Third-order Analysis of Channel Coding in the Moderate Deviations Regime

Recep Can Yavas
Joint work with Victoria Kostina and Michelle Effros

California Institute of Technology (Caltech)

2022 IEEE International Symposium on Information Theory (ISIT)
Espoo, Finland, 30 June 2022

This work was supported in part by the NSF under grant CCF-1817241 and CCF-1956386.

Channel coding problem

- Goal: Find the maximum achievable message set size $M^*(n, \epsilon)$ given a DMC $P_{Y|X}$, blocklength n, and average error probability ϵ

- Problem: A very hard problem and $M^*(n, \epsilon)$ unknown in general

- Tractable approaches: Analyze $M^*(n, \epsilon)$ as $n \to \infty$

Error prob. regimes	**Error prob.**	**Rate** $R = \frac{\log M^*(n,\epsilon)}{n}$	**Ref.**
Central Limit Theorem (CLT)	$\epsilon \in (0,1)$	$C - O\left(\frac{1}{\sqrt{n}}\right)$	Strassen 62', PPV 10'
Large Deviations (LD)	$e^{-nE(R)}$	$C - O(1)$	Gallager 65'

- Goal: Find the maximum achievable message set size $M^*(n,\epsilon)$ given a DMC $P_{Y|X}$, blocklength n, and average error probability ϵ

- Problem: A very hard problem and $M^*(n,\epsilon)$ unknown in general

- Tractable approaches: Analyze $M^*(n,\epsilon)$ as $n \to \infty$

Error prob. regimes	**Error prob.**	**Rate** $R = \frac{\log M^*(n,\epsilon)}{n}$	**Ref.**
Central Limit Theorem (CLT)	$\epsilon \in (0,1)$	$C - O\left(\frac{1}{\sqrt{n}}\right)$	Strassen 62', PPV 10'
Large Deviations (LD)	$e^{-nE(R)}$	$C - O(1)$	Gallager 65'
Moderate Deviations (MD)	$e^{-o(n)}$	$C - o(1)$	Altuğ-Wagner 14'

- Advantage of MD approximation: can be more accurate for a wider range of (n,ϵ).

Let $X_1, \ldots, X_n$ be i.i.d. and $\mathbb{E}[X_1] = 0$. Let $S_n = \sum_{i=1}^{n} X_i$ be the sum, x be a positive constant.

1. **CLT**: $\mathbb{P}[S_n \geq \sqrt{n}x] \approx \text{constant}$
2. **LD**: $\mathbb{P}[S_n \geq nx] = e^{-nE}$
3. **MD**: $\mathbb{P}[S_n \geq g(n)x] \to 0$ but slower than exponential decay, where $\sqrt{n} \ll g(n) \ll n$.

- Probabilities in the MD regime appear in the analysis of achievability and converse bounds!

1. **CLT**: Polyanskiy et al. 10' (achievability) and Tomamichel-Tan 13' (converse): for nonsingular channels

$$\log M^*(n, \epsilon) = nC - \sqrt{nV} Q^{-1}(\epsilon) + \frac{1}{2} \log n + O(1)$$

2. **CLT**: Moulin 17': for nonsingular channels with non-lattice information density

$$\log M^*(n, \epsilon) \geq nC - \sqrt{nV} Q^{-1}(\epsilon) + \frac{1}{2} \log n + \underline{S}\, Q^{-1}(\epsilon)^2 + \underline{B} + o(1)$$

$$\log M^*(n, \epsilon) \leq nC - \sqrt{nV} Q^{-1}(\epsilon) + \frac{1}{2} \log n + \overline{S}\, Q^{-1}(\epsilon)^2 + \overline{B} + o(1)$$

3. **MD**: Altuğ and Wagner 14':

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

- Normal approximation without $O(\log n)$ term is still accurate in the MD regime
- **Question:** How does the third-order term scale in the MD regime?

Non-Gaussianity and Channel skewness

- Non-Gaussianity:

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

- A new quantity: Channel skewness

$$S \triangleq \lim_{\epsilon\to 0} \limsup_{n\to\infty} \frac{\zeta(n,\epsilon) - \frac{1}{2}\log n}{Q^{-1}(\epsilon)^2}$$

- Skewness of information density dominates the behavior of channel skewness S

$$\mathrm{Sk}(\imath(X;Y)) = \frac{\mathbb{E}\left[(\imath(X;Y) - C)^3\right]}{V^{3/2}}$$

- A channel is singular if there exists a single non-zero value in each column.
- Singular channels are generalized erasure channels.

  (e.g., input/output symbols can be merged, input alphabet size > 2, or channels transition matrices can be combined as in the second figure, where each of P_i is a transition probability of an erasure channel.)
- We exclude singular channels in our results.

