletion channel to their limits.

## V. CONVERGENCE FOR MULTI-LETTER CHANNELS AND SOME EXAMPLES

Recall that Theorem III.1 presented the convergence of mutual information and channel dispersion for $d$-view DMCs $W^{(d)}$, where the channel $W$ is specified by *single-letter* transition probabilities $\{W(y|x) : x \in \mathcal{X}, y \in \mathcal{Y}\}$. In this section, we first extend the statement of this theorem and its proof to multi-letter channels. Consider a multi-letter channel $W_{n,\nu}$, for any $n, \nu \geq 0$ that are independent of $d$, specified by a channel law $\{W(y^\nu|x^n) = P_{Y^\nu|X^n}(y^\nu|x^n) : x^n \in \mathcal{X}^n, y^\nu \in \mathcal{Y}^\nu\}$. Here, the input alphabet is $\mathcal{X}^n$ and the output alphabet is $\mathcal{Y}^\nu$. We work with the $d$-view multi-letter channel $W_{n,\nu}^{(d)}$ defined by the channel law

$$P(Y^{d\nu} = y^{d\nu} \mid X^n = x^n) = \prod_{i=1}^{d} W_{n,\nu}(y^\nu|x^n). \tag{57}$$

Recall here our assumption that $|\mathcal{X}|$ and $|\mathcal{Y}|$ do not depend on $d$. Similar to the earlier definition, we let

$$I_{n,\nu}^{(d)} := I(X^n; Y^{d\nu}) \tag{58}$$

to be the mutual information rate over $W_{n,\nu}^{(d)}$. Here, we assume as before that the input distribution $P_{X^n}$ is arbitrary, but fixed; in particular, it is $d$-independent. Likewise, let the channel dispersion be

$$V_{n,\nu}^{(d)} := \mathbb{E}\left[\left(\iota(X^n; Y^{d\nu}) - I(X^n; Y^{d\nu})\right)^2\right], \tag{59}$$

where $\iota(x; y^d) := \log \frac{P(y^{d\nu}|x^n)}{P(y^{d\nu})}$. The following corollary of Theorem III.1 holds:

**Corollary V.1.** *We have that*

$$I_{n,\nu}^{(d)} = \mathsf{H}(X^n) - \exp\left(-d\rho_{n,\nu} + \Theta(\log d) + \Theta(n \cdot \log |\mathcal{X}|)\right), \tag{60}$$

*and*

$$\left|V_{n,\nu}^{(d)} - \mathsf{V}(X^n)\right| = \exp\left(-d\rho_{n,\nu} + \Theta(\log d) + \Theta(n \cdot \log |\mathcal{X}|)\right), \tag{61}$$

*where*

$$\rho_{n,\nu} := \min_{x^n, \tilde{x}^n : x^n \neq \tilde{x}^n} \mathsf{C}(P_{Y^\nu|x^n}, P_{Y^\nu|\tilde{x}^n}). \tag{62}$$

Note that the $\Theta(n \cdot \log |\mathcal{X}|)$ terms above arise due to the fact that the input alphabet size is now $|\mathcal{X}|^n$, instead of $|\mathcal{X}|$, as in Theorem III.1. Observe also, from the earlier proofs, that only the $\Theta(\log |\mathcal{X}|)$ term gets scaled by a factor of $n$ and not the $\Theta(\log d)$ term.

Now, consider the special case when $\nu = n$ and the channel $W_{n,n}$ corresponds to an $n$-letter DMC; more precisely, suppose that the multi-letter channel law obeys

$$W_{n,n}(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i). \tag{63}$$

Now, for the *single-letter* channel $W$ as above, let $\rho = \min_{x, \tilde{x} : x \neq \tilde{x}} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})$. We then have the following corollary:

**Corollary V.2.** *For an n-letter DMC $W_{n,n}$,*

$$I_{n,n}^{(d)} = \mathsf{H}(X^n) - \exp\left(-dn\rho + \Theta(\log d) + n \cdot \Theta(\log |\mathcal{X}|)\right), \tag{64}$$

*and*

$$\left| V_{n,n}^{(d)} - \mathsf{V}(X^n) \right| = \exp\left(-dn\rho + \Theta(\log d) + n \cdot \Theta(\log |\mathcal{X}|)\right). \tag{65}$$

*Proof.* The proof will follow from Corollary V.1 if we can show that $\rho_{n,n} = n\rho$, for an $n$-letter DMC $W_{n,n}$. To this end, let $u, v \in \mathcal{X}$ be such that $\rho = \mathsf{C}(P_{Y|u}, P_{Y|v})$. Further, observe that for any two inputs $x^n \neq \tilde{x}^n$, we have $\mathsf{C}(P_{Y^n|x^n}, P_{Y^n|\tilde{x}^n}) = \sum_{i=1}^{n} \mathsf{C}(P_{Y|x_i}, P_{Y|\tilde{x}_i})$. Hence,

$$\rho_{n,n} = \min_{x^n, \tilde{x}^n : x^n \neq \tilde{x}^n} \sum_{i=1}^{n} \mathsf{C}(P_{Y|x_i}, P_{Y|\tilde{x}_i}) \tag{66}$$

$$\geq \sum_{i=1}^{n} \min_{x_i, \tilde{x}_i : x_i \neq \tilde{x}_i} \mathsf{C}(P_{Y|x_i}, P_{Y|\tilde{x}_i}) = n\rho, \tag{67}$$

where equality in the second step is obtained when $x^n = (u, u, \ldots, u)$ and $\tilde{x}^n = (v, v, \ldots, v)$. $\qquad\square$

While Corollary V.1 treats $n$ and $v$ as fixed constants, one can generalize the definitions of the $d$-view mutual information and channel dispersion to the setting where $v$ is a *random variable* too. In particular, let $I_n^{(d)}$ and $V_n^{(d)}$ be defined the same way as in (58) and (59), respectively, but where the expectations are taken with respect to the randomness in the length $v$ of the outputs too. Furthermore, let

$$f_\lambda(x^n, \tilde{x}^n) := \sum_{m \geq 0} \Pr[v = m] \sum_{y^m} P^{1-\lambda}(y^m | x^n) P^\lambda(y^m | \tilde{x}^n). \tag{68}$$

We then define

$$\rho_n := \min_{x^n \neq \tilde{x}^n} \max_{\lambda \in [0,1]} -\log f_\lambda(x^n, \tilde{x}^n). \tag{69}$$

In what follows, we present some estimates of $\rho_n$ for the deletion channel, where the length $v \leq n$ of the outputs is random.

### A. Upper Bounds on $\rho_n$ for the Deletion Channel

In this subsection, we consider the multi-letter deletion channel, which is an important example of a synchronization channel [3]. Recall that for $\delta \in (0, 1)$, the deletion channel $\text{Del}(\delta)$ deletes each input bit $x_i \in \mathcal{X}$, $1 \leq i \leq n$, with probability $\delta$.

For $\text{Del}(\delta)$, for any fixed $x^n, \tilde{x}^n$ with $x^n \neq \tilde{x}^n$, and for any fixed $\lambda \in (0, 1)$, we have

$$f_\lambda(x^n, \tilde{x}^n) = \sum_{m=0}^{n} \delta^{n-m} (1-\delta)^m \sum_{y^m} N^{1-\lambda}(x^n \to y^m) \cdot N^\lambda(\tilde{x}^n \to y^m), \tag{70}$$

where $N(u^n \to v^m)$, for $u^n \in \{0,1\}^n$, $v^m \in \{0,1\}^m$, is the number of occurrences of $v^m$ as a (potentially non-contiguous) subsequence in $u^n$.

In what follows, we present three upper bounds on $\rho_n$ for $\text{Del}(\delta)$, and via these upper bounds, arrive at some interesting properties of $\rho_n$ and comparisons with the rates $\rho_{n,n}$ of the $\text{BEC}(\delta)$ and $\text{BSC}(\delta)$. The upper bounds are obtained by fixing specific sequences $x^n, \tilde{x}^n \in \mathcal{X}^n$ and using $\max_{\lambda \in [0,1]} -\log f_\lambda(x^n, \tilde{x}^n)$, evaluated at these chosen sequences, as an upper bound on $\rho_n$. The first of our upper bounds is naïve and will be subsumed by the later bounds.

**Lemma V.1.** *For* $\text{Del}(\delta)$*, we have that* $\rho_n \leq -n \log \delta$.

*Proof.* Consider *any* fixed $x^n, \tilde{x}^n$ with $x^n \neq \tilde{x}^n$. For any $\lambda \in [0,1]$, $f_\lambda(x^n, \tilde{x}^n) \geq \delta^n$, from (70), since the empty string occurs exactly once as a subsequence of both $x^n$ and $\tilde{x}^n$. Hence, we have that

$$\rho_n = \min_{x^n \neq \tilde{x}^n} \max_{\lambda \in [0,1]} -\log f_\lambda(x^n, \tilde{x}^n) \leq -n \log \delta. \tag{71}$$

$\square$

As expected, from the above lemma and from Corollary V.2, we see that the rate $\rho_n$ for the deletion channel is *smaller* than the rate $\rho_{n,n}$ for the $n$-letter erasure channel with erasure probability $\delta$. Indeed, recall that for the $\text{BEC}(\delta)$, we have that $\rho = -\log \delta$, and hence, by Corollary V.2, $\rho_{n,n} = -n \log \delta$. Note also that by an application of the data processing inequality, it follows that the capacity of $\text{Del}(\delta)$ is at most the capacity of the $\text{BEC}(\delta)$, for $\delta \in [0,1]$ [3, Sec. 5.2].

