Error probability bounds in information theory:

Role of structure, performance criteria and decision rules

Eli Haim

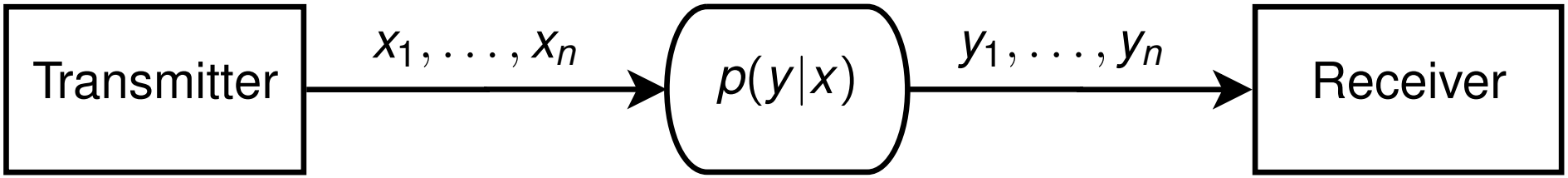
Tel Aviv University

January 21, 2018

Outline

- Introduction: Error exponent for single user channel
- Overview of linear codes in network problems
- Contribution I: Distributed expurgation using structured codes for network problems – terminals use **different** linear codes
- Contribution II: Distributed hypothesis testing using structured codes – terminals use **same** linear code

Single-User Channel



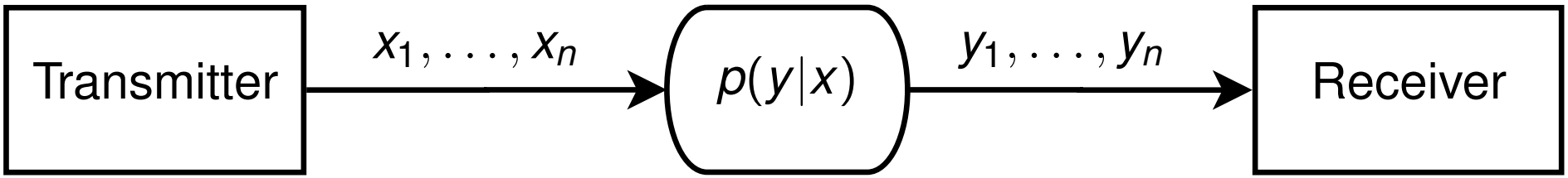
Memoryless channel

$$p\left(y_1, \ldots, y_n | x_1, \ldots, x_n\right) = \prod_{t=1}^{n} p\left(y_t | x_t\right)$$

Basic definitions:

- Blocklength n: number of channel uses
- Codebook $\mathcal{C}$: a set of $M = 2^{nR}$ codewords (vectors of length n)
- Average Error Probability: $P_e = \mathbb{P}\left(\hat{\mathbf{C}} \neq \mathbf{C}\right)$, where $\mathbf{C} \sim \text{Uniform}\left(\mathcal{C}\right)$

Single-User Channel



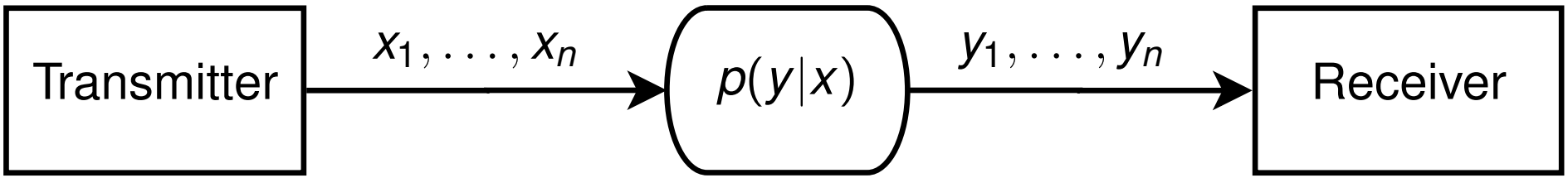
Memoryless channel

$$p\left(y_1, \ldots, y_n | x_1, \ldots, x_n\right) = \prod_{t=1}^{n} p\left(y_t | x_t\right)$$

