Complexity and the 2nd-Order Term of Capacity-Achieving Codes

Hsin-Po WANG

Department of Mathematics, University of Illinois at Urbana–Champaign

2020-10-05 CSL SINE Group Seminar

PDF available at `https://SINE.symbol.codes/`

Noisy channel

The sender inputs $X_1^{32} =$
`11001001 00001111 11011010 10100010`.

The channel output $Y_1^{32} =$
`1--01-01 ----1--- -101---0 --0--0-0`.

Noisy channel

Sender inputs $X_1^{32} \in \mathbb{F}_q^{32}$, where $\mathbb{F}_q$ is input alphabet.
[My idea] we may assume $\mathbb{F}_q$ is a finite field.

Channel outputs Y_1^{32} according to stochastic matrix
$\mathbb{P}\{Y_i = y \mid X_i = x\} = W(y|x)$ independently for each i.

Noisy channel coding

The sender inputs $X_1^{32} \in \mathcal{B} \subsetneq \mathbb{F}_q^{32}$.
$\mathcal{B}$ is a block code (codebook) of block length $N = 32$.

The channel output Y_1^{32} according to $W(y|x)$.

The receiver maximize the a posterior probability
$\hat{X}_1^{32} = \underset{x_1^{32} \in \mathcal{B}}{\text{do-my-best}} \mathbb{P}\{X_1^{32} = x_1^{32} \mid Y_1^{32}\}$.

Noisy channel coding theorem

Channel capacity $C := \sup_{X \sim Q} I(X\,;Y)$ (mutual information).

Block length is N.
Error probability is $P_{\mathrm{e}} := \mathbb{P}\{\hat{X}_1^N \neq X_1^N\}$.
Code rate is $R := \log|\mathscr{B}|/N \log q$ (recall that $\mathscr{B} \subset \mathbb{F}_q^N$).

[Shannon 1948] *One can find block code $\mathscr{B}$ such that $P_{\mathrm{e}} \to 0$ and $R \to C$ as $N \to \infty$.*
(And C is the greatest number that makes this hold.)

2nd-order term of the theorem

How fast do error probability P_e and code rate R converge to 0 and C as block length $N \to \infty$? Characterize them as functions "$P_\mathrm{e}(N)$" and "$R(N)$".

When R is fixed, $P_\mathrm{e} \approx e^{-N}$, that is, $-\log P_\mathrm{e} \approx N$.
When P_e is fixed, $R \approx C - N^{-1/2}$, that is, $(C - R)^{-2} \approx N$.

When both R and P_e vary, $(-\log P_\mathrm{e})(C - R)^{-2} \approx N$.

2nd-order term analysis

This is two-sided bound:
A code $\mathscr{B}$ exists such that $(-\log P_\mathrm{e})(C-R)^{-2} \approx N$.
$\mathscr{B}$ does not exist such that $(-\log P_\mathrm{e})(C-R)^{-2} \gg N$.

Block length N is your income;
invest error probability P_e or code rate R or both.

2nd-order term analog

Paradigm	Random variable
law of large numbers	$\bar{X} \to \mu$
large deviations principle	$\mathbb{P}\{\bar{X} - \mu > x\} \approx e^{-nI(x)}$
central limit theorem	$\bar{X} - \mu \sim \mathcal{N}(0, \sigma\sqrt{n})$
moderate deviations principle	$\frac{-\log \mathbb{P}\{\bar{X}-\mu>\varepsilon_n x\}}{\varepsilon_n^2} \approx nI(x)$

2nd-order term analog

P.	Random variable	Random code
LLN	$\bar{X} \to \mu$	$(P_{\mathrm{e}}, R) \to (0, C)$
LDP	$\mathbb{P}\{\bar{X} - \mu > x\} \approx e^{-nI(x)}$	$P_{\mathrm{e}} \approx e^{-N}$
CLT	$\bar{X} - \mu \sim \mathcal{N}(0, \sigma\sqrt{n})$	$C - R \approx N^{-1/2}$
MDP	$\frac{-\log \mathbb{P}\{\bar{X} - \mu > \varepsilon_n x\}}{\varepsilon_n^2} \approx nI(x)$	$\frac{-\log P_{\mathrm{e}}}{(C-R)^2} \approx N$

However...

The achievability bound for random code $\mathcal{B}$ assumes exponential complexity due to $\operatorname*{argmax}_{x_1^{32}\in\mathcal{B}}$.

Goal: Comparable performance,
but with a low-complexity decoder $\operatorname*{do\text{-}my\text{-}best}_{x_1^{32}\in\mathcal{B}}$.

