A Journey in Information Theory: From Second-Order Asymptotics to Common Information

Vincent Y. F. Tan

Department of Mathematics and
Department of Electrical and Computer Engineering
National University of Singapore (NUS)

Plenary at ISITA (Taipei)
Nov 2024

Outline

Outline

Outline

Outline

From Ph.D. work to Information Theory

From Ph.D. work to Information Theory

- Ph.D. (2011) work was in learning Markov fields/graphical models.

A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree Structures

Vincent Y. F. Tan, *Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, Lang Tong, *Fellow, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—The problem of maximum-likelihood (ML) estimation of discrete tree-structured distributions is considered. Chow and Liu established that ML-estimation reduces to the construction of a maximum-weight spanning tree using the empirical mutual information quantities as the edge weights. Using the theory of large-deviations, we analyze the exponent associated with the error probability of the event that the ML-estimate of the Markov tree structure differs from the true tree structure, given a set of independently drawn samples. By exploiting the fact that the output of ML-estimation is a tree, we establish that the error exponent is equal to the exponential rate of decay of a single dominant *crossover* event. We prove that in this dominant crossover event, a non-neighbor node pair replaces a true edge**

distributions. In this respect, graphical models [2] provide a significant simplification of joint distribution as the distribution can be factorized according to a graph defined on the set of nodes. Many specialized algorithms [3]–[9] exist for exact and approximate learning of graphical models Markov on sparse graphs.

There are many applications of learning graphical models, including clustering and dimensionality reduction. Suppose we have d genetic variables and we would like to group the ones that are similar together. Then the construction of a graphical model provides a visualization of the relationship between genes. Those genes that have high degree are highly correlated

Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures

Vincent Y. F. Tan, *Student Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—The problem of learning tree-structured Gaussian graphical models from independent and identically distributed (i.i.d.) samples is considered. The influence of the tree structure and the parameters of the Gaussian distribution on the learning rate as the number of samples increases is discussed. Specifically, the error exponent corresponding to the event that the estimated tree structure differs from the actual unknown tree structure of the distribution is analyzed. Finding the error exponent reduces to a least-squares problem in the very noisy learning regime. In this regime, it is shown that the extremal tree structure that minimizes the error exponent is the star for any fixed set of correlation coefficients on the edges of the tree. If the magnitudes of all the correlation coefficients are less than 0.63, it is also shown that the tree structure that maximizes the error exponent is the Markov chain. In other words, the star and the**

parameters and are tractable for learning [5] and statistical inference [1], [4].

The problem of maximum-likelihood (ML) learning of a Markov tree distribution from i.i.d. samples has an elegant solution, proposed by Chow and Liu in [5]. The ML tree structure is given by the maximum-weight spanning tree (MWST) with empirical mutual information quantities as the edge weights. Furthermore, the ML algorithm is *consistent* [6], which implies that the error probability in learning the tree structure decays to zero with the number of samples available for learning.

While consistency is an important qualitative property, there is substantial motivation for additional and more quantitative characterization of performance. One such measure, which we

Alan Willsky

From Ph.D. work to Information Theory

- Ph.D. (2011) work was in learning Markov fields/graphical models.

A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree Structures

Vincent Y. F. Tan, *Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, Lang Tong, *Fellow, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—The problem of maximum-likelihood (ML) estimation of discrete tree-structured distributions is considered. Chow and Liu established that ML-estimation reduces to the construction of a maximum-weight spanning tree using the empirical mutual information quantities as the edge weights. Using the theory of large-deviations, we analyze the exponent associated with the error probability of the event that the ML-estimate of the Markov tree structure differs from the true tree structure, given a set of independently drawn samples. By exploiting the fact that the output of ML-estimation is a tree, we establish that the error exponent is equal to the exponential rate of decay of a single dominant *crossover* event. We prove that in this dominant crossover event, a non-neighbor node pair replaces a true edge**

distributions. In this respect, graphical models [2] provide a significant simplification of joint distribution as the distribution can be factorized according to a graph defined on the set of nodes. Many specialized algorithms [3]–[9] exist for exact and approximate learning of graphical models Markov on sparse graphs.

There are many applications of learning graphical models, including clustering and dimensionality reduction. Suppose we have d genetic variables and we would like to group the ones that are similar together. Then the construction of a graphical model provides a visualization of the relationship between genes. Those genes that have high degree are highly correlated

Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures

Vincent Y. F. Tan, *Student Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—The problem of learning tree-structured Gaussian graphical models from independent and identically distributed (i.i.d.) samples is considered. The influence of the tree structure and the parameters of the Gaussian distribution on the learning rate as the number of samples increases is discussed. Specifically, the error exponent corresponding to the event that the estimated tree structure differs from the actual unknown tree structure of the distribution is analyzed. Finding the error exponent reduces to a least-squares problem in the very noisy learning regime. In this regime, it is shown that the extremal tree structure that minimizes the error exponent is the star for any fixed set of correlation coefficients on the edges of the tree. If the magnitudes of all the correlation coefficients are less than 0.63, it is also shown that the tree structure that maximizes the error exponent is the Markov chain. In other words, the star and the**

parameters and are tractable for learning [5] and statistical inference [1], [4].

The problem of maximum-likelihood (ML) learning of a Markov tree distribution from i.i.d. samples has an elegant solution, proposed by Chow and Liu in [5]. The ML tree structure is given by the maximum-weight spanning tree (MWST) with empirical mutual information quantities as the edge weights. Furthermore, the ML algorithm is *consistent* [6], which implies that the error probability in learning the tree structure decays to zero with the number of samples available for learning.

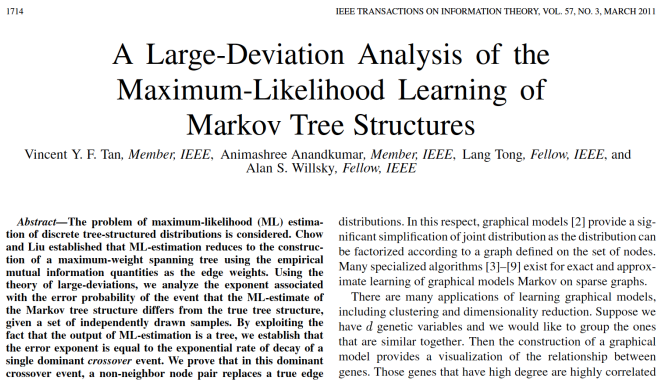
While consistency is an important qualitative property, there is substantial motivation for additional and more quantitative characterization of performance. One such measure, which we

Alan Willsky

- Graphical models and machine learning were very popular then.

From Ph.D. work to Information Theory

- Ph.D. (2011) work was in learning Markov fields/graphical models.



1714 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 57, NO. 3, MARCH 2011

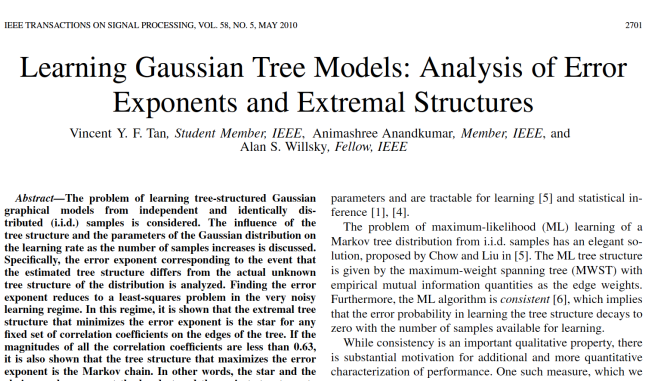
A Large-Deviation Analysis of the Maximum-Likelihood Learning of Markov Tree Structures

Vincent Y. F. Tan, *Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, Lang Tong, *Fellow, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—The problem of maximum-likelihood (ML) estimation of discrete tree-structured distributions is considered. Chow and Liu established that ML-estimation reduces to the construction of a maximum-weight spanning tree using the empirical mutual information quantities as the edge weights. Using the theory of large-deviations, we analyze the exponent associated with the error probability of the event that the ML-estimate of the Markov tree structure differs from the true tree structure, given a set of independently drawn samples. By exploiting the fact that the output of ML-estimation is a tree, we establish that the error exponent is equal to the exponential rate of decay of a single dominant *crossover* event. We prove that in this dominant crossover event, a non-neighbor node pair replaces a true edge**

distributions. In this respect, graphical models [2] provide a significant simplification of joint distribution as the distribution can be factorized according to a graph defined on the set of nodes. Many specialized algorithms [3]–[9] exist for exact and approximate learning of graphical models Markov on sparse graphs.

There are many applications of learning graphical models, including clustering and dimensionality reduction. Suppose we have d genetic variables and we would like to group the ones that are similar together. Then the construction of a graphical model provides a visualization of the relationship between genes. Those genes that have high degree are highly correlated



IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 58, NO. 5, MAY 2010 2701

Learning Gaussian Tree Models: Analysis of Error Exponents and Extremal Structures

Vincent Y. F. Tan, *Student Member, IEEE*, Animashree Anandkumar, *Member, IEEE*, and Alan S. Willsky, *Fellow, IEEE*

***Abstract*—The problem of learning tree-structured Gaussian graphical models from independent and identically distributed (i.i.d.) samples is considered. The influence of the tree structure and the parameters of the Gaussian distribution on the learning rate as the number of samples increases is discussed. Specifically, the error exponent corresponding to the event that the estimated tree structure differs from the actual unknown tree structure of the distribution is analyzed. Finding the error exponent reduces to a least-squares problem in the very noisy learning regime. In this regime, it is shown that the extremal tree structure that minimizes the error exponent is the star for any fixed set of correlation coefficients on the edges of the tree. If the magnitudes of all the correlation coefficients are less than 0.63, it is also shown that the tree structure that maximizes the error exponent is the Markov chain. In other words, the star and the**

parameters and are tractable for learning [5] and statistical inference [1], [4].

The problem of maximum-likelihood (ML) learning of a Markov tree distribution from i.i.d. samples has an elegant solution, proposed by Chow and Liu in [5]. The ML tree structure is given by the maximum-weight spanning tree (MWST) with empirical mutual information quantities as the edge weights. Furthermore, the ML algorithm is *consistent* [6], which implies that the error probability in learning the tree structure decays to zero with the number of samples available for learning.

While consistency is an important qualitative property, there is substantial motivation for additional and more quantitative characterization of performance. One such measure, which we

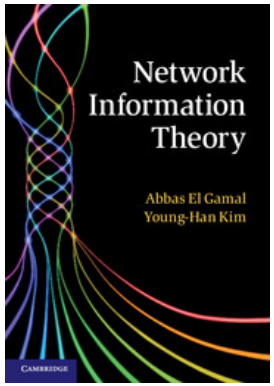
Alan Willsky

- Graphical models and machine learning were very popular then.
- What got me into information theory?

The Real Start of my Foray into Information Theory

The Real Start of my Foray into Information Theory

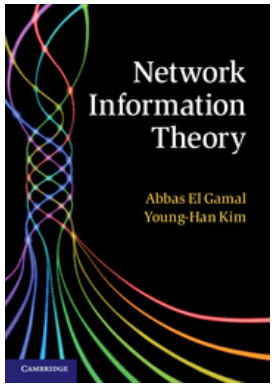
- Summer of 2010



Network Information Theory
by A. El Gamal and Y.-H. Kim

The Real Start of my Foray into Information Theory

- Summer of 2010



Network Information Theory
by A. El Gamal and Y.-H. Kim

- Came across the paper by Ingber and Kochman (DCC 2011)

The Dispersion of Lossy Source Coding

Amir Ingber
Dept. of EE-Systems, TAU
Tel Aviv 69978, Israel
Email: ingber@eng.tau.ac.il

Yuval Kochman
EECS Dept., MIT
Cambridge, MA 02139, USA
Email: yuvalko@mit.edu

Abstract

In this work we investigate the behavior of the minimal rate needed in order to guarantee a given probability that the distortion exceeds a prescribed threshold, at some fixed finite quantization block length. We show that the excess coding rate above the rate-distortion function is inversely proportional (to the first order) to the square root of the block length. We give an explicit expression for the proportion constant, which is given by the inverse Q-function of the allowed excess distortion probability, times the square root of a constant, termed the *excess distortion dispersion*. This result is the dual of a corresponding channel coding result, where the dispersion above is the dual of the channel dispersion. The work treats discrete memoryless sources, as well as the quadratic-Gaussian case.

$$\frac{\sqrt{n}\big(R(\hat{P}_{X^n}, D) - R(P, D)\big)}{\sqrt{V(P, D)}} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

Extension to NIT: The Slepian–Wolf problem

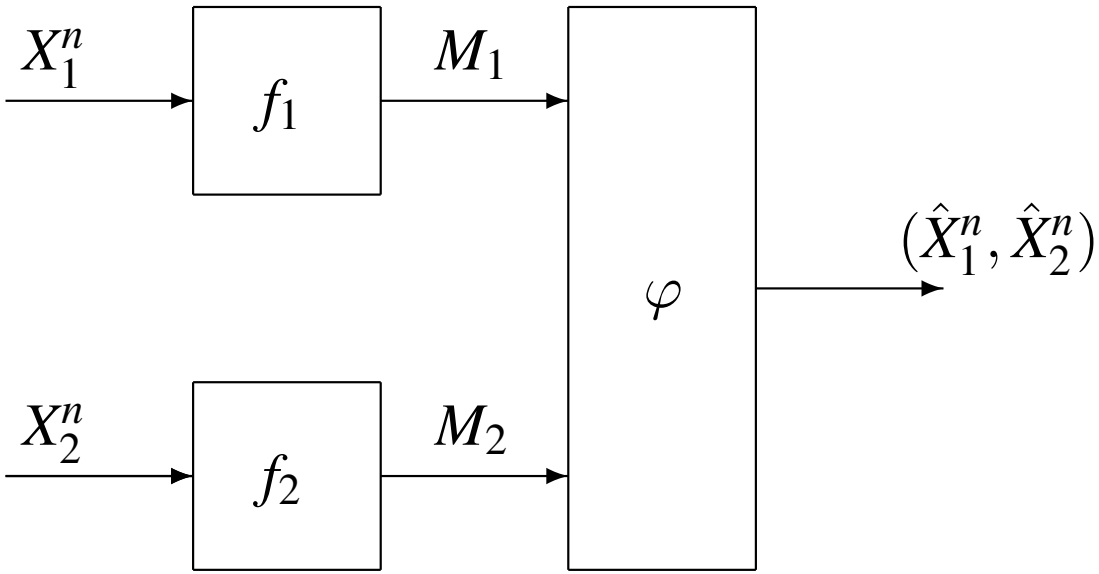


Illustration of the Slepian–Wolf problem.

Extension to NIT: The Slepian–Wolf problem

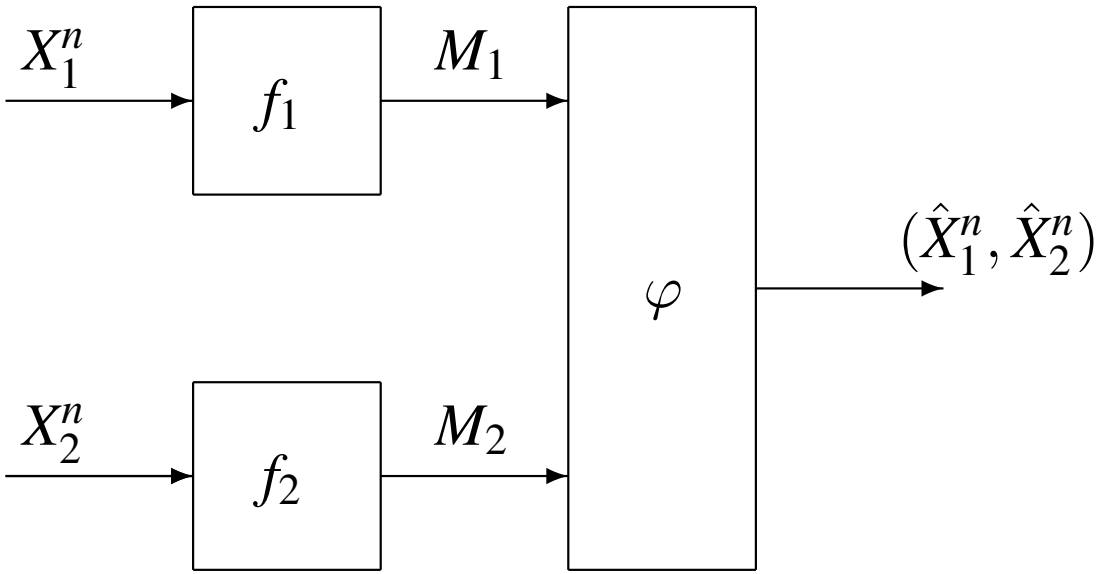


Illustration of the Slepian–Wolf problem.

- Two correlated sources $(X_1^n, X_2^n) \sim \prod_{i=1}^n P_{X_1X_2}(x_{1i}, x_{2i})$.

Extension to NIT: The Slepian–Wolf problem

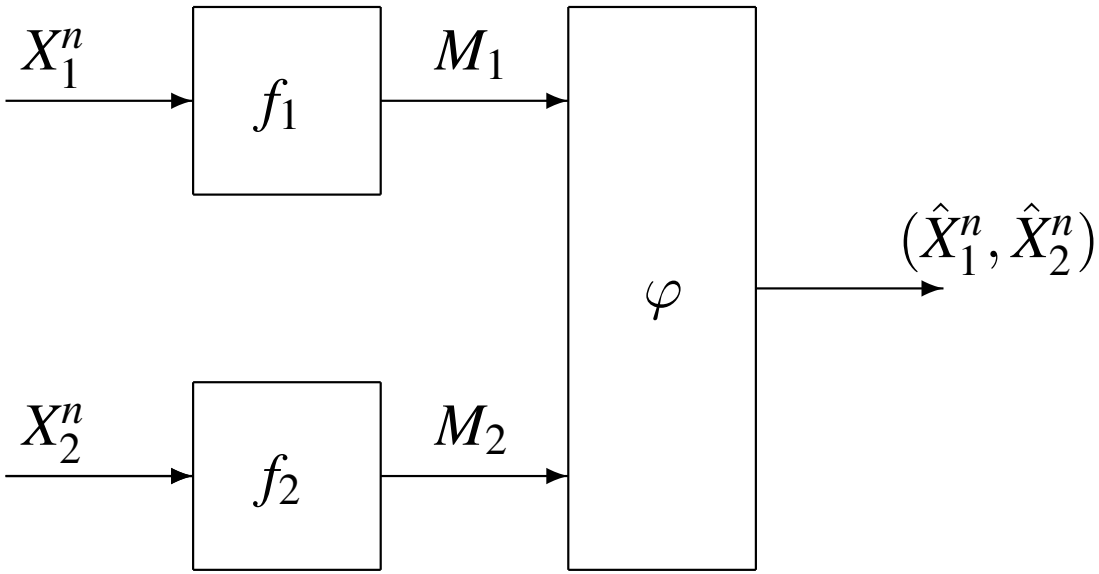


Illustration of the Slepian–Wolf problem.

