My Little Hammers and Screwdrivers for Analyzing Code Ensemble Performance

Neri Merhav

The Viterbi Faculty of Electrical & Computer Engineering
Technion—Israel Institute of Technology
Haifa, Israel

ISIT 2023, Taipei, Taiwan, June 26, 2023
In memory of Jacob Ziv,
a shining star in the sky of information theory
and a great inspiration to me and to many others,
for years to come.
A Very Quick Historical Overview

- Shannon (’48): random coding as a simple tool for proving \( \exists \) good codes.
- Elias (’55,’56); Fano (’61); Gallager (’65, ’68): exponential error bounds.
- Shannon, Gallager, Berlekamp (’67): lower bounds: SP, SLB.
- Csiszár & Körner (’81): the method of types.
- Many: extensions, improvements; ensembles of structured codes.

*Random coding – a paradigm on its own right.*
Traditional Bounding Techniques

- $P_e(\text{ML decoder}) \leq P_e(\text{another (easier) decoder})$.
- Jensen’s inequality: $E Z^\rho \leq (E Z)^\rho$, $0 \leq \rho \leq 1$ (Gallager–style bounds).
- $I\{P(y|x_j) \geq P(y|x_i)\} \leq [P(y|x_j)/P(y|x_i)]^\lambda$ (Chernoff bound).
- Simple union bound.
- Union bound with truncation: $P[\cup_j A_j] \leq \min\{1, \sum_j P[A_j]\}$.
- Union bound with a power parameter: $P[\cup_j A_j] \leq (\sum_j P[A_j])^\rho$, $0 \leq \rho \leq 1$.
- Union bound with intersection: $P[\cup_j A_j] \leq \sum_j P[A_j \cap G] + P[G^c]$.
- “Power distribution” inequality: $(\sum_i a_i)^s \leq \sum_i a_i^s$, $0 \leq s \leq 1$ (Forney ’68).
All these tools facilitate the analysis a great deal but at the risk of compromising exponential tightness.

Main message of this talk: It is often possible to preserve exponential tightness by bypassing some of the above inequalities.
My Little Hammers and Screwdrivers

- Type class enumeration (on top of the MoT).

- Analogue of the MoT for infinite alphabets.

- The saddle-point method – assessing probabilities and volumes.

- Integral representations of some functions (with I. Sason).

- “Jensen-like” inequalities.
Many derivations are associated with summations of exponentially many terms, e.g.,

\[ P_e \leq \sum_y E \left\{ P(y|X)^{1/(1+\rho)} \right\} \cdot E \left[ \sum_m P(y|X_m)^{1/(1+\rho)} \right]^\rho, \]

\[ P_c = \frac{1}{M} E \left\{ \sum_y \max_m P(y|X_m) \right\} = \frac{1}{M} \lim_{\beta \to \infty} \sum_y E \left\{ \left[ \sum_m P(y|X_m)^\beta \right]^{1/\beta} \right\}. \]

In some situations (e.g., the BC, the IFC, the GPC, the wiretap channel, erasure/list decoding), the optimal likelihood function = sum of exponentially many terms,

Broadcast channel: \( \text{score}_i = \sum_m P(y|x_{m,i}) \)

Interference channel: \( \text{score}_i = \sum_m P(y|x_{i,m}) \)
The idea:

\[
\sum_m P(y|X_m)^\beta = \sum_Q N_y(Q) \cdot P(y|x_Q)^\beta = \sum_Q N_y(Q) \cdot e^{n\beta f(Q)},
\]

where

\[N_y(Q) = \text{number of } X_m \text{ in a given type } Q \text{ of } x \text{ given } y.\]

What have we gained?

- \(\sum\) of exponentially many terms \(\rightarrow\) \(\sum\) of polynomially few terms.
- \(N_y(Q) \sim \text{Binomial } (e^{nR}, e^{-nI(Q)})\) – easy to handle.
- Marginals of \(\{N_y(Q)\}\) almost always suffice; Pairs are \(\sim\) independent.
Consequence: Avoiding the Use of Jensen’s Inequality

\[
E \left\{ \left[ \sum_m P(y|X_m)^\beta \right]^{1/\beta} \right\} = E \left[ \sum_Q N_y(Q) \cdot e^{n\beta f(Q)} \right]^{1/\beta} = E \left[ \max_Q N_y(Q) \cdot e^{n\beta f(Q)} \right]^{1/\beta} = E \left\{ \max_Q [N_y(Q)]^{1/\beta} \cdot e^{nf(Q)} \right\} = E \left\{ \sum_Q [N_y(Q)]^{1/\beta} \cdot e^{nf(Q)} \right\} = \sum_Q E\{[N_y(Q)]^{1/\beta}\} \cdot e^{nf(Q)}.
\]

- We just have to know how to assess moments of \( N_y(Q) \).
- Equivalently, deal with the large deviations behavior.
Properties of $N \sim \text{Binomial}(e^{nA}, e^{-nB})$

Drastic difference between $A > B$ and $A < B$: phase transition at $A = B$.