	$P_{Y\|X}$	$\mathcal{Y}$: 0	1	2	e_1	e_2
	0	$1-\delta$	0	0	δ_1	δ_2
$\mathcal{X}$	1	0	$0.5-\delta_1$	$0.5-\delta_2$	δ_1	δ_2
redundant	2	$1-\delta$	0	0	δ_1	δ_2

$1'$ can be merged (outputs 1 and 2)

e with erasure prob. $\delta = \delta_1 + \delta_2$ (outputs e_1 and e_2)

$P_{Y\|X}$	$\mathcal{Y}$		
	P_1	0	0
$\mathcal{X}$	0	P_2	0
	0	0	P_3

Theorem 1

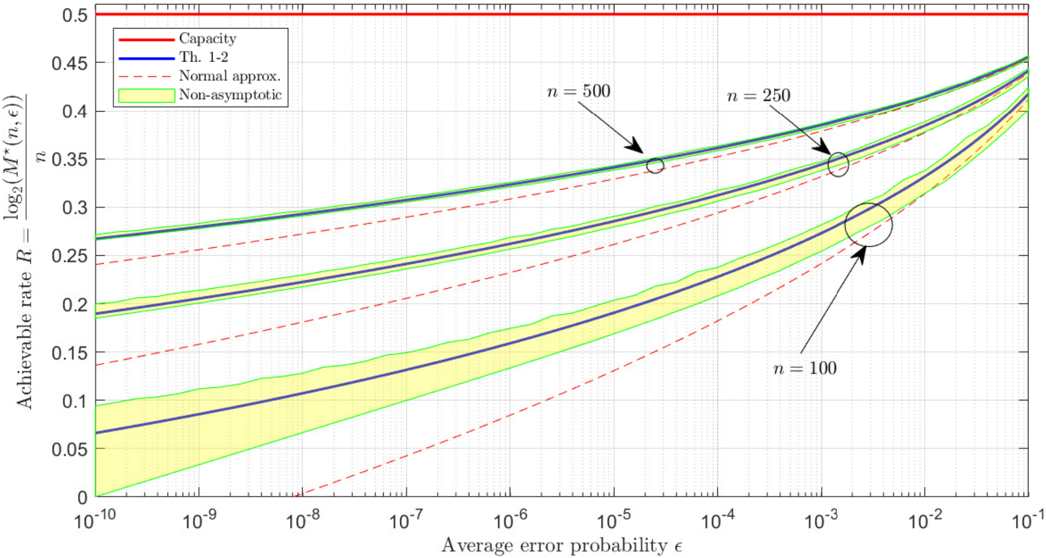
MD: *for nonsingular DMCs*

$$\zeta(n, \epsilon_n) \geq \frac{1}{2}\log n + \underline{S}\, Q^{-1}(\epsilon_n)^2 + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right) + O(1)$$

$$\zeta(n, \epsilon_n) \leq \frac{1}{2}\log n + \overline{S}\, Q^{-1}(\epsilon_n)^2 + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right) + O(1)$$

- $\underline{S}$ and $\overline{S}$ are constants that depend only on the channel $P_{Y|X}$, equal to Moulin's constants in the CLT regime.
- $\underline{S}$ and $\overline{S}$ give lower and upper bounds for channel skewness S.
- The third-order term $\frac{1}{2}\log n + O(1)$ is no longer accurate because as $n \to \infty$, $\epsilon_n \to 0$ and $Q^{-1}(\epsilon_n)^2 \to \infty$.

BSC(0.11)



0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Capacity
Th. 1-2
Normal approx.
Non-asymptotic
$n = 500$
$n = 250$
$n = 100$
Achievable rate $R = \frac{\log_2(M^*(n,\epsilon))}{n}$
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}
Average error probability ϵ

Formulas for lower and upper bounds

$$\underline{S} = \max_{P_X^*} \left(\frac{\mathrm{Sk}(P_X^*)\sqrt{V}}{6} + \frac{1-\eta(P_X^*)}{2(1+\eta(P_X^*))} + A_0(P_X^*) \right)$$

$$\overline{S} = \max_{P_X^*} \left(\frac{\mathrm{Sk}(P_X^*)\sqrt{V}}{6} + \frac{1}{2} + A_0(P_X^*) - A_1(P_X^*) \right)$$

- **Easy to compute quantities:** constants depend on the moments of the information density $\imath(X;Y)$.
- Maximization is over all P_X^* that achieve C and V.
- $\eta(P_X^*) \in [0,1]$: the singularity parameter, taking the value 1 iff the channel is singular.
- The bound is achieved using i.i.d. random codewords drawn from

$$P_X = P_X^* + \frac{Q^{-1}(\epsilon_n)}{2\sqrt{nV}} \nabla^2 I(P_X^*) \nabla V(P_X^*)$$