The bound that follows improves on this naïve upper bound, for a certain range of $\delta$ values, and for large enough $n$. For simplicity, in the second part of the next lemma, we assume that $n$ is even; the case of odd $n$ can be handled with minor modifications to the exisiting proof.

**Lemma V.2.** *For* $\text{Del}(\delta)$*, we have that* $\limsup_{n \to \infty} \rho_n = 0$*, if* $\delta > 1/2$.

*Proof.* Fix $\delta > 1/2$. We pick $x^n, \tilde{x}^n$ to be, respectively, the alternating sequences $01010\ldots$ and $10101\ldots$, each having length $n$, for the purposes of this lemma.

Now, for any $y^m \in \{0,1\}^m$, with $m \leq n$, we see that $N(x^n \to y^m) = N(\tilde{x}^n \to y^m)$. Furthermore, $N(x^n \to y^m) > 0$ iff $w(y^m) \in [\max\{0, m - n/2\}, n/2]$; we then call $y^m$ "admissible". For such admissible $y^m$, we have that

$$N(x^n \to y^m) = \binom{n/2}{w(y^m)} \cdot \binom{n/2}{m - w(y^m)} = N(\tilde{x}^n \to y^m). \tag{72}$$

The above expression is obtained by choosing $w(y^m)$ locations among the $n/2$ 1s in $x^n$ (resp. $\tilde{x}^n$) as the locations of the 1s in $y^m$, and likewise, $n - w(y^m)$ locations among the $n/2$ 0s in $x^n$ (resp. $\tilde{x}^n$) as the locations of the 0s in $y^m$. Furthermore, in the above expression, we implicitly assume that the binomial coefficient $\binom{k}{t} = 0$ if $t < 0$ or $t > k$.

Thus, for $\lambda \in (0, 1)$,

$$f_\lambda(x^n, \tilde{x}^n) = \sum_{m=0}^{n} \delta^{n-m}(1-\delta)^m \sum_{y^m \text{ admissible}} \binom{n/2}{w(y^m)}\binom{n/2}{m - w(y^m)} \tag{73}$$

$$= \sum_{m=0}^{n} \delta^{n-m}(1-\delta)^m \sum_{w=\max\{m-n/2,0\}}^{\min\{m,n/2\}} \binom{m}{w}\binom{n/2}{w}\binom{n/2}{m-w}. \tag{74}$$

Let us call the inner summation above as $\beta(m)$. Now, note that

$$\beta(m) = \begin{cases} \sum_{w=0}^{m} \binom{m}{w}\binom{n/2}{w}\binom{n/2}{m-w}, & \text{if } m \leq n/2, \\ \sum_{w=m-n/2}^{n/2} \binom{m}{w}\binom{n/2}{w}\binom{n/2}{m-w}, & \text{if } m > n/2. \end{cases} \tag{75}$$

After some algebra, we can write

$$\beta(m) = \begin{cases} \sum_{w=0}^{m} \binom{m}{w}\binom{n/2}{w}\binom{n/2}{m-w}, & \text{if } m \leq n/2, \\ \sum_{w=0}^{n-m} \binom{m}{n/2-w}\binom{n/2}{w}\binom{n/2}{m-(n/2-w)}, & \text{if } m > n/2. \end{cases} \tag{76}$$

Let

$$\alpha_1(m) := \sum_{w=0}^{m} \binom{m}{w}\binom{n/2}{w}\binom{n/2}{m-w} \tag{77}$$

and

$$\alpha_2(m) := \sum_{w=0}^{n-m} \binom{m}{n/2-w}\binom{n/2}{w}\binom{n/2}{m-(n/2-w)}. \tag{78}$$

Consider $\alpha_1(m)$, for the case when $m \leq n/2$. We then have that

$$\alpha_1(m) \geq \min_{0 \leq \omega \leq m} \binom{n/2}{m-\omega} \cdot \sum_{w=0}^{m} \binom{m}{w}\binom{n/2}{w} \tag{79}$$

$$= \binom{n/2}{0}\binom{n/2+m}{m} = \binom{n/2+m}{m}, \tag{80}$$

where the first equality follows from the Chu-Vandermonde identity (see, e.g., [29, p. 42]). Likewise, consider $\alpha_2(m)$, for the case when $m \geq n/2$. Note that

$$\alpha_2(m) = \sum_{w=0}^{n-m} \binom{m}{n/2-w}\binom{n/2}{w}\binom{n/2}{m-(n/2-w)} \tag{81}$$

$$= \sum_{w=0}^{n-m} \binom{m}{n/2-w}\binom{n/2}{w}\binom{n/2}{n-m-w} \tag{82}$$

$$\geq \sum_{w=0}^{n-m} \binom{n/2}{w}\binom{n/2}{n-m-w} = \binom{n}{m}, \tag{83}$$

where the last equality above again uses the Chu-Vandermonde identity. Thus, we obtain that

$$\beta(m) \geq \begin{cases} \binom{n/2+m}{m}, & \text{if } m \leq n/2, \\ \binom{n}{m}, & \text{if } m > n/2. \end{cases} \tag{84}$$

Plugging this back into (74), we see that

$$f_\lambda(x^n, \tilde{x}^n) \geq \sum_{m=0}^{n/2} \delta^{n-m}(1-\delta)^m \cdot \binom{n/2+m}{m} + \sum_{m=n/2+1}^{n} \delta^{n-m}(1-\delta)^m \cdot \binom{n}{m}. \tag{85}$$

Call the first summation above as $\gamma_1$ and the second as $\gamma_2$. Now, observe that $\gamma_2$ is precisely the probability $\Pr[L > n/2]$, for a binomial random variable $L \sim \mathrm{Bin}(n, \delta)$. It is well-known (see, e.g., [19, Example 1.6.2]) that for $\delta > 1/2$, this probability is at least

$$1 - \exp(-n \cdot D(\mathrm{Ber}(1/2)\|\mathrm{Ber}(\delta))) = 1 - Z(\delta)^n, \tag{86}$$

where $Z(\delta) = 2\sqrt{\delta(1-\delta)}$ is the Bhattacharya parameter of the $\mathrm{BSC}(\delta)$. Furthermore, note that

$$\gamma_1 \geq (\delta(1-\delta))^{n/2} \cdot \binom{n}{n/2} \tag{87}$$

$$= (\delta(1-\delta))^{n/2} \cdot 2^{n-O(\log n)} \tag{88}$$

$$= Z(\delta)^n \cdot 2^{-O(\log n)}. \tag{89}$$

where the first equality follows from [30, Eq. (5.38)].

Hence, we obtain that for $\lambda \in (0, 1)$,

$$f_\lambda(x^n, \tilde{x}^n) \geq 1 - Z(\delta)^n \cdot (1 - 2^{-O(\log n)}). \tag{90}$$

Using the previous inequality, we obtain that when $\delta > 1/2$,

$$\limsup_{n \to \infty} \rho_n \leq \limsup_{n \to \infty} \max_{\lambda \in [0,1]} -\log f_\lambda(x^n, \tilde{x}^n) \tag{91}$$

$$\leq \limsup_{n} \max\{0, -\log(1 - Z(\delta)^n \cdot (1 - 2^{-O(\log n)}))\} \tag{92}$$

$$= 0, \tag{93}$$

where the last inequality uses the fact that for $\lambda \in \{0, 1\}$, we have $f_\lambda(x^n, \tilde{x}^n) = 1$. $\qquad\square$

Next, we show that for all $\delta \in (0, 1)$ (and hence for $\delta \leq 1/2$ too), the rate of convergence $\rho_n$ can in fact be bounded from above by a constant, for large enough $n$. Before we do so, we state and prove a useful lemma.

**Lemma V.3.** *For* $\mathrm{Del}(\delta)$*, where* $\delta \in (0, 1)$*, we have that*

$$\rho_n \leq \max_{\lambda \in [0,1]} \lambda \log n - \log \mathbb{E}_{L \sim \mathrm{Bin}(n, \delta)}\left[L^\lambda\right]. \tag{94}$$

*Proof.* As before, we pick specific sequences $x^n, \tilde{x}^n$ and seek to upper bound $\rho_{n,\nu}$ by $\max_{\lambda \in [0,1]} -\log f_\lambda(x^n, \tilde{x}^n)$, for this choice of $x^n, \tilde{x}^n$. Here, we choose $x^n = 0^n$ and $\tilde{x}^n = 0^{n-1}1$.