2nd-order term goal

P.	Random code	Low-complexity code
LLN	$(P_\mathrm{e}, R) \to (0, C)$	$(P_\mathrm{e}, R) \to (0, C)$
LDP	$P_\mathrm{e} \approx e^{-N}$	$P_\mathrm{e} \approx e^{-N^\pi}$
CLT	$C - R \approx N^{-1/2}$	$C - R \approx N^{-\rho}$
MDP	$\frac{-\log P_\mathrm{e}}{(C-R)^2} \approx N$	$(P_\mathrm{e}, C - R) \approx (e^{-N^\pi}, N^{-\rho})$

$(0 < \pi, \rho$ and $\pi + 2\rho < 1)$

Polar coding

[Arıkan 2009] invented polar coding.

It is a mechanism to produce practical codes as well as to estimate, mathematically, their P_{e} and R.

TOC: channel transformation; channel tree; channel parameter; channel process; channel polarization.

Channel transformation

Channel $W = (X \mid Y)$; input X; output Y.
Make i.i.d. copies $(X_1 \mid Y_1)$ and $(X_2 \mid Y_2)$.

$W^{(1)} := (X_1 - X_2 \mid Y_1^2)$;
$W^{(2)} := (X_2 \mid (X_1 - X_2)Y_1^2)$ (juxtaposition is tupling).

Channel transformation (other kernel)

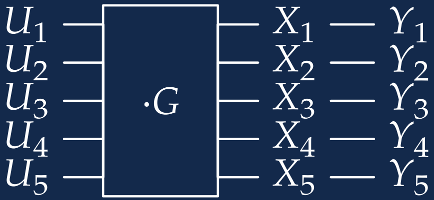
U_1^2 two free variables; G a 2×2 matrix (called kernel);
$X_1^2 := U_1^2 \cdot G$; channels generate Y_1^2.

$W^{(1)} := (U_1 \mid Y_1^2)$;
$W^{(2)} := (U_2 \mid U_1 Y_1^2)$ (juxtaposition is tupling).

Channel transformation (larger kernel)

U_1^ℓ this many free variables; G an $\ell \times \ell$ kernel matrix;
$X_1^\ell := U_1^\ell \cdot G$; channels generate Y_1^ℓ.

$$
\begin{aligned}
W^{(1)} &:= (U_1 \mid Y_1^\ell);\\
W^{(2)} &:= (U_2 \mid U_1 Y_1^\ell);\\
W^{(3)} &:= (U_3 \mid U_1^2 Y_1^\ell);\\
&\vdots\\
W^{(\ell-1)} &:= (U_\ell \mid U_1^{\ell-2} Y_1^\ell);\\
W^{(\ell)} &:= (U_\ell \mid U_1^{\ell-1} Y_1^\ell).
\end{aligned}
$$

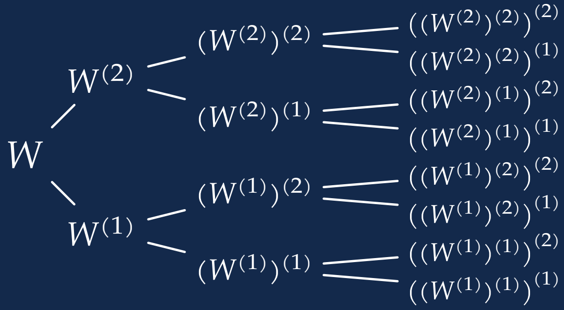


Channel tree

Channel W grows $W^{(1)}, W^{(2)}, \ldots, W^{(\ell)}$.

For each i, channel $W^{(i)}$ grows $(W^{(i)})^{(1)}, \ldots, (W^{(i)})^{(\ell)}$.

For each j, channel $(W^{(i)})^{(j)}$ grows $((W^{(i)})^{(j)})^{(1)}, \ldots$.

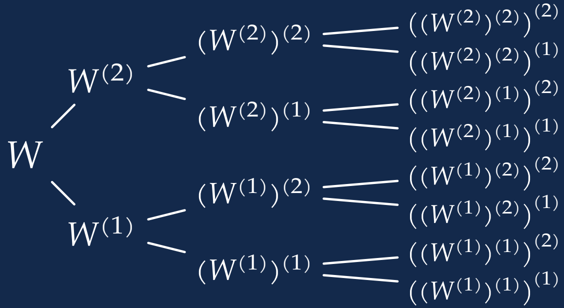


Dynamic kernel [my idea*]

Channel W grows $W^{(1)}, W^{(2)}, \ldots, W^{(\ell)}$ using G.

Channel $W^{(i)}$ grows $(W^{(i)})^{(1)}, \ldots, (W^{(i)})^{(\ell)}$ using $G^{(i)}$.

Channel $(W^{(i)})^{(j)}$ grows $((W^{(i)})^{(j)})^{(1)}, \ldots$ using $G^{(ij)}$.