- $A_0(P_X^*)$ and $A_1(P_X^*)$ depend on the Hessian of $I(P_X^*)$ and the gradient of $V(P_X^*)$, and capture the benefit of employing P_X instead of P_X^*

A special case

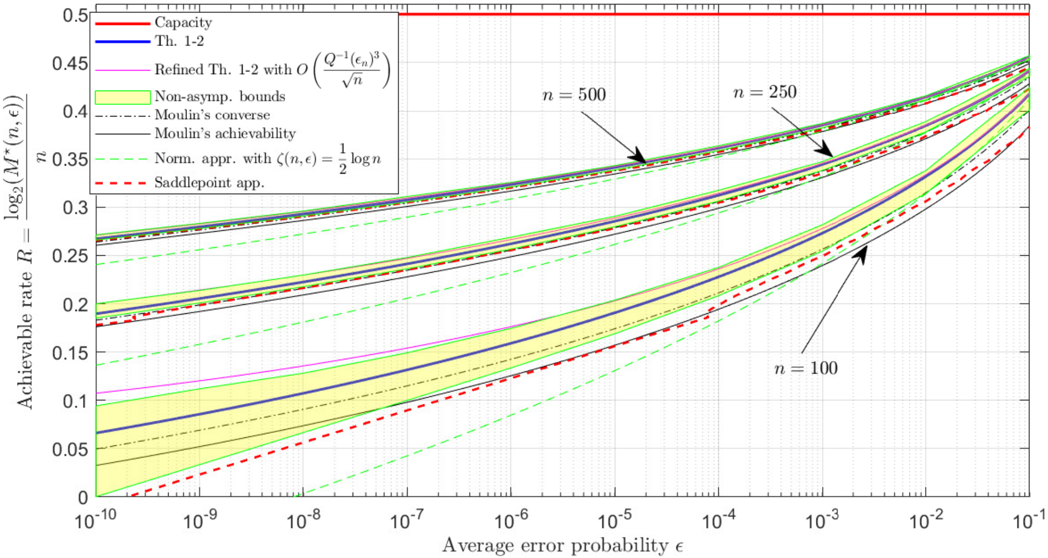
- **Cover-Thomas symmetry:** All rows (resp. columns) of the channel transition matrix are permutations of each other. Example: BSC.

- For Cover-Thomas symmetric channels, the lower and upper bounds on S match:

$$S = \frac{\mathrm{Sk}(P_X^*)\sqrt{V}}{6} + \frac{1}{2}$$

- $P_X^* = \mathrm{Unif}(\mathcal{X})$ achieves S.

BSC(0.11)



Capacity
Th. 1-2
Refined Th. 1-2 with $O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}}\right)$
Non-asymp. bounds
Moulin's converse
Moulin's achievability
Norm. appr. with $\zeta(n,\epsilon) = \frac{1}{2}\log n$
Saddlepoint app.
$n = 500$
$n = 250$
$n = 100$
Achievable rate $R = \frac{\log_2(M^\star(n,\epsilon))}{n}$
Average error probability ϵ
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
10^{-10}
10^{-9}
10^{-8}
10^{-7}
10^{-6}
10^{-5}
10^{-4}
10^{-3}
10^{-2}
10^{-1}

Theorem 2 (Petrov)

Let $X_1, \ldots, X_n$ be i.i.d., $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}[X] = \sigma^2$. Let $x = o(\sqrt{n})$, and $x \geq 0$. Under Cramér's condition, i.e., $\mathbb{E}\left[e^{tX}\right] < \infty$ in the neighborhood of zero,

$$\mathbb{P}\left[\sum_{i=1}^{n} X_i \geq \sqrt{n\sigma^2}x\right] = Q(x)\exp\left(\frac{x^3\mathrm{Sk}(X_1)}{6\sqrt{n}} + O\left(\frac{x^4}{n}\right)\right)\left(1 + O\left(\frac{1+x}{\sqrt{n}}\right)\right)$$

- Gives a multiplicative correction term to normality.
- Skewness dominates the correction term.

Achievability proof

- Following the same steps as Polyanskiy's random coding union (RCU) bound:

$$P_{\text{error}} \leq \mathbb{P}\left[\imath(X^n; Y^n) < \tau\right] + (M-1)\mathbb{P}\left[\imath(\bar{X}^n; Y^n) \geq \imath(X^n; Y^n) \geq \tau\right]$$

  where $\bar{X}^n$ is a sample from the codebook that is independent of Y^n.
- This upper bound on error probability is a relaxation of the ML decoder.