Moments:

$$E\{N^s\} = \begin{cases} \exp\{n s (A - B)\} & A > B \\ \exp\{n (A - B)\} & A < B \end{cases}$$

Intuition:

- $A > B$: double-exponential concentration of $N$ around its mean $e^{n(A-B)}$.
- $A < B$: $E\{N^s\} = \sum_{n \geq 1} n^s P[N = n] \approx 1^s P[N = 1] \approx e^{n(A-B)}$. 
Properties of $N \sim \text{Binomial}(e^nA, e^{-nB})$ (Cont’d)

Large deviations behavior:

$$\Pr\{N \geq e^{\lambda n}\} \doteq e^{-nE},$$

with

$$E = \begin{cases} 
[B - A]_+ & [A - B]_+ \geq \lambda \\
\infty & \text{elsewhere}
\end{cases}$$

Intuition – “interesting” for $A < B$ and $\lambda \leq 0$: $P[N \geq 1] \doteq e^{-n(B - A)}$.

$$\Pr\{N \leq e^{\lambda n}\} \doteq \begin{cases} 
1 & A \leq B + [\lambda]_+ \\
0 & \text{elsewhere}
\end{cases}$$
Example—Exponentially Tight Evaluation of $\overline{P}_c$

Consider the BSC with crossover probability $p$. Using the relation

$$E\{N_y(Q)^{1/\beta}\} \doteq \begin{cases} \exp\{n[R - I_Q(X;Y)]/\beta\} & R > I_Q(X;Y) \\ \exp\{n[R - I_Q(X;Y)]\} & R < I_Q(X;Y) \end{cases}$$

plugging it to the expression of $\overline{P}_c$, and using the MoT, we get

$$\overline{P}_c \doteq \exp\{-nD(\delta_{GV}(R)\|p)\}$$

$$= \exp\left\{-n \left[ \delta_{GV}(R) \ln \frac{1}{p} + (1 - \delta_{GV}(R)) \ln \frac{1}{1 - p} - h_2(\delta_{GV}(R)) \right]\right\}$$

where $\delta_{GV}(R)$ is the (smaller) solution to the equation

$$\ln 2 - h_2(\delta) = R.$$
It is interesting to compare it to the result of using Jensen’s inequality:

\[
\overline{P_c} = \frac{1}{M} \lim_{\beta \to \infty} \sum_y E \left\{ \left[ \sum_m P(y|X_m)^\beta \right]^{1/\beta} \right\}
\]

\[
\leq \frac{1}{M} \lim_{\beta \to \infty} \sum_y \left[ E \sum_m P(y|X_m)^\beta \right]^{1/\beta}
\]

\[
= \exp \left( -n \left[ \min \left\{ \ln \frac{1}{p}, \ln \frac{1}{1-p} \right\} - h_2(\delta_{GV}(R)) \right] \right)
\]

Reminder: the red expression should be compared to

\[
\delta_{GV}(R) \ln \frac{1}{p} + (1 - \delta_{GV}(R)) \ln \frac{1}{1-p}
\]

of the exponentially tight evaluation of the previous slide.
Consider the process of random binning:

Each \( x \in \mathcal{X}^n \) is randomly assigned to a bin \( z = f(x) \sim \text{Unif}\{1, \ldots, e^{nR}\} \).