Clearly, for any $m \leq n$, we have that $y^m \in \{0, 1\}^n$ is a subsequence of $x^n$ iff $y^m = 0^m$. For $y^m = 0^m$, hence, it follows that

$$N(x^n \to y^m) = \binom{n}{m} \quad \text{and} \quad N(\tilde{x}^n \to y^m) = \binom{n-1}{m}. \tag{95}$$

Thus, for $\lambda \in (0, 1)$,

$$f_\lambda(x^n, \tilde{x}^n) = \sum_{m=0}^{n} \delta^{n-m}(1-\delta)^m \cdot \binom{n}{m}^{1-\lambda} \cdot \binom{n-1}{m}^\lambda \tag{96}$$

$$= \sum_{m=0}^{n} \delta^{n-m}(1-\delta)^m \cdot \binom{n}{m} \cdot \left(\frac{n-m}{n}\right)^\lambda \tag{97}$$

$$= \mathbb{E}_{L \sim \mathrm{Bin}(n, 1-\delta)} \left[ \left(1 - \frac{L}{n}\right)^\lambda \right] \tag{98}$$

$$= \frac{1}{n^\lambda} \cdot \mathbb{E}_{L \sim \mathrm{Bin}(n, \delta)} \left[ L^\lambda \right], \tag{99}$$

and the claim follows by noting that $f_\lambda(x^n, \tilde{x}^n) = 1$, for $\lambda \in \{0, 1\}$. $\qquad \square$

We next use Lemma V.3 to show that for large enough $n$, $\rho_n$ can be bounded from above by a *constant*. We thus obtain the following corollary:

**Corollary V.3.** *For* $\mathrm{Del}(\delta)$*, where* $\delta \in (0, 1)$*, we have that* $\limsup_{n \to \infty} \rho_n \leq -\log \delta$*.*

*Proof.* The proof proceeds by obtaining estimates of the fractional moments of a binomial random variable used in Lemma V.3. Let $\beta, \eta > 0$ be small real numbers. We then have that for $L \sim \mathrm{Bin}(n, \delta)$, and $\lambda \in (0, 1)$,

$$\mathbb{E}[L^\lambda] \geq (n(\delta - \beta))^\lambda \cdot \Pr[L \geq n(\delta - \beta)] \tag{100}$$

$$\geq (n(\delta - \beta))^\lambda \cdot (1 - \exp(-n \cdot D(\mathrm{Ber}(\delta - \beta) || \mathrm{Ber}(\delta)))) \tag{101}$$

$$\geq (1 - \eta) \cdot (n(\delta - \beta))^\lambda, \tag{102}$$

for $n$ large enough.

Plugging (102) into Lemma V.3, we obtain that for $n$ large enough,

$$\rho_n \leq \max_{\lambda \in [0,1]} \lambda \log n - \log \mathbb{E}_{L \sim \mathrm{Bin}(n, \delta)} \left[ L^\lambda \right] \tag{103}$$

$$\leq \max \left\{ 0, \max_{\lambda \in (0,1)} \lambda \log n - \lambda \log(n(\delta - \beta)) - \log(1 - \eta) \right\} \tag{104}$$

$$= \max \left\{ 0, \max_{\lambda \in (0,1)} -\lambda \log(\delta - \beta) - \log(1 - \eta) \right\} \tag{105}$$

$$\leq -(1 - \zeta) \log \delta, \tag{106}$$

for small $\zeta > 0$. Thus, taking $\limsup_{n \to \infty}$ on both sides of the inequality above and letting $\zeta \downarrow 0$, we obtain the statement of the corollary. $\qquad \square$

Hence, putting together Lemma V.2 and Corollary V.3, we see that for the deletion channel $\mathrm{Del}(\delta)$, the rate $\rho_n$, for large enough $n$, is bounded above by a fixed constant, for $\delta \leq 1/2$, and by an arbitrarily small constant, for $\delta > 1/2$.

In the next section, we take a detour from the asymptotic regime of large $d$ considered in the rest of the paper, and seek to obtain tight, non-asymptotic bounds on the capacity of a special multi-view DMC, which holds for all values of $d$.

## VI. Poisson Approximation Channel

In this section, we consider the special case of the $d$-view $\mathrm{BSC}(p)$, denoted by $\mathrm{BSC}^{(d)}(p)$. Recall that the capacity of this channel equals the capacity of the binomial channel $\mathrm{Bin}_d(p)$, which is given in (26). Our objective is to show that the Poisson approximation channel $\mathrm{Poi}_d(p)$ introduced in Section III has capacity that is a very good lower bound on the capacity of $\mathrm{BSC}^{(d)}(p)$, via the proof of Theorem III.3. Before we proceed with the proof, we state and prove a simple lemma about the concavity of the mutual information $I^{(d)} = I(X; Y^d)$, for any $d$-view DMC $W^{(d)}$ and input distribution $P_X$.

**Lemma VI.1.** *For any $d$-view DMC $W^{(d)}$ and input distribution $P_X$, we have that $I^{(d)}$ is concave over the positive integers, i.e.,*

$$I^{(d)} - I^{(d-1)} \geq I^{(d+1)} - I^{(d)}, \tag{107}$$

*for all $d \geq 1$.*

*Proof.* We first rewrite what we intend to prove differently: since $I^{(m)} = \mathsf{H}(X) + \mathsf{H}(Y^m) - \mathsf{H}(X, Y^m)$, for any $m \geq 0$, with $I^{(0)} := 0$, we would like to show that

$$\mathsf{H}(Y^d) - \mathsf{H}(Y^{d-1}) + \mathsf{H}(X, Y^{d-1}) - \mathsf{H}(X, Y^d)$$
$$\geq \mathsf{H}(Y^{d+1}) - \mathsf{H}(Y^d) + \mathsf{H}(X, Y^d) - \mathsf{H}(X, Y^{d+1}). \tag{108}$$

However, observe that for any $m \geq 0$, $\mathsf{H}(X, Y^{m+1}) - \mathsf{H}(X, Y^m) = \mathsf{H}(Y_{m+1}|X, Y^m) = \mathsf{H}(Y_{m+1}|X)$, by the channel law of $W^{(d)}$. Hence, (108) reduces to showing that

$$\mathsf{H}(Y^d) - \mathsf{H}(Y^{d-1}) \geq \mathsf{H}(Y^{d+1}) - \mathsf{H}(Y^d). \tag{109}$$

Now, by the submodularity of entropy [24, Thm. 1.6], we have that

$$\mathsf{H}(Y^{d-1}, Y_{d+1}) - \mathsf{H}(Y^{d-1}) \geq \mathsf{H}(Y^{d+1}) - \mathsf{H}(Y^d). \tag{110}$$

The above implies (109), since the distributions of $(Y^{d-1}, Y_{d+1})$ and $Y^d$ are equal, by the stationarity of the DMC $W^{(d)}$. $\square$

For clarity, we first write down the capacity $C(\mathsf{Poi}_d(p))$ of the Poisson approximation channel. As argued in Section III, the capacity is achieved when the inputs $X \sim P_X = \mathrm{Ber}(1/2)$, yielding

$$C(\mathsf{Poi}_d(p)) = \mathsf{H}(X) - \mathsf{H}(X|R_1, R_2) \tag{111}$$

$$= 1 + \sum_{\substack{r_1 \geq 0, \\ r_2 \geq 0}} \frac{e^{-dp}(dp)^{r_1}}{r_1!} \cdot \frac{e^{-d(1-p)}(d(1-p))^{r_2}}{r_2!} \cdot \log \frac{p^{r_1}(1-p)^{r_2}}{p^{r_1}(1-p)^{r_2} + p^{r_2}(1-p)^{r_1}}. \tag{112}$$

We now prove Theorem III.3.

*Proof of Thm. 3.* Consider the mutual information between the input and outputs of the $\mathsf{Poi}_d(p)$ channel when $P_X$ is uniform over the inputs. Since the number of outputs $N = R_1 + R_2$ is known to the decoder, we have that

$$C(\mathsf{Poi}_d(p)) = I(X; R_1, R_2) \tag{113}$$

$$= I(X; R_1, R_2 \mid N) \tag{114}$$

$$= \sum_{n=0}^{\infty} \mathbb{P}[N = n] \cdot I(X; R_1, R_2 \mid N = n). \tag{115}$$

Now, observe that since the sum of a $\mathrm{Poi}(dp)$ random variable and a $\mathrm{Poi}(d(1-p))$ random variable is distributed according to $\mathrm{Poi}(d)$, we have by symmetry that $N \sim \mathrm{Poi}(d)$. Further, it is easy to check that the conditional distribution $P_{R_1,R_2|N=n}$ obeys $P_{R_1,R_2|N=n}(r_1, r_2 \mid n) = \binom{n}{r_1} p^{r_1}(1-p)^{n-r_1}$, for all $r_1, r_2 \geq 0$ such that $r_1 + r_2 = n$. Hence, we have that $I(X; R_1, R_2 \mid N = n) = C(\mathsf{Bin}_n(p))$.

