ISIT'2013 1569714399

# Refined Bounds on the Empirical Distribution of Good Channel Codes via Concentration Inequalities

Maxim Raginsky
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA.
E-mail: maxim@illinois.edu

Igal Sason
Department of Electrical Engineering
Technion-Israel Institute of Technology
Haifa 32000, Israel.
E-mail: sason@ee.technion.ac.il

***Abstract*—We derive sharpened inequalities on the empirical output distribution of good channel codes with deterministic encoders and with non-vanishing maximal probability of decoding error. These inequalities refine recent bounds of Polyanskiy and Verdú by identifying closed-form expressions for certain asymptotic terms, which facilitates their calculation for finite blocklengths. The analysis relies on concentration-of-measure inequalities, specifically on McDiarmid's method of bounded differences and its close ties to transportation inequalities for weighted Hamming metrics. An operational implication of the new bounds is addressed.**

## I. Introduction

The importance of sharp concentration-of-measure inequalities for characterizing fundamental limits of coding schemes in information theory is evident from the recent flurry of activity on *finite-blocklength* analysis of source and channel codes (cf., e.g., [6], [9], [10], [12]). Theory and applications of concentration of measure in information theory, communications and coding have been recently surveyed in [11].

A recent application of concentration-of-measure inequalities to information theory has to do with characterizing stochastic behavior of output sequences of good channel codes. For capacity-achieving sequences of codes with asymptotically vanishing probability of error, Shamai and Verdú proved the following remarkable statement [14, Theorem 2]: given a DMC $T : \mathcal{X} \to \mathcal{Y}$, any capacity-achieving sequence of channel codes with asymptotically vanishing probability of error (maximal or average) has the property that

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) = 0 \tag{1}$$

where, for each $n$, $P_{Y^n}$ denotes the output distribution on $\mathcal{Y}^n$ induced by the code (assuming that the messages are equiprobable), while $P^*_{Y^n}$ is the product of $n$ copies of the single-letter capacity-achieving output distribution (see below for a more detailed exposition). In fact, the convergence in (1) holds not just for DMC's, but for arbitrary channels satisfying the condition $C = \lim_{n\to\infty} \frac{1}{n} \sup_{P_{X^n} \in \mathcal{P}(\mathcal{X}^n)} I(X^n; Y^n)$. In a recent preprint [10], Polyanskiy and Verdú extended the results of [14] for codes with *nonvanishing* probability of error, provided the maximal probability of error criterion and deterministic encoders are used.

In this paper, we present some refinements of the results from [10] in the context of the material covered in [11]. To keep things simple, we will only focus on channels with finite input and output alphabets. Thus, let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, and consider a DMC $T : \mathcal{X} \to \mathcal{Y}$ with capacity $C = \max_{P_X \in \mathcal{P}(\mathcal{X})} I(X;Y)$ (all information quantities in this paper are measured in nats). Let $P^*_X \in \mathcal{P}(\mathcal{X})$ be any capacity-achieving input distribution (there may be several). It can be shown [15] that the corresponding output distribution $P^*_Y \in \mathcal{P}(\mathcal{Y})$ is unique, and that for any $n \in \mathbb{N}$, the product distribution $P^*_{Y^n} \equiv (P^*_Y)^{\otimes n}$ has the key property

$$D(P_{Y^n|X^n=x^n} \| P^*_{Y^n}) \le nC, \qquad \forall x^n \in \mathcal{X}^n \tag{2}$$

where $P_{Y^n|X^n=x^n}$ denotes the product distribution $T^n(\cdot|x^n)$. A key consequence of (2) is that, for any input distribution $P_{X^n}$, the corresponding output distribution $P_{Y^n}$ satisfies

$$D(P_{Y^n} \| P^*_{Y^n}) \le nC - I(X^n; Y^n).$$

Given $n, M \in \mathbb{N}$, an $(n, M)$-*code* is a pair $\mathcal{C} = (f_n, g_n)$ consisting of an *encoder* $f_n : \{1, \ldots, M\} \to \mathcal{X}^n$ and a *decoder* $g_n : \mathcal{Y}^n \to \{1, \ldots, M\}$. Given $0 < \varepsilon \le 1$, $\mathcal{C}$ is an $(n, M, \varepsilon)$-*code* if $\max_{1\le i\le M} \mathbb{P}\big(g_n(Y^n) \ne i \big| X^n = f_n(i)\big) \le \varepsilon$.