At the decoder

\[
\hat{x}(y, z) = \arg \max_{x \in f^{-1}(z)} P(x|y)
\]

Then,

\[
\overline{P_e} = \Pr \bigcup_{x' \neq x} \{f(x') = f(x), P(x'|y) \geq P(x|y)\}
\]

\[
= \sum_{xy} P(x, y) \sum_{Q_{X'Y} \in \mathcal{E}} \Pr \{N(Q_{X'Y}, f(x)) \geq 1\}
\]

where \( \mathcal{E} \) is the class of all \( \{Q_{X'Y}\} \) with \( \mathbf{E}_{Q'} \ln P(X'|Y) \geq \mathbf{E}_Q \ln P(X|Y) \) and where the type class enumerator

\[
N(Q_{X'Y}, z) = |\mathcal{I}(Q_{X'Y}|y) \cap f^{-1}(z)| \sim \text{Binomial}(|\mathcal{I}(Q_{X'Y}|y)|, e^{-nR}).
\]
Avoid Bounding Indicator Functions by Chernoff Bounds

Consider the error + erasure event a la Forney ('68): Instead of

\[
\Pr\{E_1\} = \Pr\left\{ \frac{P(y|x_m)}{\sum_{m' \neq m} P(y|X_{m'})} < e^{nT} \right\} \leq e^{nsT} \mathbb{E}\left\{ \left( \sum_{m' \neq m} \frac{P(y|X_{m'})}{P(y|x_m)} \right)^s \right\},
\]

use:

\[
\Pr\{E_1\} = \Pr\left\{ \sum_{m' \neq m} P(y|X_{m'}) > e^{-nT} P(y|x_m) \right\}
\]

\[
= \Pr\left\{ \sum_{Q} N_y(Q)e^{nf(Q)} > e^{-nT} e^{nf(Q_m)} \right\}
\]

\[
= \Pr\left\{ \max_{Q} N_y(Q)e^{nf(Q)} > e^{-nT} e^{nf(Q_m)} \right\}
\]

\[
= \Pr\bigcup_{Q} \left\{ N_y(Q)e^{nf(Q)} > e^{n[f(Q_m)-T]} \right\}
\]

\[
= \max_{Q} \Pr \left\{ N_y(Q) > e^{n[f(Q_m)-f(Q)-T]} \right\}
\]

and now the large deviations properties of a single \(N_y(Q)\) are invoked..
What if Those Sums Appear Also in the Denominator?

Consider the likelihood decoder that randomly selects \( \hat{m} \) under the posterior:

\[
P_{e|m=0} = E \left\{ \frac{\sum_{m=1}^{M-1} P(Y|X_m)}{\sum_{m=0}^{M-1} P(Y|X_m)} \right\}.
\]

\[
E \left\{ \frac{\sum_{m=1}^{M-1} P(y|X_m)}{P(y|x_0) + \sum_{m=1}^{M-1} P(y|X_m)} \right\}
\]

\[
= \int_0^1 ds \cdot Pr \left\{ \frac{\sum_{m=1}^{M-1} P(y|X_m)}{P(y|x_0) + \sum_{m=1}^{M-1} P(y|X_m)} \geq s \right\}
\]

\[
= n \cdot \int_0^\infty d\theta e^{-n\theta} Pr \left\{ \frac{\sum_{m=1}^{M-1} P(y|X_m)}{P(y|x_0) + \sum_{m=1}^{M-1} P(y|X_m)} \geq e^{-n\theta} \right\}
\]

\[
= \int_0^\infty d\theta e^{-n\theta} Pr \left\{ \sum_{m=1}^{M-1} P(y|X_m) \geq e^{-n\theta} P(y|x_0) \right\}
\]

and the rest is as before.
Sometimes random denominators can be handled using transform methods. For example, let $X_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$, be independent. Then,

$$\mathbb{E}\left\{ \frac{1}{\sum_{i=1}^{n} X_i^2} \right\} = ???$$
Sometimes random denominators can be handled using \textit{transform methods}. For example, let $X_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$, be independent. Then,

$$
\mathbb{E}\left\{ \frac{1}{\sum_{i=1}^{n} X_i^2} \right\} = \mathbb{E}\left\{ \int_{0}^{\infty} dt \cdot \exp \left[ -t \sum_{i=1}^{n} X_i^2 \right] \right\}
= \int_{0}^{\infty} dt \cdot \mathbb{E}\left\{ \exp \left[ -t \sum_{i=1}^{n} X_i^2 \right] \right\}
= \int_{0}^{\infty} \frac{dt}{(1 + 2\sigma^2 t)^{n/2}}
= \begin{cases} 
\infty & n \leq 2 \\
\frac{1}{(n-2)\sigma^2} & n > 2 
\end{cases}
$$
Analogue of the MoT for Infinite Alphabets

In the memoryless finite–alphabet (FA) case, we usually think of the type class of a given \( x \) as the set of all \( x' \)
- with the same empirical distribution as \( x \),
- that are permutations of \( x \).