Thus, substituting in (115), we get that

$$C(\mathsf{Poi}_d(p)) = \mathbb{E}_{N \sim \mathrm{Poi}(d)} \left[ C(\mathsf{Bin}_N(p)) \right]. \tag{116}$$

Next, using the above pleasing expression, we show that $C(\mathsf{Poi}_d(p))$ is indeed a lower bound on $C(\mathsf{Bin}_d(p))$. From Lemma VI.1 we have that $C(\mathsf{Bin}_n(p))$ is concave over the positive integers. Therefore, applying Jensen's inequality, we get that

$$C(\mathsf{Poi}_d(p)) = \mathbb{E}_{N \sim \mathrm{Poi}(d)} \left[ C(\mathsf{Bin}_N(p)) \right] \tag{117}$$

$$\leq C(\mathsf{Bin}_{\mathbb{E}N}(p)) = C(\mathsf{Bin}_d(p)). \tag{118}$$

Furthermore, we have that

$$C(\mathsf{Bin}_d(p)) - C(\mathsf{Poi}_d(p)) = C(\mathsf{Bin}_d(p)) - \mathbb{E}_{N \sim \mathrm{Poi}(d)} \left[ C(\mathsf{Bin}_N(p)) \right] \tag{119}$$

$$\leq C(\mathsf{Bin}_d(p)) - 1 + \mathbb{E}_{N \sim \mathrm{Poi}(d)} [Z(p)^N] \tag{120}$$

$$\leq \exp(-d(1 - Z(p))) - Z(p)^{2d}, \tag{121}$$

where the two inequalities make use of Corollary IV.1 and the last inequality uses the well-known expression for the moment-generating function of a Poisson random variable. $\qquad \square$

*Remark.* For the binomial channel $\overline{\mathsf{Bin}}_n$, whose input is a real number in $[0, 1]$, with $n$ output trials, that is considered in [31] too, a suitable Poisson approximation channel $\overline{\mathsf{Poi}}_n$ can be designed. This channel takes as input $X \in [0, 1]$ and outputs the pair $(R_1, R_2)$ such that $P_{R_1|X} = \mathrm{Poi}(nX)$ and $P_{R_2|X} = \mathrm{Poi}(n(1 - X))$. For $\overline{\mathsf{Poi}}_n$, as above, we will have that $C(\overline{\mathsf{Poi}}_n) = \mathbb{E}_{N \sim \mathrm{Poi}(n)} \left[ C(\overline{\mathsf{Bin}}_N) \right]$. However, the proof of the concavity of the capacity $C(\overline{\mathsf{Bin}}_n)$ as in Lemma VI.1 will not hold, owing to the failure of the conditional independence assumption that is crucially employed in the proof. We are hence unable to obtain a straightforward lower bound on $C(\overline{\mathsf{Bin}}_n)$ via $C(\overline{\mathsf{Poi}}_n)$, using the previous techniques.

## VII. PROOFS

In this section, we provide a detailed proof of our main result, Theorem III.1.

### A. Convergence of Mutual Information for General DMCs

Given a general DMC $W$ with input alphabet $\mathcal{X}$ (not necessarily binary) and output alphabet $\mathcal{Y}$, consider the $d$-view DMC $W^{(d)}$. Let the input to $W^{(d)}$ be $X$, with input distribution being $P_X$, and let its outputs be $Y_1, \ldots, Y_d \in \mathcal{Y}^d$. The joint distribution of the inputs and outputs is such that $P_{X,Y^d}(x, y^d) = P_X(x) \prod_{i=1}^{d} W(y_i|x)$. Much like the proof of Corollary IV.1, the inequality

$$\mathsf{H}(X|Y^d) \leq Z_g(W)^d, \tag{122}$$

can be shown to hold (see [26, Prop. 4.8]), where

$$Z_g(W) := \sum_{x \neq x'} \sum_{y \in \mathcal{Y}} \sqrt{P(x)W(y|x)P(x')W(y|x')} \tag{123}$$

is a scaled version of the Bhattacharya parameter for general DMCs. This then results in $\rho(I) \geq -\log Z_g(W)$, thereby showing that for well-behaved DMCs, the rate of convergence of $I^{(d)}$ to $\mathsf{H}(X)$ is exponential in $d$. We shall now take up the question of the exact speed of this convergence.

To this end, we will be next interested in evaluating, up to exponential tightness, the behaviour of

$$\mathsf{H}(X|Y^d) = \mathbb{E}\left[ \log \frac{1}{P_{X|Y^d}(X \mid Y^d)} \right] \tag{124}$$

$$= \sum_{x \in \mathcal{X}} P_X(x) \mathbb{E}\left[ \log \frac{1}{P_{X|Y^d}(x \mid Y^d)} \middle| X = x \right]. \tag{125}$$

Now, fix an $x \in \mathcal{X}$ and focus on the inner term in (125). Then,

$$\mathbb{E}\left[ \log \frac{1}{P_{X|Y^d}(x \mid Y^d)} \middle| X = x \right] = \int_0^\infty \mathbb{P}\left[ \log \frac{1}{P_{X|Y^d}(x \mid Y^d)} \geq t \middle| X = x \right] \mathrm{d}t. \tag{126}$$

The inner probability, to wit,

$$p_x(t) := \mathbb{P}\left[ -\log P_{X|Y^d}(x \mid Y^d) \geq t | X = x \right] \tag{127}$$

is bounded as follows:

**Lemma VII.1.** *We have that*

$$\frac{p_x(t)}{(|\mathcal{X}| - 1)} \leq \max_{\tilde{x} \neq x} \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-(t - \log(|\mathcal{X}| - 1))}}{1 - e^{-t}}\middle| X = x\right] \tag{128}$$

*and*

$$p_x(t) \geq \max_{\tilde{x} \neq x} \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right]. \tag{129}$$

*Proof.* First, observe that

$$p_x(t) = \mathbb{P}\left[\frac{P_{XY^d}(x, Y^d)}{P_{Y^d}(Y^d)} \leq e^{-t}\middle| X = x\right] \tag{130}$$

$$= \mathbb{P}\left[P(x)P(Y^d \mid x) \leq e^{-t} \sum_{\tilde{x}} P(\tilde{x})P(Y^d \mid \tilde{x})\middle| X = x\right] \tag{131}$$

$$= \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{\sum_{\tilde{x} \neq x} P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right]. \tag{132}$$

We first prove the upper bound in the lemma. The above probability obeys

$$\mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{\sum_{\tilde{x} \neq x} P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right] \leq \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{(|\mathcal{X}| - 1) \max_{\tilde{x} \neq x} P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right] \tag{133}$$

$$= \mathbb{P}\left[\frac{1}{|\mathcal{X}| - 1} \cdot \min_{\tilde{x} \neq x} \frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right] \tag{134}$$

$$\leq \sum_{\tilde{x} \neq x} \mathbb{P}\left[\frac{1}{|\mathcal{X}| - 1} \cdot \frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right] \tag{135}$$

$$\leq (|\mathcal{X}| - 1) \cdot \max_{\tilde{x} \neq x} \mathbb{P}\left[\frac{1}{|\mathcal{X}| - 1} \frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}}\middle| X = x\right], \tag{136}$$

thereby showing the upper bound. Further, for any $\tilde{x}$, we have that

$$p_x(t) = \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{\sum_{\tilde{x} \neq x} P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}} \mid X = x\right] \tag{137}$$

$$\geq \max_{\tilde{x} \neq x} \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-t}}{1 - e^{-t}} \mid X = x\right], \tag{138}$$

thereby proving the lower bound. □

Since we are interested mainly in the behaviour of $\mathsf{H}(X|Y^d)$ up to first order in the exponent $d$, it suffices for us to understand the behaviour of $p_x(t)$ upto first order in the exponent $d$ with the expectation that we will accrue an additional $\Theta(\log d)$ term in the exponent, in the final expression (see Eq. (126) above and [27, Sec. 4.2] for intuition). To this end, following Lemma VII.1, for fixed $z \in [0, \infty)$ and $c > 0$, we shall focus on the term

$$\Gamma_{x, \tilde{x}}(z) := \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{ce^{-z}}{1 - e^{-z}}\middle| X = x\right], \tag{139}$$

with the reasoning that the rate of convergence of $\mathsf{H}(X|Y^d)$ will be dictated by the smallest rate of $\Gamma_{x,\tilde{x}}(z)$, as $z$ increases from 0 to $\infty$.

Before we proceed, we require some more notation. Let us denote the log-likelihood ratio by

$$L_{x,\tilde{x}}^y := \log \frac{P_{Y|X}(y \mid x)}{P_{Y|X}(y \mid \tilde{x})}. \tag{140}$$

For a given $y^d \in \mathcal{Y}^d$ and for $b \in \mathcal{Y}$, let $\hat{Q}_{y^d}(b) := \frac{1}{d} \sum_{i=1}^d \mathbb{1}\{y_i = b\}$ be the empirical frequency of the letter $b$ in any output sequence. Further, we define

$$L_{x,\tilde{x}}(\hat{Q}) := \sum_{b \in \mathcal{Y}} \hat{Q}(b) L_{x,\tilde{x}}^b. \tag{141}$$

Finally, let

$$v_{x,\tilde{x}} \equiv v_{x,\tilde{x}}(z) := \frac{1}{d} \left( \log \left( \frac{ce^{-z}}{1 - e^{-z}} \right) - \log \frac{P(x)}{P(\tilde{x})} \right) \tag{142}$$

and let $z(v_{x,\tilde{x}}) := \log \left( 1 + \frac{cP(\tilde{x})}{P(x)} \cdot e^{-dv_{x,\tilde{x}}} \right)$ be its inverse function.