Channel parameter ($\ell = 2$ and $n = 3$)

Block length $N = \ell^n = 2^3 = 8$.

Select indices $\mathcal{I} := \{212, 221, 222\} \in \{1,2\}^3$.
Code rate $R = |\mathcal{I}|/N = 3/8$ (nontrivial).

Error probability $P_\mathrm{e} \leq \sum_{ijk \in \mathcal{I}} H\Big(\big((W^{(i)})^{(j)}\big)^{(k)}\Big)$ (nontrivial);
$H(X \mid Y)$ is conditional entropy (base to be specify).

It suffices to understand

$$H(W), H(W^{(i)}), H((W^{(i)})^{(j)}), H\Big(((W^{(i)})^{(j)})^{(k)}\Big), \ldots.$$

Block length N will be $\ell^{\text{where we stop}}$.

Code rate R will be the fraction of small H-values.

Error probability P_{e} will be $\sum_{\text{those}}$ small H-values.

Channel process (syntax candy)

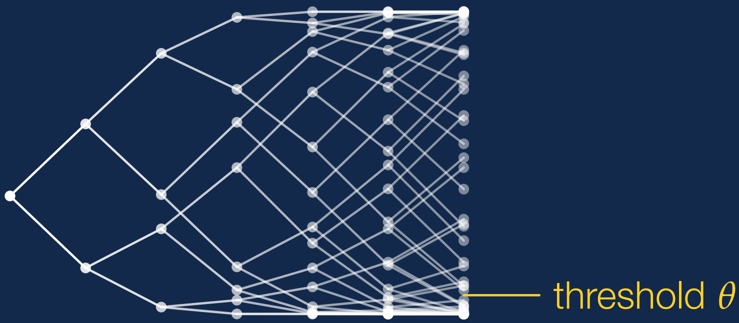
$W_0 := W$.
$W_{n+1} := W_n^{(K_{n+1})}$, where $K_{n+1} \in \{1, 2, \ldots, \ell\}$ i.i.d. uniform.

$H_n := H(W_n)$.

Decide depth n, then block length $N = \ell^n$.
Decide threshold θ, then code rate $R = \mathbb{P}\{H_n < \theta\}$.
Error probability $P_{\mathrm{e}} < \sum$ small $H_n < \sum \theta = RN\theta \leq N\theta$.

Channel polarization

$H_n := H(W_n)$ is a martingale. (Invoke the Doob's.)
$H_n \to H_\infty$ a.e. as $n \to \infty$; it turns out $H_\infty \in \{0, 1\}$.



It suffices to understand

$$\mathbb{P}\{H_n < \text{threshold}\} > C - \text{gap}.$$

Goal: $\mathbb{P}\{H_n < e^{-\ell^{\pi n}}\} > C - \ell^{-\rho n}$, where $\pi + 2\rho < 1$.
Then $N = \ell^n$ and $P_e < Ne^{-N^\pi}$ and $R > C - N^{-\rho}$.

Proof outline

Local LDP behavior: $Z(W^{(k)}) \leq \ell e^{qZ(W)\ell}(qZ(W))^{\lceil k^2/3\ell \rceil}$.
(Never heard Bhattacharyya parameter? $Z := H$.)

Local CLT behavior: $\sum_{k=1}^{\ell} f(H(W^{(k)})) < 4\ell^{1/2+\alpha}$,
where $\alpha = \log\log\ell/\log\ell$ and $f(z) := \min(z, 1-z)^{\alpha}$.

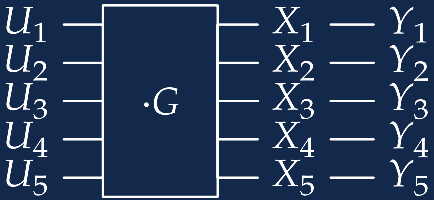
Global MDP behavior: $\mathbb{P}\{H_n < e^{-\ell^{\pi n}}\} > C - \ell^{-\rho n}$, where $\pi + 2\rho < 1$, given local LDP and local CLT behaviors.

Local LDP behavior 1/3

Want to prove $Z(W^{(k)}) \leq \ell e^{qZ(W)\ell}(qZ(W))^{\lceil k^2/3\ell \rceil}$.
Let $z := Z(W)$; want $Z(W^{(k)}) \leq \ell e^{qz\ell}(qz)^{\lceil k^2/3\ell \rceil}$.