- Two correlated sources $(X_1^n, X_2^n) \sim \prod_{i=1}^n P_{X_1X_2}(x_{1i}, x_{2i})$.
- Separately encoded

Extension to NIT: The Slepian–Wolf problem

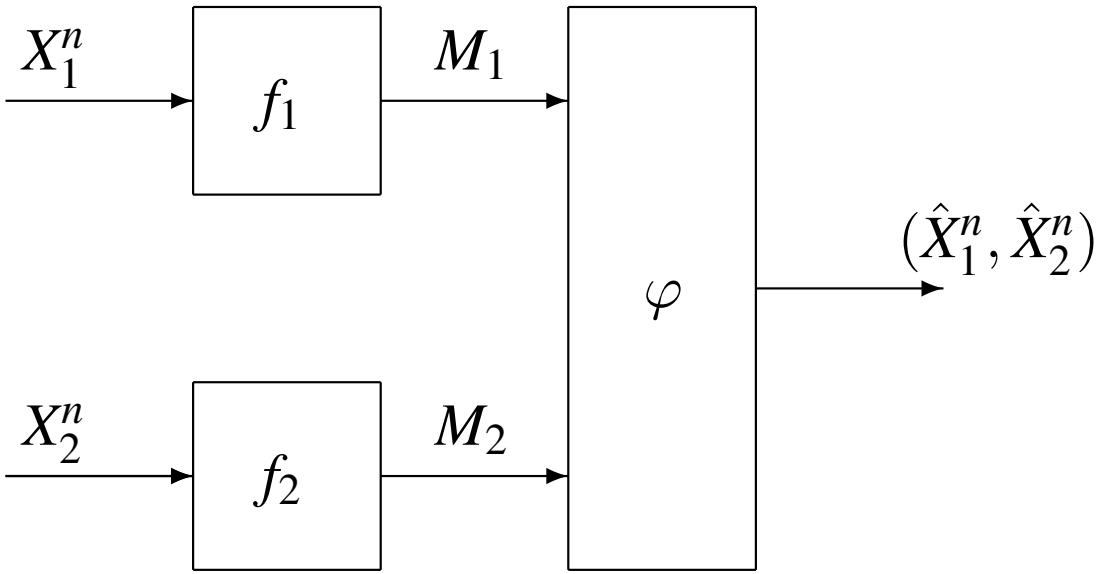


Illustration of the Slepian–Wolf problem.

- Two correlated sources $(X_1^n, X_2^n) \sim \prod_{i=1}^n P_{X_1X_2}(x_{1i}, x_{2i})$.
- Separately encoded
- Both to be decoded at destination

The Slepian–Wolf theorem

- Sources to be compressed to nR_1 and nR_2 bits respectively.

The Slepian–Wolf theorem

- Sources to be compressed to nR_1 and nR_2 bits respectively.
- (R_1, R_2) achievable $\iff$ $\exists$ a sequence of $(2^{nR_1}, 2^{nR_2}, n)$-codes

$$\lim_{n\to\infty} \Pr\left((\hat{X}_1^n, \hat{X}_2^n) \neq (X_1^n, X_2^n)\right) = 0.$$

- $\mathcal{R}(P_{X_1X_2})$ is the set of all achievable (R_1, R_2) pairs.

The Slepian–Wolf theorem

- Sources to be compressed to nR_1 and nR_2 bits respectively.
- (R_1, R_2) achievable $\iff$ $\exists$ a sequence of $(2^{nR_1}, 2^{nR_2}, n)$-codes

$$\lim_{n\to\infty} \Pr\left((\hat{X}_1^n, \hat{X}_2^n) \neq (X_1^n, X_2^n)\right) = 0.$$

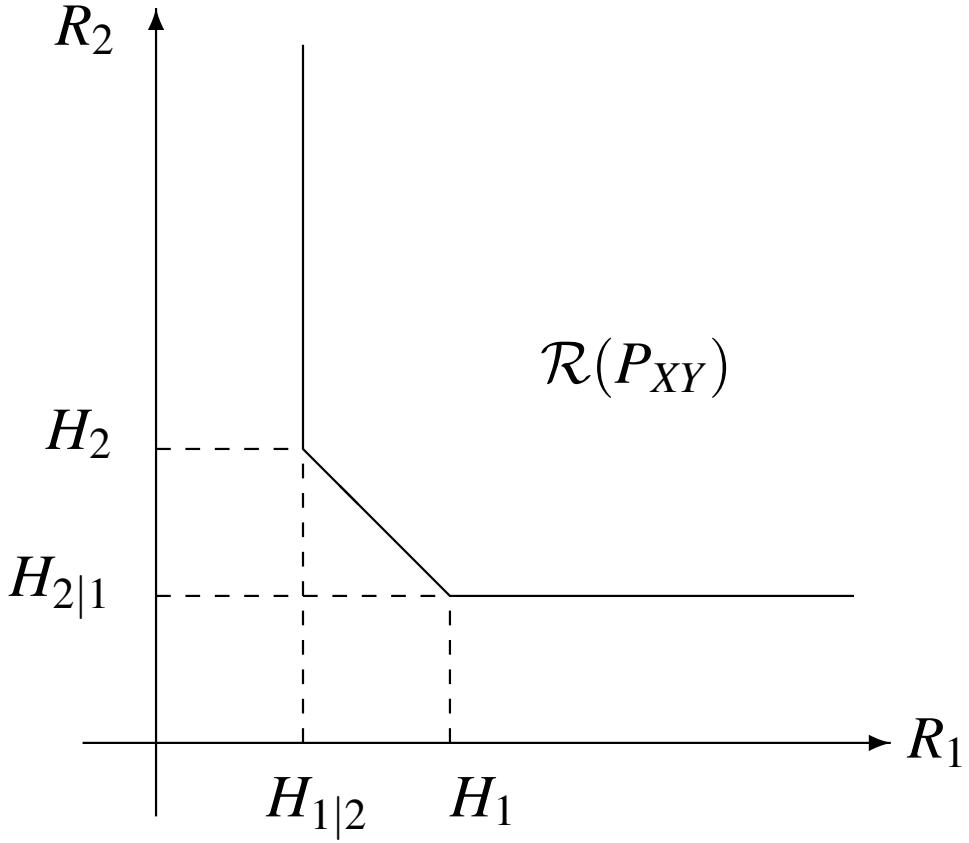
- $\mathcal{R}(P_{X_1X_2})$ is the set of all achievable (R_1, R_2) pairs.
- Slepian and Wolf (1973)

$$\mathcal{R}(P_{XY}) = \{R_1 \geq H(X_1|X_2), R_2 \geq H(X_2|X_1), R_1 + R_2 \geq H(X_1, X_2)\}$$

D. Slepian

J. Wolf

The Slepian-Wolf region



$$R_1 \geq H(X_1|X_2)$$
$$R_2 \geq H(X_2|X_1)$$
$$R_1 + R_2 \geq H(X_1, X_2)$$

Second-Order Asymptotics for Slepian–Wolf

Joint work with Oliver Kosut (postdoc at MIT at that time)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 60, NO. 2, FEBRUARY 2014 881

On the Dispersions of Three Network Information Theory Problems

Vincent Y. F. Tan, *Member, IEEE*, and Oliver Kosut, *Member, IEEE*

***Abstract*— We analyze the dispersions of distributed lossless source coding (the Slepian–Wolf problem), the multiple-access channel, and the asymmetric broadcast channel. For the two-encoder Slepian–Wolf problem, we introduce a quantity known as the entropy dispersion matrix, which is analogous**

While the characterization of capacity regions is a difficult problem in general, there have been positive results for several special classes of networks such as the multiple-access channel [2], [3] and the asymmetric [4] or degraded

Second-Order Asymptotics for Slepian–Wolf

Joint work with Oliver Kosut (postdoc at MIT at that time)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 60, NO. 2, FEBRUARY 2014 881

On the Dispersions of Three Network Information Theory Problems

Vincent Y. F. Tan, *Member, IEEE*, and Oliver Kosut, *Member, IEEE*

***Abstract*— We analyze the dispersions of distributed lossless source coding (the Slepian–Wolf problem), the multiple-access channel, and the asymmetric broadcast channel. For the two-encoder Slepian–Wolf problem, we introduce a quantity known as the entropy dispersion matrix, which is analogous**

While the characterization of capacity regions is a difficult problem in general, there have been positive results for several special classes of networks such as the multiple-access channel [2], [3] and the asymmetric [4] or degraded

- Also contains partial results for MAC and degraded BC.

Second-Order Asymptotics for Slepian–Wolf

Joint work with Oliver Kosut (postdoc at MIT at that time)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 60, NO. 2, FEBRUARY 2014 881

On the Dispersions of Three Network Information Theory Problems

Vincent Y. F. Tan, *Member, IEEE*, and Oliver Kosut, *Member, IEEE*

***Abstract*— We analyze the dispersions of distributed lossless source coding (the Slepian–Wolf problem), the multiple-access channel, and the asymmetric broadcast channel. For the two-encoder Slepian–Wolf problem, we introduce a quantity known as the entropy dispersion matrix, which is analogous**

While the characterization of capacity regions is a difficult problem in general, there have been positive results for several special classes of networks such as the multiple-access channel [2], [3] and the asymmetric [4] or degraded

- Also contains partial results for MAC and degraded BC.
- O. Kosut made a recent major breakthrough for the MAC.

Second-Order Asymptotics for Slepian–Wolf

Joint work with Oliver Kosut (postdoc at MIT at that time)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 60, NO. 2, FEBRUARY 2014 881

On the Dispersions of Three Network Information Theory Problems

Vincent Y. F. Tan, *Member, IEEE*, and Oliver Kosut, *Member, IEEE*

***Abstract*— We analyze the dispersions of distributed lossless source coding (the Slepian–Wolf problem), the multiple-access channel, and the asymmetric broadcast channel. For the two-encoder Slepian–Wolf problem, we introduce a quantity known as the entropy dispersion matrix, which is analogous**

While the characterization of capacity regions is a difficult problem in general, there have been positive results for several special classes of networks such as the multiple-access channel [2], [3] and the asymmetric [4] or degraded

- Also contains partial results for MAC and degraded BC.
- O. Kosut made a recent major breakthrough for the MAC.

3552 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 68, NO. 6, JUNE 2022

A Second-Order Converse Bound for the Multiple-Access Channel via Wringing Dependence

Oliver Kosut, *Member, IEEE*

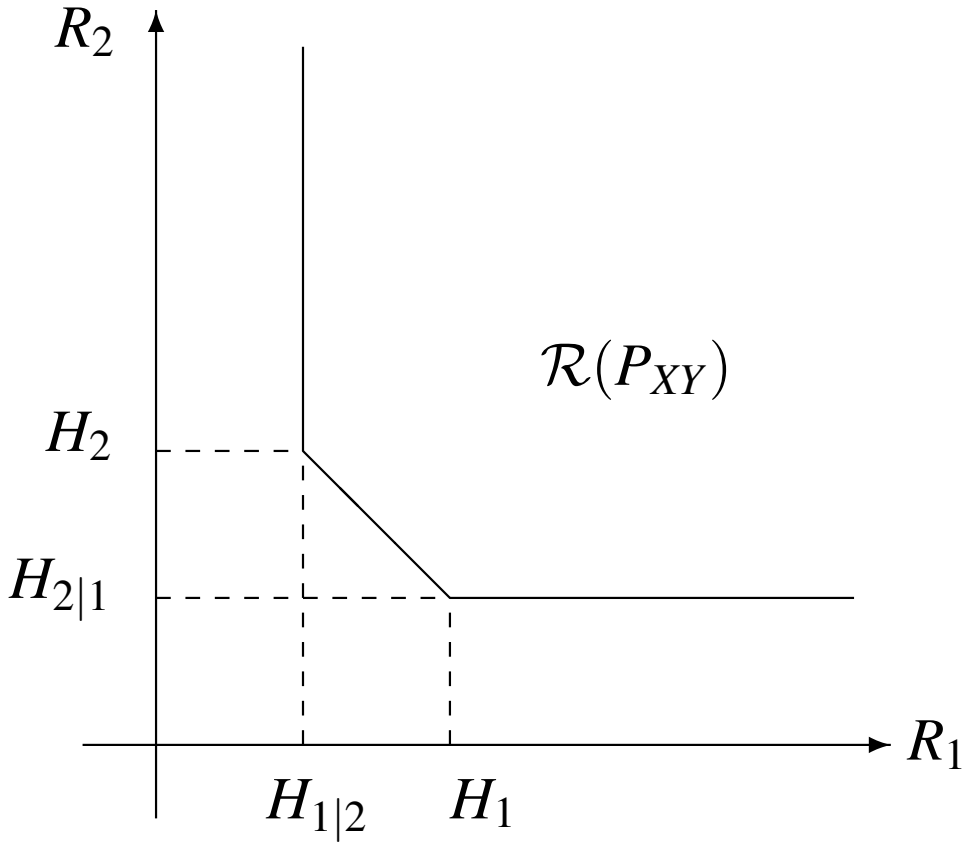
Setup for Second-Order Asymptotics for Slepian–Wolf

Setup for Second-Order Asymptotics for Slepian–Wolf

Note that we're operating on the boundary of the SW region!

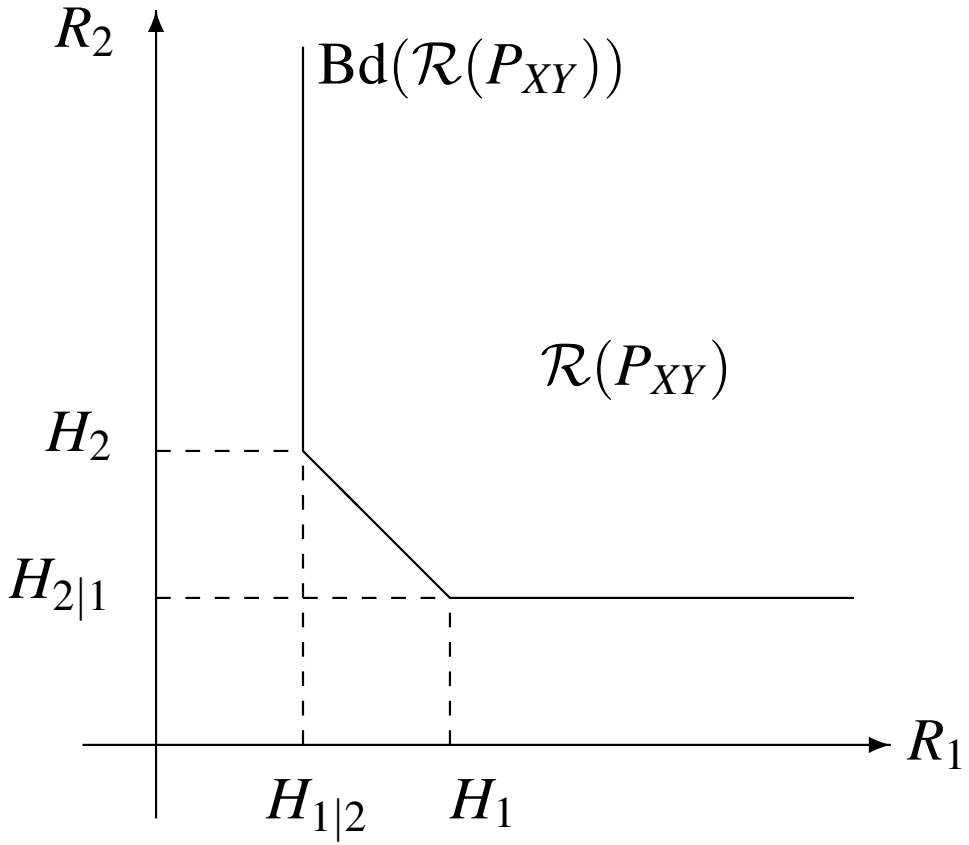
Setup for Second-Order Asymptotics for Slepian–Wolf

Note that we're operating on the boundary of the SW region!



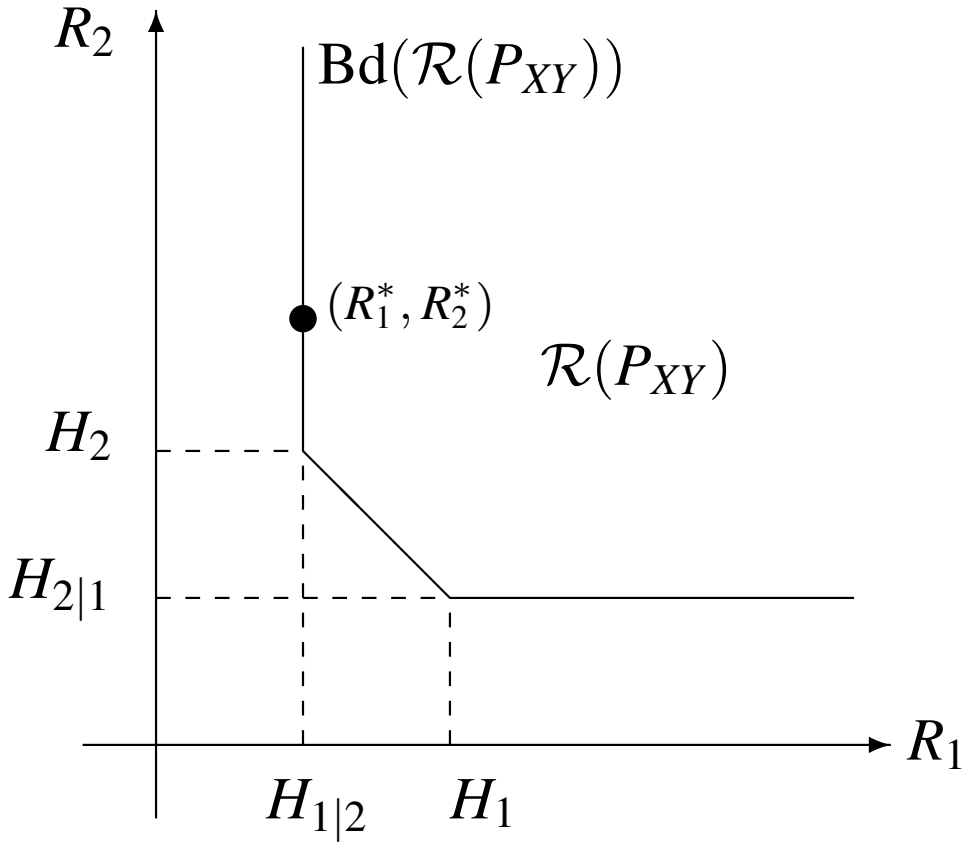
Setup for Second-Order Asymptotics for Slepian–Wolf

Note that we're operating on the boundary of the SW region!



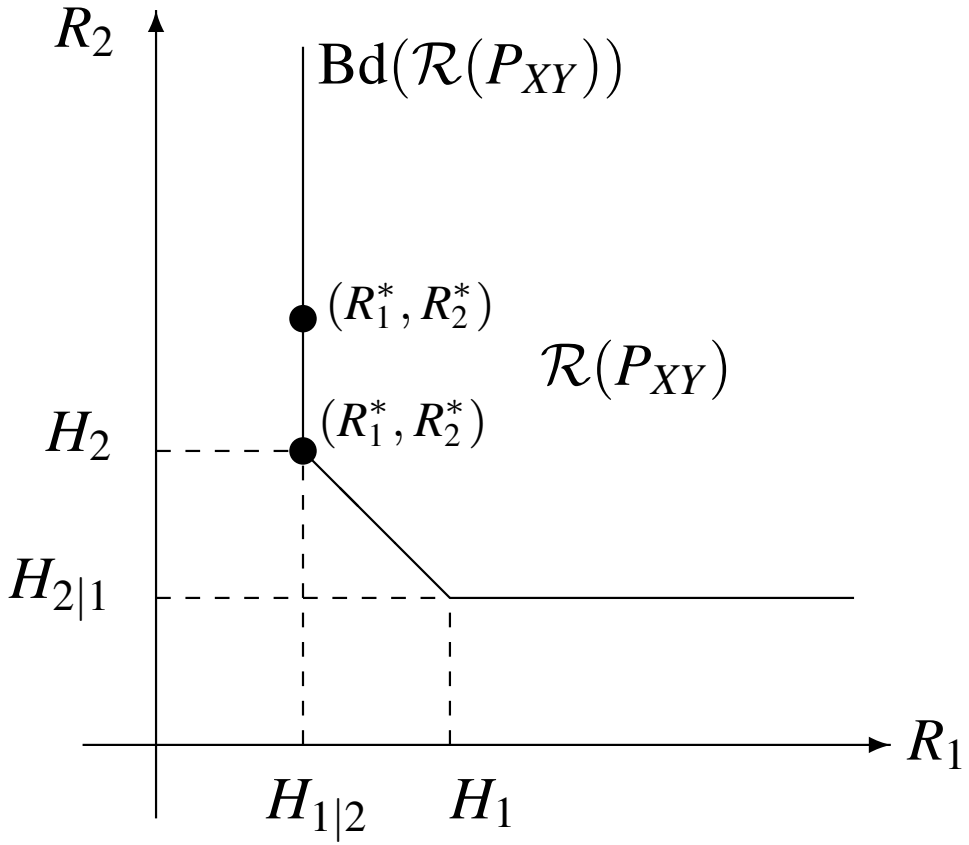
Setup for Second-Order Asymptotics for Slepian–Wolf

Note that we're operating on the boundary of the SW region!