Consider any $(n, M)$-code $\mathcal{C} = (f_n, g_n)$ for $T$, and let $J$ be a random variable uniformly distributed on $\{1, \ldots, M\}$. We can think of any $1 \le i \le M$ as one of $M$ equiprobable messages to be transmitted over $T$. Let $P^{(\mathcal{C})}_{X^n}$ denote the distribution of $X^n = f_n(J)$, and let $P^{(\mathcal{C})}_{Y^n}$ denote the corresponding output distribution. The central result of [10] is that the output distribution $P^{(\mathcal{C})}_{Y^n}$ of any $(n, M, \varepsilon)$-code satisfies

$$D\big(P^{(\mathcal{C})}_{Y^n} \big\| P^*_{Y^n}\big) \le nC - \ln M + o(n); \tag{3}$$

moreover, the $o(n)$ term was refined in [10, Theorem 5] to $O(\sqrt{n})$ for any DMC, except those that have zeroes in their transition matrix. In one the following results, we present a sharpened bound with a modified proof, in which we specify an explicit form for the term that scales like $O(\sqrt{n})$. For the case where there are zeroes in the transition matrix of the DMC, an upper bound on the relative entropy is derived with an explicit closed-form expression for the $o(n)$ term in (3), which scales like $O\big(\sqrt{n}\,(\ln n)^{3/2}\big)$; this forms a sharpened version of [10, Theorem 6].

## II. Preliminaries

### A. Concentration inequalities

Let $U$ be a random variable taking values in some space $\mathcal{U}$, and fix a function $f : \mathcal{U} \to \mathbb{R}$ with $\mathbb{E}[f(U)] = 0$. We are interested in tight upper bounds on the *deviation probabilities* $\mathbb{P}(f(U) \geq r)$ for $r \geq 0$. We will need the following facts (we refer the reader to [11] for details): (1) If $f$ is such that $\ln \mathbb{E}\left[\exp(tf(U))\right] \leq \kappa t^2/2, \forall t \geq 0$ with some $\kappa > 0$, then

$$\mathbb{P}\left(f(U) \geq r\right) \leq \exp\left(-r^2/2\kappa\right), \quad \forall\, r \geq 0. \tag{4}$$

(2) Suppose the space $\mathcal{U}$ is equipped with a metric $d$. The $L^1$ *Wasserstein distance* between $\mu, \nu \in \mathcal{P}(\mathcal{U})$ is defined as

$$W_1(\mu, \nu) \triangleq \inf_{U \sim \mu, V \sim \nu} \mathbb{E}[d(U, V)].$$

We say that $\mu$ satisfies an $L^1$ *transportation cost inequality* with constant $c > 0$, or a $T_1(c)$ inequality for short, if

$$W_1(\mu, \nu) \leq \sqrt{2c\, D(\nu \| \mu)}, \qquad \forall \nu \in \mathcal{P}(\mathcal{U}) \tag{5}$$

The *Lipschitz constant* of $f : \mathcal{U} \to \mathbb{R}$ is defined by

$$\|f\|_{\mathrm{Lip}} \triangleq \sup_{u \neq v} \frac{|f(u) - f(v)|}{d(u, v)}.$$

A key result due to Bobkov and Götze ([4], see also [11, Theorem 36]) says that $\mu$ satisfies $T_1(c)$ if and only if the bound $\ln \mathbb{E}\left[\exp\left(tf(U)\right)\right] \leq ct^2/2$ holds for all $t \geq 0$ and for all $f : \mathcal{U} \to \mathbb{R}$ with $\mathbb{E}[f(U)] = 0$ and $\|f\|_{\mathrm{Lip}} \leq 1$. In that case, the bound (4) holds for any $f : \mathcal{U} \to \mathbb{R}$ with $\kappa = c\|f\|^2_{\mathrm{Lip}}$.

We are interested in the case when $\mathcal{U}$ is a product space $\mathcal{Z}^n$ with the *weighted Hamming metric*

$$d(z^n, z'^n) = \sum_{i=1}^{n} c_i 1_{\{z_i \neq z'_i\}}, \tag{6}$$

for some fixed $c_1, \ldots, c_n > 0$. A function $f : \mathcal{Z}^n \to \mathbb{R}$ has $\|f\|_{\mathrm{Lip}} \leq 1$ if and only if it has *bounded differences*, i.e.,

$$|f(z_1^{i-1}, z_i, z_{i+1}^n) - f(z_1^{i-1}, z'_i, z_{i+1}^n)| \leq c_i \tag{7}$$

for all $i \in \{1, \ldots, n\}$, $z^n \in \mathcal{Z}^n$, $z'_i \in \mathcal{Z}$. Let $U = (Z_1, \ldots, Z_n)$ be a tuple of independent $\mathcal{Z}$-valued random variables. Then, the conditions of the Bobkov–Götze theorem are met with $c = (1/4)\sum_{i=1}^n c_i^2$ [11, Theorem 28]. Therefore, for any function $f$ that satisfies (7) we have the inequality

$$\mathbb{P}\left(f(Z^n) - \mathbb{E}[f(Z^n)] \geq r\right) \leq e^{-\frac{2r^2}{\sum_{i=1}^n c_i^2}}, \quad \forall\, r \geq 0 \tag{8}$$

originally obtained by McDiarmid [7], [8] using martingales.