**Proposition VII.1.** *We have that for any fixed $\delta > 0$,*

*1) If $v_{x,\tilde{x}} \geq L_{x,\tilde{x}}(P_{Y|x}) + \delta$, then $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = \Theta(1)$.*

*2) If $v_{x,\tilde{x}} \leq L_{x,\tilde{x}}(P_{Y|x}) - \delta$, then $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = e^{-dE(v_{x,\tilde{x}}) + \Theta(\log d)}$, where*

$$E(v_{x,\tilde{x}}) := \begin{cases} \max_{\lambda \geq 0} \left[ -\log \left( \sum_y P(y|x)^{1-\lambda} P(y|x')^\lambda \right) - \lambda v_{x,\tilde{x}} \right], & \text{if } v_{x,\tilde{x}} \geq \min_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b, \\ \infty, & \text{otherwise.} \end{cases} \tag{143}$$

*Proof.* Observe that for any $z \in [0, \infty)$,

$$\Gamma_{x,\tilde{x}}(z) = \mathbb{P} \left[ \frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{ce^{-z}}{1 - e^{-z}} \middle| X = x \right] \tag{144}$$

$$= \mathbb{P} \left[ \sum_{i=1}^d \log \frac{P_{Y|X}(Y_i \mid x)}{P_{Y|X}(Y_i \mid \tilde{x})} \leq \log \left( \frac{ce^{-z}}{1 - e^{-z}} \right) - \log \frac{P(x)}{P(\tilde{x})} \middle| X = x \right] \tag{145}$$

$$= \mathbb{P} \left[ \sum_{i=1}^d L_{x,\tilde{x}}^{Y_i} \leq \log \left( \frac{ce^{-z}}{1 - e^{-z}} \right) - \log \frac{P(x)}{P(\tilde{x})} \middle| X = x \right] \tag{146}$$

$$= \mathbb{P} \left[ \mathbb{E}_{\hat{Q}} \left[ L_{x,\tilde{x}}^{Y^d} \right] \leq v_{x,\tilde{x}}(z) \middle| X = x \right]. \tag{147}$$

Writing this differently, we have that for any $v_{x,\tilde{x}} \in \mathbb{R}$,

$$\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = \mathbb{P} \left[ \mathbb{E}_{\hat{Q}} \left[ L_{x,\tilde{x}}^{Y^d} \right] \leq v_{x,\tilde{x}} \middle| X = x \right]. \tag{148}$$

Consider first the case when $v_{x,\tilde{x}} \geq L_{x,\tilde{x}}(P_{Y|x}) + \delta$. Note that $L_{x,\tilde{x}}(P_{Y|x}) = D(P_{Y|x} || P_{Y|\tilde{x}}) > 0$. We split this case into two subcases.

- If $v_{x,\tilde{x}} \geq \max_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b$ then $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = 1$.

- Otherwise, consider the case where $L_{x,\tilde{x}}(P_{Y|x}) + \delta \le v_{x,\tilde{x}} \le \max_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b$. Now, observe that

$$1 \ge \Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) \ge \mathbb{P}\left[\mathbb{E}_{\hat{Q}}\left[L_{x,\tilde{x}}^{Y^d}\right] \le L_{x,\tilde{x}}(P_{Y|x}) + \delta \Big| X = x\right]. \tag{149}$$

Furthermore, by a simple application of Markov's inequality, we get that since $L_{x,\tilde{x}}(P_{Y|x}) = \mathbb{E}_{Y^d \sim P_{Y^d|x}}[L_{x,\tilde{x}}^{Y^d}]$, the following inequality holds:

$$\mathbb{P}\left[\mathbb{E}_{\hat{Q}}\left[L_{x,\tilde{x}}^{Y^d}\right] \le L_{x,\tilde{x}}(P_{Y|x}) + \delta \Big| X = x\right] \ge \frac{\delta}{L_{x,\tilde{x}}(P_{Y|x}) + \delta}. \tag{150}$$

Hence, we have that $1 \ge \Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) \ge \frac{\delta}{L_{x,\tilde{x}}(P_{Y|x}) + \delta}$.

Hence, overall, if $v_{x,\tilde{x}} \ge L_{x,\tilde{x}}(P_{Y|x}) + \delta$, we get that $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = \Theta(1)$.

Next, consider the case when $v_{x,\tilde{x}} \le L_{x,\tilde{x}}(P_{Y|x}) - \delta$. We again split this case into two subcases.

- If $v_{x,\tilde{x}} \le \min_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b$, then $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = 0$.

- If $\min_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b \le v_{x,\tilde{x}} \le L_{x,\tilde{x}}(P_{Y|X=x})$, then this is a large-deviations event. The probability $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}}))$ thus decays exponentially in $d$, i.e., $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}})) = e^{-dE_1(v_{x,\tilde{x}})+\Theta(\log d)}$, for some $E_1 : \mathbb{R} \to [0, \infty)$, where by Sanov's theorem (see [17, Sec. 11.4]), we have that

$$E_1(v_{x,\tilde{x}}) = \min_{Q : L_{x,\tilde{x}}(Q) \le v_{x,\tilde{x}}} D(Q \| P_{Y|X=x}). \tag{151}$$

We now aim to characterize this function $E_1$.

From convex duality (see [32, Sec. 5.2]), we obtain that

$$E_1(v_{x,\tilde{x}}) = \max_{\lambda \ge 0} \min_Q D(Q \| P_{Y|X=x}) + \lambda(L_{x,\tilde{x}}(Q) - v_{x,\tilde{x}}). \tag{152}$$

First, let us solve the inner minimization. Fix $\lambda \ge 0$. Observe that the term $f(\lambda, Q) := D(Q \| P_{Y|X=x}) + \lambda L_{x,\tilde{x}}(Q)$ obeys

$$f(\lambda, Q) = \sum_y Q(y) \left( \log \frac{Q(y)}{P_{Y|X=x}(y)} + \lambda L_{x,\tilde{x}}(y) \right) \tag{153}$$

$$= -\sum_y Q(y) \left( \log \frac{P_{Y|X=x}(y)}{Q(y)e^{\lambda L_{x,\tilde{x}}(y)}} \right) \tag{154}$$

$$\ge -\log \sum_y \frac{P_{Y|X=x}(y)}{e^{\lambda L_{x,\tilde{x}}(y)}}, \tag{155}$$

using Jensen's inequality. Also note that equality holds when

$$Q(y) = Q^*(y) = \frac{1}{Z_{x,\tilde{x}}(\lambda)} P_{Y|X=x}(y) e^{-\lambda L_{x,\tilde{x}}(y)}, \tag{156}$$

where $Z_{x,\tilde{x}}(\lambda) := \sum_{y \in \mathcal{Y}} P_{Y|X=x}(y)e^{-\lambda L_{x,\tilde{x}}(y)}$ is the partition function (normalization constant). Substituting $Q^*$ back into the function results

$$f(\lambda, Q^*) = -\log Z_{x,\tilde{x}}(\lambda) - \lambda v_{x,\tilde{x}} \tag{157}$$

$$= -\log\left(\sum_y P_{Y|X=x}(y)e^{-\lambda L_{x,\tilde{x}}(y)}\right) - \lambda v_{x,\tilde{x}} \tag{158}$$

$$= -\log\left(\sum_y P_{Y|X=x}^{1-\lambda}(y \mid x)P_{Y|X}^{\lambda}(y \mid \tilde{x})\right) - \lambda v_{x,\tilde{x}}. \tag{159}$$

Plugging back into (152), we get that

$$E_1(v_{x,\tilde{x}}) = \max_{\lambda \geq 0}\left[-\log\left(\sum_y P_{Y|X=x}^{1-\lambda}(y \mid x)P_{Y|X}^{\lambda}(y \mid \tilde{x})\right) - \lambda v_{x,\tilde{x}}\right]. \tag{160}$$

Note that by setting $E(v_{x,\tilde{x}}) = E_1(v_{x,\tilde{x}})$ when $\min_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b \leq v_{x,\tilde{x}} \leq L_{x,\tilde{x}}(P_{Y|X=x})$ and $E(v_{x,\tilde{x}}) = \infty$ when $v_{x,\tilde{x}} \leq \min_{b \in \mathcal{Y}} L_{x,\tilde{x}}^b$, we obtain a tight characterization of the rate of exponential decay of $\Gamma_{x,\tilde{x}}(z(v_{x,\tilde{x}}))$, in these ranges of $v_{x,\tilde{x}}$. $\qquad\square$

We now state a prove a lemma that will allow us to identify the exact rate of exponential decay of $\mathsf{H}(X|Y^d)$.