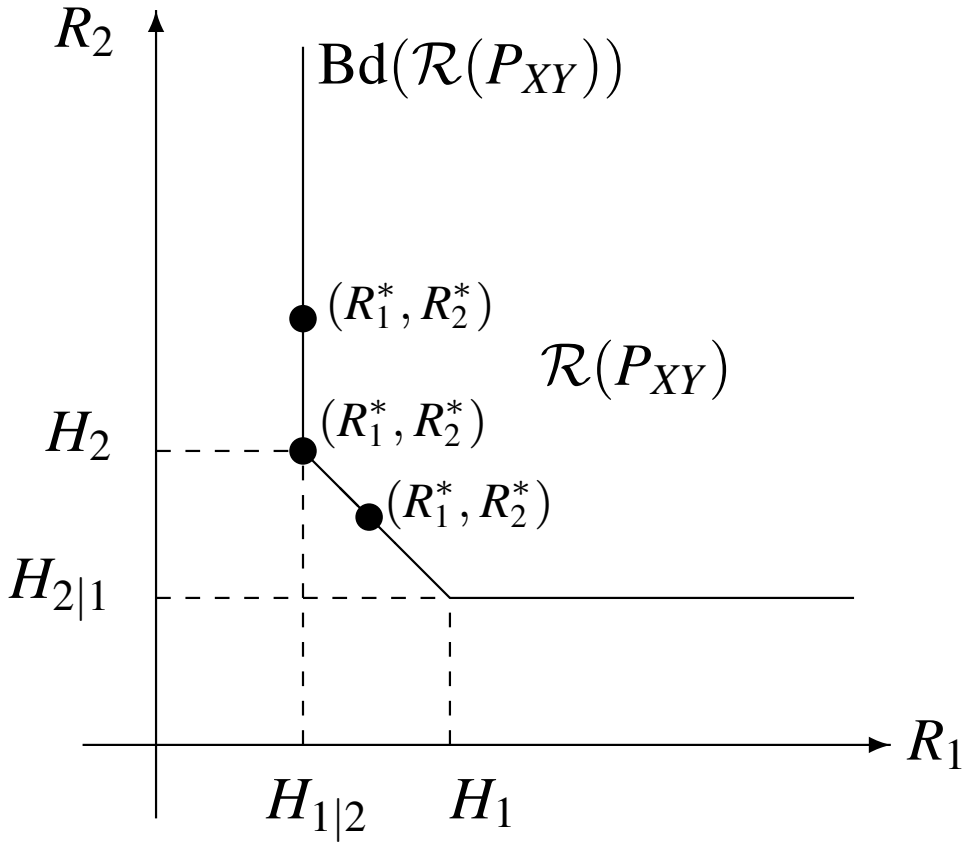
Setup for Second-Order Asymptotics for Slepian–Wolf

Note that we're operating on the boundary of the SW region!



Setup for Second-Order Asymptotics for Slepian–Wolf

Note that we're operating on the boundary of the SW region!



Definitions for Second-Order Asymptotics for SW

Definitions for Second-Order Asymptotics for SW

- $(L_1, L_2) \in \mathbb{R}^2$ is $(R_1^*, R_2^*, \varepsilon)$-achievable

Definitions for Second-Order Asymptotics for SW

- $(L_1, L_2) \in \mathbb{R}^2$ is $(R_1^*, R_2^*, \varepsilon)$-achievable
  $\iff$ sequence of $(n, M_{1n}, M_{2n}, \varepsilon_n)$-codes such that

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}}(\log M_{1n} - nR_1^*) \le L_1$$

$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}}(\log M_{2n} - nR_2^*) \le L_2$$

and

$$\limsup_{n\to\infty} \varepsilon_n \le \varepsilon.$$

Definitions for Second-Order Asymptotics for SW

- $(L_1, L_2) \in \mathbb{R}^2$ is $(R_1^*, R_2^*, \varepsilon)$-achievable
  $\iff$ sequence of $(n, M_{1n}, M_{2n}, \varepsilon_n)$-codes such that

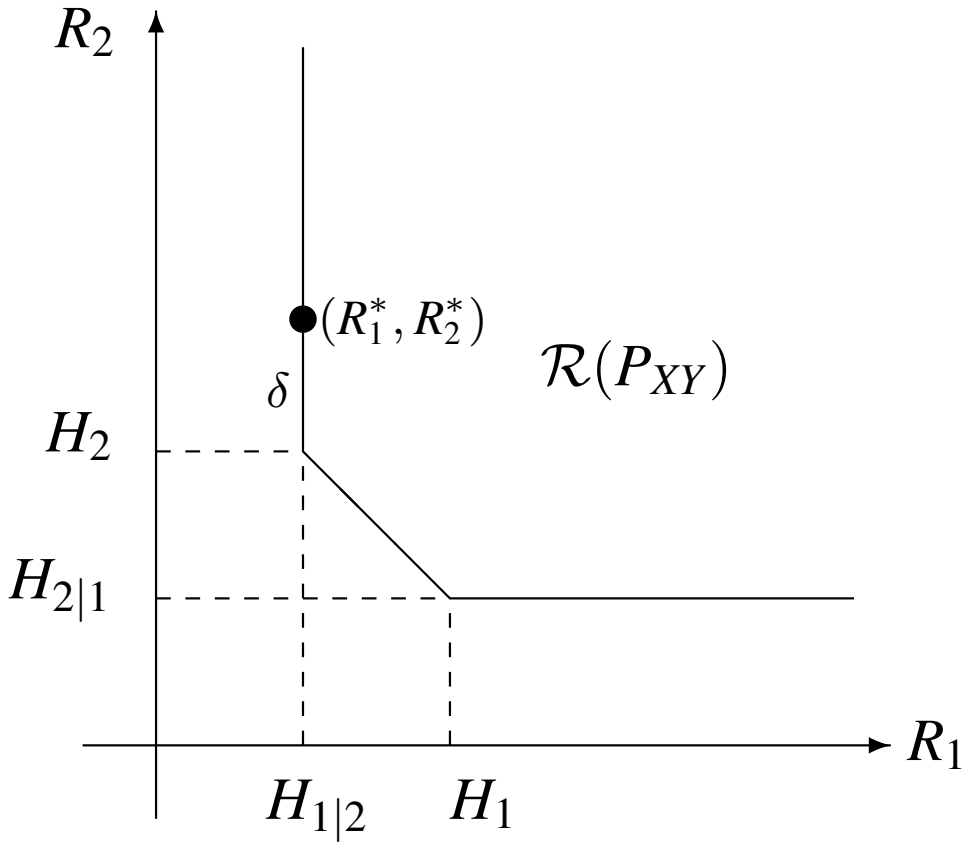
$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}}(\log M_{1n} - nR_1^*) \le L_1$$
$$\limsup_{n\to\infty} \frac{1}{\sqrt{n}}(\log M_{2n} - nR_2^*) \le L_2$$

and

$$\limsup_{n\to\infty} \varepsilon_n \le \varepsilon.$$

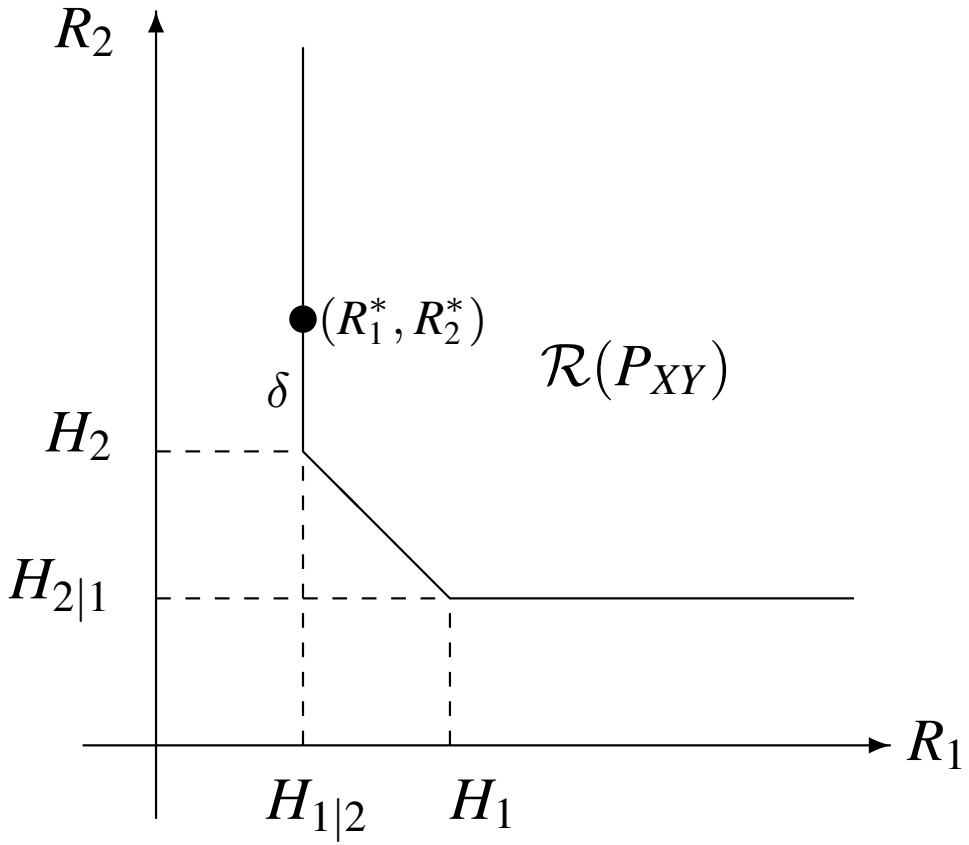
- $\mathcal{L}(\varepsilon; R_1^*, R_2^*)$ is the set of all $(R_1^*, R_2^*, \varepsilon)$-achievable (L_1, L_2) pairs

Case: Non-Corner Point



$$R_1 \geq H(X_1|X_2) \quad R_2 \geq H(X_2|X_1) \quad R_1 + R_2 \geq H(X_1, X_2)$$

Case: Non-Corner Point



$$R_1 \geq H(X_1|X_2) \quad R_2 \geq H(X_2|X_1) \quad R_1 + R_2 \geq H(X_1, X_2)$$

Suppose $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2) + \delta)$ for some $\delta > 0$.

Second-Order Asymptotics for Non-Corner Point Case

Theorem (Tan–Kosut (2014), Nomura–Han (2015))

When $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2) + \delta)$ *for some* $\delta > 0$*, the set*

$$\mathcal{L}(\varepsilon; R_1^*, R_2^*) = \left\{(L_1, L_2) : L_1 \geq \sqrt{V(X_1|X_2)}\Phi^{-1}(1-\varepsilon)\right\}$$

Second-Order Asymptotics for Non-Corner Point Case

Theorem (Tan–Kosut (2014), Nomura–Han (2015))

When $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2) + \delta)$ *for some* $\delta > 0$*, the set*

$$\mathcal{L}(\varepsilon; R_1^*, R_2^*) = \left\{(L_1, L_2) : L_1 \geq \sqrt{V(X_1|X_2)}\Phi^{-1}(1-\varepsilon)\right\}$$

- Same as lossless source coding with full decoder side information

Second-Order Asymptotics for Non-Corner Point Case

Theorem (Tan–Kosut (2014), Nomura–Han (2015))

When $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2) + \delta)$ *for some* $\delta > 0$*, the set*

$$\mathcal{L}(\varepsilon; R_1^*, R_2^*) = \left\{(L_1, L_2) : L_1 \geq \sqrt{V(X_1|X_2)}\Phi^{-1}(1-\varepsilon)\right\}$$

- Same as lossless source coding with full decoder side information
- No constraints on $L_2 \in \mathbb{R}$

Second-Order Asymptotics for Non-Corner Point Case

Theorem (Tan–Kosut (2014), Nomura–Han (2015))

When $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2) + \delta)$ *for some* $\delta > 0$*, the set*

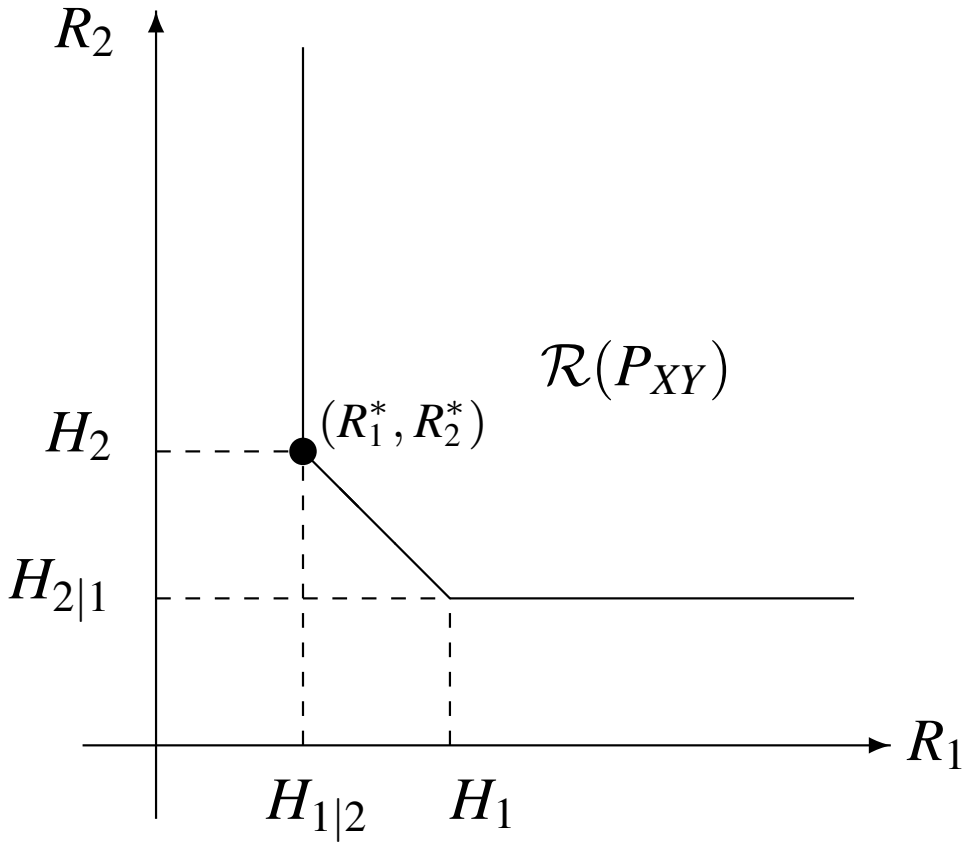
$$\mathcal{L}(\varepsilon; R_1^*, R_2^*) = \left\{(L_1, L_2) : L_1 \geq \sqrt{V(X_1|X_2)}\Phi^{-1}(1-\varepsilon)\right\}$$

- Same as lossless source coding with full decoder side information
- No constraints on $L_2 \in \mathbb{R}$
- Intuition is that R_2^* is strictly above $H(X_2)$ which means

$$\Pr(\hat{X}_2^n \neq X_2^n) \leq \exp(-nc(\delta))$$

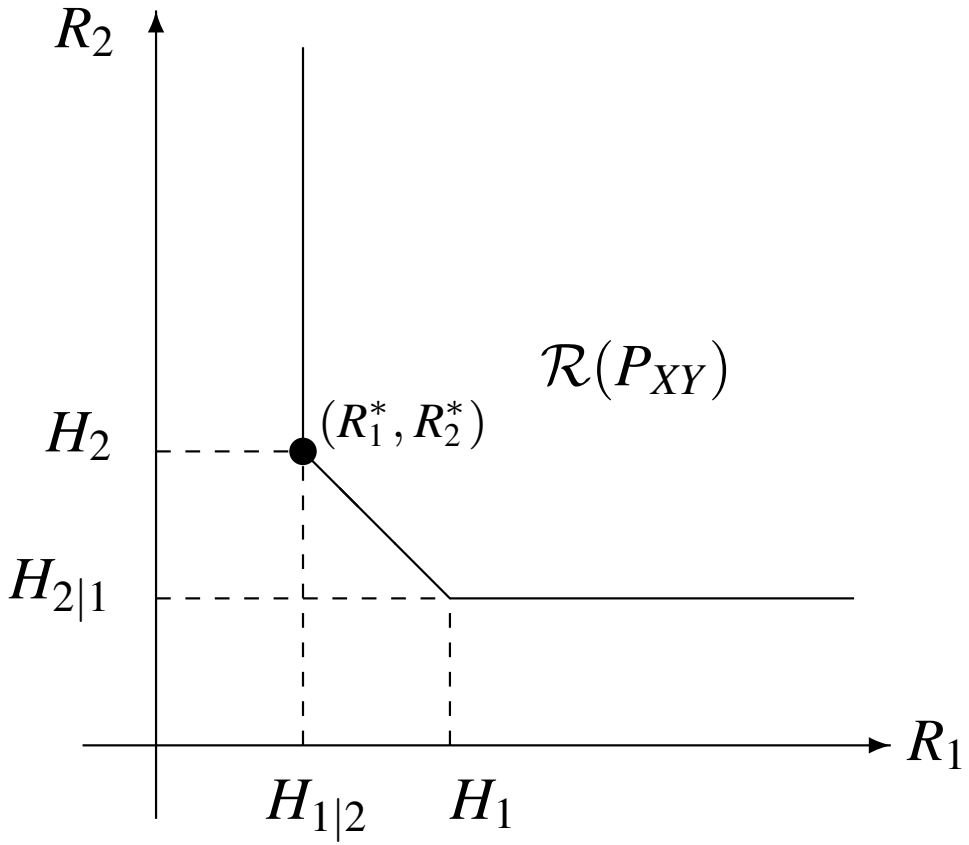
for some $c(\delta) > 0$. Doesn't affect second-order asymptotics.

Corner-Point Case



$$R_1 \geq H(X_1|X_2) \quad R_2 \geq H(X_2|X_1) \quad R_1 + R_2 \geq H(X_1, X_2)$$

Corner-Point Case



$$R_1 \geq H(X_1|X_2) \quad R_2 \geq H(X_2|X_1) \quad R_1 + R_2 \geq H(X_1, X_2)$$

Suppose $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2))$, i.e., near the corner point.