The well-known "blowing up lemma" (see, e.g., [5, Lemma 1.5.4] or [11, Section 3.6.1]) can be derived from (8) as follows. Let $c_1 = \ldots = c_n = 1$. Fix an arbitrary set $A \subset \mathcal{Z}^n$ and consider the function

$$f(z^n) = d(z^n, A) \triangleq \min_{z'^n \in A} d(z^n, z'^n).$$

Then $\|f\|_{\mathrm{Lip}} \leq 1$. Applying (8) to $f$ and $r = \mathbb{E}[f(Z^n)]$,

$$P_{Z^n}(A) = \mathbb{P}\left(d(Z^n, A) \leq 0\right) \leq \exp\left[-\frac{2\left(\mathbb{E}[d(Z^n, A)]\right)^2}{n}\right],$$

which gives $\mathbb{E}[d(Z^n, A)] \leq \sqrt{\frac{n}{2}\ln\frac{1}{P_{Z^n}(A)}}$. This and (8) give

$$P_{Z^n}([A]_r) \geq 1 - \exp\left[-\frac{2}{n}\left(r - \sqrt{\frac{n}{2}\ln\frac{1}{P_{Z^n}(A)}}\right)^2\right] \tag{9}$$

for all $r \geq \sqrt{\frac{n}{2}\ln\frac{1}{P_{Z^n}(A)}}$, where $[A]_r \triangleq \{z^n : d(z^n, A) \leq r\}$ denotes the *$r$-blowup* of $A$.

### B. Augustin's strong converse

Just as in [10], the proof of (3) with the $O(\sqrt{n})$ term uses the following strong converse for channel codes due to Augustin [3] (see also [10, Theorem 1] and [2, Section 2]):

*Theorem 1 (Augustin):* Let $S : \mathcal{U} \to \mathcal{V}$ be a DMC, and let $P_{V|U}$ be the transition probability induced by $S$. For any $M \in \mathbb{N}$ and $\varepsilon \in (0, 1]$, let $f : \{1, \ldots, M\} \to \mathcal{U}$ and $g : \mathcal{V} \to \{1, \ldots, M\}$ be two mappings, such that

$$\max_{1 \leq i \leq M} \mathbb{P}\big(g(V) \neq i \big| U = f(i)\big) \leq \varepsilon.$$

Let $Q_V \in \mathcal{P}(\mathcal{V})$ be an auxiliary output distribution, and fix an arbitrary mapping $\gamma : \mathcal{U} \to \mathbb{R}$. Then

$$M \leq \frac{\exp\big\{\mathbb{E}[\gamma(U)]\big\}}{\inf\limits_{u \in \mathcal{U}} P_{V|U=u}\left(\ln\dfrac{\mathrm{d}P_{V|U=u}}{\mathrm{d}Q_V} < \gamma(u)\right) - \varepsilon}, \tag{10}$$

provided the denominator is strictly positive. The expectation in the numerator is taken w.r.t. the distribution of $U = f(J)$ with $J \sim \mathrm{Uniform}\{1, \ldots, M\}$.

## III. Refined Bounds on the Empirical Output Distribution

In this section, we first establish the bound (3) for the case when the DMC $T$ is such that

$$C_1 \triangleq \max_{x, x' \in \mathcal{X}} D(P_{Y|X=x} \| P_{Y|X=x'}) < \infty. \tag{11}$$

Note that $C_1 < \infty$ if and only if the transition matrix of $T$ does not have any zeroes. Consequently,

$$c(T) \triangleq 2 \max_{x, x' \in \mathcal{X}} \max_{y, y' \in \mathcal{Y}} \left|\ln \frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')}\right| < \infty. \tag{12}$$

We establish the following sharpened version of the bound in [10, Theorem 5]:

*Theorem 2:* Let $T : \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ satisfying (11). Then, any $(n, M, \varepsilon)$-code $\mathcal{C}$ for $T$ with $0 < \varepsilon < 1/2$ satisfies

$$D\big(P_{Y^n}^{(\mathcal{C})} \big\| P_{Y^n}^*\big) \leq nC - \ln M + \ln\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}}. \tag{13}$$

*Remark 1:* As shown in [10], the restriction to codes with deterministic encoders and to the maximal probability of error criterion is necessary for both this theorem and the next one.

*Proof:* Fix an input sequence $x^n \in \mathcal{X}^n$ and consider the function $h_{x^n} : \mathcal{Y}^n \to \mathbb{R}$ defined by

$$h_{x^n}(y^n) \quad \triangleq \ln \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(y^n).$$

Then $\mathbb{E}[h_{x^n}(Y^n)|X^n=x^n] = D(P_{Y^n|X^n=x^n}\|P_{Y^n}^{(\mathcal{C})})$. For any $i\in\{1,\ldots,n\}$ and $y^n\in\mathcal{Y}^n$, let $\overline{y}^i\in\mathcal{Y}^{n-1}$ denote the $(n-1)$-tuple obtained by deleting the $i$th coordinate from $y^n$. Then, for any $i\in\{1,\ldots,n\}$, $y^n\in\mathcal{Y}^n$ and $y_i'\in\mathcal{Y}$, we have