Lemma: $Z(W^{(k)}) \leq \sum_{u_{k+1}^{\ell} \in \mathbb{F}_q^{\ell-k}} z^{\mathrm{wt}(0_1^{k-1} 1_k u_{k+1}^{\ell} \cdot G)}$;
RHS is weight enumerator of a coset code.

$$W^{(1)} := (U_1 \mid Y_1^{\ell});$$
$$W^{(2)} := (U_2 \mid U_1 Y_1^{\ell});$$
$$\vdots$$
$$W^{(\ell)} := (U_\ell \mid U_1^{\ell-1} Y_1^{\ell}).$$



Local LDP behavior 2/3

Want $\sum_{u_{k+1}^\ell} z^{\mathrm{wt}(0_1^{k-1} 1_k u_{k+1}^\ell \cdot G)} \leq \ell e^{qz\ell} (qz)^{\lceil k^2/3\ell \rceil}$ for some G.

G random; $\mathbb{E}\mathrm{LHS} = q^{-k}(1 + (q-1)z)^\ell \leq q^{-k}(1+qz)^\ell$.

Compare $(qz)^w$-coefficients: $q^{-k}\binom{\ell}{w}$ vs $\ell \frac{\ell^{w - \lceil k^2/3\ell \rceil}}{(w - \lceil k^2/3\ell \rceil)!}$.

Simplify: $2^{-k}\binom{\ell}{\lceil k^2/3\ell \rceil}\binom{\ell - \lceil k^2/3\ell \rceil}{w - \lceil k^2/3\ell \rceil}$ vs $\ell \binom{\ell}{w - \lceil k^2/3\ell \rceil}$.

Local LDP behavior 3/3

Boils down to $2^{-k}\binom{\ell}{\lceil k^2/3\ell\rceil}$ vs ℓ; ignore/cancel ⌈⌉ and ℓ.

$\binom{\ell}{d}$ is about $2^{\ell h_2(d/\ell)}$ for $d = \Theta(\ell)$. (Large deviations!)
Hence k vs $h_2(k^2/3\ell^2)$, which becomes $\sqrt{3x}$ vs $h_2(x)$.

zoom →

Local CLT behavior 1/4

Want to prove $\sum_{k=1}^{\ell} f(H(W^{(k)})) < 4\ell^{1/2+\alpha}$.

Break into segments $\begin{cases} \sum_{k=H(W)+\ell^{-1/2+\alpha}}^{\ell} < \ell^{1/2+\alpha}, \\ \sum_{k=H(W)-\ell^{-1/2+\alpha}}^{H(W)+\ell^{-1/2+\alpha}} < 2\ell^{1/2+\alpha}, \\ \sum_{k=1}^{H(W)-\ell^{-1/2+\alpha}} < \ell^{1/2+\alpha}. \end{cases}$

Local CLT behavior 2/4

Want to show $\displaystyle\sum_{k=H(W)+\ell^{-1/2+\alpha}}^{\ell} f(H(W^{(k)})) < \ell^{1/2+\alpha}$.

Jensen LHS: $(\ell-m)f\left(\frac{1}{\ell-m}\displaystyle\sum_{k=m+1}^{\ell} H(W^{(k)})\right) < \ell^{1/2+\alpha}$,
where $m = H(W) + \ell^{-1/2+\alpha} - 1$.

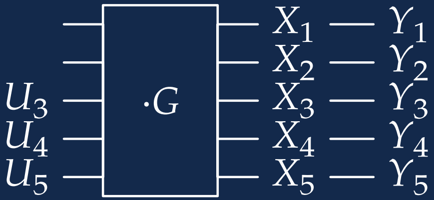
$$\vdots$$
$$W^{(\ell-2)} := (U_{\ell-2} \mid U_1^{\ell-3}Y_1^{\ell}),$$
$$W^{(\ell-1)} := (U_{\ell-1} \mid U_1^{\ell-2}Y_1^{\ell}),$$
$$W^{(\ell)} := (U_{\ell} \mid U_1^{\ell-1}Y_1^{\ell}).$$

$$\sum_{k=m+1}^{\ell} = H(U_{m+1}^{\ell} \mid U_1^{m}Y_1^{\ell}).$$

Local CLT behavior 3/4

$H(U_{m+1}^{\ell} \mid U_1^m Y_1^{\ell})$ is what? ($m = H(W) + \ell^{-1/2+\alpha} - 1$)

The conditional entropy of noisy channel coding.



Gallager has good bounds.

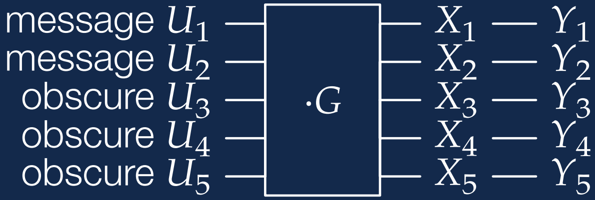
Local CLT behavior 4/4

The other segment: $\sum_{k=1}^{H(W)-\ell^{-1/2+\alpha}} f(H(W^{(k)})) < 4\ell^{1/2+\alpha}$.