Second-Order Asymptotics for Corner-Point Case

Theorem (Tan–Kosut (2014), Nomura–Han (2015))

When $(R_1^*, R_2^*) = (H(X_1|X_2), H(X_2))$,

$$\mathcal{L}(\varepsilon; R_1^*, R_2^*) = \left\{(L_1, L_2) : \Psi(L_1, L_1 + L_2; [\mathbf{V}]_{1,12}) \geq 1 - \varepsilon\right\}$$

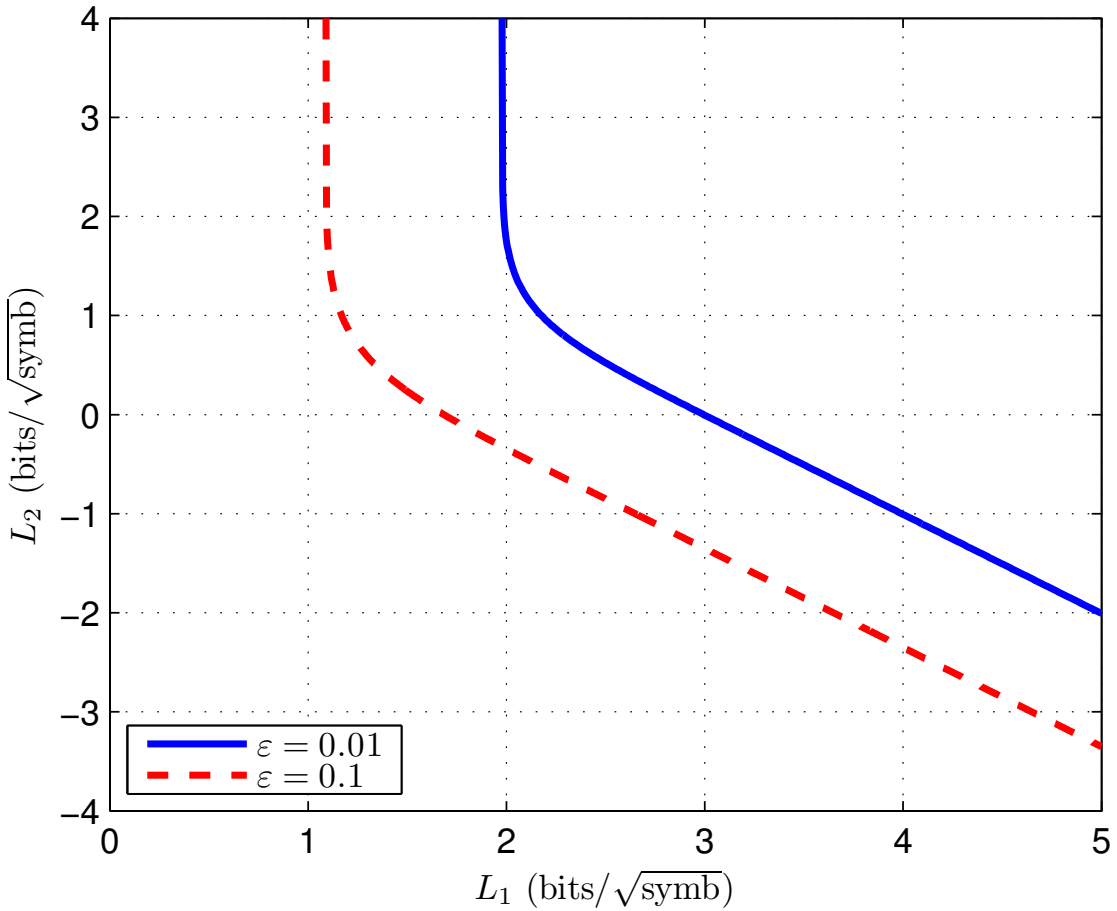
where

$$\Psi(a, b; \mathbf{V}) = \int_{-\infty}^{a} \int_{-\infty}^{b} \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{V})\, d\mathbf{x}.$$

and

$$\mathbf{V} = \mathrm{Cov} \begin{bmatrix} -\log P_{X_1|X_2}(X_1|X_2) \\ -\log P_{X_2|X_1}(X_2|X_1) \\ -\log P_{X_1X_2}(X_1, X_2) \end{bmatrix}.$$

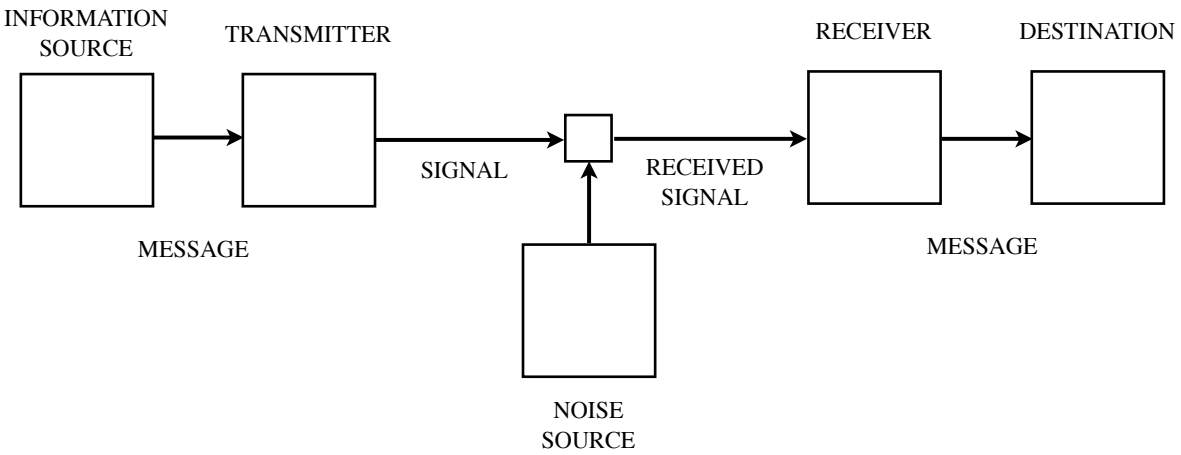
Second-Order Asymptotics for Corner-Point Case

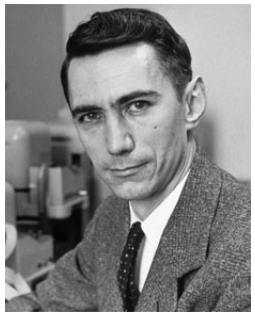


$\mathcal{L}(\varepsilon; R_1^*, R_2^*)$ for the source $\begin{bmatrix} 0.7 & 0.1 \\ 0.1 & 0.1 \end{bmatrix}$ with $\varepsilon = 0.01, 0.1$.

Outline

Transmission of Information

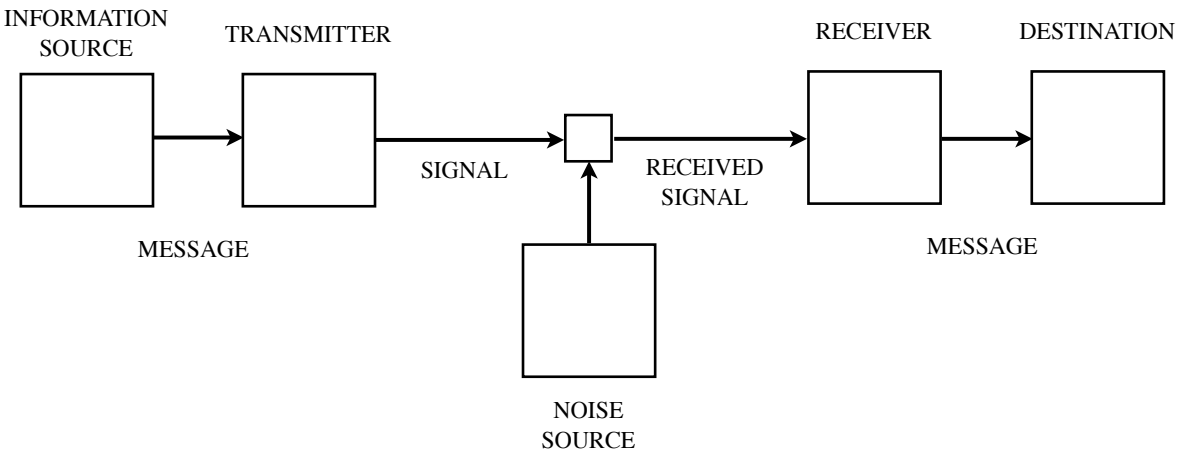


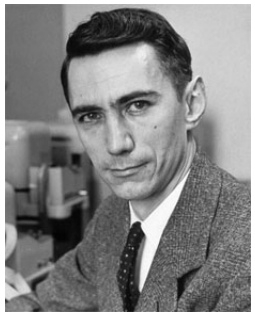


Shannon's Figure 1

- Information theory $\equiv$ Finding fundamental limits for reliable information transmission

Transmission of Information

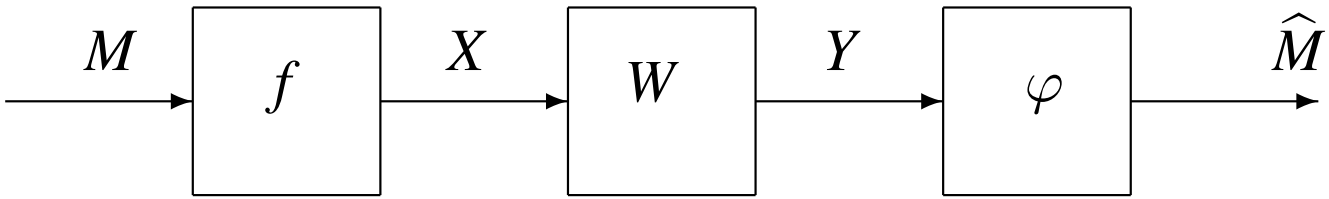




Shannon's Figure 1

- Information theory $\equiv$ Finding fundamental limits for reliable information transmission
- Channel coding: Concerned with the maximum rate of communication in bits per channel use

Channel Coding (One-Shot)

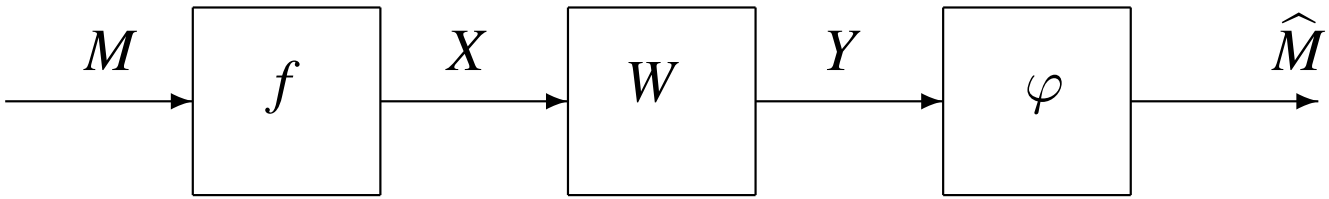


- A code is an triple $\mathcal{C} = \{\mathcal{M}, f, \varphi\}$ where $\mathcal{M}$ is the message set
- The average error probability $p_{\text{err}}(\mathcal{C})$ is

$$p_{\text{err}}(\mathcal{C}) \ := \ \Pr\left(\widehat{M} \neq M\right)$$

where M is uniform on $\mathcal{M}$.

Channel Coding (One-Shot)



- A code is an triple $\mathcal{C} = \{\mathcal{M}, f, \varphi\}$ where $\mathcal{M}$ is the message set
- The average error probability $p_{\text{err}}(\mathcal{C})$ is

$$p_{\text{err}}(\mathcal{C}) \ := \ \Pr\left(\widehat{M} \neq M\right)$$

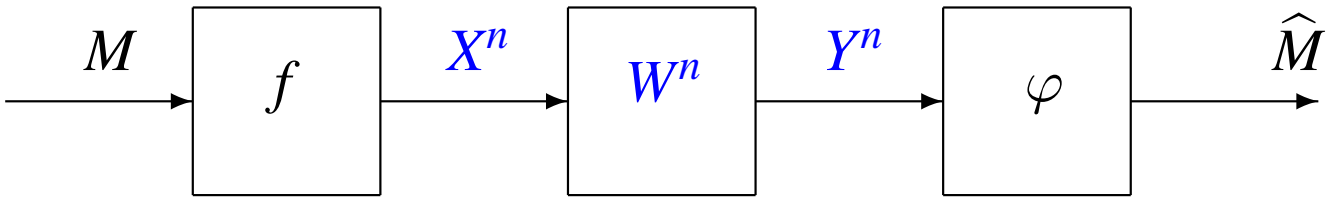
  where M is uniform on $\mathcal{M}$.

- A non-asymptotic fundamental limit can be defined as

$$M^*(W, \varepsilon) := \max\left\{m \in \mathbb{N} \ : \ \exists \mathcal{C} \ \text{ s.t. } \ m = |\mathcal{M}|, \, p_{\text{err}}(\mathcal{C}) \leq \varepsilon\right\}$$

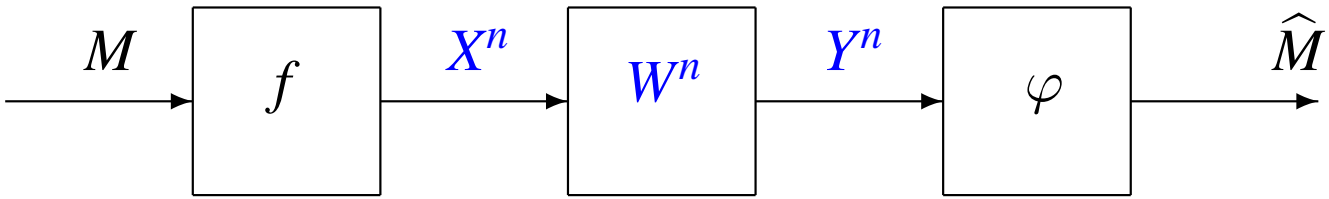
- Central problem in information theory to characterize $M^*(W, \varepsilon)$.

Channel Coding (n-Shot)



- n independent uses of a discrete memoryless channel (DMC) W^n

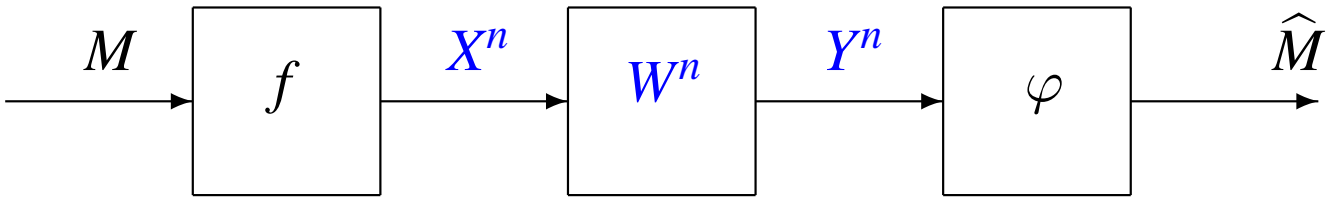
Channel Coding (n-Shot)



- n independent uses of a discrete memoryless channel (DMC) W^n
- Channel law is

$$W^n(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i)$$

Channel Coding (n-Shot)



- n independent uses of a discrete memoryless channel (DMC) W^n
- Channel law is

$$W^n(y^n|x^n) = \prod_{i=1}^{n} W(y_i|x_i)$$

- Seek non-asymptotic fundamental limit for n uses of W

$$\log M^*(W^n, \varepsilon)$$

Second-Order Asymptotics

Theorem (Strassen '62, Hayashi '09, Polyanskiy–Poor–Verdú '10)

For all "well-behaved" DMCs and Gaussian channels,

$$\log M^*(W^n, \varepsilon) = nC + \sqrt{nV}\Phi^{-1}(\varepsilon) + O(\log n)$$

where C and V are the capacity and dispersion.

Second-Order Asymptotics

Theorem (Strassen '62, Hayashi '09, Polyanskiy–Poor–Verdú '10)

For all "well-behaved" DMCs and Gaussian channels,

$$\log M^*(W^n, \varepsilon) = nC + \sqrt{nV}\Phi^{-1}(\varepsilon) + O(\log n)$$

where C and V are the capacity and dispersion.

V. Strassen

M. Hayashi

Polyanskiy–Poor–Verdú

Tight Third-Order Term

Theorem (Tomamichel–Tan (2013))

If W is a DMC with positive ε-dispersion,

$$\rho_n := \log M^*(W^n, \varepsilon) - \left[nC + \sqrt{nV}\Phi^{-1}(\varepsilon)\right] \leq \frac{1}{2}\log n + O(1)$$

Tight Third-Order Term

Theorem (Tomamichel–Tan (2013))

If W is a DMC with positive ε-dispersion,

$$\rho_n := \log M^*(W^n, \varepsilon) - \left[nC + \sqrt{nV}\Phi^{-1}(\varepsilon)\right] \leq \frac{1}{2}\log n + O(1)$$

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 59, NO. 11, NOVEMBER 2013 7041

A Tight Upper Bound for the Third-Order Asymptotics for Most Discrete Memoryless Channels

Marco Tomamichel, *Member, IEEE*, and Vincent Y. F. Tan, *Member, IEEE*

***Abstract*—This paper shows that the logarithm of the ε-error capacity (average error probability) for n uses of a discrete memoryless channel (DMC) is upper bounded by the normal approximation plus a third-order term that does not exceed $\frac{1}{2}\log n + O(1)$ if the ε-dispersion of the channel is positive. This matches a lower bound by Y. Polyanskiy (2010) for DMCs with positive reverse dispersion. If the ε-dispersion vanishes, the logarithm of the ε-error capacity is upper bounded by n times the capacity plus a constant term except for a small class of DMCs and $\varepsilon \geq \frac{1}{2}$.**

where $\rho_n = O(\log n)$, V_ε is the *ε-channel dispersion* [4], [5], and $\Phi(\cdot)$ is the Gaussian cumulative distribution function.[1] These quantities will be defined precisely in Section II-A. In fact, this asymptotic expansion also holds for $M^*(W^n, \varepsilon)$ [4, (284)–(286)] and implies that if an error probability of ε is tolerable, the backoff from channel capacity C at finite blocklength n is roughly $\sqrt{V_\varepsilon/n}\,\Phi^{-1}(\varepsilon)$. There have been several recent refinements to and extensions of Strassen's normal

Marco Tomamichel

Tight Third-Order Term

Theorem (Tomamichel–Tan (2013))

If W is a DMC with positive ε-dispersion,

$$\rho_n := \log M^*(W^n, \varepsilon) - \left[nC + \sqrt{nV}\Phi^{-1}(\varepsilon)\right] \leq \frac{1}{2}\log n + O(1)$$

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 59, NO. 11, NOVEMBER 2013 7041

A Tight Upper Bound for the Third-Order Asymptotics for Most Discrete Memoryless Channels

Marco Tomamichel, *Member, IEEE*, and Vincent Y. F. Tan, *Member, IEEE*

***Abstract*—This paper shows that the logarithm of the ε-error capacity (average error probability) for n uses of a discrete memoryless channel (DMC) is upper bounded by the normal approximation plus a third-order term that does not exceed $\frac{1}{2}\log n + O(1)$ if the ε-dispersion of the channel is positive. This matches a lower bound by Y. Polyanskiy (2010) for DMCs with positive reverse dispersion. If the ε-dispersion vanishes, the logarithm of the ε-error capacity is upper bounded by n times the capacity plus a constant term except for a small class of DMCs and $\varepsilon \geq \frac{1}{2}$.**

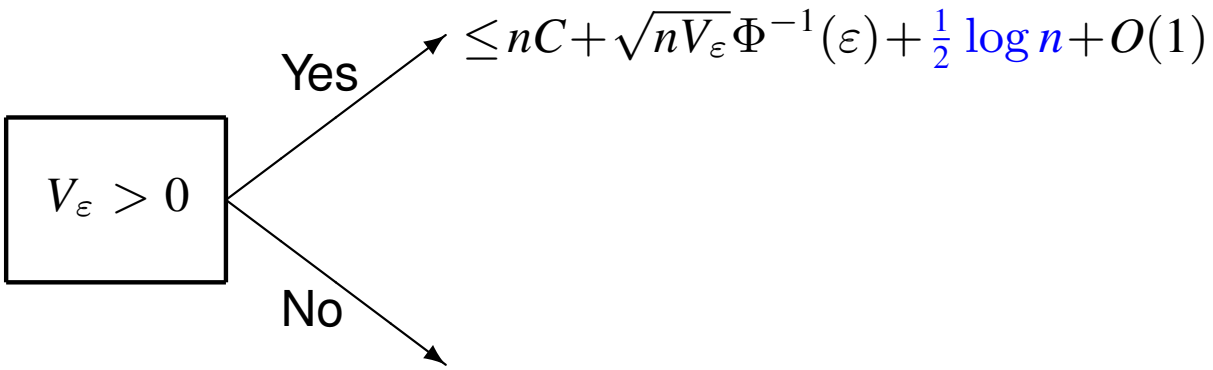
where $\rho_n = O(\log n)$, V_ε is the *ε-channel dispersion* [4], [5], and $\Phi(\cdot)$ is the Gaussian cumulative distribution function.[1] These quantities will be defined precisely in Section II-A. In fact, this asymptotic expansion also holds for $M^*(W^n, \varepsilon)$ [4, (284)–(286)] and implies that if an error probability of ε is tolerable, the backoff from channel capacity C at finite blocklength n is roughly $\sqrt{V_\varepsilon/n}\,\Phi^{-1}(\varepsilon)$. There have been several recent refinements to and extensions of Strassen's normal

Marco Tomamichel

$\rho_n \geq \frac{1}{2}\log n + O(1)$ for non-singular DMCs (e.g., BSCs) [Polyanskiy '10].