These definitions are specific to the FA memoryless case.

An alternative definition that lends itself to extensions:

\[
\mathcal{T}(x) = \{ x' : \quad P(x') = P(x) \text{ for every memoryless source } P \}.
\]

For a general parametric family of sources \( \{ P_\theta, \theta \in \Theta \} \):

\[
\mathcal{T}(x) = \{ x' : \quad P_\theta(x') = P_\theta(x) \text{ for every } \theta \in \Theta \}.
\]
Analogue of the MoT for Infinite Alphabets (Cont’d)

If \( \{ P_\theta, \theta \in \Theta \} \) is an exponential family:

\[
P_\theta(x) = \frac{\exp \left\{ - \sum_{i=1}^{k} \theta_i \phi_i(x) \right\}}{Z(\theta)},
\]

then

\[
\mathcal{I}(x) = \{ x' : \phi_i(x') = \phi_i(x), \ i = 1, 2, \ldots, k \}.
\]

FA memoryless: \( \phi_i(x) = \sum_{t=1}^{n} \mathcal{I}\{x_t = i\} \)

FA Markov: \( \phi_{ij}(x) = \sum_{t=1}^{n} \mathcal{I}\{x_t = i, x_{t+1} = j\} \)

Gaussian memoryless: \( \phi_1(x) = \sum_{t=1}^{n} x_t; \ \phi_2(x) = \sum_{t=1}^{n} x_t^2. \)

Zero–mean, Gaussian AR(\( p \)): \( \phi_i(x) = \sum_{t=1}^{n} x_t x_{t+i}, \ i = 0, 1, \ldots, k \)
Analogue of the MoT for Infinite Alphabets (Cont’d)

The main building blocks (just like in the ordinary MoT):

- A computable expression for $|\mathcal{T}(x)|$, or $\text{Vol}\{\mathcal{T}(x)\}$.
- Make sure that number of different types is not too large.

If $\mathcal{X} = \mathbb{R}$ (say, the Gaussian case), we have two problems:

- $\text{Vol}\{\mathcal{T}(x)\} = 0$.
- The space is unbounded $\rightarrow$ infinitely many types.

First problem – allow some tolerance $\epsilon$:

$$\mathcal{T}_\epsilon(x) = \{x' : |\phi_i(x') - \phi_i(x)| \leq \epsilon, \ i = 1, 2, \ldots, k\}.$$ 

But this still does not resolve the second problem.

Second problem—confining attention to a bounded region in $\mathbb{R}^n$ (say, a sphere), outside of which the probability decays with a large enough exponent.
Analogue of the MoT for Infinite Alphabets (Cont’d)

To assess the exponent of $\text{Vol}\{\mathcal{T}(x)\}$:

$$1 \geq \int_{\mathcal{T}_\epsilon(x)} dx' \cdot P_\theta(x') = \text{Vol}\{\mathcal{T}_\epsilon(x)\} \cdot P_\theta(x),$$

leading to

$$\text{Vol}\{\mathcal{T}_\epsilon(x)\} \leq \frac{1}{P_\theta(x)} = \exp \left\{ \ln Z(\theta) + \sum_{i=1}^{k} \theta_i \phi_i(x) \right\}$$

and since this is $\forall \theta : \text{Vol}\{\mathcal{T}_\epsilon(x)\} \leq \min_{\theta} \exp \left\{ \ln Z(\theta) + \sum_{i=1}^{k} \theta_i \phi_i(x) \right\}$.

Exponentially tight as the minimizer $\theta^*$ assigns $P_{\theta^*}\{\mathcal{T}_\epsilon(x)\} \approx 1$ (WLLN).

The same idea applies to assess volumes to conditional types:

$$\mathcal{T}_\epsilon(x|y) = \{x' : |\phi_i(x', y) - \phi_i(x, y)| \leq \epsilon, \ i = 1, 2, \ldots, k\}.$$

Here one defines an exponential family of channels.
A challenge (relevant to ISI channels) is to assess the volume of a conditional type defined by both $\sum_t x_t y_t$ and $\sum_{t=1}^n x_t x_{t-j}$, $j = 0, 1, \ldots, k$. For example, the volume of

$$\mathcal{I}(\phi, \psi, \mu | y) = \left\{ x : \sum_{t=1}^n x_t^2 = n\phi, \sum_{t=1}^n x_t x_{t-1} = n\psi, \sum_{t=1}^n x_t y_t = n\mu \right\}$$

is

$$\int_{\mathbb{R}^n} dx \delta \left( \sum_{t=1}^n x_t^2 - n\phi \right) \delta \left( \sum_{t=1}^n x_t x_{t-1} - n\psi \right) \delta \left( \sum_{t=1}^n x_t y_t - n\mu \right).$$