**Lemma VII.2.** *We have that*

$$\int_0^\infty \max_{\tilde{x} \neq x} \Gamma_{x,\tilde{x}}(z)\mathrm{d}z = e^{\Theta(\log d|\mathcal{X}|)} \cdot e^{-d \cdot \min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})}. \tag{161}$$

*Proof.* Consider an integral of the form

$$\bar{I} = \int_0^\infty \max_{\tilde{x} \neq x} \Gamma_{x,\tilde{x}}(z)\mathrm{d}z. \tag{162}$$

First, observe that

$$\bar{I} \leq \int_0^\infty \sum_{\tilde{x} \neq x} \Gamma_{x,\tilde{x}}(z)\mathrm{d}z \tag{163}$$

$$\leq |\mathcal{X}|^2 \cdot \max_{\tilde{x} \neq x} \int_0^\infty \Gamma_{x,\tilde{x}}(z)\mathrm{d}z \tag{164}$$

$$= e^{\Theta(\log|\mathcal{X}|)} \cdot \max_{\tilde{x} \neq x} \int_0^\infty \Gamma_{x,\tilde{x}}(z)\mathrm{d}z. \tag{165}$$

Further,

$$\bar{I} \geq \max_{\tilde{x} \neq x} \int_0^\infty \Gamma_{x,\tilde{x}}(z)\mathrm{d}z. \tag{166}$$

Hence, overall, we have that $\bar{I} = \exp(\Theta(\log|\mathcal{X}|)) \cdot \max_{\tilde{x} \neq x} \int_0^\infty \Gamma_{x,\tilde{x}}(z)\mathrm{d}z$. We next focus on the integral

$$I = \int_0^\infty \Gamma_{x,\tilde{x}}(z)\mathrm{d}z, \tag{167}$$

for a fixed $x, \tilde{x}$.

By the definition

$$v_{x,\tilde{x}} = v_{x,\tilde{x}}(z) = \frac{1}{d}\left(\log\left(\frac{ce^{-z}}{1-e^{-z}}\right) - \log\frac{P(x)}{P(x')}\right), \tag{168}$$

and from Prop. VII.1, we see that the integral $I$ can be split into two integrals for its evaluation: integral $I_1$, when

$$z \leq \log\left(1 + \frac{cP(\tilde{x})}{P(x)} \cdot e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}\right) := u(d), \tag{169}$$

or equivalently, when $v_{x,\tilde{x}} \geq L_{x,\tilde{x}}(P_{Y|x}) + \delta$, and integral $I_2$, when

$$z \geq \log\left(1 + \frac{cP(\tilde{x})}{P(x)} \cdot e^{-d(L_{x,\tilde{x}}(P_{Y|x})-\delta)}\right) := \ell(d), \tag{170}$$

or equivalently, when $v_{x,\tilde{x}} \leq L_{x,\tilde{x}}(P_{Y|x}) - \delta$, with $I = I_1 + I_2$. Here, we pick $\delta < L_{x,\tilde{x}}(P_{Y|x})$.

Consider first the integral $I_1$. We have that

$$I_1 = \int_{z \leq u(d)} \Gamma_{x,\tilde{x}}(z)\mathrm{d}z \tag{171}$$

$$= \Theta(1) \cdot u(d). \tag{172}$$

Now, we claim that $u(d) = e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}$. To see this, let $\tilde{c} := \frac{cP(\tilde{x})}{P(x)}$ and recall that $\frac{t}{1+t} \leq \log(1+t) \leq t$, for $t > -1$. Hence, we have that

$$u(d) \leq \tilde{c}e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}, \tag{173}$$

and

$$u(d) \geq \frac{\tilde{c}e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}}{1 + \tilde{c}e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}} \geq \frac{\tilde{c}}{2} \cdot e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}, \tag{174}$$

since $L_{x,\tilde{x}}(P_{Y|x}) = D(P_{Y|x}||P_{Y|\tilde{x}}) > 0$. Hence, overall, we have that for any fixed $\delta > 0$,

$$I_1 = e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)+\Theta(1)}. \tag{175}$$

Taking the limit as $\delta \downarrow 0$, we get that

$$I_1 = e^{-dL_{x,\tilde{x}}(P_{Y|x})+\Theta(1)} = e^{-d \cdot D(P_{Y|x}||P_{Y|\tilde{x}})+\Theta(1)}. \tag{176}$$

Consider now the integral $I_2$. Note that from case 2) in Prop. VII.1, when $z \geq \log\left(1 + \tilde{c}e^{-d \cdot \min_b L_{x,\tilde{x}}^b}\right) := U(d)$, we have that $I_2 = 0$. Hence, we can write

$$I_2 = \int_{\ell(d) \leq z \leq U(d)} \Gamma_{x,\tilde{x}}(z)\mathrm{d}z, \tag{177}$$

where, for this interval $\ell(d) \leq z \leq U(d)$ of interest,

$$\Gamma_{x,\tilde{x}}(z) = e^{-dE(v_{x,\tilde{x}}(z))+\Theta(\log d)}, \tag{178}$$

with

$$E(v_{x,\tilde{x}}(z)) = \max_{\lambda \geq 0}\left[-\log\left(\sum_y P(y|x)^{1-\lambda}P(y|x')^{\lambda}\right) - \lambda v_{x,\tilde{x}}\right], \tag{179}$$

from Prop. VII.1.

It is clear that $E(v_{x,\tilde{x}}(z))$ increases as $v_{x,\tilde{x}}$ decreases. In the integral $I_2$, we thus make the change of variables

$$v = \frac{1}{d}\left(\log\left(\frac{ce^{-z}}{1-e^{-z}}\right) - \log\frac{P(x)}{P(x')}\right), \tag{180}$$

with $dz = \frac{-dce^{-dv}}{1+ce^{-dv}}dv$. Thus,

$$I_2 = dc\int_{v=\min_b L_{x,\tilde{x}}^b}^{L_{x,\tilde{x}}(P_{Y|x})-\delta} e^{-dE(v)} \cdot \frac{e^{-dv}}{1+ce^{-dv}}dv \tag{181}$$

$$= dc\int_{v=0}^{L_{x,\tilde{x}}(P_{Y|x})-\delta} e^{-dE(v)} \cdot \frac{e^{-dv}}{1+ce^{-dv}}dv + dc\int_{v=\min_b L_{x,\tilde{x}}^b}^{0} e^{-dE(v)} \cdot \frac{e^{-dv}}{1+ce^{-dv}}dv. \tag{182}$$

Let us call the first integral in the expansion above as $I_{2,a}$ and the second as $I_{2,b}$ (note that $\min_b L_{x,\tilde{x}}^b \leq 0$). Now, observe that

$$I_{2,a} = dc\int_{v=0}^{L_{x,\tilde{x}}(P_{Y|x})-\delta} e^{-dE(v)} \cdot e^{-dv+\Theta(\log d)}dv \tag{183}$$

$$= \exp(-d(E(L_{x,\tilde{x}}(P_{Y|x})) + L_{x,\tilde{x}}(P_{Y|x}) - \delta) + \Theta(\log d)). \tag{184}$$

Here, in the first eqauility, we argue as previously that

$$\frac{e^{-dv}}{1+ce^{-dv}} = e^{-dv+\Theta(1)}. \tag{185}$$

Furthermore, the second equation follows from Laplace's method (e.g., see [27, Sec. 4.2]), using the fact that both $e^{-dE(v)}$ and $e^{-dv}$ increase as $v$ increases. Taking $\delta \downarrow 0$, we obtain that

$$I_{2,a} = e^{-d(E(L_{x,\tilde{x}}(P_{Y|x}))+L_{x,\tilde{x}}(P_{Y|x}))+\Theta(\log d)}. \tag{186}$$

Moreover,

$$I_{2,b} = e^{\Theta(\log d)} \cdot \int_{v=\min_b L_{x,\tilde{x}}^b}^{0} e^{-dE(v)}dv \tag{187}$$

$$= e^{-dE(0)+\Theta(\log d)} = e^{-d\cdot C(P_{Y|x},P_{Y|\tilde{x}})+\Theta(\log d)}. \tag{188}$$

Here, for the first equality, we use the fact that when $v \leq 0$,

$$\frac{1}{2} \leq \frac{e^{-dv}}{1+ce^{-dv}} \leq \frac{1}{c}, \tag{189}$$

and hence $\frac{e^{-dv}}{1+ce^{-dv}} = \Theta(1)$. The second equality holds again by Laplace's method, using the fact that $e^{-dE(v)}$ increases with $v$.

Putting everything together, in order to nail down the exponential behaviour of integral $I$, we need to reconcile Equations (176), (186), and (188), to understand which among $I_1$, $I_{2,a}$, and $I_{2,b}$, is the largest. First, observe that since

$$C(P_{Y|x}, P_{Y|\tilde{x}}) \leq D(P_{Y|x}||P_{Y|\tilde{x}}), \tag{190}$$

we have that $I_{2,b}$ dominates $I_1$.

Therefore, the exponential behaviour of integral $I$ is governed by the rate

$$\overline{E} = \min \left\{ E(L_{x,\tilde{x}}(P_{Y|x})) + L_{x,\tilde{x}}(P_{Y|x}), \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) \right\}. \tag{191}$$

Now note that the maximum over $\lambda \geq 0$ in the expressions for $E$ and $\mathsf{C}$ can be rewritten as a maximum over $\lambda \in [0,1]$ (owing to the concavity of $-\log\left(\sum_y P(y|x)^{1-\lambda} P(y|x')^{\lambda}\right)$ and the fact that it equals 0 at $\lambda = 0, 1$). We then get that

$$E(L_{x,\tilde{x}}(P_{Y|x})) + L_{x,\tilde{x}}(P_{Y|x}) \geq \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}), \tag{192}$$

implying that $\overline{E} = \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})$. Therefore,

$$I = e^{-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d)}, \tag{193}$$

and hence that

$$\overline{I} = \max_{\tilde{x} \neq x} \exp(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d) + \Theta(\log|\mathcal{X}|)). \tag{194}$$

$\square$

We are now ready to prove Theorem III.1 for the convergence of the mutual information.