Basic tradeoff:

Tradeoff between number of codewords, blocklength and average error probability

Single-User Channel



Memoryless channel

$$p\left(y_1, \ldots, y_n | x_1, \ldots, x_n\right) = \prod_{t=1}^{n} p\left(y_t | x_t\right)$$

First-Order (Capacity): asymptotics in blocklength

Capacity C: Highest achievable rate with *vanishing* P_e as $n \to \infty$

Shannon Theory

- Random Code: Symbol-wise (and codeword-wise) i.i.d. $p(x)$
- Information Density:

$$i(X;Y) \stackrel{\triangle}{=} \log \frac{p(X,Y)}{p(X)p(Y)}$$

- Mutual Information:

$$I(X;Y) \stackrel{\triangle}{=} \mathbb{E}i(X;Y)$$

Shannon's Channel Coding Theorem ['48] (first-order characterization)

$$C = \max_{p(x)} I(X;Y)$$

- maximization over all input distributions $p(x)$

Tradeoff: Refined Analysis

- There is a long history of finite blocklength bounds: Elias, Feinstein, Gallager, ...
- Polyanskiy et al. [2010] gave two simple achievability bounds (DT & RCU). Disturbing point: neither dominates
- We have resolved this issue (but not in this talk...)
- Asymptotic analysis: the error event amounts to (except for low rates)

$$i(X^n; Y^n) \triangleq \frac{1}{n} \sum_{k=1}^{n} i(X_k; Y_k) < R$$

Asymptotic Bounds on the Information Density

The following asymptotics are with respect to the blocklength (for high rates):

- Central Limit Theorem (CLT): good for high P_e, dispersion [Strassen 1962, Polyanskiy et al. 2010]
- We have derived results regarding the extension to network problems (but not in this talk...)
- Large Deviations Principle (LDP): good for low P_e, exponent

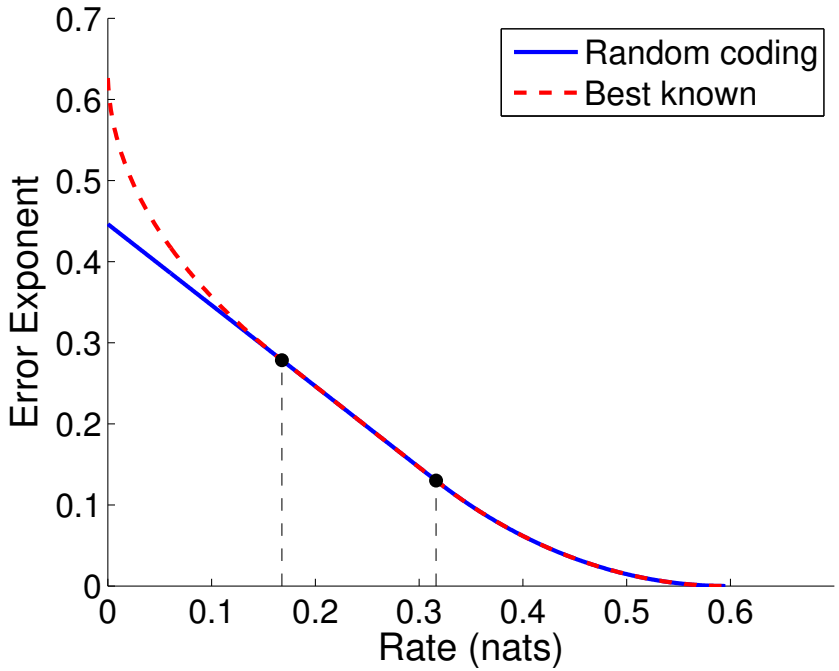
$$\Pr\{i(X^n; Y^n) < R\} \leq \exp\{-nE(R)\}$$