Next, represent $\delta(A) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\{i\omega A\} d\omega$, $i = \sqrt{-1}$

then interchange the integrations, and finally, use the saddle–point method. Such a derivation is doable since this is a Gaussian integral (Huleihel, Salamatian, Merhav & Médard, 2017).
The logarithmic function
Consider the identity,

$$\ln x = \int_0^\infty \frac{e^{-u} - e^{-ux}}{u} \, du, \quad x > 0$$

which implies

$$\mathbb{E}\{\ln X\} = \int_0^\infty \frac{e^{-u} - \mathbb{E}\{e^{-uX}\}}{u} \, du.$$

A frequently encountered situation is when $X = \sum_i Y_i$, for i.i.d. $\{Y_i\}$:

$$\mathbb{E}\{\ln(Y_1 + \ldots + Y_n)\} = \int_0^\infty \frac{e^{-u} - \left[\mathbb{E}\{e^{-uY_1}\}\right]^n}{u} \, du.$$ 

Application examples include the calculations of the:

- differential entropy of a generalized multivariate Cauchy distribution;
- ergodic capacity of the Rayleigh SIMO channel;
- redundancy of universal source codes;
- moments of the empirical entropy.
The power function
Consider the identity,

\[ x^\rho = 1 + \frac{\rho}{\Gamma(1 - \rho)} \int_0^\infty \frac{e^{-u} - e^{-ux}}{u^{1+\rho}} \, du, \quad x \geq 0, \quad 0 \leq \rho \leq 1 \]

which implies

\[ \mathbb{E}\{X^\rho\} = 1 + \frac{\rho}{\Gamma(1 - \rho)} \int_0^\infty \frac{e^{-u} - \mathbb{E}\{e^{-ux}\}}{u^{1+\rho}} \, du. \]

Application examples include the calculations of:
- moments of guesswork;
- moments of parameter estimation error;
- Rényi entropy of the generalized multivariate Cauchy density;
- mutual information for channels with jammers.
Jensen-Like Inequalities

In the proof of Jensen’s inequality,

\[ \mathbb{E}\{f(X)\} \geq \sup_a \mathbb{E}\{f(a) + f'(a)(X - a)\} = f(\mathbb{E}\{X\}) \text{ attained by } a^* = \mathbb{E}\{X\} \]

But what if \( f \) is just part of a more complicated expression, e.g., \( \mathbb{E}\{f(X)g(X)\}, \mathbb{E}\{g[f(X)]\}, \mathbb{E}\{h[f(X)] \cdot g(X)\}, \) etc.?

The optimal value of \( a \) is generally different.

\[
\begin{align*}
\text{Graph of } f(x) &= 10 - 0.5x^2 \\
\text{Graph of } g(x) &= x
\end{align*}
\]
Just A Few Examples of Jensen-Like Inequalities

\[
\mathbb{E}\{-X \ln X\} \geq -\mathbb{E}\{X\} \cdot \ln(\mathbb{E}\{X\}) - \mathbb{E}\{X\} \cdot \ln \left(1 + \frac{\text{Var}\{X\}}{\mathbb{E}^2\{X\}}\right)
\]

\[
\mathbb{E}\{X^s\} \geq \mathbb{E}^s\{X\} \cdot \left(1 + \frac{\text{Var}\{X\}}{\mathbb{E}^2\{X\}}\right)^{s-1} \quad s \notin (1, 2)
\]

\[
\mathbb{E}\{\ln^2(1 + X)\} \leq \ln(1 + \mathbb{E}\{X\}) \cdot \ln \left(1 + \frac{\mathbb{E}\{X\} \ln(1 + \mathbb{E}\{X^2\}/\mathbb{E}\{X\})}{\ln(1 + \mathbb{E}\{X\})}\right).
\]

- Bounds in terms of: (i) first two moments, and (ii) MGF and its derivative.
- In many cases, easy to optimize in closed-form.
- Reverse Jensen inequalities.
- Bounds for functions that are neither convex nor concave.
- Extend easily to multivariate convex functions.
- Applicable to many information-theoretic analyses.
Some Results . . .
Example 1: List Decoding (IT, Nov. 2014)

- A code $\mathcal{C} = \{x_0, x_1, \ldots, x_{M-1}\}$, $M = e^{nR}$, is selected at random.