$$\begin{aligned}
&\left|h_{x^n}(y_1^{i-1},y_i,y_{i+1}^n) - h_{x^n}(y_1^{i-1},y_i',y_{i+1}^n)\right| \\
&\le \left|\ln P_{Y^n|X^n=x^n}(y^{i-1},y,y_{i+1}^n) - \ln P_{Y^n|X^n=x^n}(y^{i-1},y',y_{i+1}^n)\right| \\
&\quad + \left|\ln P_{Y^n}^{(\mathcal{C})}(y^{i-1},y,y_{i+1}^n) - \ln P_{Y^n}^{(\mathcal{C})}(y^{i-1},y',y_{i+1}^n)\right| \\
&\le \left|\ln\frac{P_{Y_i|X_i=x_i}(y)}{P_{Y_i|X_i=x_i}(y')}\right| + \left|\ln\frac{P_{Y_i|\overline{Y}^i}^{(\mathcal{C})}(y|\overline{y}^i)}{P_{Y_i|\overline{Y}^i}^{(\mathcal{C})}(y'|\overline{y}^i)}\right| \\
&\le 2\max_{x,x'\in\mathcal{X}}\max_{y,y'\in\mathcal{Y}}\left|\ln\frac{P_{Y|X}(y|x)}{P_{Y|X}(y'|x')}\right| \qquad (14)\\
&= c(T) < \infty \qquad (15)
\end{aligned}$$

where the inequality in (14) is proved in [11, Appendix 3.D]. Hence, for each $x^n\in\mathcal{X}^n$, the function $h_{x^n}:\mathcal{Y}^n\to\mathbb{R}$ satisfies the bounded differences condition (7) with $c_1,\ldots,c_n = c(T)$. Then, from (8), it follows that for any $r\ge 0$,

$$P_{Y^n|X^n=x^n}\left(\ln\frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(Y^n) \ge D(P_{Y^n|X^n=x^n}\|P_{Y^n}^{(\mathcal{C})}) + r\right) \le e^{-\frac{2r^2}{nc^2(T)}} \quad (16)$$

(In fact, the above derivation goes through for any possible output distribution $P_{Y^n}$, not necessarily one induced by a code.) This is where we have departed from the original proof by Polyanskiy and Verdú [10]: we have used McDiarmid's inequality (8) to control the deviation probability for the "conditional" information density $h_{x^n}$ directly, whereas they bounded the *variance* of $h_{x^n}$ and then derived a bound on the deviation probability using Chebyshev's inequality. The sharp concentration inequality (16) allows us to explicitly identify the constant multiplying $\sqrt{n}$ in (13).

We are now in a position to apply Theorem 1. To that end, we let $\mathcal{U}=\mathcal{X}^n$, $\mathcal{V}=\mathcal{Y}^n$, and consider the DMC $S=T^n$ with an $(n,M,\varepsilon)$-code $(f,g)=(f_n,g_n)$. Furthermore, let

$$\zeta_n = \zeta_n(\varepsilon) \triangleq c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}} \qquad (17)$$

and take $\gamma(x^n) = D(P_{Y^n|X^n=x^n}\|P_{Y^n}^{(\mathcal{C})})+\zeta_n$. Using (10) with the auxiliary distribution $Q_V = P_{Y^n}^{(\mathcal{C})}$, we get

$$M \le \frac{\exp\{\mathbb{E}[\gamma(X^n)]\}}{\inf_{x^n\in\mathcal{X}^n} P_{Y^n|X^n=x^n}\left(\ln\frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}} < \gamma(x^n)\right) - \varepsilon} \qquad (18)$$

where $\mathbb{E}[\gamma(X^n)] = D\big(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})}\,|\,P_{X^n}^{(\mathcal{C})}\big)+\zeta_n$. The concentration inequality in (16) with $\zeta_n$ in (17) therefore gives

$$P_{Y^n|X^n=x^n}\left(\ln\frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}} \ge \gamma(x^n)\right) \le 1-2\varepsilon,\ \forall x^n.$$

From this and (18) it follows that

$$M \le \frac{1}{\varepsilon}\,\exp\Big(D\big(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})}\,|\,P_{X^n}^{(\mathcal{C})}\big)+\zeta_n\Big).$$

Taking logarithms on both sides of the last inequality, rearranging terms, and using (17), we get

$$\begin{aligned}
D(P_{Y^n|X^n}\|P_{Y^n}^{(\mathcal{C})}\,|\,P_{X^n}^{(\mathcal{C})}) &\ge \ln M + \ln\varepsilon - \zeta_n \\
&= \ln M + \ln\varepsilon - c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}}. \qquad (19)
\end{aligned}$$

We are now ready to derive (13):

$$\begin{aligned}
&D\big(P_{Y^n}^{(\mathcal{C})}\big\|P_{Y^n}^*\big) \\
&= D\big(P_{Y^n|X^n}\big\|P_{Y^n}^*\big|P_{X^n}^{(\mathcal{C})}\big) - D\big(P_{Y^n|X^n}\big\|P_{Y^n}^{(\mathcal{C})}\big|P_{X^n}^{(\mathcal{C})}\big) \qquad (20)\\
&\le nC - \ln M + \ln\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\ln\frac{1}{1-2\varepsilon}} \qquad (21)
\end{aligned}$$

where (20) uses the chain rule for divergence, while (21) uses (2) and (19). This completes the proof of Theorem 2. ∎