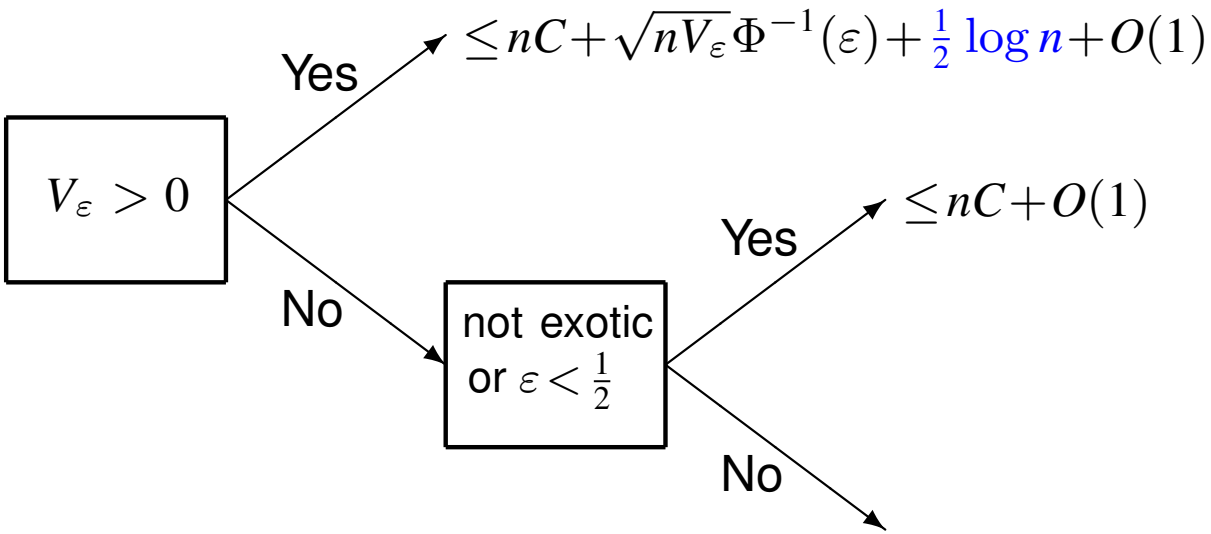
Main Result: Tight Third-Order Term

All cases are covered



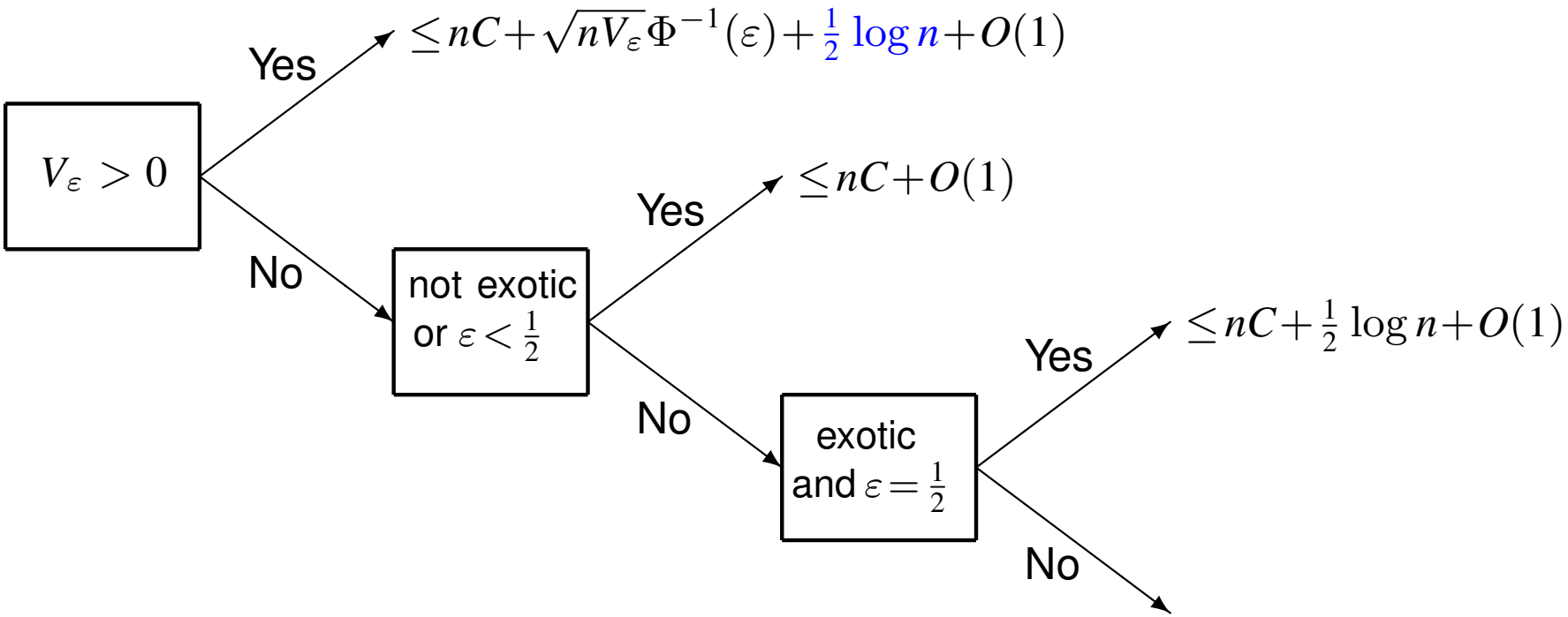
Main Result: Tight Third-Order Term

All cases are covered



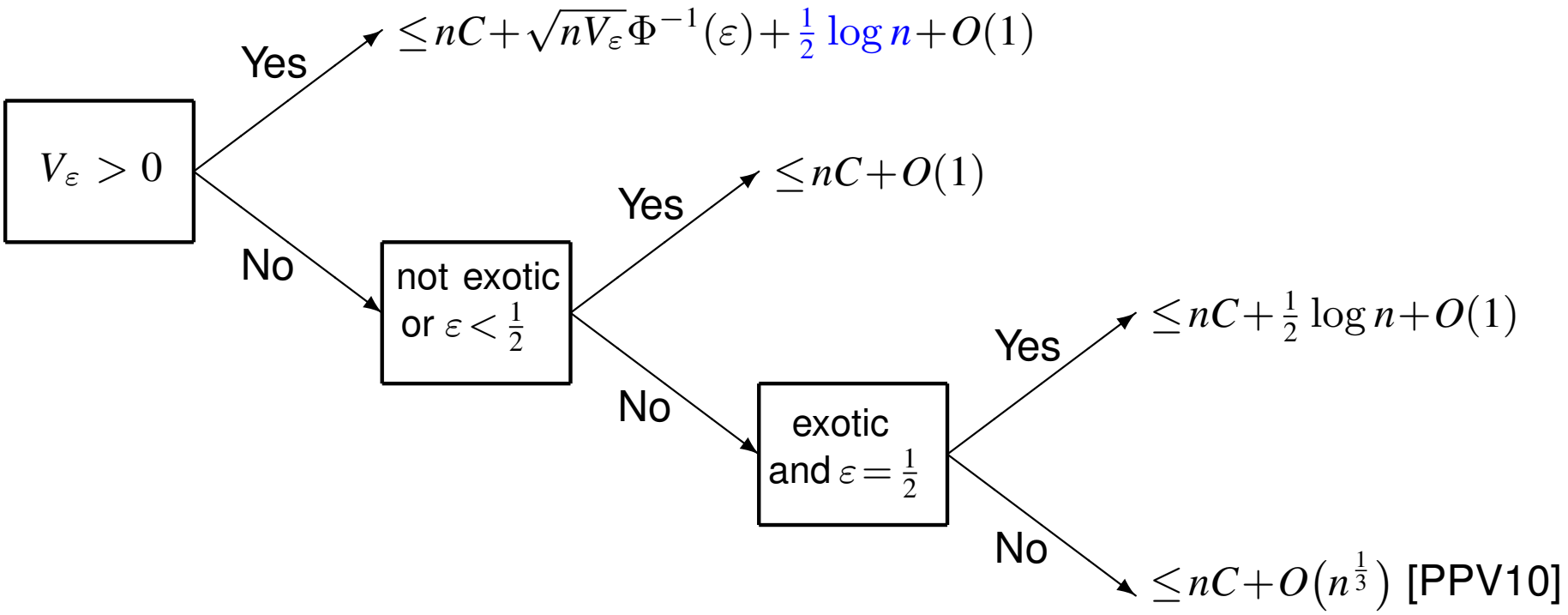
Main Result: Tight Third-Order Term

All cases are covered



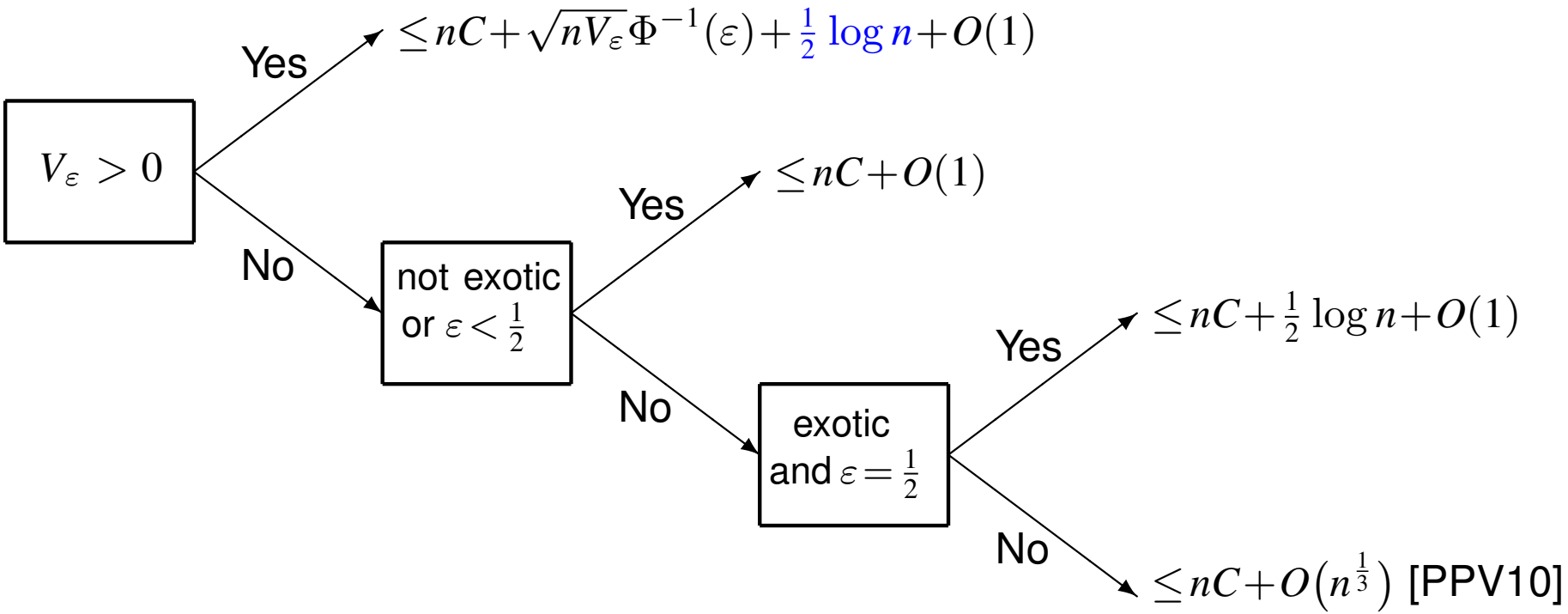
Main Result: Tight Third-Order Term

All cases are covered



Main Result: Tight Third-Order Term

All cases are covered



W is exotic if $V_{\max}(W) = 0$ and $\exists\, x_0 \in \mathcal{X}$ such that

$$D(W(\cdot|x_0) \,\|\, Q^*) = C, \quad \text{and} \quad V(W(\cdot|x_0) \,\|\, Q^*) > 0.$$

Proof Technique for Tight Third-Order Term

- For the regular case, $\rho_n \leq \frac{1}{2}\log n + O(1)$.

Proof Technique for Tight Third-Order Term

- For the regular case, $\rho_n \leq \frac{1}{2} \log n + O(1)$.
- Type-counting trick and standard upper bounds on $\log M_P^*(W^n, \varepsilon)$ are not sufficiently tight.

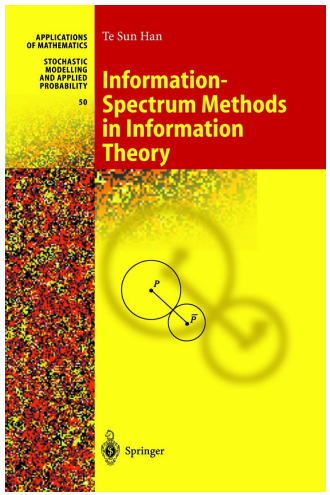
Proof Technique for Tight Third-Order Term

- For the regular case, $\rho_n \leq \frac{1}{2}\log n + O(1)$.
- Type-counting trick and standard upper bounds on $\log M_P^*(W^n, \varepsilon)$ are not sufficiently tight.
- Need a convenient converse bound for general DMCs.

Proof Technique for Tight Third-Order Term

- For the regular case, $\rho_n \leq \frac{1}{2}\log n + O(1)$.
- Type-counting trick and standard upper bounds on $\log M_P^*(W^n, \varepsilon)$ are not sufficiently tight.
- Need a convenient converse bound for general DMCs.
- Information spectrum divergence

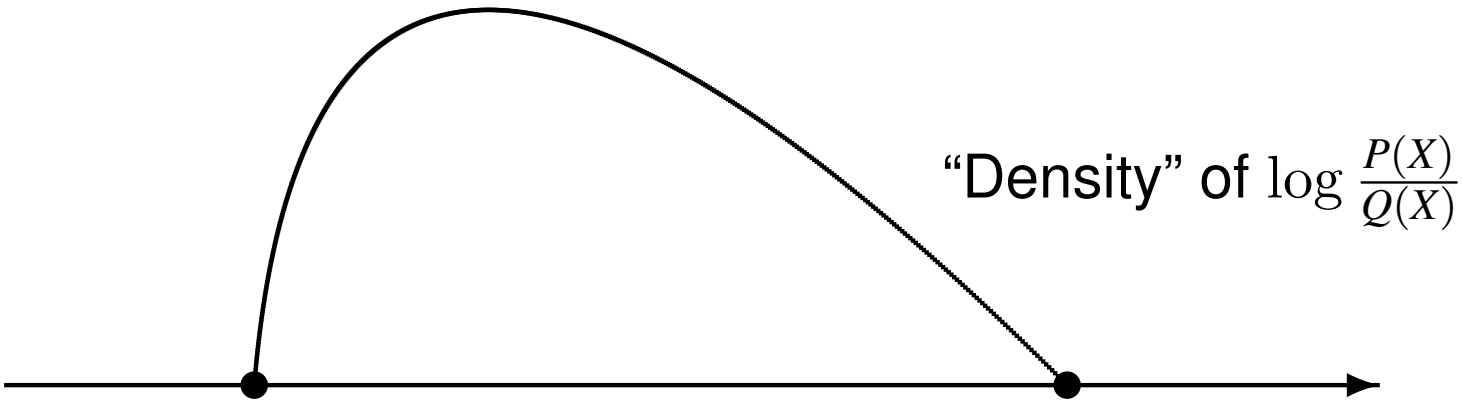
$$D_s^{\varepsilon}(P \,\|\, Q) := \sup\left\{R \geq 0 \,:\, P\left\{x : \log \frac{P(x)}{Q(x)} \leq R\right\} \leq \varepsilon\right\}$$



Te Sun Han

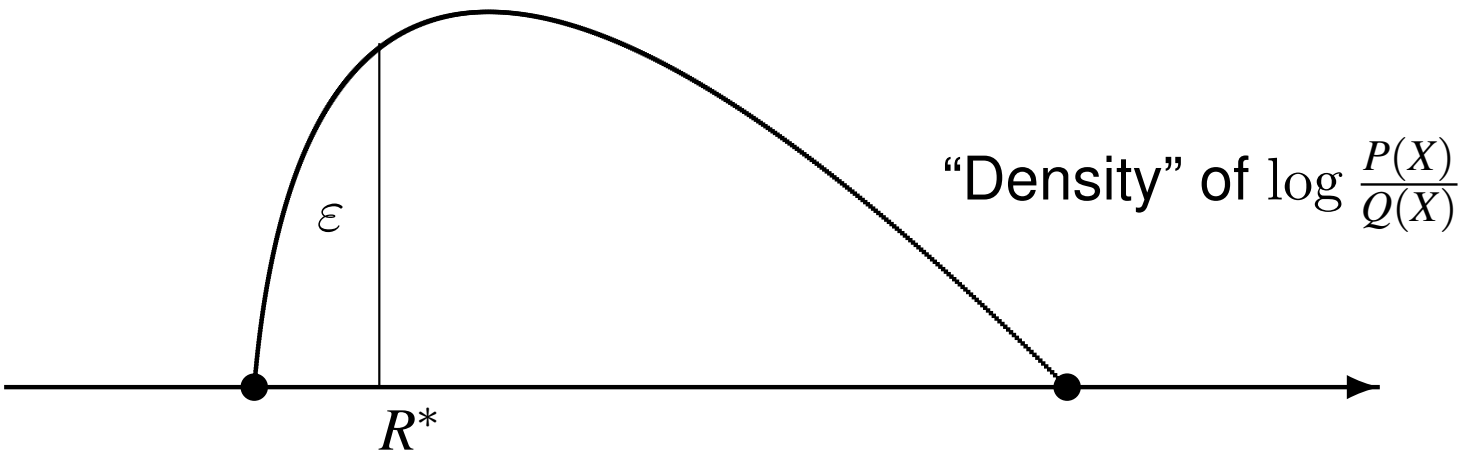
Proof Technique: Information Spectrum Divergence

$$D_s^{\varepsilon}(P \,\|\, Q) := \sup \left\{ R \geq 0 \,:\, P\left\{ x : \log \frac{P(x)}{Q(x)} \leq R \right\} \leq \varepsilon \right\}$$



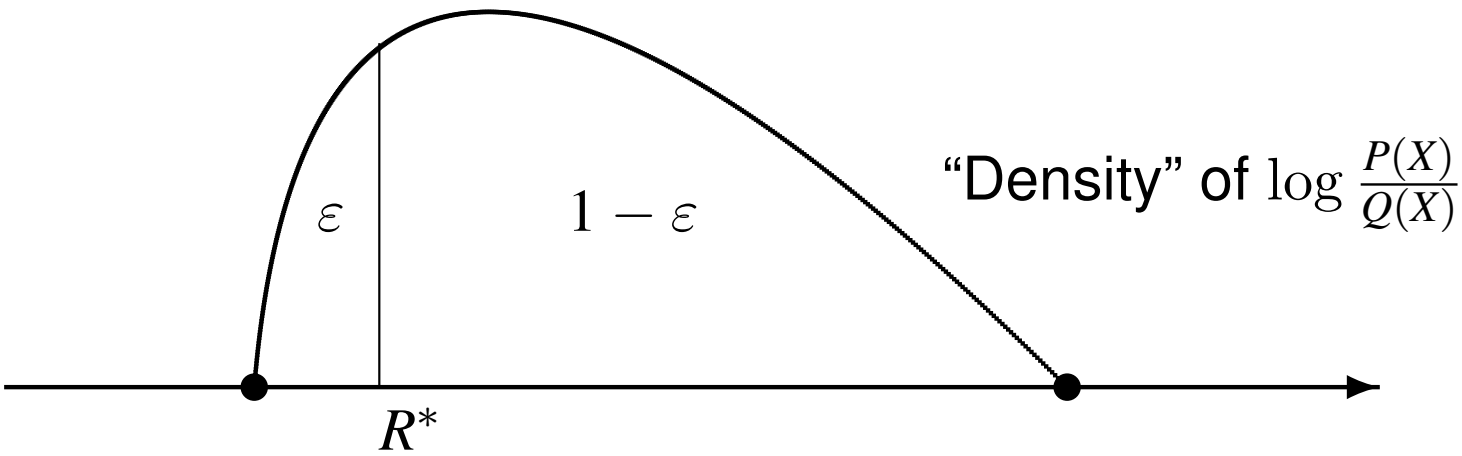
Proof Technique: Information Spectrum Divergence

$$D_s^{\varepsilon}(P \,\|\, Q) := \sup \left\{ R \geq 0 \,:\, P\left\{ x : \log \frac{P(x)}{Q(x)} \leq R \right\} \leq \varepsilon \right\}$$