- The marginal of each codeword $x_i \in \mathcal{X}^n$ is $\text{Unif}\{\mathcal{Y}(Q)\}$.

- The channel $P(y|x)$ is a DMC.

- The index $I$ of the transmitted message $x_I$ is $\text{Unif}\{0, 1, \ldots, M - 1\}$.

- The decoder outputs the indices of the $L$ most likely messages.

- Error event: $I$ is not on the list.

- Regimes: fixed list size (FLS) and exponential list size (ELS).
Example 1: List Decoding (Cont’d)

A general, non–asymptotic bound:

**Theorem:** The average probability of list error, $\overline{P_e}$, associated with the optimal list decoder, is upper bounded by

$$
\overline{P_e} \leq \sum_{x,y} P(x)P(y|x) \exp \left\{-nL \left[ \hat{I}_{xy}(X;Y) + \frac{\ln L}{n} - R - O \left( \frac{\log n}{n} \right) \right]_+ \right\},
$$

where $P(x)$ is the uniform distribution over $\mathcal{F}(Q)$ and $\hat{I}_{xy}(X;Y)$ is the empirical mutual information induced by $(x,y)$.

The proof is by a large deviations analysis of the binomial RV

$$
N(x, y) = \sum_{m=1}^{M-1} \mathcal{I}\{P(y|X_m) \geq P(y|x)\}.
$$
Example 1: List Decoding (Cont’d)

The dependence on $L$ appears twice:

$$\overline{P_e} \leq \sum_{x,y} P(x)P(y|x) \exp \left\{ -nL \begin{bmatrix} \text{FLS} & \hat{I}_{xy}(X;Y) + \frac{\ln L}{n} - R - O \left( \frac{\log n}{n} \right) \\ \text{ELS} & + \right\} \right),$$

In the FLS regime, $\frac{\ln L}{n} \to 0$, and averaging $\exp\left\{-nL[\hat{I}_{xy}(X;Y) - R]_+\right\}$ yields

$$\overline{P_e} \leq e^{-nE(R,L,Q)}, \quad \text{where}$$

$$E(R, L, Q) \triangleq \min_{\tilde{P}_Y|X} \left\{ D(\tilde{P}_Y|X || P_Y|X|Q) + L \cdot [\tilde{I}(X;Y) - R]_+ \right\},$$

The best exponent is obtained by maximizing over $Q$ to yield

$$E(R, L) = \max_Q E(R, L, Q).$$
Example 1: List Decoding (Cont’d)

\[
\overline{P_e} \leq \sum_{x, y} P(x)P(y|x) \exp \left\{ -nL \left[ \hat{I}_{xy}(X;Y) + \frac{\ln L}{n} - R - O \left( \frac{\log n}{n} \right) \right]_+ \right\},
\]

In the ELS regime, \( \frac{\ln L}{n} = \lambda \). By defining

\[
\mathcal{E} = \left\{ (x, y) : \hat{I}_{xy}(X;Y) + \lambda - R \geq \epsilon \right\},
\]

we see that the contribution of \( \mathcal{E} \) is \( \leq \exp(-n \epsilon e^{\lambda n}) = e^{-n\infty} \), and so,

\[
\overline{P_e} \cdot \Pr\{\mathcal{E}^c\} = \exp \left\{ -n \min_{\{\tilde{P}_Y|X : \hat{I}(X;Y) \leq R-\lambda\}} D(\tilde{P}_Y|X \parallel P_Y|X|Q) \right\}
\]

\[
\triangleq \exp\{-nE_{sp}(R - \lambda, Q)\}
\]

which, for the optimum \( Q \), becomes \( \exp\{-nE_{sp}(R - \lambda)\} \) — meeting the converse bound of Shannon–Gallager–Berlekamp (’67).
Example 2: Erasure/List S–W Decoding (2014)

Let \((X, Y) \sim \prod_{i=1}^{n} P(x_i, y_i)\).

- \(x\) – source to be encoded.
- \(y\) – side info @ decoder.