For an arbitrary DMC $T$ with nonzero capacity and zeroes in its transition matrix, we have the following result which forms a sharpened version of the bound in [10, Theorem 6]:

*Theorem 3:* Let $T:\mathcal{X}\to\mathcal{Y}$ be a DMC with $C>0$. Then, for any $0<\varepsilon<1$, any $(n,M,\varepsilon)$-code $\mathcal{C}$ for $T$ satisfies

$$\begin{aligned}
&D\big(P_{Y^n}^{(\mathcal{C})}\big\|P_{Y^n}^*\big) \le nC - \ln M \\
&+\sqrt{2n}\,(\ln n)^{3/2}\left(1+\sqrt{\frac{1}{\ln n}\,\ln\left(\frac{1}{1-\varepsilon}\right)}\right)\left(1+\frac{\ln|\mathcal{Y}|}{\ln n}\right) \\
&+3\ln n + \ln\big(2|\mathcal{X}||\mathcal{Y}|^2\big). \qquad (22)
\end{aligned}$$

*Proof:* Given an $(n,M,\varepsilon)$-code $\mathcal{C}=(f_n,g_n)$, let $c_1,\ldots,c_M\in\mathcal{X}^n$ be its codewords, and let $\widetilde{D}_1,\ldots,\widetilde{D}_M\subset\mathcal{Y}^n$ be the corresponding decoding regions:

$$\widetilde{D}_i = g_n^{-1}(i) \equiv \{y^n\in\mathcal{Y}^n : g_n(y^n)=i\},\quad i=1,\ldots,M.$$

If we choose

$$\delta_n = \delta_n(\varepsilon) = \frac{1}{n}\left\lceil n\left(\sqrt{\frac{\ln n}{2n}}+\sqrt{\frac{1}{2n}\ln\frac{1}{1-\varepsilon}}\right)\right\rceil \qquad (23)$$

(note that $n\delta_n$ is an integer), then by (9) the "blown-up" decoding regions $D_i \triangleq \left[\widetilde{D}_i\right]_{n\delta_n}$ satisfy

$$P_{Y^n|X^n=c_i}(D_i^c) \le 1/n\,, \qquad \forall\, i\in\{1,\ldots,M\}. \quad (24)$$

We now complete the proof by a random coding argument along the lines of [1]. For

$$N \triangleq \left\lceil \frac{M}{n\binom{n}{n\delta_n}|\mathcal{Y}|^{n\delta_n}} \right\rceil, \qquad (25)$$

let $U_1,\ldots,U_N$ be independent random variables, each uniformly distributed on the set $\{1,\ldots,M\}$. For each realization $V=U^N$, let $P_{X^n(V)}\in\mathcal{P}(\mathcal{X}^n)$ denote the induced distribution of $X^n(V)=f_n(c_J)$, where $J$ is uniformly

distributed on the set $\{U_1, \ldots, U_N\}$, and let $P_{Y^n(V)}$ denote the corresponding output distribution of $Y^n(V)$:

$$P_{Y^n(V)} = \frac{1}{N} \sum_{i=1}^{N} P_{Y^n|X^n=c_{U_i}}. \tag{26}$$

It is easy to show that $\mathbb{E}\left[P_{Y^n(V)}\right] = P_{Y^n}^{(\mathcal{C})}$, the output distribution of the original code $\mathcal{C}$, where the expectation is w.r.t. the distribution of $V = U^N$. Now, for $V = U^N$ and for every $y^n \in \mathcal{Y}^n$, let $\mathcal{N}_V(y^n)$ denote the list of all those indices in $\{U_1, \ldots, U_N\}$ such that $y^n \in D_{U_j}$, so $\mathcal{N}_V(y^n) \triangleq \left\{j : y^n \in D_{U_j}\right\}$. Consider the list decoder $Y^n \mapsto \mathcal{N}_V(Y^n)$, and let $\varepsilon(V)$ denote its conditional decoding error probability: $\varepsilon(V) \triangleq P(J \notin \mathcal{N}_V(Y^n)|V)$. Then, for each realization of $V$,

$$\begin{aligned}
&D\big(P_{Y^n(V)}\big\|P_{Y^n}^*\big) \\
&= D\big(P_{Y^n(V)|X^n(V)}\big\|P_{Y^n}^*\big|P_{X^n(V)}\big) - I(X^n(V); Y^n(V)) && (27)\\
&\le nC - I(X^n(V); Y^n(V)) && (28)\\
&\le nC - I(J; Y^n(V)) && (29)\\
&\le nC - \ln N + (1 - \varepsilon(V))\,\mathbb{E}[\ln|\mathcal{N}_V(Y^n)|] \\
&\quad + n\varepsilon(V)\ln|\mathcal{X}| + \ln 2 && (30)
\end{aligned}$$

where:

- (27) is by the chain rule for divergence;
- (28) is by (2);
- (29) is by the data processing inequality and the fact that $J \to X^n(V) \to Y^n(V)$ is a Markov chain; and
- (30) is by Fano's inequality for list decoding (see [11, Appendix 3.C]), and also since (i) $N \le |\mathcal{X}|^n$, (ii) $J$ is uniformly distributed on $\{U_1, \ldots, U_N\}$, so $H(J|U_1, \ldots, U_N) = \ln N$ and $H(J) \ge \ln N$.