*Proof of Thm. III.1 for mutual information.* Recall from Lemma VII.1 that

$$p_x(t) \leq (|\mathcal{X}| - 1) \cdot \max_{\tilde{x} \neq x} \Gamma_{x,\tilde{x}}(t), \tag{195}$$

with the constant $c$ in the definition of $\Gamma_{x,\tilde{x}}(t)$ taken to be $|\mathcal{X}| - 1$. Thus, from Lemma VII.2 and (126), we get that

$$\mathbb{E}\left[\log \frac{1}{P_{X|Y^d}(x \mid Y^d)} \middle| X = x\right] = \int_0^\infty p_x(t)\mathrm{d}t \tag{196}$$

$$\leq \exp\left(-d \cdot \min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right), \tag{197}$$

where the inequality holds by the Laplace method [27, Sec. 4.2]. Similarly, since

$$p_x(t) \geq \max_{\tilde{x} \neq x} \Gamma_{x,\tilde{x}}(t), \tag{198}$$

with the constant $c$ in the definition of $\Gamma_{x,\tilde{x}}(t)$ taken to be 1, we get that

$$\mathbb{E}\left[\log \frac{1}{P_{X|Y^d}(x \mid Y^d)} \middle| X = x\right] \geq \exp\left(-d \cdot \min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{199}$$

Thus,

$$\mathbb{E}\left[\log \frac{1}{P_{X|Y^d}(x \mid Y^d)} \middle| X = x\right] = \exp\left(-d \cdot \min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{200}$$

Plugging back in (125), we obtain that

$$\mathsf{H}(X|Y^d) = \max_{x,\tilde{x}:\tilde{x}\neq x} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right), \tag{201}$$

yielding the required result. $\square$

## B. Convergence of Channel Dispersion of DMCs

We now shift our attention to proving the convergence of the dispersion $V^{(d)} = \mathbb{E}\left[\left(\iota(X;Y^d) - I(X;Y^d)\right)^2\right]$, to the varentropy $\mathsf{V}(X)$ of the input distribution $P_X$, for general DMCs $W$.

First, we write

$$V^{(d)} = \mathbb{E}\left[\left(\log \frac{P(X|Y^d)}{P(X)} - \left(\mathsf{H}(X) - \mathsf{H}(X|Y^d)\right)\right)^2\right] \tag{202}$$

$$= \mathbb{E}\left[\left(\log \frac{1}{P(X)} - \mathsf{H}(X)\right)^2\right] + \mathbb{E}\left[\left(\log P(X|Y^d) + \mathsf{H}(X|Y^d)\right)^2\right]$$

$$+ 2 \cdot \mathbb{E}\left[\left(\log P(X|Y^d) + \mathsf{H}(X|Y^d)\right) \cdot \left(\log \frac{1}{P(X)} - \mathsf{H}(X)\right)\right] \tag{203}$$

$$= \mathsf{V}(X) + \left(\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right] - \mathsf{H}(X|Y^d)^2\right) + \theta_d, \tag{204}$$

where we use the fact that $\mathbb{E}\left[-\log P(X|Y^d)\right] = \mathsf{H}(X|Y^d)$, and the cross-term $\theta_d$ is

$$2 \cdot \mathbb{E}\left[\left(\log P(X|Y^d) + \mathsf{H}(X|Y^d)\right) \cdot \left(\log \frac{1}{P(X)} - \mathsf{H}(X)\right)\right]. \tag{205}$$

Further, observe that when $P_X = \mathrm{Unif}(\mathcal{X})$, the cross term vanishes, since for any $x \in \mathcal{X}$, we have that $\log \frac{1}{P(x)} = \log |\mathcal{X}| = \mathsf{H}(X)$. Our proof strategy hence is to first characterize (up to exponential tightness) the behaviour of $\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right]$ and then handle the cross-term $\theta_d$.

**Lemma VII.3.** *We have that*

$$\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right] = e^{\Theta(\log d|\mathcal{X}|)} \cdot \max_{x,\tilde{x}:\tilde{x}\neq x} e^{-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})}. \tag{206}$$

*Proof.* This proof is entirely analogous to the proof of convergence of $I^{(d)}$ to $\mathsf{H}(X)$, for general DMCs, in Section VII-A, and we only highlight the key steps.

We first consider the term $\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right]$. As in (125), we first expand this term as:

$$\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right] = \sum_{x \in \mathcal{X}} P(x)\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2 \middle| X = x\right], \tag{207}$$

where for any $x \in \mathcal{X}$, we have

$$\mathbb{E}\left[\left(\log P(x|Y^d)\right)^2 \middle| X = x\right] = \int_{t=0}^{\infty} \Pr\left[\left(\log P(x|Y^d)\right)^2 \geq t \middle| X = x\right] dt \tag{208}$$

$$= \int_{t=0}^{\infty} \mathbb{P}\left[\log P(x|Y^d) \geq \sqrt{t} \middle| X = x\right] dt + \int_{t=0}^{\infty} \mathbb{P}\left[\log P(x|Y^d) \leq -\sqrt{t} \middle| X = x\right] dt \tag{209}$$

$$= \int_{t=0}^{\infty} \mathbb{P}\left[\log P(x|Y^d) \leq -\sqrt{t} \middle| X = x\right] dt. \tag{210}$$

Here, the last equality holds since $P(x|Y^d) \leq 1$, for any $Y^d \in \mathcal{Y}^d$. As earlier, our task reduces to understanding the exponential behaviour of

$$q_x(t) := \Pr\left[\log P(x|Y^d) \leq -\sqrt{t} \middle| X = x\right]. \tag{211}$$

Exactly as in Lemma VII.1, we have that

$$\frac{q_x(t)}{(|\mathcal{X}|-1)} \leq \max_{\tilde{x} \neq x} \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-(\sqrt{t}-\log(|\mathcal{X}|-1))}}{1 - e^{-\sqrt{t}}}\middle| X = x\right], \tag{212}$$

and

$$q_x(t) \geq \max_{\tilde{x} \neq x} \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{e^{-\sqrt{t}}}{1 - e^{-\sqrt{t}}}\middle| X = x\right]. \tag{213}$$

Consider, for a fixed $c > 0$ and $z \in [0, \infty)$, the expression

$$\Delta_{x,\tilde{x}}(z) := \mathbb{P}\left[\frac{P(x)P(Y^d \mid x)}{P(\tilde{x})P(Y^d \mid \tilde{x})} \leq \frac{ce^{-\sqrt{z}}}{1 - e^{-\sqrt{z}}}\middle| X = x\right]. \tag{214}$$

As argued in the case of the mutual information, the rate of exponential decay of $\Delta_{x,\tilde{x}}(z)$, as a function of $z$, dictates the rate of exponential decay of $q_x(t)$, and hence of $\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right]$. Once again, letting

$$v_{x,\tilde{x}} = v_{x,\tilde{x}}(z) = \frac{1}{d}\left(\log\left(\frac{ce^{-z}}{1 - e^{-z}}\right) - \log\frac{P(x)}{P(x')}\right) \tag{215}$$

and setting $(z(v_{x,\tilde{x}}))^{1/2} := \log\left(1 + \frac{cP(\tilde{x})}{P(x).}e^{-dv_{x,\tilde{x}}}\right)$, we obtain a result similar to Proposition VII.1, with the only modification being the replacement of $\Gamma_{x,\tilde{x}}$ by $\Delta_{x,\tilde{x}}$.

Next, as in Lemma VII.2, consider the integral

$$\overline{I}^V = \int_0^\infty \max_{\tilde{x} \neq x} \Delta_{x,\tilde{x}}(z)\mathrm{d}z. \tag{216}$$

We claim that

$$\overline{I}^V = e^{\Theta(\log d) + \Theta(\log|\mathcal{X}|)} \cdot e^{-d\min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})}. \tag{217}$$

To obtain this, we argue as in the proof of Lemma VII.2 that $\overline{I}^V = e^{\Theta(\log|\mathcal{X}|)} \cdot \max_{\tilde{x} \neq x} I^V$, where

$$I^V = \int_0^\infty \Delta_{x,\tilde{x}}(z)\mathrm{d}z, \tag{218}$$

for a fixed $x, \tilde{x}$.