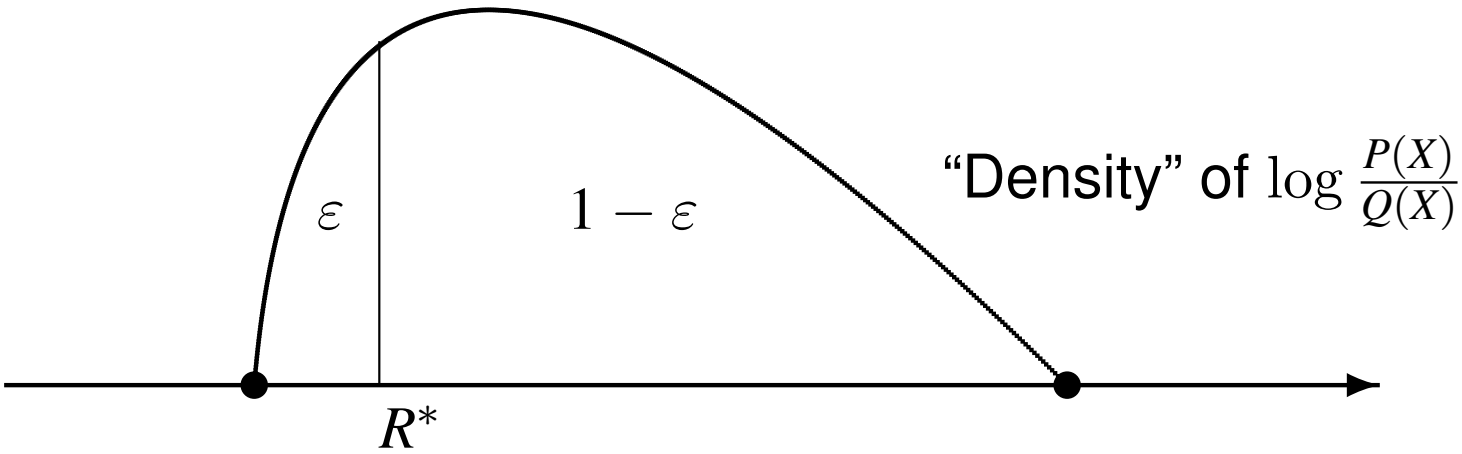
Proof Technique: Information Spectrum Divergence

$$D_s^{\varepsilon}(P \,\|\, Q) := \sup\left\{R \geq 0 \,:\, P\left\{x : \log \frac{P(x)}{Q(x)} \leq R\right\} \leq \varepsilon\right\}$$



Proof Technique: Information Spectrum Divergence

$$D_s^{\varepsilon}(P \,\|\, Q) := \sup \left\{ R \geq 0 \,:\, P\left\{ x : \log \frac{P(x)}{Q(x)} \leq R \right\} \leq \varepsilon \right\}$$



If X^n is i.i.d. P, the Berry–Esseen theorem yields

$$D_s^{\varepsilon}(P^n \,\|\, Q^n) = nD(P \,\|\, Q) + \sqrt{nV(P \,\|\, Q)}\Phi^{-1}(\varepsilon) + O(1)$$

Proof Technique: Symbol-Wise Converse Bound

Lemma (Tomamichel–Tan (2013))

For every channel W, every $\varepsilon \in (0,1)$ and $\delta \in (0, 1-\varepsilon)$, we have

$$\log M^*(W,\varepsilon) \leq \min_{Q \in \mathcal{P}(\mathcal{Y})} \max_{x \in \mathcal{X}} D_s^{\varepsilon+\delta}(W(\cdot|x) \,\|\, Q) + \log \frac{1}{\delta}$$

Proof Technique: Symbol-Wise Converse Bound

Lemma (Tomamichel–Tan (2013))

For every channel W, every $\varepsilon \in (0,1)$ and $\delta \in (0, 1-\varepsilon)$, we have

$$\log M^*(W,\varepsilon) \le \min_{Q \in \mathcal{P}(\mathcal{Y})} \max_{x \in \mathcal{X}} D_s^{\varepsilon+\delta}(W(\cdot|x) \,\|\, Q) + \log \frac{1}{\delta}$$

- When DMC is used n times,

$$\log M^*(W^n,\varepsilon) \le \min_{Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)} \left(\max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) \right) + \log \frac{1}{\delta}$$

Proof Technique: Symbol-Wise Converse Bound

Lemma (Tomamichel–Tan (2013))

For every channel W, every $\varepsilon \in (0,1)$ and $\delta \in (0, 1-\varepsilon)$, we have

$$\log M^*(W,\varepsilon) \le \min_{Q \in \mathcal{P}(\mathcal{Y})} \max_{x \in \mathcal{X}} D_s^{\varepsilon+\delta}(W(\cdot|x) \,\|\, Q) + \log \frac{1}{\delta}$$

- When DMC is used n times,

$$\log M^*(W^n,\varepsilon) \le \min_{Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)} \left(\max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) \right) + \log \frac{1}{\delta}$$

- Choose $\delta = n^{-\frac{1}{2}}$ so $\log \frac{1}{\delta} = \frac{1}{2} \log n$

Proof Technique: Symbol-Wise Converse Bound

Lemma (Tomamichel–Tan (2013))

For every channel W, every $\varepsilon \in (0,1)$ and $\delta \in (0, 1-\varepsilon)$, we have

$$\log M^*(W,\varepsilon) \le \min_{Q \in \mathcal{P}(\mathcal{Y})} \max_{x \in \mathcal{X}} D_s^{\varepsilon+\delta}(W(\cdot|x) \,\|\, Q) + \log \frac{1}{\delta}$$

- When DMC is used n times,

$$\log M^*(W^n,\varepsilon) \le \min_{Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)} \left(\max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) \right) + \log \frac{1}{\delta}$$

- Choose $\delta = n^{-\frac{1}{2}}$ so $\log \frac{1}{\delta} = \frac{1}{2} \log n$
- Since all $\mathbf{x}$ within a type class result in the same $D_s^{\varepsilon+\delta}$ (if $Q^{(n)}$ is permutation invariant), it's really a $\max$ over types $P_{\mathbf{x}} \in \mathcal{P}_n(\mathcal{X})$

Proof Technique: Choice of Output Distribution

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) + \log \frac{1}{\delta}, \qquad \forall\, Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$$

- $Q^{(n)}(\mathbf{y})$: invariant to permutations of the n channel uses

Proof Technique: Choice of Output Distribution

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) + \log \frac{1}{\delta}, \qquad \forall\, Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$$

- $Q^{(n)}(\mathbf{y})$: invariant to permutations of the n channel uses

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}) + \frac{1}{2} \sum_{P \in \mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|} (PW)^n(\mathbf{y})$$

Proof Technique: Choice of Output Distribution

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) + \log \frac{1}{\delta}, \qquad \forall\, Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$$

- $Q^{(n)}(\mathbf{y})$: invariant to permutations of the n channel uses

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}) + \frac{1}{2} \sum_{P \in \mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|} (PW)^n(\mathbf{y})$$

- First term: $Q_{\mathbf{k}}$'s and $\lambda(\mathbf{k})$'s designed to form an $n^{-\frac{1}{2}}$-cover of $\mathcal{P}(\mathcal{Y})$:

$$\forall\, Q \in \mathcal{P}(\mathcal{Y}), \quad \exists\, \mathbf{k} \in \mathcal{K} \quad \text{s.t.} \quad \|Q - Q_{\mathbf{k}}\|_2 \leq n^{-\frac{1}{2}}.$$

Proof Technique: Choice of Output Distribution

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) + \log \frac{1}{\delta}, \qquad \forall\, Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$$

- $Q^{(n)}(\mathbf{y})$: invariant to permutations of the n channel uses

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}) + \frac{1}{2} \sum_{P \in \mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|} (PW)^n(\mathbf{y})$$

- First term: $Q_{\mathbf{k}}$'s and $\lambda(\mathbf{k})$'s designed to form an $n^{-\frac{1}{2}}$-cover of $\mathcal{P}(\mathcal{Y})$:

$$\forall\, Q \in \mathcal{P}(\mathcal{Y}), \quad \exists\, \mathbf{k} \in \mathcal{K} \quad \text{s.t.} \quad \|Q - Q_{\mathbf{k}}\|_2 \leq n^{-\frac{1}{2}}.$$

- Second term: Uniform mixture over output distributions induced by input types [Hayashi (2009)]

Proof Technique: Choice of Output Distribution

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^n} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \,\|\, Q^{(n)}) + \log \frac{1}{\delta}, \qquad \forall\, Q^{(n)} \in \mathcal{P}(\mathcal{Y}^n)$$

- $Q^{(n)}(\mathbf{y})$: invariant to permutations of the n channel uses

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}) + \frac{1}{2} \sum_{P \in \mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|} (PW)^n(\mathbf{y})$$

- First term: $Q_{\mathbf{k}}$'s and $\lambda(\mathbf{k})$'s designed to form an $n^{-\frac{1}{2}}$-cover of $\mathcal{P}(\mathcal{Y})$:

$$\forall\, Q \in \mathcal{P}(\mathcal{Y}), \quad \exists\, \mathbf{k} \in \mathcal{K} \quad \text{s.t.} \quad \|Q - Q_{\mathbf{k}}\|_2 \leq n^{-\frac{1}{2}}.$$

- Second term: Uniform mixture over output distributions induced by input types [Hayashi (2009)]
- Take cares of "bad input types" (i.e., types $P \in \mathcal{P}_n(\mathcal{X})$ such that PW is far from Q^*)

Proof Technique: Novel Choice of Output Distribution

- First term is

$$\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}), \qquad \lambda(\mathbf{k}) = \frac{\exp\left(-\gamma\|\mathbf{k}\|_2^2\right)}{Z}, \qquad \sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) = 1,$$

and $\mathbf{k} = \{k_y\}_{y\in\mathcal{Y}} \in \mathbb{Z}^{|\mathcal{Y}|}$ indexes distance to the capacity-achieving output distribution (CAOD). Note $Z < \infty$.

Proof Technique: Novel Choice of Output Distribution

- First term is

$$\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}), \qquad \lambda(\mathbf{k}) = \frac{\exp\left(-\gamma\|\mathbf{k}\|_2^2\right)}{Z}, \qquad \sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) = 1,$$

  and $\mathbf{k} = \{k_y\}_{y\in\mathcal{Y}} \in \mathbb{Z}^{|\mathcal{Y}|}$ indexes distance to the capacity-achieving output distribution (CAOD). Note $Z < \infty$.

- Choose each $Q_{\mathbf{k}}$ as follows:

$$Q_{\mathbf{k}}(y) := Q^*(y) + \frac{k_y}{\sqrt{n\zeta}},$$

  where $\mathcal{K} := \left\{\mathbf{k} \in \mathbb{Z}^{|\mathcal{Y}|} : \sum_{y\in\mathcal{Y}} k_y = 0, k_y \geq -Q^*(y)\sqrt{n\zeta}, \forall\, y \in \mathcal{Y}\right\}$

Proof Technique: Novel Choice of Output Distribution

- First term is

$$\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^n(\mathbf{y}), \qquad \lambda(\mathbf{k}) = \frac{\exp\left(-\gamma\|\mathbf{k}\|_2^2\right)}{Z}, \qquad \sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) = 1,$$

  and $\mathbf{k} = \{k_y\}_{y\in\mathcal{Y}} \in \mathbb{Z}^{|\mathcal{Y}|}$ indexes distance to the capacity-achieving output distribution (CAOD). Note $Z < \infty$.

- Choose each $Q_{\mathbf{k}}$ as follows:

$$Q_{\mathbf{k}}(y) := Q^*(y) + \frac{k_y}{\sqrt{n\zeta}},$$

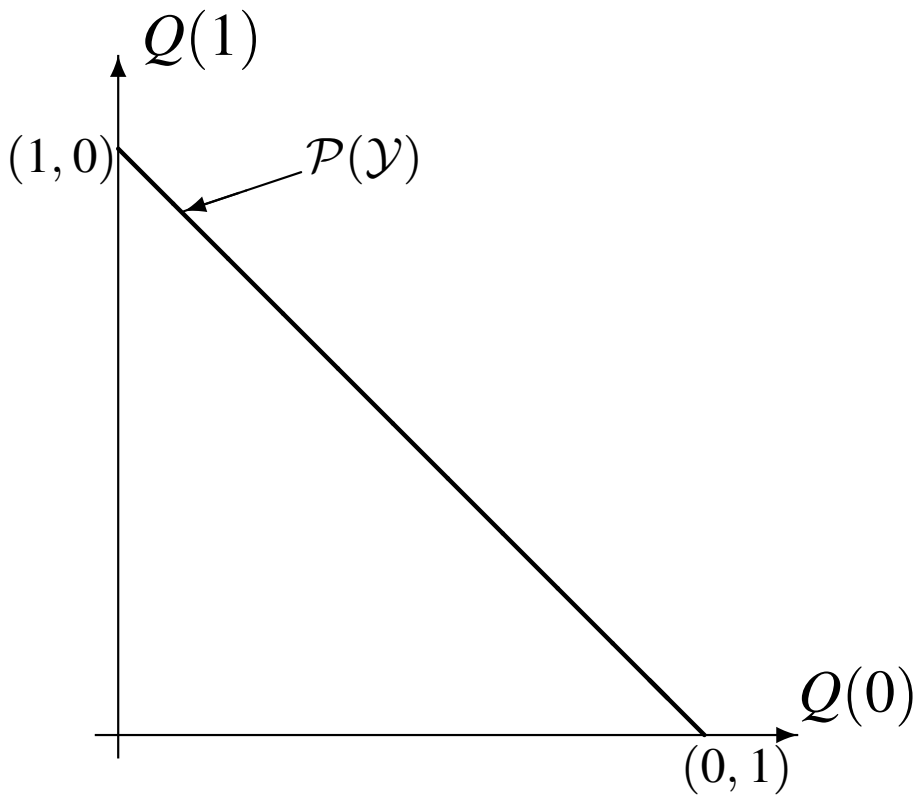
  where $\mathcal{K} := \left\{\mathbf{k} \in \mathbb{Z}^{|\mathcal{Y}|} : \sum_{y\in\mathcal{Y}} k_y = 0, k_y \geq -Q^*(y)\sqrt{n\zeta}, \forall\, y \in \mathcal{Y}\right\}$

- Construction ensures that

$$\forall\, Q \in \mathcal{P}(\mathcal{Y}), \quad \exists\, \mathbf{k} \in \mathcal{K}, \quad \text{s.t.} \quad \|Q - Q_{\mathbf{k}}\|_2 \leq \frac{1}{\sqrt{n}}.$$

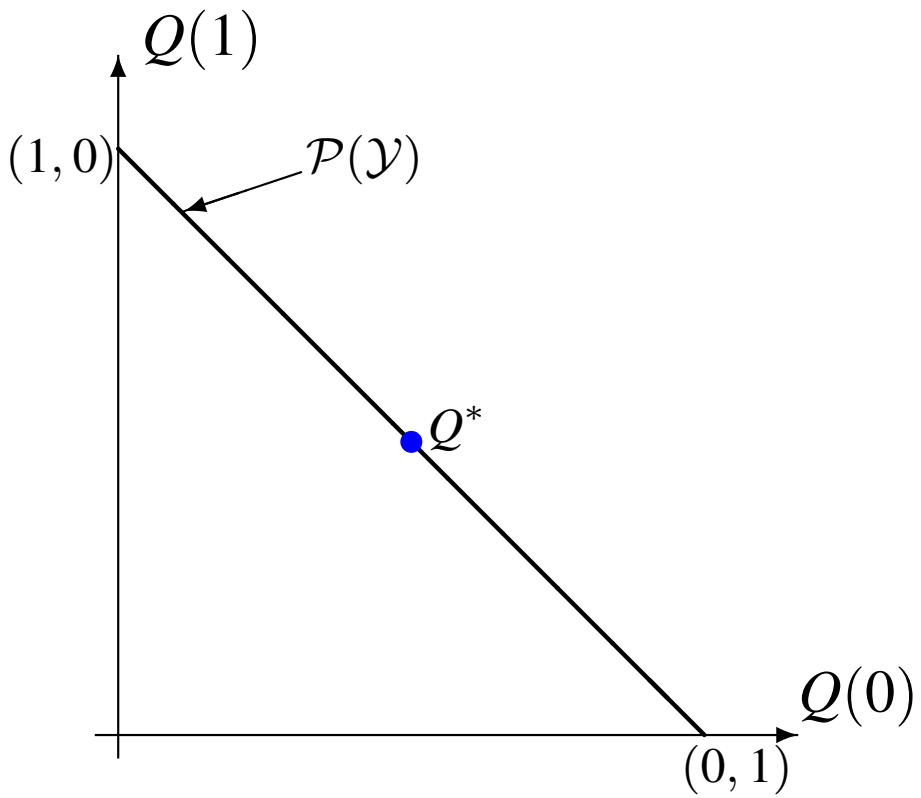
Proof Technique: Novel Choice of Output Distribution

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^{n}(\mathbf{y}) + \frac{1}{2}\sum_{P\in\mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|}(PW)^n(\mathbf{y})$$



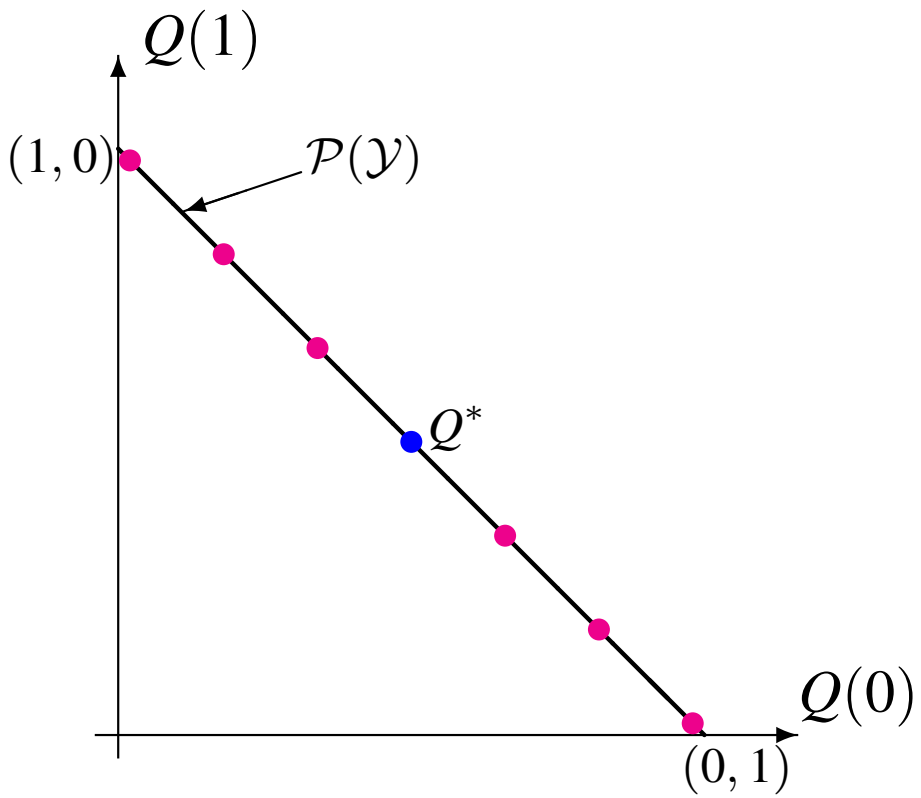
Proof Technique: Novel Choice of Output Distribution