(Note that all the quantities indexed by $V$ in the above chain of estimates are actually random variables, since they depend on the realization $V = U^N$.) Now, from (25), it follows that

$$\begin{aligned}
\ln N &\ge \ln M - \ln n - \ln\binom{n}{n\delta_n} - n\delta_n \ln|\mathcal{Y}| \\
&\ge \ln M - \ln n - n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right) && (31)
\end{aligned}$$

where the last inequality uses the simple inequality[1] $\binom{n}{k} \le n^k$ for $k \le n$ with $k \triangleq n\delta_n$. Moreover, each $y^n \in \mathcal{Y}^n$ can belong to at most $\binom{n}{n\delta_n}|\mathcal{Y}|^{n\delta_n}$ blown-up decoding sets, so

$$\ln|\mathcal{N}_V(Y^n = y^n)| \le n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right), \quad \forall\, y^n \in \mathcal{Y}^n. \tag{32}$$

Substituting (31) and (32) into (30), we get

$$\begin{aligned}
D\big(P_{Y^n(V)}\big\|P_{Y^n}^*\big) &\le nC - \ln M + \ln n \\
&\quad + 2n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right) + n\varepsilon(V)\ln|\mathcal{X}| + \ln 2. && (33)
\end{aligned}$$

Using the fact that $\mathbb{E}\left[P_{Y^n(V)}\right] = P_{Y^n}^{(\mathcal{C})}$, convexity of the relative entropy, and (33), we get

$$\begin{aligned}
D\big(P_{Y^n}^{(\mathcal{C})}\big\|P_{Y^n}^*\big) &\le nC - \ln M + \ln n + 2n\delta_n\left(\ln n + \ln|\mathcal{Y}|\right) \\
&\quad + n\,\mathbb{E}\left[\varepsilon(V)\right]\ln|\mathcal{X}| + \ln 2. && (34)
\end{aligned}$$

[1]Note that the gain in using instead the inequality $\binom{n}{n\delta_n} \le \exp\big(n\,h(\delta_n)\big)$ is marginal, and it does not have any advantage asymptotically for large $n$.

To finish the proof and get (22), we use the fact that

$$\mathbb{E}\left[\varepsilon(V)\right] \le \max_{1\le i\le M} P_{Y^n|X^n=c_i}\left(D_i^c\right) \le \frac{1}{n},$$

which follows from (24), as well as the substitution of (23) in (34). This completes the proof of Theorem 3. ∎

We are now ready to examine some consequences of Theorems 2 and 3. To start with, consider a sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$, where each $\mathcal{C}_n = (f_n, g_n)$ is an $(n, M_n, \varepsilon)$-code for a DMC $T : \mathcal{X} \to \mathcal{Y}$ with $C > 0$. We say that $\{\mathcal{C}_n\}_{n=1}^{\infty}$ is *capacity-achieving* if $\lim_{n\to\infty} \frac{1}{n}\ln M_n = C$. Then, from Theorems 2 and 3, it follows that any such sequence satisfies (1) with $P_{Y^n} = P_{Y^n}^{(\mathcal{C}_n)}$ for all $n$. This is discussed in detail in [10].

Another remarkable fact that follows from the above theorems is that a broad class of functions evaluated on the output of a good code concentrate sharply around their expectations with respect to the capacity-achieving output distribution. Specifically, we have the following version of [10, Proposition 10] (again, we have streamlined the statement and the proof to relate them to the material in Section II-A):

*Theorem 4:* Let $T : \mathcal{X} \to \mathcal{Y}$ be a DMC with $C > 0$ and $C_1 < \infty$. Let $d : \mathcal{Y}^n \times \mathcal{Y}^n \to \mathbb{R}_+$ be a metric, and suppose that there exists a constant $c > 0$, such that the conditional probability distributions $P_{Y^n|X^n=x^n}$, $x^n \in \mathcal{X}^n$, as well as $P_{Y^n}^*$ satisfy the $T_1(c)$ inequality on the metric space $(\mathcal{Y}^n, d)$. Then, for any $\varepsilon \in (0, 1/2)$, any $(n, M, \varepsilon)$-code $\mathcal{C}$ for $T$, and any function $f : \mathcal{Y}^n \to \mathbb{R}$ we have

$$\begin{aligned}
&P_{Y^n}^{(\mathcal{C})}\Big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r\Big) \\
&\le \frac{4}{\varepsilon}\,\exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|_{\mathrm{Lip}}^2}\right),\ \forall\, r \ge 0 && (35)
\end{aligned}$$

where $\mathbb{E}[f(Y^{*n})]$ designates the expected value of $f(Y^n)$ w.r.t. the capacity-achieving output distribution $P_{Y^n}^*$, $\|f\|_{\mathrm{Lip}}$ is the Lipschitz constant of $f$ w.r.t. the metric $d$, and

$$a \triangleq c(T)\sqrt{\frac{1}{2}\ln\frac{1}{1-2\varepsilon}}. \tag{36}$$

*Remark 2:* Our sharpening of the bound in [10, Proposition 10] gives an explicit constant in front of $\sqrt{n}$ in the bound (35).