- Similar lower bounds are known

Error Exponent: Code Structure May Matter

- High rates: typical error due to a "bad" channel $i(X^n; Y^n) < R$. Random coding achieves the exponent
- Low rates: typical error due to "bad" codewords (e.g. for BSC, minimum distance dominates)
- Can be solved by expurgation of random codes, or (almost all) linear codes
- Who cares about expurgation? For almost noiseless (binary input) channels

$$\frac{R_{\text{ex}}}{C} \xrightarrow[c \to 1]{} 1$$



Outline

- Introduction: Error exponent for single user channel
- Overview of linear codes in network problems
- Contribution I: Distributed expurgation using structured codes for network problems – terminals use **different** linear codes
- Contribution II: Distributed hypothesis testing using structured codes – terminals use **same** linear code

But First: Why Linear Codes in Single-User Channel?

- Whenever uniform distribution is optimal, linear codes achieve capacity, exponents, dispersion
- But no theoretical gain
- Historically, interest was due to practical (complexity) advantages

Why Linear Codes in Networks?

Contribution II (in this talk...)

- Recent interest, reviving a theme introduced by Körner-Marton 1979: first-order (capacity) advantage in some network settings (Nazer & Gastpar, Wilson et al., Philisof et al., . . .)
- In this work: distributed hypothesis testing
- Terminals use **the same** linear code

Contribution I (in this talk...)

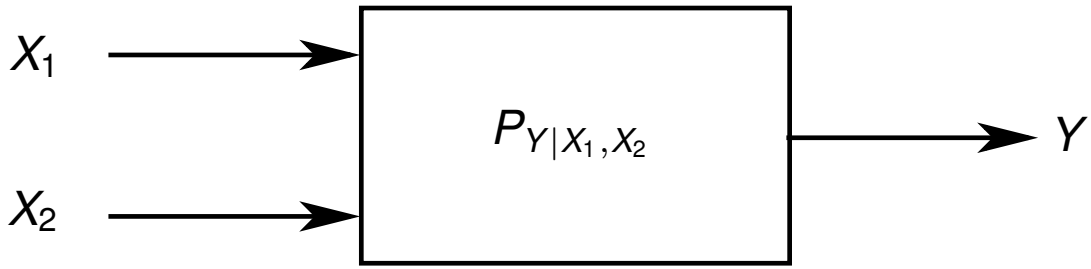
- Error-probability advantage in network settings (even when no first-order gain) – multiple-access (MAC) channel
- Terminals use **different** linear codes
- The prospect for such an improvement was hinted to in a distributed source coding context by Csiszár [1982, "Linear Codes for Sources and Source Networks: Error Exponents, Universal Coding"]

Outline

- Introduction: Error exponent for single user channel
- Overview of linear codes in network problems
- Contribution I: Distributed expurgation using structured codes for network problems – terminals use **different** linear codes
- Contribution II: Distributed hypothesis testing using structured codes – terminals use **same** linear code