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^{n}(\mathbf{y}) + \frac{1}{2}\sum_{P\in\mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|}(PW)^n(\mathbf{y})$$



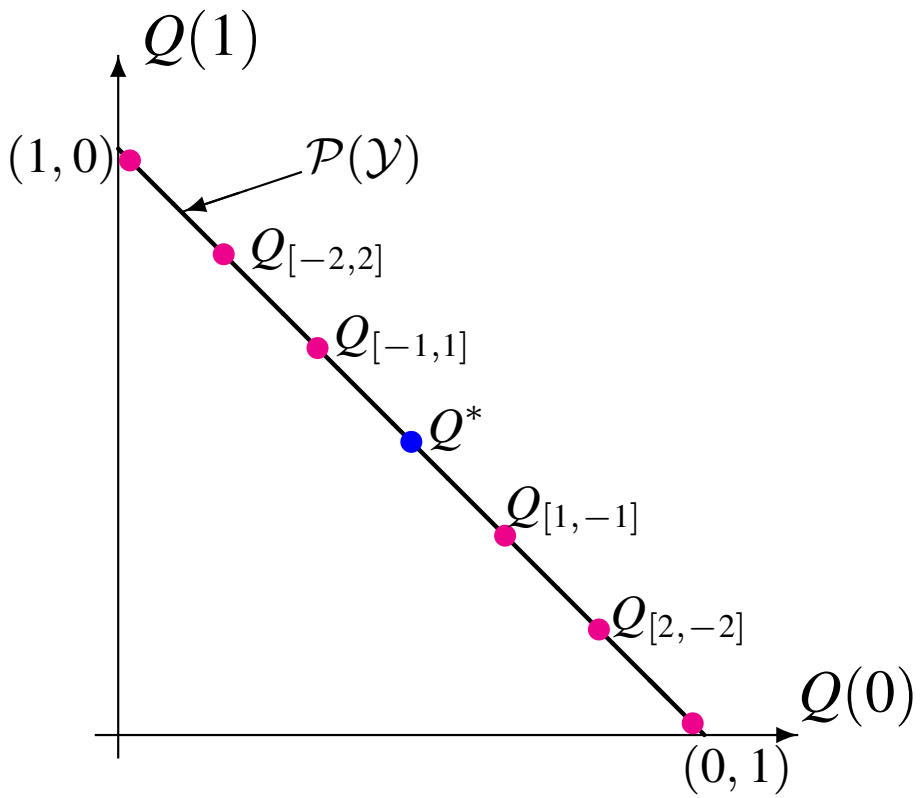
Proof Technique: Novel Choice of Output Distribution

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^{n}(\mathbf{y}) + \frac{1}{2}\sum_{P\in\mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|}(PW)^n(\mathbf{y})$$



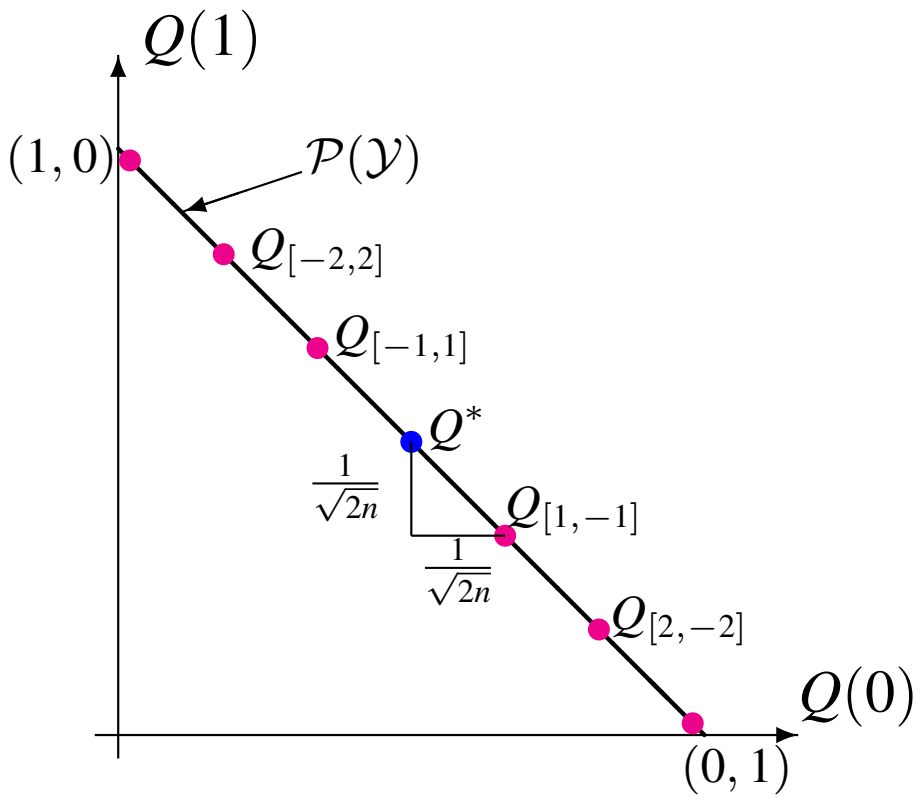
Proof Technique: Novel Choice of Output Distribution

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^{n}(\mathbf{y}) + \frac{1}{2}\sum_{P\in\mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|}(PW)^n(\mathbf{y})$$

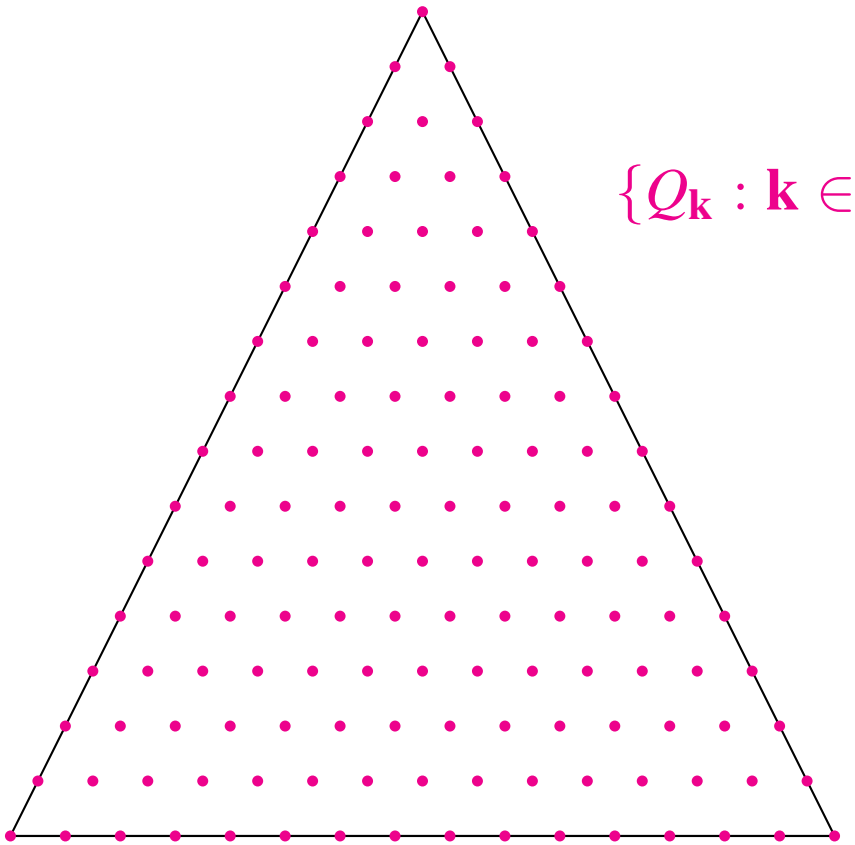


Proof Technique: Novel Choice of Output Distribution

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2}\sum_{\mathbf{k}\in\mathcal{K}} \lambda(\mathbf{k}) Q_{\mathbf{k}}^{n}(\mathbf{y}) + \frac{1}{2}\sum_{P\in\mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|}(PW)^n(\mathbf{y})$$

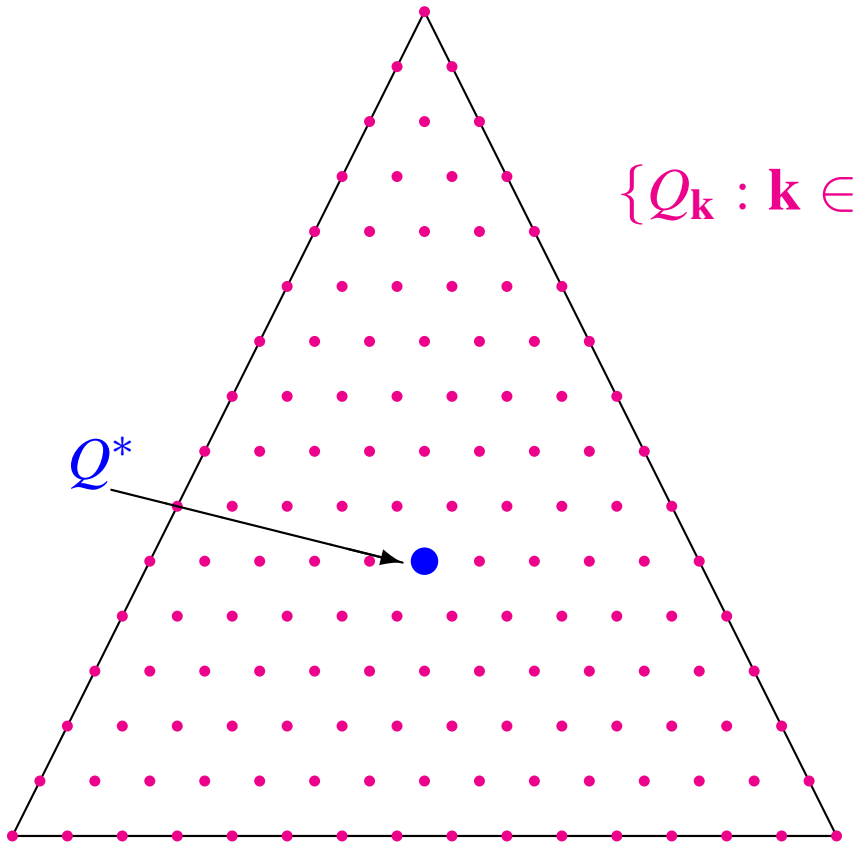


Proof Technique: Novel Choice of Output Distribution

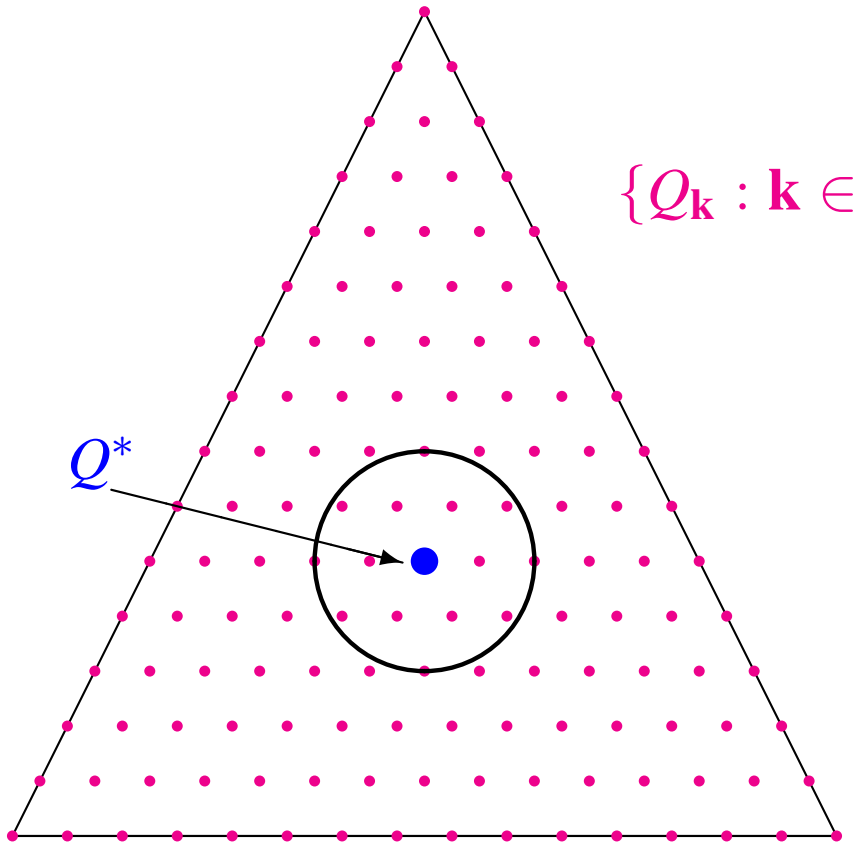


$$\{Q_{\mathbf{k}} : \mathbf{k} \in \mathcal{K}\} \subset \mathcal{P}(\mathcal{Y})$$

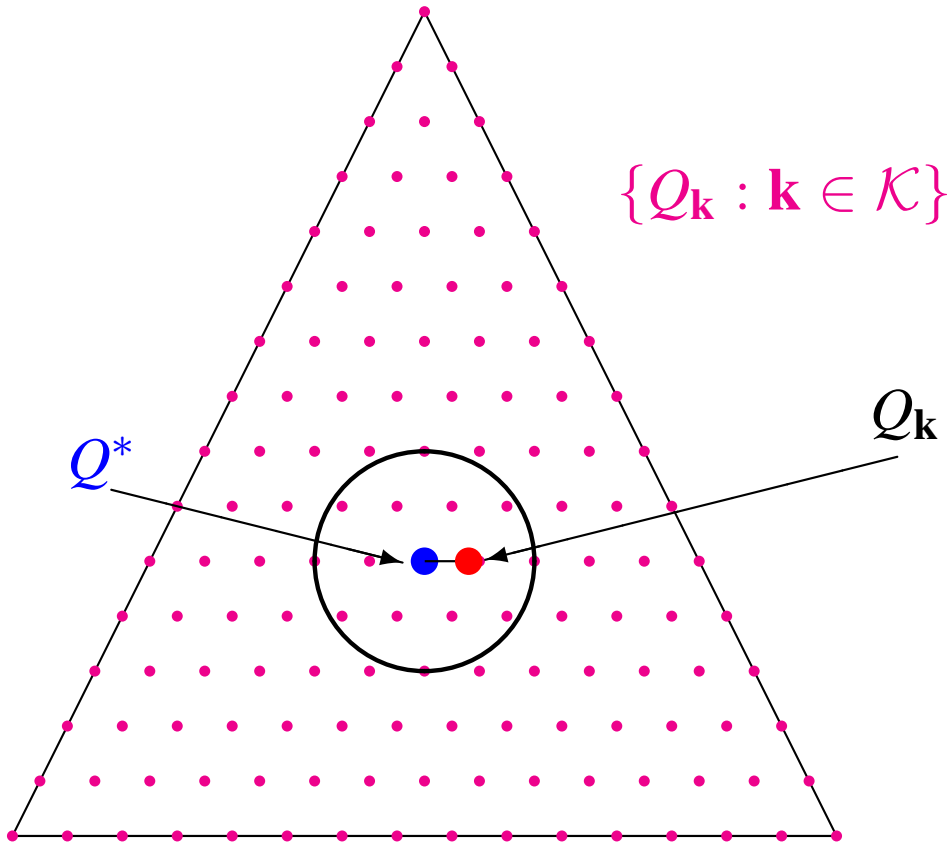
Proof Technique: Novel Choice of Output Distribution



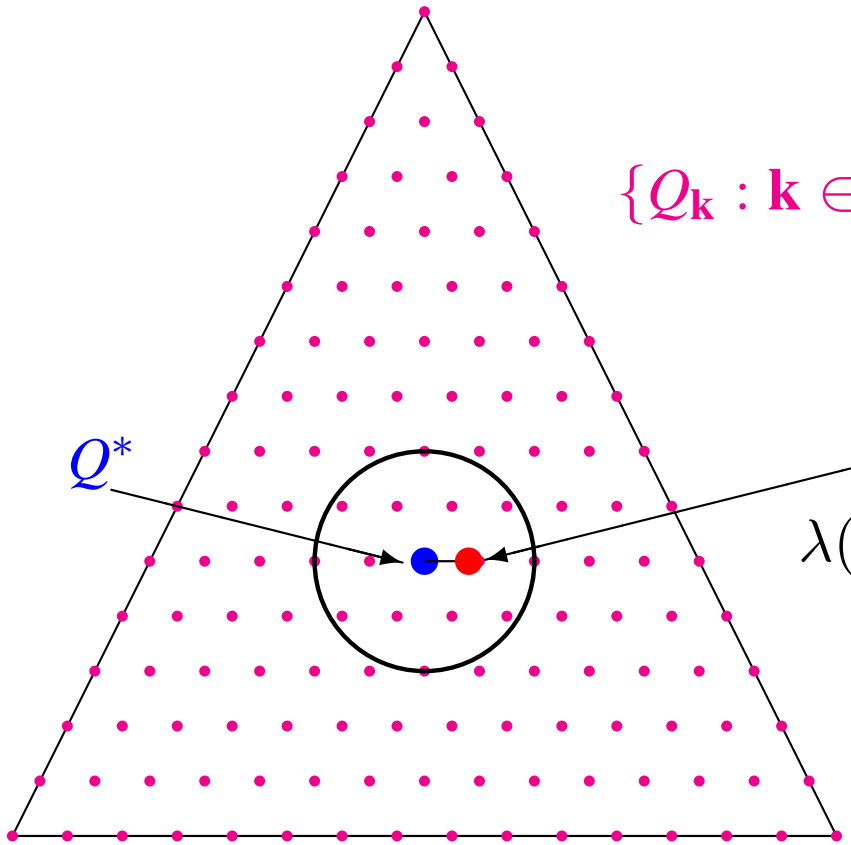
Proof Technique: Novel Choice of Output Distribution



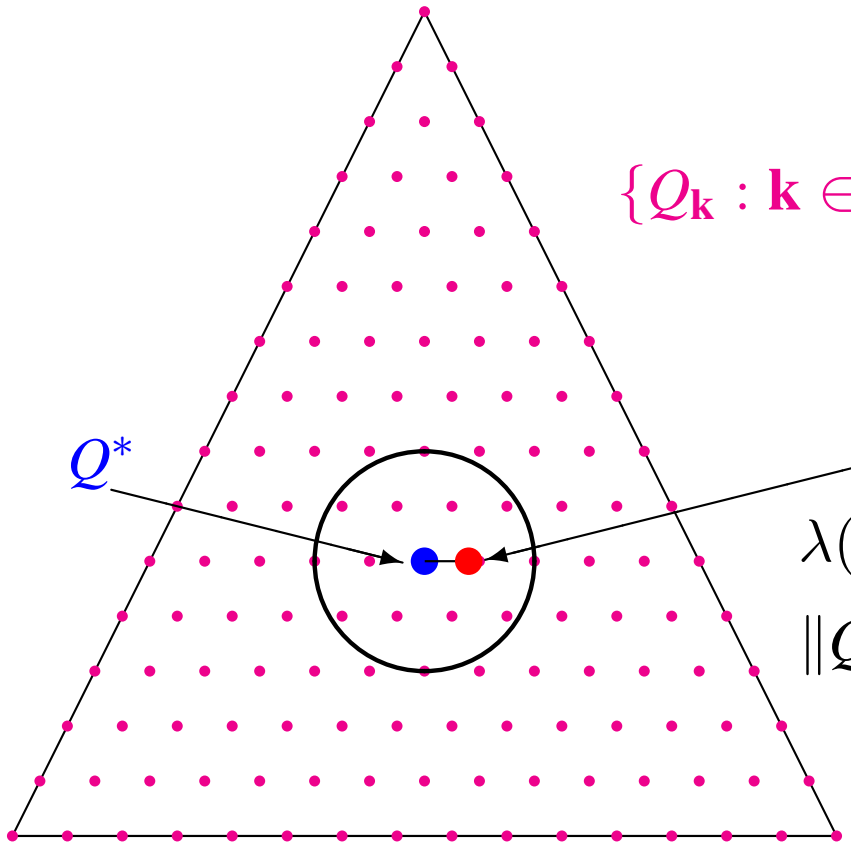
Proof Technique: Novel Choice of Output Distribution