Jensen inequality: $mf\left(\frac{1}{m}\sum_{k=1}^{m+1} H(W^{(k)})\right) < 4\ell^{1/2+\alpha}$.

Chain rule: $H(U_1^m \mid Y_1^\ell)$, what is this? Guess?

[My idea]
wiretap channel;
Hayashi has
good bounds.



A calculus machinery [my idea]

local LDP behavior: $Z(W^{(k)}) \leq \ell e^{qZ(W)\ell}(qZ(W))^{\lceil k^2/3\ell \rceil}$.

Local CLT behavior: $\sum_{k=1}^{\ell} f(H(W^{(k)})) < 4\ell^{1/2+\alpha}$.

eigen: $\mathbb{E}[f(H_{n+1}) \mid H_0, \ldots, H_n] \leq \ell^{-1/2+3\alpha} f(H_n)$.

en23: $\mathbb{P}\{Z_n < e^{-n^{2/3}}\} > C - \ell^{(-1/2+4\alpha)n}$.

een13: $\mathbb{P}\left\{Z_n < \exp(-e^{n^{1/3}})\right\} > C - \ell^{(-1/2+4\alpha)n}$.

elpin: $\mathbb{P}\{Z_n < e^{-\ell^{\pi n}}\} > C - \ell^{-\rho n}$.

Summary so far

For all $\pi + 2\rho < 1$, there exist codes with error probability $P_{\mathrm{e}} < e^{-N^{\pi}}$ and code rate $R > C - N^{-\rho}$.

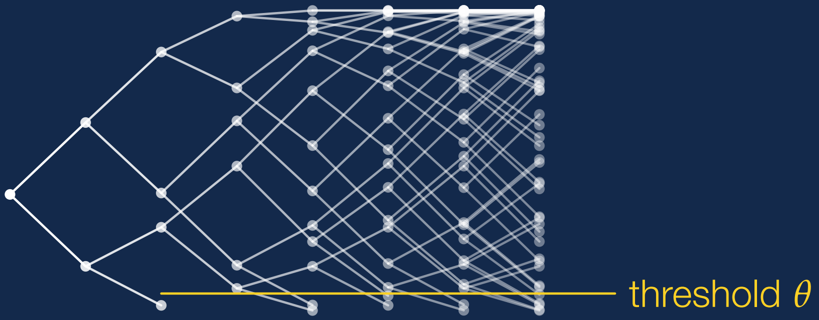
When only 2×2 kernel are allowed, at least $\pi, \rho > 0$.

It happens that they have complexity $O(\log N)$ per bit.

Can we reduce the complexity further
(at the expense of worse performance etc)?

Pruning

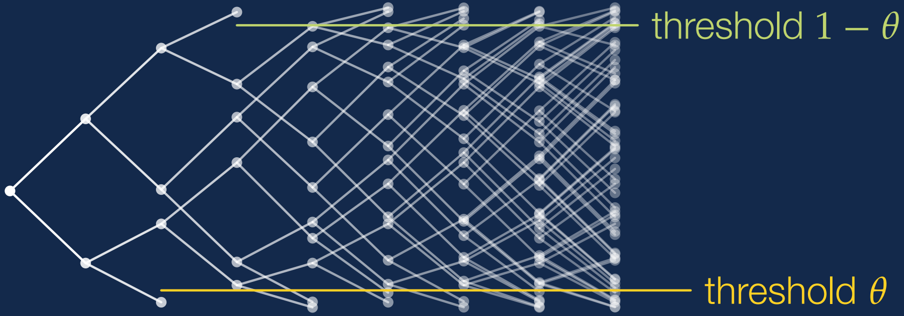
The bottom channel is good enough before we reach our favorite n.



Why do we apply transform any further? (Ans: don't!)

Pruning

The top channel is too bad. Do we expect any of its descendants to be good enough?



We don't.

Stopping time

Channel H_i needs transformation if $\theta < H_i < 1 - \theta$.

Set $\theta = N^{-10}$; assume $i > O(\log \log N)$, then $e^{-2^{\pi i}} < \theta$.

Then $\mathbb{P}\{H_i \leq \theta\} > \mathbb{P}\{H_i \leq e^{-2^{\pi i}}\} \geq C - \ell^{-\rho i}$ and
$\mathbb{P}\{1 - \theta \leq H_i\} > \mathbb{P}\{1 - e^{-2^{\pi i}} \leq H_i\} \geq 1 - C - \ell^{-\rho i}$

That is, $\mathbb{P}\{\theta < H_i < 1 - \theta\} \leq 2\ell^{-\rho i}$.