**Encoder:** \(f : X^n \rightarrow \{0, 1, \ldots, M - 1\}, \ M = e^{nR}\).

\[
z = f(x).
\]

Random binning:
For every \(x \in X^n\), \(z\) is selected independently at random from \(\{0, 1, \ldots, M - 1\}\).
Example 2: Erasure/List S–W Decoding (Cont’d)

Erasure/list decoder: Given \( y \in \mathcal{Y}^n \) and \( z \), calculate for all \( \hat{x} \in f^{-1}(z) \):

\[
P(\hat{x}, y) \sum_{x' \in f^{-1}(z) \setminus \{\hat{x}\}} P(x', y).
\]

If \( \geq e^{nT} \), \( \hat{x} \) is a candidate.

- If there are no candidates – an erasure is declared.
- If there is exactly one candidate – ordinary decoding: \( \hat{x} \) = candidate.
- If there is more than one candidate – a list is of all candidates is created.

Define \( \mathcal{E}_1 \) as the event where the real \( x \) is not a candidate.

Let \( E_1(R, T) = \) exponent of \( \Pr\{\mathcal{E}_1\} \). The other exponent

\[
E_2(R, T) = \left\{ \begin{array}{ll}
\text{decoding error exp} & \text{erasure mode} \\
\text{expected list size exp} & \text{list mode}
\end{array} \right. = E_1(R, T) + T.
\]
Example 2: Erasure/List S–W Decoding (Cont’d)

Model: A double–BSS with a BSC\((p)\) in between.

\[ E_{\text{tce}}^1(R, T) \geq E_{\text{Forney}}^1(R, T) \text{ always.} \]

For some regions in the plane \(R—T\), \(E_{\text{tce}}^1(R, T)\) may be larger than \(E_{\text{Forney}}^1(R, T)\) by an arbitrarily large factor!

1. For \(R > h(p)\) and \(T < \ln \frac{p}{1-p} \):

\[ E_{\text{Forney}}^1(R, T) \leq R + |T| < \infty; \quad E_{\text{tce}}^1(R, T) = \infty. \]

2. Consider the case of very weakly correlated sources, i.e., \(p = \frac{1}{2} - \epsilon, \ |\epsilon| \ll 1\).

For \(R \in [h(p), \ln 2]\) and \(T = -\tau \epsilon^2 \) with \(\tau > 4\):

\[ E_{\text{Forney}}^1(R, T) \leq (\tau + 2) \epsilon^2, \quad E_{\text{tce}}^1(R, T) \geq \left[ \frac{\tau(\tau + 8)}{16} - 1 \right] \epsilon^2. \]
Example 3: Typical Random Codes (2017)

While traditional random coding error exponents are defined as

$$E_r(R) = \lim_{n \to \infty} \left[ -\ln E P_e(C_n) / n \right],$$

typical-code error exponents are defined as

$$E_{\text{typ}}(R) = \lim_{n \to \infty} \left[ -\text{E} \ln P_e(C_n) / n \right].$$

- By Jensen’s inequality, $E_{\text{typ}}(R) \geq E_r(R)$.
- $E_r(R)$ — dominated by bad codes; $E_{\text{typ}}(R)$ dominated by typical codes.

Let $\mathcal{G}_E = \{C : P_e(C) = e^{-nE}\}$.

$$P_e(C) \doteq \sum_E P(\mathcal{G}_E) \cdot e^{-nE} \doteq P(\mathcal{G}^*_E) \cdot e^{-nE^*},$$

whereas $E_{\text{typ}}(R) = E_0$, where $P[\mathcal{G}_{E_0}] \to 1$. 

We derive the exact typical–code error exponent for a class of stochastic decoders,

\[ P(\hat{m} = m | y) \propto \exp\{ng(\hat{P_{x_m}|y})\}. \]

and show that

\[ E_{\text{typ}}(R) = E_{\text{ex}}(2R) + R, \]

Extending Barg & Forney (2002) in several directions:

- General DMC is considered, not merely the BSC.
- Covering a wider family of decoders.
- Ensemble of constant composition codes – optimal PI distribution.
- Relation to expurgated exponent – for all \( R \) and a general decoder.
- The analysis technique is applicable also to more general scenarios.
Example 3: Typical Random Codes (Cont’d)

Particularizing to ML decoding, the error exponent formula includes minimization subject to the constraint,

\[ E_Q \ln W(Y|X') \geq \max \{ E_Q \ln W(Y|X), D(R, Q_Y) \}, \]

\[ D(R, Q_Y) = \sup \{ E_Q \ln W(Y|X'') : I_Q(X'';Y) \leq R, (Q_Y \times Q_{X''|Y})X = Q_X \}, \]

being the typical highest score of an incorrect message.