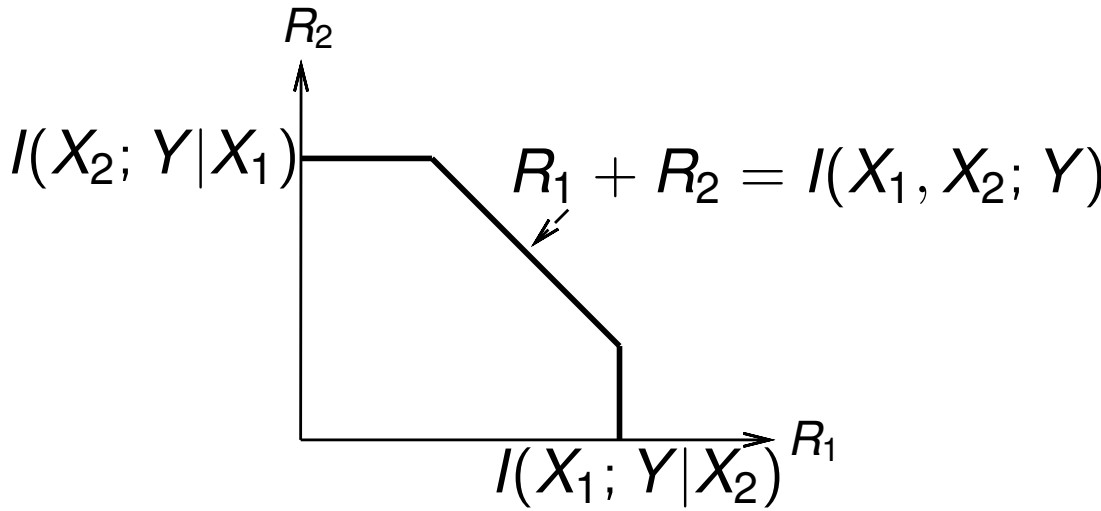
MAC Channel

- For simplicity 2 users



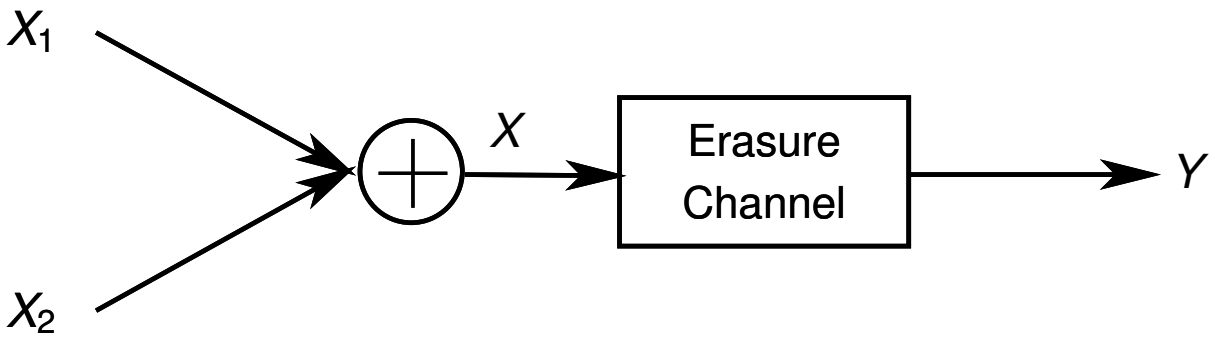
- Capacity region: the closure of the convex-hull of all (R_1, R_2) satisfying:

$$R_1 \leq I(X_1; Y|X_2)$$
$$R_2 \leq I(X_2; Y|X_1)$$
$$R_1 + R_2 \leq I(X_1, X_2; Y),$$



over some product distribution $p(x_1, x_2) = p(x_1)p(x_2)$

Toy Example: Erasure-Additive MAC Channel



Obvious bounds on P_e

- Lower bound: single-user erasure channel
- Upper bound: same with half blocklength (time sharing)

Is any of these bounds tight?

What Can Be Achieved Using Random Codes?

- Slepian & Wolf [’73], Gallager [’85]
- Receiver’s perspective: sum of codebooks, $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$
- For random codes: summation preserves pairwise independence, thus most standard bounds (RCU, DT, dispersion, random exponent) hold
- Codebook structure (e.g. minimum distance) is *not* preserved
- But recall that minimum distance dictates error exponent at low rates
- Expurgation attempts recently by Nazari et al.: expurgate one user (even for MAC channel with many users)