Geometric complexity

Complexity = #transformations = $\sum_{i=0}^{n} \mathbb{P}\{\theta < H_i < 1-\theta\}$.

$\sum_{i=O(\log\log N)}^{n} \mathbb{P}\{\theta < H_i < 1-\theta\} \leq \sum 2\ell^{-\rho i} = O(1)$;
$\sum_{i=0}^{O(\log\log N)} \mathbb{P}\{\theta < H_i < 1-\theta\} \leq \sum 1 = O(\log\log N)$.

Complexity is $O(\log\log N)$ per bit.

Summary

There exist codes with complexity $O(\log \log N)$ per bit, error probability $P_\mathrm{e} < N^{-9}$, and code rate $R = C - N^{-\rho}$.

There exist codes with complexity $O(\log N)$ per bit, error probability $P_\mathrm{e} < e^{-N^{\pi}}$, and code rate $R > C - N^{-\rho}$.

Are there codes in between?

Summary

MDP code taken from (with Duursma)
Polar Codes' Simplicity, Random Codes' Durability
https://arxiv.org/abs/1912.08995.

Log-log code taken from (with Duursma)
Log-logarithmic Time Pruned Polar Coding
https://arxiv.org/abs/1905.13340.

Question?

PDF available at `https://SINE.symbol.codes/`

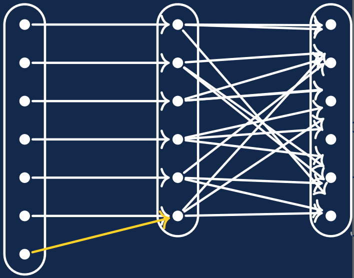
Predefined questions:
Why input alphabet is finite field? What advantage?
Definition of Bhattacharyya parameter?
References for XYZ?
What channels? Your contribution over others?
Future plan?

Code	Error	Gap	Complex	Channel
random	e^{-N^π}	$N^{-\rho}$	$\exp(N)$	DMC
RM	$\to 0$	$\to 0$	$O(N^2)$	BEC
LDPC	$\to 0$	$\to 0$	???	S. BDMC
RA family	$\to 0$	$\to 0$	$O(1)$	BEC
[my polar]	e^{-N^π}	$N^{-\rho}$	$O(\log N)$	DMC
old prune	$e^{-N^{1/2}}$	$O(1)$	$\Theta(\log N)$	S. BDMC
[my prune]	N^{-9}	$N^{-\rho}$	$O(\log\log N)$	DMC

	symmetric			asymmetric	
P.	binary	prime-ary	finite	binary	finite
LLN	[3]	[11]	[11]	[35]	W.
LDP*	[5]	[29]	[32]	[24]	W.
CLT*	[26, 28]	[9]	W.	W.	W.
MDP*	[19, 28]	[10]	W.	W.	W.
LDP	[27, 21]	W.	W.	W.	W.
CLT	[15, 20]	W.	W.	W.	W.
MDP	W.	W.	W.	W.	W.

Input alphabet [my idea]

$$\begin{bmatrix} W(y_1|1) & W(y_2|1) & W(y_3|1) & \cdots \\ W(y_1|2) & W(y_2|2) & W(y_3|2) & \cdots \\ W(y_1|3) & W(y_2|3) & W(y_3|3) & \cdots \\ W(y_1|4) & W(y_2|4) & W(y_3|4) & \cdots \\ W(y_1|5) & W(y_2|5) & W(y_3|5) & \cdots \\ W(y_1|6) & W(y_2|6) & W(y_3|6) & \cdots \\ W(y_1|6) & W(y_2|6) & W(y_3|6) & \cdots \end{bmatrix}$$



Asymmetric channels

Recall U_i is the coordinate as in $X_1^\ell := U_1^\ell \cdot G$.
The difficulty of asymmetric channels is nonuniform U_i.

Define synthetic channel $V^{(k)} := (U_i \mid U_1^{i-1})$.
Define $V^{(i)}, (V^{(i)})^{(j)}, ((V^{(i)})^{(j)})^{(k)}, \ldots$; define $\{V_n\}$.
It polarizes, and at the same pace.

High $H(V_n)$ low $H(W_n)$ vs both high vs both low.

Bhattacharyya parameter

Binary $Z(W) := \sum_{y \in \mathcal{Y}} \sqrt{W(y|0)W(y|1)}$.

Non-binary $\frac{1}{q-1} \sum_{\substack{x,x' \in \mathbb{F}_q \\ x \neq x'}} \sum_{y \in \mathcal{Y}} \sqrt{W(x,y)W(x',y)}$.

[My idea] $\max_{0 \neq d \in \mathbb{F}_q} \sum_{x \in \mathbb{F}_q} \sum_{y \in \mathcal{Y}} \sqrt{W(x,y)W(x+d,y)}$.