*Proof:* For any $f$, define

$$\mu_f^* \triangleq \mathbb{E}[f(Y^{*n})],\ \phi(x^n) \triangleq \mathbb{E}[f(Y^n)|X^n = x^n],\ \forall\, x^n \in \mathcal{X}^n.$$

Since each $P_{Y^n|X^n=x^n}$ satisfies $T_1(c)$, by the Bobkov–Götze theorem (see Section II-A) we have

$$\mathbb{P}\Big(|f(Y^n) - \phi(x^n)| \ge r\Big|X^n = x^n\Big) \le 2e^{-\frac{r^2}{2c\|f\|_{\mathrm{Lip}}^2}}, \tag{37}$$

for all $r \ge 0$. Now, given $\mathcal{C}$, consider a subcode $\mathcal{C}'$ with codewords $x^n \in \mathcal{X}^n$ satisfying $\phi(x^n) \ge \mu_f^* + r$ for $r \ge 0$. The number of codewords $M'$ of $\mathcal{C}'$ satisfies

$$M' = MP_{X^n}^{(\mathcal{C})}\left(\phi(X^n) \ge \mu_f^* + r\right). \tag{38}$$

Let $Q = P_{Y^n}^{(\mathcal{C}')}$ be the output distribution induced by $\mathcal{C}'$. Then

$$\mu_f^* + r \le \frac{1}{M'} \sum_{x^n \in \text{ codewords}(\mathcal{C}')} \phi(x^n) \tag{39}$$

$$= \mathbb{E}_Q[f(Y^n)] \tag{40}$$

$$\le \mathbb{E}[f(Y^{*n})] + \|f\|_{\text{Lip}} \sqrt{2cD(Q_{Y^n} \| P_{Y^n}^*)} \tag{41}$$

$$\le \mu_f^* + \|f\|_{\text{Lip}} \sqrt{2c\left(nC - \ln M' + a\sqrt{n} + \ln\frac{1}{\varepsilon}\right)}, \tag{42}$$

where:

- (39) and (40) are by definition of $\mathcal{C}'$ and $\phi$;
- (41) follows from the fact that $P_{Y^n}^*$ satisfies $T_1(c)$ and from the Kantorovich–Rubinstein theorem (see, e.g., [16, Theorem 1.14] or [11, Section 3.4.3]); and
- (42) holds with $a = a(T, \varepsilon) > 0$ in (36) due to Theorem 2 (see (13)) and because $\mathcal{C}'$ is an $(n, M', \varepsilon)$-code for $T$.

From this and (38), and then using the same line of reasoning with $-f$ instead of $f$, it follows that

$$P_{X^n}^{(\mathcal{C})}\Big(\big|\phi(X^n) - \mu_f^*\big| \ge r\Big) \le 2e^{nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{2c\|f\|_{\text{Lip}}^2}}. \tag{43}$$

Finally, it follows that for every $r \ge 0$,

$$P_{Y^n}^{(\mathcal{C})}\Big(\big|f(Y^n) - \mu_f^*\big| \ge r\Big) \le P_{X^n,Y^n}^{(\mathcal{C})}\Big(|f(Y^n) - \phi(X^n)| \ge r/2\Big) + P_{X^n}^{(\mathcal{C})}\Big(\big|\phi(X^n) - \mu_f^*\big| \ge r/2\Big)$$

$$\le 2e^{-\frac{r^2}{8c\|f\|_{\text{Lip}}^2}} + 2e^{nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{8c\|f\|_{\text{Lip}}^2}} \tag{44}$$

$$\le 4e^{nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} - \frac{r^2}{8c\|f\|_{\text{Lip}}^2}}, \tag{45}$$

where (44) is by (37) and (43), while (45) follows from the fact that $nC - \ln M + a\sqrt{n} + \ln\frac{1}{\varepsilon} \ge D(P_{Y^n}^{(\mathcal{C})} \| P_{Y^n}^*) \ge 0$ by Theorem 2, and from (36). This proves (35). ∎