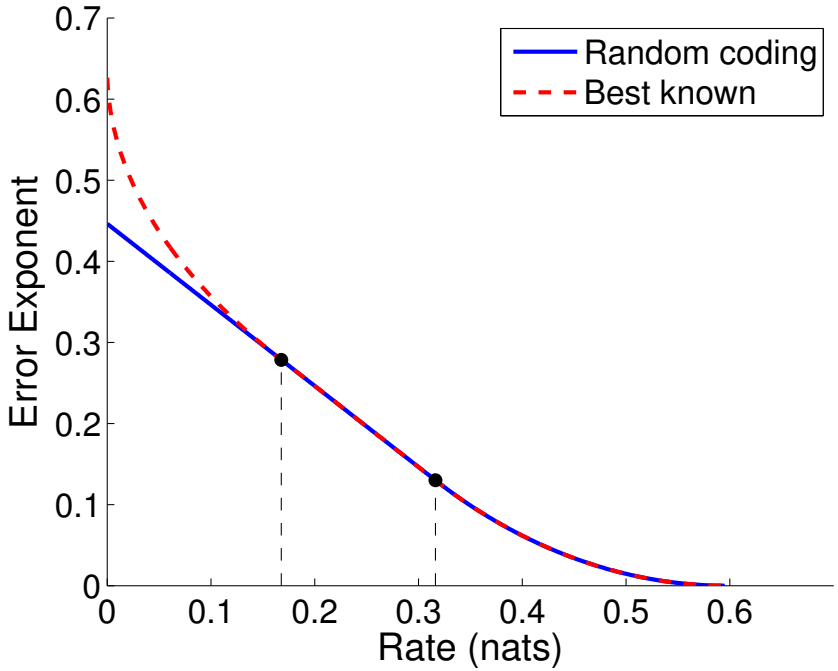
Solution: Use Linear Codes

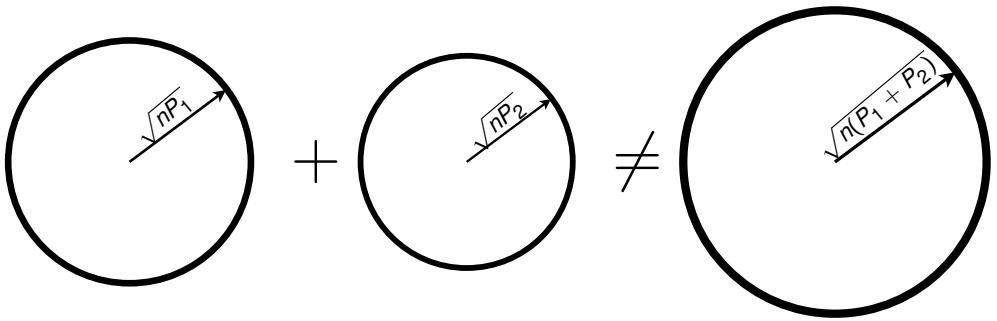
- Create a linear sum-codebook (recall: inherently expurgated)
- Simply split the generating matrix between users
- At the receiver, the summation is *indistinguishable* from a single user channel with the sum-rate
- Performance *identical* to single user with the sum rate
- Any performance that is attainable via linear codes over the single-user channel is also attainable for the considered MAC

- The generation process is equivalent to generating two different linear codes

The Error Exponent of MAC Channels

- In toy example: single-user random+expurgated exponents are achievable
- Extends to any MAC channel that is finite-field summation + single-user channel (e.g., BSC MAC)
- Advantage for any “similar” channel (by continuity)
- AWGN MAC channel - constraints are a challenge.
- For certain parameters - improving on Gallager [’85]
- General case: wide open.