- The optimal allocation of ϵ_n is (equivalently the optimal τ):

$$\underbrace{\mathbb{P}\left[\imath(X^n;Y^n) < \tau\right]}_{\epsilon_n - \epsilon_n \frac{Q^{-1}(\epsilon_n)}{\sqrt{nV(P_X)}}} + \underbrace{(M-1)\mathbb{P}\left[\imath(\bar{X}^n;Y^n) \geq \imath(X^n;Y^n) \geq \tau\right]}_{\epsilon_n \frac{Q^{-1}(\epsilon_n)}{\sqrt{nV(P_X)}}}$$

- Apply Petrov's theorem to the first term and Strong Large Deviations theorem to the second term.
- Maximize the right-hand side of

$$\log M^* \geq nI(P_X) - \sqrt{nV(P_X)}Q^{-1}(\epsilon_n) + \frac{1}{2}\log n$$
$$+ Q^{-1}(\epsilon_n)^2 \left(\frac{\mathrm{Sk}(P_X)\sqrt{V(P_X)}}{6} + \frac{1-\eta(P_X)}{2(1+\eta(P_X))} \right) + O\left(\frac{Q^{-1}(\epsilon_n)^3}{\sqrt{n}} \right) + O(1)$$

by taking Taylor series expansions w.r.t. capacity-achieving P_X^*.

Proof technique comparison

	Metric	Encoder	Conv. technique	Probability theorem
Polyanskiy's normal appr.	$\imath(X;Y)$	P_X^*	Divergence spect.	Berry-Esseen + WLD
Moulin's refined CLT appr.	$\imath(X;Y)$	$P_X^* + O\left(\frac{Q^{-1}(\epsilon)}{\sqrt{n}}\right)$	Meta-Converse	Edgeworth + SLD
Altuğ-Wagner's LD appr.	$\tilde{\imath}(X;Y)$	error exp. ach. P_X	Sphere-packing	SLD
Our MD appr.	$\imath(X;Y)$	$P_X^* + O\left(\frac{Q^{-1}(\epsilon_n)}{\sqrt{n}}\right)$	Divergence spect.	Petrov + SLD

- All achievability proofs analyze the ML decoder.
- Divergence spectrum method by Tomamichel-Tan 13' is a relaxation of the meta-converse. Converting the divergence spectrum bound to a single-letter minimax optimization is easier than the meta-converse approach.
- WLD (weak LD) from Polyanskiy et al. 10' is an upper bound on the tail probability and does not have the right prefactor. SLD (strong LD) provides the right prefactor in the tail probability.
- Tilted information density:

$$\tilde{\imath}(X;Y) = \log \frac{P_{Y|X}(Y|X)}{c\left(\sum_x P_X(x) P_{Y|X}(Y|x)^{1/1+\rho}\right)^{1+\rho}}$$

Extensions and future work

- We recently calculated the channel skewness of the Gaussian channel with maximal power constraint

$$S(P) = \frac{6 + 6P + 4P^2 + P^3}{6(P+1)^2(P+2)}.$$

- Altuğ and Wagner's tilted information density method (first introduced by Fano) could potentially improve the achievability bound for non-symmetric DMCs. Our current method is a special case when $\rho = 0$.

Summary

1. For the third-order term in the MD regime of channel coding:
   - We derive easy-to-compute lower and upper bounds.
   - For symmetric channels, the bounds match each other, and channel skewness S is governed by the skewness of information density $\imath(X;Y)$.
   - Our MD approximation is more accurate than the normal approximation of Polyanskiy, not surprisingly, especially for the lower error probabilities (e.g., $\epsilon \leq 10^{-2}$)
2. For some non-symmetric channels, capacity-achieving distribution does not achieve S.

References

1. R. C. Yavas, V. Kostina, and M. Effros, "Third-order analysis of channel coding in the moderate deviations regime," arXiv:2203.01418, Mar. 2022.
2. Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.
3. Y. Altuğ and A. B. Wagner, "Moderate deviations in channel coding," *IEEE Trans. Inf. Theory*, vol. 60, no. 8, pp. 4417–4426, Aug. 2014.
4. M. Tomamichel and V. Y. F. Tan, "A tight upper bound for the third-order asymptotics of discrete memoryless channels," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, July 2013, pp. 1536–1540.
5. P. Moulin, "The log-volume of optimal codes for memoryless channels, asymptotically within a few nats," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2278–2313, Apr. 2017.
6. Y. Altuğ and A. B. Wagner, "On exact asymptotics of the error probability in channel coding: Symmetric channels," *IEEE Trans. Inf. Theory, vol. 67, no. 2, pp. 844–868, Feb. 2021.*

Thank you!