A technical issue: handling summations of exponentially many fractions with random denominators — exploit concentration properties.

\[ E \left[ \frac{1}{M} \sum_m \sum_{m' \neq m} \sum_y P(y|X_m) \cdot \frac{P(y|X_{m'})}{P(y|X_m) + \sum_{\tilde{m} \neq m} P(y|X_{\tilde{m}})} \right]^\rho. \]
Example 4: Broadcast Channels (with R. Averbuch, 2018)

- Exact exponents for the weak and strong user with optimal decoders.
- Universal decoders for both users, achieving the same error exponents.
- Significant improvement and simplification of earlier results.
- Gallager–style lower bounds for both users.
- Expurgated exponents (joint work also with N. Weinberger, 2019).
Example 5: Channel Decoding with VQ’ed Codewords

- Rate-$R_c$ “codebook” of $y$’s, quantized versions of corresponding $x$’s.
- Motivation: biometric identification (enrollment vs. authentication).
- Objectives: ensemble performance; universal decoding.
- Difficulty: the effective channel, $\{P(z|y)\}$, is complicated:

$$P(z|y_m) = \frac{P(y_m, z)}{P(y_m)} = \frac{\sum x G(x) W(z|x) I\{f(x) = y_m\}}{\sum x G(x) I\{f(x) = y_m\}}$$
Main contributions:

- Exponentially tight bound on the ensemble performance.
- Improvement relative to Dasarathy & Draper (2011).
- Universal decoder a.g.a. ML decoder ($\forall x, z : W(z|x) > 0$).
- Also a.g.a. any decoder that depends on joint empirical statistics ($\forall W$).
- A good approximation to the channel $\{P(z|y)\}$. 
Example 5: Decoding with VQ’ed Codewords (Cont’d)

Ensemble of VQ’s:

- ∀ input type, $Q_X$, choose $Q_{Y|X}$ (s.t. compression constraints).
- Randomly draw $e^{nR_Q}$ vectors from $\mathcal{T}(Q_Y)$, with $R_Q = I_Q(X;Y) + \Delta$.
- Randomly rank all members of every $\mathcal{T}(Q_{Y|X}|x)$.
- Let $M(x, y) =$ rank of $y \in \mathcal{T}(Q_{Y|X}|x)$.
- Code ensemble: random codebook + random rank function.
- Quantize $x$ to $y \in \mathcal{T}(Q_{Y|X}|x) \cap$ code with the smallest $M(x, y)$. 
Example 5: Decoding with VQ’ed Codewords (Cont’d)

- For most codes in the ensemble, we can approximate

\[ P(y_m) = \sum_x G(x) \cdot I\{f(x) = y_m\} = \exp\{-n\alpha(\hat{P}_y)_m\}, \]  

where \(\alpha(\cdot)\) has a certain single–letter formula.

- The proposed modified MMI decoder is of the form

\[ \hat{m} = \arg\min_m \left\{ \log N(y_m|z) - n\alpha(\hat{P}_y)_m \right\}, \]  

where

\[ N(y_m|z) = \left| \mathcal{T}(y_m|z) \cap \mathcal{C} \right|, \]  

\(\mathcal{C}\) being the VQ code.
Some Other Works

- Improved bounds for erasure/list decoding (2008).
- The interference channel (w. Etkin & Ordentlich, 2010).
- The broadcast channel (w. Kaspi, 2011).
- Exact bounds for erasure/list decoding (w. Somekh–Baruch, 2011).
- Codeword or noise? (w. Weinberger, 2014).
- Optimal bin index decoding (2014).
- Correct wiretapper decoding (2014).
- Universal source/channel with SI (2016).
- Simplified erasure/list decoding (w. Weinberger, 2017).
- Improved exponents for the IFC (w. Huleihel, 2017).
Some Other Works (Cont’d)

- Exact secrecy exponents (w. Bastani-Parizi & Telatar, 2017).
- Exact exponents & universal decoding for the ABC (w. Averbuch, 2017).
- 2nd order & moderate deviations in error+erasure (Hayashi & Tan, 2015).
- Residual uncertainties under Rényi entropies (Hayashi & Tan, 2016).
Future Challenges and Open Problems

- Handling ensembles of linear/lattice/convolutional/ LDPC codes, etc.
- Further results on typical random codes (multi-user configurations).
- Simplify optimization problems (e.g., Gallager–style bounds).
- A more solid theory for the extended MoT (for exponential families).
Thank U 4 Coming & Listening!