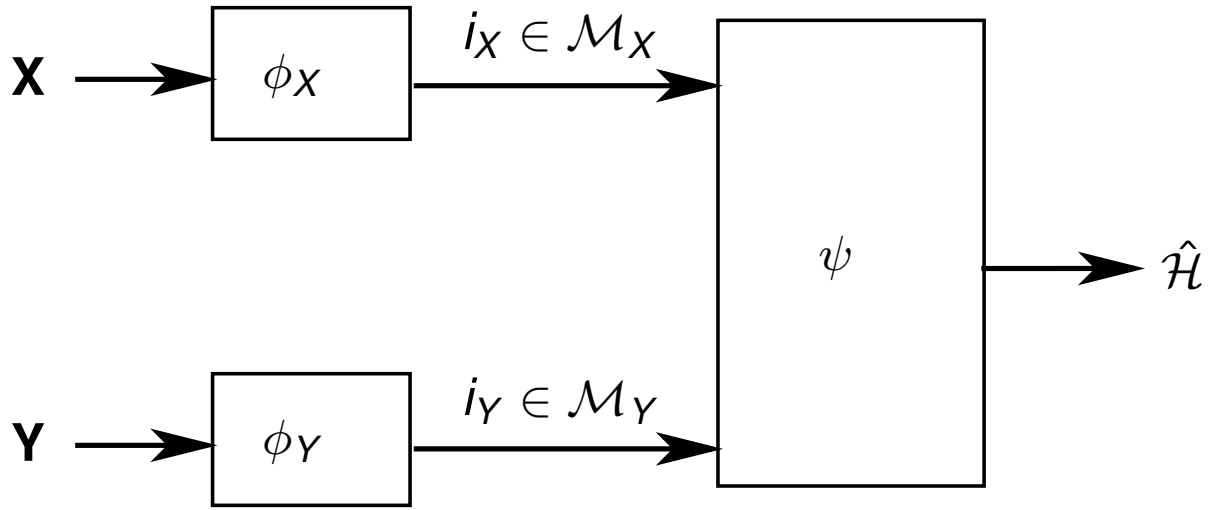




Outline

- Introduction: Error exponent for single user channel
- Overview of linear codes in network problems
- Contribution I: Distributed expurgation using structured codes for network problems – terminals use **different** linear codes
- Contribution II: Distributed hypothesis testing using structured codes – terminals use **same** linear code

Distributed Hypothesis Testing [Berger '79]



$$\mathcal{H}_0 : (\mathbf{X}, \mathbf{Y}) \sim \text{i.i.d. } P_0(x, y)$$
$$\mathcal{H}_1 : (\mathbf{X}, \mathbf{Y}) \sim \text{i.i.d. } P_1(x, y)$$

- Rates: $R_X = 1/n \cdot \log |\mathcal{M}_X|$, $R_Y = 1/n \cdot \log |\mathcal{M}_Y|$
- Error probabilities $\{\epsilon_0\}$, $\{\epsilon_1\}$ as in standard hypothesis testing
- But now, there is a tradeoff between rates, error probabilities and blocklength
- Long history: Ahlswede & Csiszár '81, '86, Han '87, Shalaby & Papamarcou '92, Shimokawa et al. '94, Han & Amari '98, Rahman & Wagner 2012...

Rate-Exponents Tradeoff

- For (a sequence of) error probabilities $\{\epsilon_0(n)\}$, $\{\epsilon_1(n)\}$, the exponential decay rates are defined as:

$$E_i = \liminf_{n\to\infty} -\frac{1}{n}\log \epsilon_i(n)$$

- Goal: Characterize the achievable region of (E_0, E_1) pairs subject to the rate constraints
- Two extreme (and natural) cases:
  - Side-information case: R_Y unconstrained
  - Symmetric rate constraints: $R_X = R_Y = R$

Binary Symmetric Case

- Under both hypotheses, (X, Y) is a doubly-symmetric source
- Noise / difference sequence: $Z = (X + Y) \bmod 2$
- $\mathcal{H}_i$: Z is Bernoulli-p_i, where $p_0 < p_1 \leq 1/2$
- The key point is that the type of $\mathbf{Z}$ is a sufficient statistic
- For $R \geq 1$, the unconstrained exponents are achievable: For any $p_0 \leq s \leq p_1$,