As an illustration, equip $\mathcal{Y}^n$ with the metric $d(y^n, v^n) = \sum_{i=1}^n 1_{\{y_i \ne v_i\}}$, i.e., the weighted Hamming metric (6) with $c_1, \ldots, c_n = 1$. Then, any function $f : \mathcal{Y}^n \to \mathbb{R}$ of the form $f(y^n) = \frac{1}{n}\sum_{i=1}^n f_i(y_i)$, $y^n \in \mathcal{Y}^n$, where $f_1, \ldots, f_n : \mathcal{Y} \to \mathbb{R}$ are Lipschitz functions on $\mathcal{Y}$, will satisfy $\|f\|_{\text{Lip}} \le \frac{L}{n}$ where $L \triangleq \max_{1 \le i \le n} \|f_i\|_{\text{Lip}}$. Any probability distribution $P$ on $\mathcal{Y}$ satisfies the $T_1(1/4)$ inequality w.r.t. the Hamming metric (this is simply Pinsker's inequality); by tensorization of transportation-cost inequalities (see, e.g., [11, Proposition 11]), any product probability measure on $\mathcal{Y}^n$ satisfies $T_1(n/4)$ w.r.t. the above metric. Consequently, for any $(n, M, \varepsilon)$-code for $T$ and any function $f$ of the above form, Theorem 4 gives

$$P_{Y^n}^{(\mathcal{C})}\big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r\big) \le \frac{4}{\varepsilon} e^{nC - \ln M + a\sqrt{n} - \frac{nr^2}{2L^2}} \tag{46}$$

for every $r \ge 0$. As pointed out in [10], concentration inequalities like (35), or its more specialized version (46), can be very useful for characterizing the performance of good channel codes without having to explicitly construct such codes: all one needs to do is to find the capacity-achieving output distribution $P_Y^*$ and evaluate $\mathbb{E}[f(Y^{*n})]$ for any $f$ of interest. Then, Theorem 4 guarantees that $f(Y^n)$ concentrates tightly around $\mathbb{E}[f(Y^{*n})]$, which is relatively easy to compute since $P_{Y^n}^*$ is a product measure.

## IV. Discussion

The new bounds presented in Theorems 2 and 3 quantify the trade-offs between the minimal blocklength required for achieving a certain gap (in rate) to capacity with a fixed block error probability, and normalized divergence between the *output distribution* induced by the code and the (unique) capacity-achieving output distribution of the channel. Moreover, these bounds sharpen the asymptotic $O(\cdot)$ terms in the results of Polyanskiy and Verdú [10] for all finite blocklengths $n$.

The results of this paper are similar in spirit to a lower bound on the rate loss with respect to fully random block codes (whose average distance spectrum is binomially distributed) in terms of the normalized divergence between the distance spectrum of a code and the binomial distribution. Specifically, a combination of [13, Eqs. (A17) and (A19)] provides a lower bound on the rate loss with respect to fully random block codes in terms of the normalized divergence between the distance spectrum of the code and the binomial distribution; the latter result refers to the empirical *input distribution* of good codes.

**Acknowledgment**: The work of M. Raginsky was supported in part by the U.S. National Science Foundation (NSF) under CAREER award no. CCF–1254041. The work of I. Sason was supported by the Israeli Science Foundation (ISF), grant number 12/12.

## References

[1] R. Ahlswede and G. Dueck, "Every bad code has a good subcode: a local converse to the coding theorem," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 34, pp. 179–182, 1976.

[2] R. Ahlswede, "An elementary proof of the strong converse theorem for the multiple-access channel," *Journal of Combinatorics, Information and System Sciences*, vol. 7, no. 3, pp. 216–230, 1982.

[3] U. Augustin, "Gedächtnisfreie Kanäle für diskrete Zeit," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 6, pp. 10–61, 1966.

[4] S. G. Bobkov and F. Götze, "Exponential integrability and transportation cost related to logarithmic Sobolev inequalities," *Journal of Functional Analysis*, vol. 163, pp. 1–28, 1999.

[5] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Cambridge University Press, 1981.

[6] V. Kostina and S. Verdú, "Fixed-length lossy compression in the finite blocklength regime," *IEEE Trans. on Info. Theory*, vol. 58, no. 6, pp. 3309–3338, June 2012.

[7] C. McDiarmid, "Concentration," *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248, Springer, 1998.

[8] —, "On the method of bounded differences," *Surveys in Combinatorics*, vol. 141, pp. 148–188, Cambridge University Press, 1989.

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

[10] Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability," preprint, December 12. See: http://people.lids.mit.edu/yp/homepage/data/optcodes_journal.pdf.

[11] M. Raginsky and I. Sason, "Concentration of Measure Inequalities in Information Theory, Communications and Coding," submitted to the *Foundations and Trends in Communications and Information Theory*, December 2012. [Online]. Available: http://arxiv.org/abs/1212.4663.

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

[13] S. Shamai and I. Sason, "Variations on the Gallager bounds, connections, and applications," *IEEE Trans. on Information Theory*, vol. 48, no. 12, pp. 3029–3051, December 2002.

[14] S. Shamai and S. Verdú, "The empirical distribution of good codes," *IEEE Trans. on Info. Theory*, vol. 43, no. 3, pp. 836–846, May 1997.

[15] F. Topsøe, "An information theoretical identity and a problem involving capacity," *Studia Scientiarum Mathematicarum Hungarica*, vol. 2, pp. 291–292, 1967.

[16] C. Villani, *Topics in Optimal Transportation*, AMS, 2003.