Now, we split the evaluation of $I^V$ into two integrals: integral $I_1^V$, when

$$z \leq \left(\log\left(1 + \frac{cP(\tilde{x})}{P(x)} \cdot e^{-d(L_{x,\tilde{x}}(P_{Y|x})+\delta)}\right)\right)^2, \tag{219}$$

and integral $I_2^V$, when

$$z \geq \left(\log\left(1 + \frac{cP(\tilde{x})}{P(x)} \cdot e^{-d(L_{x,\tilde{x}}(P_{Y|x})-\delta)}\right)\right)^2, \tag{220}$$

with $I^V = I_1^V + I_2^V$. By arguments analogous to those in the proof of Lemma VII.2, we obtain that

$$I_1^V = e^{-2d \cdot D(P_{Y|x}||P_{Y|\tilde{x}})+\Theta(\log d)}, \tag{221}$$

and

$$I_2^V = e^{-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})+\Theta(\log d)}. \tag{222}$$

Thus, we have that

$$I^V = e^{-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d)}, \tag{223}$$

and hence that

$$\overline{I}^V = \exp\left(-d \cdot \min_{\tilde{x} \neq x} \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{224}$$

Putting everything together, we get that

$$\mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right] = \max_{x,\tilde{x}:\tilde{x} \neq x} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{225}$$

$\square$

**Lemma VII.4.** *We have that*

$$\theta_d = e^{\Theta(\log d|\mathcal{X}|)} \cdot \max_{x,\tilde{x}:\tilde{x} \neq x} e^{-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}})}. \tag{226}$$

*Proof.* Recall the cross term

$$\theta_d := 2 \cdot \mathbb{E}\left[\left(\log P(X|Y^d) + \mathsf{H}(X|Y^d)\right) \cdot \left(\log \frac{1}{P(X)} - \mathsf{H}(X)\right)\right]. \tag{227}$$

We then write

$$\theta_d(x) := \left(\log \frac{1}{P(x)} - \mathsf{H}(X)\right) \cdot \left(\mathbb{E}\left[\log P(x|Y^d)\Big| X = x\right] + \mathsf{H}(X|Y^d)\right), \tag{228}$$

with $\theta_d = \sum_{x \in \mathcal{X}} P(x)\theta_d(x)$. From the capacity analysis, we know that (see (126))

$$\mathbb{E}\left[\log P(x|Y^d)\Big| X = x\right] = -\max_{x' \neq x} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{229}$$

From our previous analysis, we know that

$$\mathsf{H}(X|Y^d) = \max_{x,x':x' \neq x} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{230}$$

Hence,

$$\theta_d = e^{\Theta(\log d)} \cdot \theta_d(x) = \max_{x,x':x \neq x'} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right). \tag{231}$$

$\square$

With these lemmas in place, the proof of Theorem is easily completed.

*Proof of Thm. III.1 for dispersion.* From (204) we see that

$$V^{(d)} - \mathsf{V}(X) = \mathbb{E}\left[\left(\log P(X|Y^d)\right)^2\right] - \mathsf{H}(X|Y^d)^2 + \theta_d. \tag{232}$$

Putting together Lemmas VII.3 and VII.4 and the fact that

$$\mathsf{H}(X|Y^d) = \max_{x,x':x' \neq x} \exp\left(-d \cdot \mathsf{C}(P_{Y|x}, P_{Y|\tilde{x}}) + \Theta(\log d|\mathcal{X}|)\right) \tag{233}$$

(from the proof of Thm. III.1 for mutual information), we readily obtain the required result. $\square$

## VIII. Conclusion and Future Work

In this paper, we considered the behaviour of the information rate and channel dispersion of multi-view noisy channels, with arbitrary, but fixed, input distributions, in the regime where the number of views is very large. In the first part of the paper, we considered multi-view discrete memoryless channels (DMCs), and showed that the information rate and the channel dispersion converge exponentially quickly to the input entropy and varentropy, respectively, with an input distribution-independent rate that equals the smallest Chernoff information between two conditional distributions of the output given different inputs. We then extended these results to general multi-letter channels, and computed explicit upper bounds on the Chernoff information above for the deletion channel. Finally, for the special case of the multi-view binary symmetric channel, we also presented a close, non-asymptotic approximation to its capacity, via the capacity of the so-called Poisson approximation channel.

One interesting direction for future work would be the extension of the techniques and results in this paper to other multi-view channels with memory and general synchronization channels (see [33] and [3, Sec. VII]).

## References

[1] V. Levenshtein, "Efficient reconstruction of sequences," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 2–22, 2001.

[2] C. Shannon, R. Gallager, and E. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels. I," *Information and Control*, vol. 10, no. 1, pp. 65–103, 1967. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019995867900526

[3] M. Cheraghchi and J. Ribeiro, "An overview of capacity results for synchronization channels," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3207–3232, 2021.

[4] R. Heckel, G. Mikutis, and R. N. Grass, "A characterization of the DNA data storage channel," *Scientific Reports*, vol. 9, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4743189

[5] I. Shomorony and R. Heckel, "Information-theoretic foundations of DNA data storage," *Foundations and Trends® in Communications and Information Theory*, vol. 19, no. 1, pp. 1–106, 2022. [Online]. Available: http://dx.doi.org/10.1561/0100000117

[6] N. Weinberger, "Error probability bounds for coded-index DNA storage systems," *IEEE Transactions on Information Theory*, vol. 68, no. 11, pp. 7005–7022, 2022.

[7] N. Weinberger and N. Merhav, "The DNA storage channel: Capacity and error probability bounds," *IEEE Transactions on Information Theory*, vol. 68, no. 9, pp. 5657–5700, 2022.

[8] M. Kovačević and V. Y. F. Tan, "Codes in the space of multisets—coding for permutation channels with impairments," *IEEE Transactions on Information Theory*, vol. 64, no. 7, pp. 5156–5169, 2018.

[9] A. Lenz, P. H. Siegel, A. Wachter-Zeh, and E. Yaakobi, "Coding over sets for DNA storage," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2331–2351, 2020.

[10] R. Gabrys, H. M. Kiah, and O. Milenkovic, "Asymmetric Lee distance codes: New bounds and constructions," in *2015 IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5.

[11] M. Mitzenmacher, "On the theory and practice of data recovery with multiple versions," in *2006 IEEE International Symposium on Information Theory*, 2006, pp. 982–986.

[12] I. Land and J. Huber, "Information combining," *Foundations and Trends® in Communications and Information Theory*, vol. 3, no. 3, pp. 227–330, 2006. [Online]. Available: http://dx.doi.org/10.1561/0100000013

[13] I. Sutskover, S. Shamai, and J. Ziv, "Extremes of information combining," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1313–1325, 2005.

[14] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '04.  USA: Society for Industrial and Applied Mathematics, 2004, p. 910–918.

[15] M. Cheraghchi, R. Gabrys, O. Milenkovic, and J. Ribeiro, "Coded trace reconstruction," *IEEE Transactions on Information Theory*, vol. 66, no. 10, pp. 6084–6103, 2020.

[16] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, 2017, pp. 228–239.

[17] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed.  Wiley-India, 2010.

[18] C. E. Shannon, "A mathematical theory of communication," *Bell Systems Technical Journal*, vol. 27, no. 3, pp. 379–423, July 1948.

[19] R. G. Gallager, *Stochastic Processes: Theory for Applications*.  Cambridge: Cambridge University Press, 2013.

[20] T. Richardson and R. Urbanke, *Modern Coding Theory*.  Cambridge University Press, 2008.

[21] A. Guillen i Fabregas, I. Land, and A. Martinez, "Extremes of error exponents," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2201–2207, 2013.

[22] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[23] V. Y. F. Tan, "Asymptotic estimates in information theory with non-vanishing error probabilities," *Foundations and Trends® in Communications and Information Theory*, vol. 11, no. 1-2, pp. 1–184, 2014. [Online]. Available: http://dx.doi.org/10.1561/0100000086

[24] Y. Polyanskiy and Y. Wu, *Information Theory: From Coding to Learning*.  Cambridge University Press, 2023+ (*book draft*).

[25] I. Land, S. Huettinger, P. Hoeher, and J. Huber, "Bounds on information combining," *IEEE Transactions on Information Theory*, vol. 51, no. 2, pp. 612–619, 2005.

[26] E. Şaşoğlu, "Polarization and polar codes," *Foundations and Trends® in Communications and Information Theory*, vol. 8, no. 4, pp. 259–381, 2012. [Online]. Available: http://dx.doi.org/10.1561/0100000041

[27] N. Merhav, "Statistical physics and information theory," *Foundations and Trends® in Communications and Information Theory*, vol. 6, no. 1–2, pp. 1–212, 2010. [Online]. Available: http://dx.doi.org/10.1561/0100000052

[28] Z. Zhou and P. Li, "Non-asymptotic Chernoff lower bound and its application to community detection in stochastic block model," *arXiv e-prints*, p. arXiv:1812.11269, Dec. 2018.

[29] W. Koepf, *Hypergeometric Summation An Algorithmic Approach to Summation and Special Function Identities*.  Vieweg, 1998.

[30] J. Spencer, *Asymptopia*.  American Mathematical Society, 2014.

[31] R. D. Wesel, E. E. Wesel, L. Vandenberghe, C. Komninakis, and M. Medard, "Efficient binomial channel capacity computation with an application to molecular communication," in *2018 Information Theory and Applications Workshop (ITA)*, 2018, pp. 1–5.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*.  Cambridge: Cambridge University Press, 2004.

[33] B. Haeupler and M. Mitzenmacher, "Repeated deletion channels," in *2014 IEEE Information Theory Workshop (ITW 2014)*, 2014, pp. 152–156.