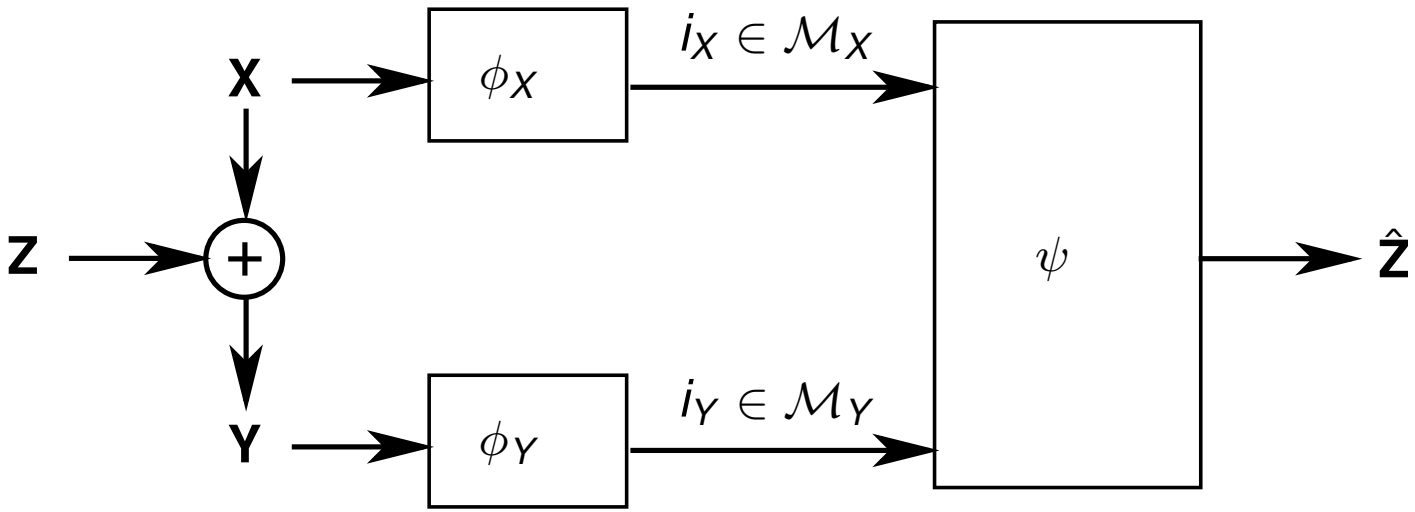
$$E_0(s) = D_b(s\|p_0)$$
$$E_1(s) = D_b(s\|p_1)$$

where $D_b(\cdot)$ is the binary KL divergence

Side-Information Case: Random Binning [Shimokawa et al. '94]

- Base on Slepian-Wolf coding (random binning)
- Decoder recovers the sources first (decoding similar to BSC decoder with "channel" noise $\mathbf{Z}$)
- Key observation: under a binning error, typically the reconstruction will *not* fall in the vicinity of $\mathbf{Y}$
- This gives a non-trivial exponent pair
- Can be improved by using quantization
- We have further improvements using geometric analysis (but not in this talk...)
- But what about the symmetric constraints case?

Körner-Marton Reminder



- Setting: Suppose we wish to compress the difference $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$ ($\mathbf{X}$ and $\mathbf{Y}$ BSS pair) in a distributed manner
  - Using SW (first reconstructing $\mathbf{X}, \mathbf{Y}$) requires: $R_X = H(\mathbf{Z}), R_Y = H(\mathbf{Y})$
  - But KM showed that it suffices to require: $R_X = H(\mathbf{Z}), R_Y = H(\mathbf{Z})$
- Again: linear codes are the way to go

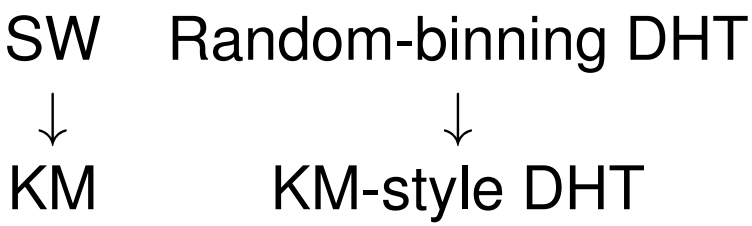
Körner-Marton Coding Scheme (crash course)

- Let H be a parity-check matrix of a linear code of rate R
- $\phi_X(\mathbf{X}) = H\mathbf{X}$, $\phi_Y(\mathbf{Y}) = H\mathbf{Y}$ have rate $1 - R$
- The decoder evaluates $H\mathbf{X} + H\mathbf{Y} = H\mathbf{Z}$
- Finally, a syndrome decoder is used
- $\hat{\mathbf{Z}} = \mathbf{Z}$ if and only if $\mathbf{Z}$ is inside the basic "Voronoi" cell
- Same error event as in the side-information (SW coding) case

Main Result

Achievable tradeoff for symmetric constraints

- We can leverage KM coding to the distributed hypothesis problem
- (Essentially the) same exponents are therefore achievable, as in the side-information case



Thank you for your attention!