Random codes references

LDP: [14, 16, 33, 18, 17, 8, 6, 25, 13]

CLT: [37, 36, 12, 34, 7, 22, 30]

MDP: [1, 31, 2, 4, 23]

[1] Y. Altuğ and A. B. Wagner.
Moderate deviation analysis of channel coding: Discrete memoryless case.
In *2010 IEEE International Symposium on Information Theory*, pages 265–269, June 2010.
doi:10.1109/ISIT.2010.5513319.

[2] Y. Altuğ and A. B. Wagner.
Moderate deviations in channel coding.
IEEE Transactions on Information Theory, 60(8):4417–4426, Aug 2014. doi:10.1109/TIT.2014.2323418.

[3] E. Arikan.
Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels.
IEEE Transactions on Information Theory, 55(7):3051–3073, July 2009. doi:10.1109/TIT.2009.2021379.

[4] E. Arikan.
A packing lemma for polar codes.
In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2441–2445, June 2015.
doi:10.1109/ISIT.2015.7282894.

[5] E. Arikan and E. Telatar.
On the rate of channel polarization.
In *2009 IEEE International Symposium on Information Theory*, pages 1493–1495, June 2009.
doi:10.1109/ISIT.2009.5205856.

[6] A. Barg and G. D. Forney.
Random codes: minimum distances and error exponents.
IEEE Transactions on Information Theory, 48(9):2568–2573, Sep. 2002. doi:10.1109/TIT.2002.800480.

[7] D. Baron, M. A. Khojastepour, and R. G. Baraniuk.
How quickly can we approach channel capacity?
In *Conference Record of the Thirty-Eighth Asilomar Conference on Signals, Systems and Computers, 2004.*, volume 1, pages 1096–1100 Vol.1, Nov 2004. doi:10.1109/ACSSC.2004.1399310.

[8] R. Blahut.
Hypothesis testing and information theory.
IEEE Transactions on Information Theory, 20(4):405–417, July 1974. doi:10.1109/TIT.1974.1055254.

[9] Jarosław Błasiok, Venkatesan Guruswami, Preetum Nakkiran, Atri Rudra, and Madhu Sudan.
General strong polarization.
In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2018, pages 485–492, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3188745.3188816.

[10] Jaroslaw Blasiok, Venkatesan Guruswami, and Madhu Sudan.
Polar Codes with Exponentially Small Error at Finite Block Length.
In Eric Blais, Klaus Jansen, José D. P. Rolim, and David Steurer, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2018)*, volume 116 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 34:1–34:17, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. URL: `http://drops.dagstuhl.de/opus/volltexte/2018/9438`, `doi:10.4230/LIPIcs.APPROX-RANDOM.2018.34`.

[11] E. Şaşoğlu, E. Telatar, and E. Arikan.
Polarization for arbitrary discrete memoryless channels.
In *2009 IEEE Information Theory Workshop*, pages 144–148, Oct 2009. `doi:10.1109/ITW.2009.5351487`.

[12] R. L. Dobrushin.
Mathematical problems in the shannon theory of optimal coding of information.
In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 211–252, Berkeley, Calif., 1961. University of California Press. URL: `https://projecteuclid.org/euclid.bsmsp/1200512168`.

[13] Y. Domb, R. Zamir, and M. Feder.
The random coding bound is tight for the average linear code or lattice.
IEEE Transactions on Information Theory, 62(1):121–130, Jan 2016. `doi:10.1109/TIT.2015.2496308`.

[14] R.M. Fano.
Transmission of Information: A Statistical Theory of Communications.
M.I.T. Press, 1961. URL: `https://books.google.com.tw/books?id=VSYIAQAAIAAJ`.

[15] A. Fazeli, H. Hassani, M. Mondelli, and A. Vardy.
Binary linear codes with optimal scaling: Polar codes with large kernels.
In *2018 IEEE Information Theory Workshop (ITW)*, pages 1–5, Nov 2018. `doi:10.1109/ITW.2018.8613428`.

[16] R. Gallager.
A simple derivation of the coding theorem and some applications.
IEEE Transactions on Information Theory, 11(1):3–18, January 1965. `doi:10.1109/TIT.1965.1053730`.

[17] R. Gallager.
The random coding bound is tight for the average code (corresp.).
IEEE Transactions on Information Theory, 19(2):244–246, March 1973. `doi:10.1109/TIT.1973.1054971`.

[18] Robert G. Gallager.
Information Theory and Reliable Communication.
John Wiley & Sons, Inc., New York, NY, USA, 1968. URL:
https://www.wiley.com/en-us/Information+Theory+and+Reliable+Communication-p-9780471290483.

[19] V. Guruswami and P. Xia.
Polar codes: Speed of polarization and polynomial gap to capacity.
IEEE Transactions on Information Theory, 61(1):3–16, Jan 2015. doi:10.1109/TIT.2014.2371819.

[20] Venkatesan Guruswami, Andrii Riazanov, and Min Ye.
Arikan meets shannon: Polar codes with near-optimal convergence to channel capacity.
In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pages 552–564, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3357713.3384323.

[21] S. H. Hassani, R. Mori, T. Tanaka, and R. L. Urbanke.
Rate-dependent analysis of the asymptotic behavior of channel polarization.
IEEE Transactions on Information Theory, 59(4):2267–2276, April 2013. doi:10.1109/TIT.2012.2228295.

[22] M. Hayashi.
Information spectrum approach to second-order coding rate in channel coding.
IEEE Transactions on Information Theory, 55(11):4947–4966, Nov 2009. doi:10.1109/TIT.2009.2030478.

[23] M. Hayashi and V. Y. F. Tan.
Erasure and undetected error probabilities in the moderate deviations regime.
In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1821–1825, June 2015. doi:10.1109/ISIT.2015.7282770.

[24] J. Honda and H. Yamamoto.
Polar coding without alphabet extension for asymmetric models.
IEEE Transactions on Information Theory, 59(12):7829–7838, Dec 2013. doi:10.1109/TIT.2013.2282305.

[25] A. G. i. Fàbregas, I. Land, and A. Martinez.
Extremes of random coding error exponents.
In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2896–2898, July 2011. doi:10.1109/ISIT.2011.6034105.

[26] S. B. Korada, A. Montanari, E. Telatar, and R. Urbanke.
An empirical scaling law for polar codes.
In *2010 IEEE International Symposium on Information Theory*, pages 884–888, June 2010. doi:10.1109/ISIT.2010.5513579.

[27] S. B. Korada, E. Sasoglu, and R. Urbanke.
Polar codes: Characterization of exponent, bounds, and constructions.
IEEE Transactions on Information Theory, 56(12):6253–6264, Dec 2010. doi:10.1109/TIT.2010.2080990.

[28] M. Mondelli, S. H. Hassani, and R. L. Urbanke.
Unified scaling of polar codes: Error exponent, scaling exponent, moderate deviations, and error floors.
IEEE Transactions on Information Theory, 62(12):6698–6712, Dec 2016. doi:10.1109/TIT.2016.2616117.

[29] R. Mori and T. Tanaka.
Source and channel polarization over finite fields and reed-solomon matrices.
IEEE Transactions on Information Theory, 60(5):2720–2736, May 2014. doi:10.1109/TIT.2014.2312181.

[30] Y. Polyanskiy, H. V. Poor, and S. Verdu.
Channel coding rate in the finite blocklength regime.
IEEE Transactions on Information Theory, 56(5):2307–2359, May 2010. doi:10.1109/TIT.2010.2043769.

[31] Y. Polyanskiy and S. Verdú.
Channel dispersion and moderate deviations limits for memoryless channels.
In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1334–1339, Sept 2010. doi:10.1109/ALLERTON.2010.5707068.

[32] Eren Sasoglu.
Polar Coding Theorems for Discrete Systems.
PhD thesis, École Polytechnique Fédérale de Lausanne, Lausanne, 2011. URL: http://infoscience.epfl.ch/record/168993, doi:10.5075/epfl-thesis-5219.

[33] C.E. Shannon, R.G. Gallager, and E.R. Berlekamp.
Lower bounds to error probability for coding on discrete memoryless channels. i.
Information and Control, 10(1):65 – 103, 1967. URL: http://www.sciencedirect.com/science/article/pii/S0019995867900526, doi:https://doi.org/10.1016/S0019-9958(67)90052-6.

[34] V. Strassen.
Asymptotische abschätzungen in shannons informationstheorie.
In *Transactions of the Third Prague Conference on Information Theory*, pages 689–723. Publishing House of the Czechoslovak Academy of Sciences, 1962. URL: https://www.math.cornell.edu/~pmlut/strassen.pdf.

[35] D. Sutter, J. M. Renes, F. Dupuis, and R. Renner.
Achieving the capacity of any dmc using only polar codes.
In *2012 IEEE Information Theory Workshop*, pages 114–118, Sep. 2012. doi:10.1109/ITW.2012.6404638.

[36] LIONEL WEISS.
On the strong converse of the coding theorem for symmetric channels without memory.
Quarterly of Applied Mathematics, 18(3):209–214, 1960. URL: http://www.jstor.org/stable/43636330.

[37] J. Wolfowitz.
The coding of messages subject to chance errors.
Illinois J. Math., 1(4):591–606, 12 1957. doi:10.1215/ijm/1255380682.