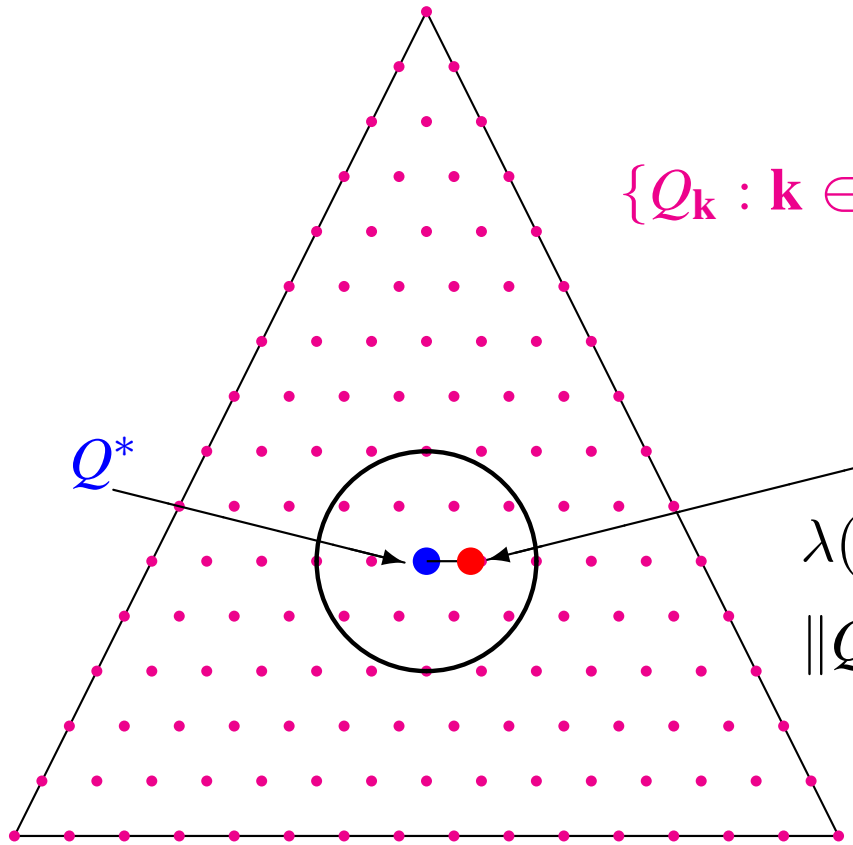
Proof Technique: Novel Choice of Output Distribution



Proof Technique: Novel Choice of Output Distribution



Proof Technique: Novel Choice of Output Distribution

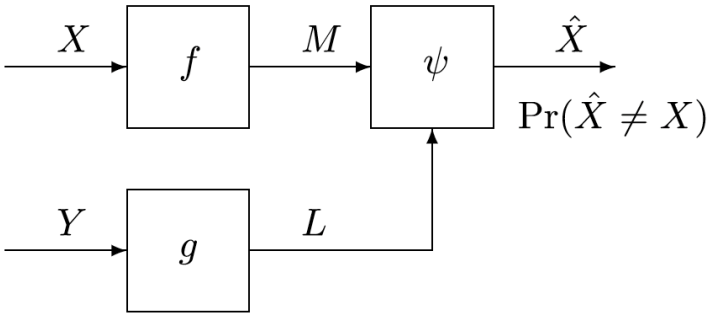


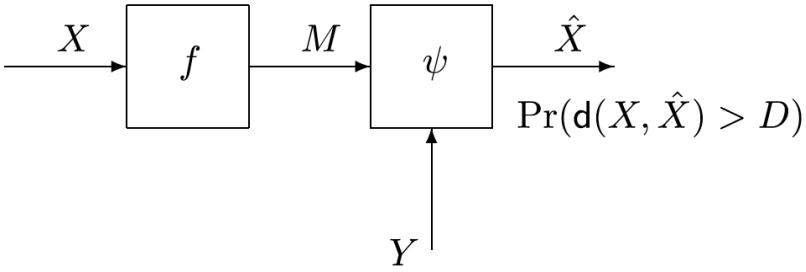
For most DMCs,

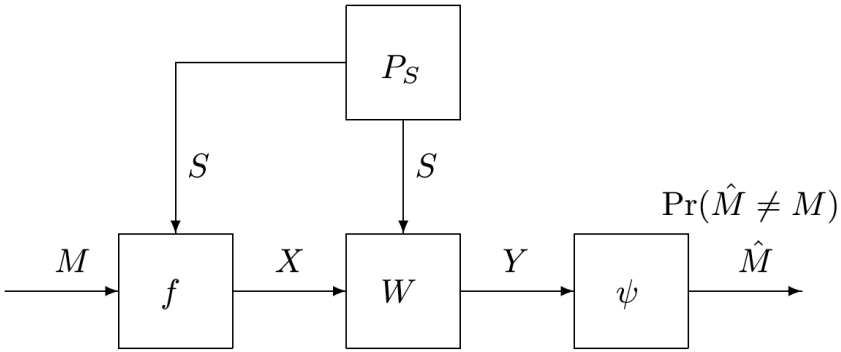
$$\log M^*(W^n, \varepsilon) - \left[nC + \sqrt{nV}\Phi^{-1}(\varepsilon)\right] = \frac{1}{2}\log n + O(1).$$

Other Contributions to Second- and Third-Order

Side Information with S. Watanabe and S. Kuzuoka







Nonasymptotic and Second-Order Achievability Bounds for Coding With Side-Information

Shun Watanabe, *Member, IEEE*, Shigeaki Kuzuoka, *Member, IEEE*, and Vincent Y. F. Tan, *Member, IEEE*

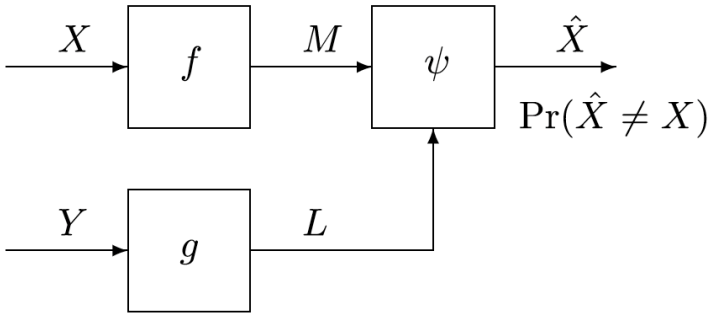
***Abstract*— We present a novel nonasymptotic or finite blocklength achievability bounds for three side-information problems in network information theory. These include: 1) the Wyner-Ahlswede-Körner (WAK) problem of almost-lossless source coding with rate-limited side-information; 2) the Wyner–Ziv (WZ) problem of lossy source coding with side-information at the decoder; and 3) the Gel'fand–Pinsker (GP) problem of channel coding with noncausal state information available at the encoder. The bounds are proved using ideas from channel simulation and channel resolvability. Our bounds for all three problems improve on all previous nonasymptotic**

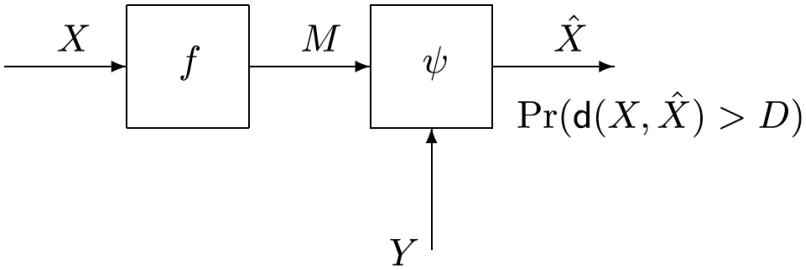
problems whose asymptotic rate characterizations are well known. These include

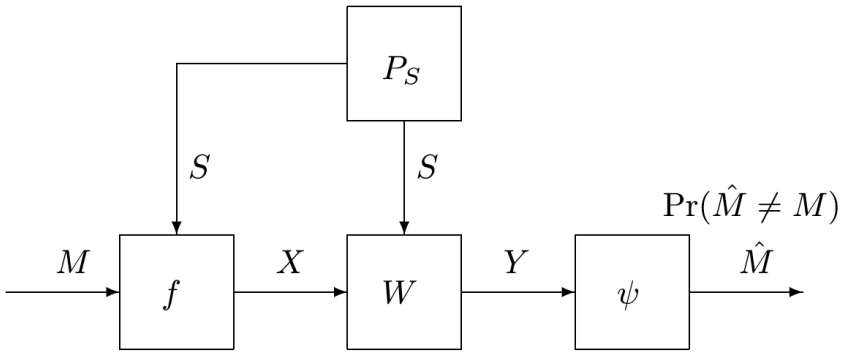
- The *Wyner-Ahlswede-Körner* (WAK) problem of almost-lossless source coding with rate-limited (aka coded) side-information [2], [3],
- The *Wyner-Ziv* (WZ) problem of lossy source coding with side-information at the decoder [4], and
- The *Gel'fand-Pinsker* (GP) problem of channel coding with noncausal state information at the encoder [5].

Other Contributions to Second- and Third-Order

Side Information with S. Watanabe and S. Kuzuoka







1574 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 61, NO. 4, APRIL 2015

Nonasymptotic and Second-Order Achievability Bounds for Coding With Side-Information

Shun Watanabe, *Member, IEEE*, Shigeaki Kuzuoka, *Member, IEEE*, and Vincent Y. F. Tan, *Member, IEEE*

***Abstract*— We present a novel nonasymptotic or finite blocklength achievability bounds for three side-information problems in network information theory. These include: 1) the Wyner-Ahlswede-Körner (WAK) problem of almost-lossless source coding with rate-limited side-information; 2) the Wyner–Ziv (WZ) problem of lossy source coding with side-information at the decoder; and 3) the Gel'fand–Pinsker (GP) problem of channel coding with noncausal state information available at the encoder. The bounds are proved using ideas from channel simulation and channel resolvability. Our bounds for all three problems improve on all previous nonasymptotic**

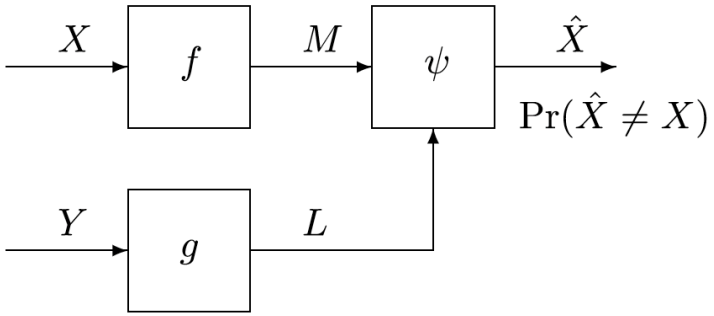
problems whose asymptotic rate characterizations are well known. These include

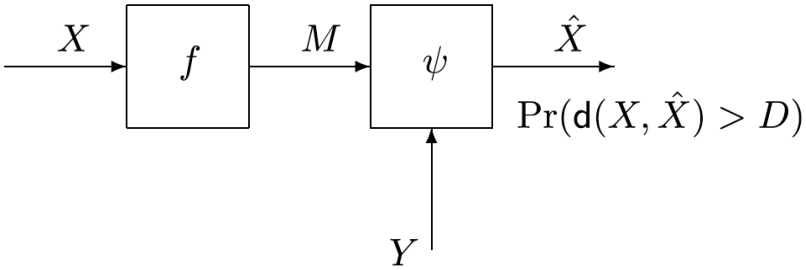
- The *Wyner-Ahlswede-Körner* (WAK) problem of almost-lossless source coding with rate-limited (aka coded) side-information [2], [3],
- The *Wyner-Ziv* (WZ) problem of lossy source coding with side-information at the decoder [4], and
- The *Gel'fand-Pinsker* (GP) problem of channel coding with noncausal state information at the encoder [5].

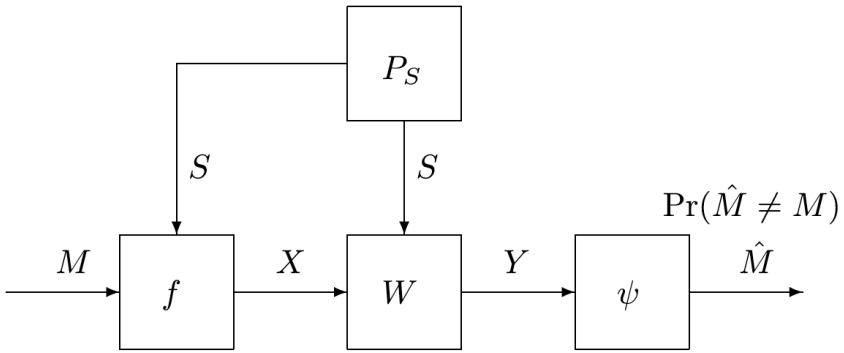
$$\varepsilon \lessapprox \Pr\Big(\text{covering error} \cup \text{packing error}\Big)$$

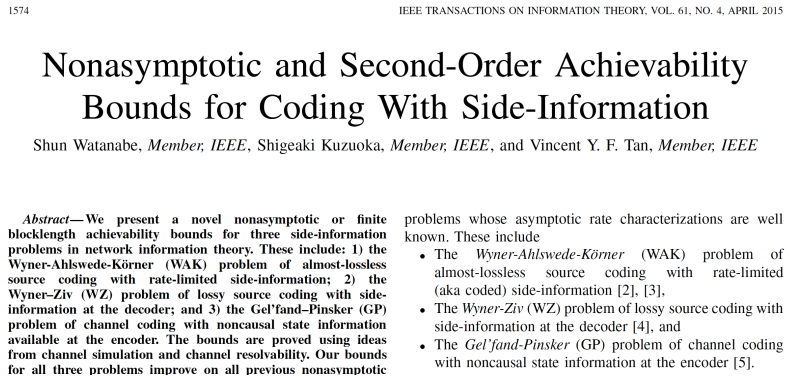
Other Contributions to Second- and Third-Order

Side Information with S. Watanabe and S. Kuzuoka









1574 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 61, NO. 4, APRIL 2015

Nonasymptotic and Second-Order Achievability Bounds for Coding With Side-Information

Shun Watanabe, *Member, IEEE*, Shigeaki Kuzuoka, *Member, IEEE*, and Vincent Y. F. Tan, *Member, IEEE*

***Abstract*— We present a novel nonasymptotic or finite blocklength achievability bounds for three side-information problems in network information theory. These include: 1) the Wyner-Ahlswede-Körner (WAK) problem of almost-lossless source coding with rate-limited side-information; 2) the Wyner–Ziv (WZ) problem of lossy source coding with side-information at the decoder; and 3) the Gel'fand–Pinsker (GP) problem of channel coding with noncausal state information available at the encoder. The bounds are proved using ideas from channel simulation and channel resolvability. Our bounds for all three problems improve on all previous nonasymptotic**

problems whose asymptotic rate characterizations are well known. These include

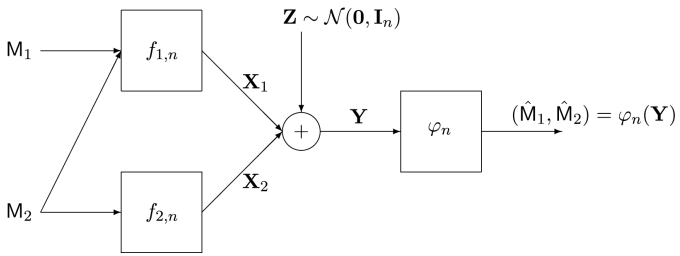
- The *Wyner-Ahlswede-Körner* (WAK) problem of almost-lossless source coding with rate-limited (aka coded) side-information [2], [3],
- The *Wyner-Ziv* (WZ) problem of lossy source coding with side-information at the decoder [4], and
- The *Gel'fand-Pinsker* (GP) problem of channel coding with noncausal state information at the encoder [5].

$$\varepsilon \lessapprox \Pr\Big(\text{covering error} \cup \text{packing error}\Big)$$

Also see Jingbo Liu's work for a converse for WAK (IEEE T-IT '20)

Other Contributions to Second- and Third-Order

Gaussian MAC with degraded message sets with J. Scarlett



6700 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 61, NO. 12, DECEMBER 2015

Second-Order Asymptotics for the Gaussian MAC With Degraded Message Sets

Jonathan Scarlett, *Member, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— This paper studies the second-order asymptotics of the Gaussian multiple-access channel with degraded message sets. For a fixed average error probability $\varepsilon \in (0, 1)$ and an arbitrary point on the boundary of the capacity region, we characterize the speed of convergence of rate pairs that converge to that boundary point for codes that have asymptotic error probability no larger than ε. As a stepping stone to this local notion of the second-order asymptotics, we study a global notion, and establish relationships between the two. We provide a numerical example**

given by the set of rate pairs (R_1, R_2) satisfying

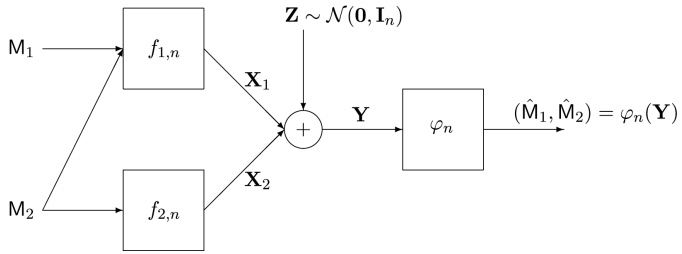
$$R_1 \le \mathsf{C}\big((1-\rho^2)S_1\big) \quad (1)$$

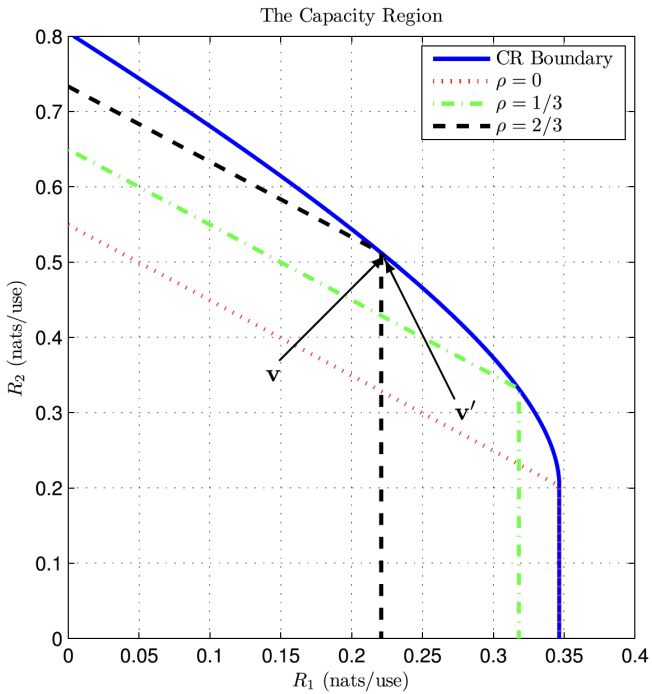
$$R_1 + R_2 \le \mathsf{C}\big(S_1 + S_2 + 2\rho\sqrt{S_1 S_2}\big) \quad (2)$$

for some $\rho \in [0, 1]$, where S_1 and S_2 are the admissible transmit powers, and $\mathsf{C}(x) := \frac{1}{2}\log(1+x)$ is the Gaussian capacity function. The capacity region $\mathcal{C}$ does not depend on

Other Contributions to Second- and Third-Order

Gaussian MAC with degraded message sets with J. Scarlett





Second-Order Asymptotics for the Gaussian MAC With Degraded Message Sets

Jonathan Scarlett, *Member, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— This paper studies the second-order asymptotics of the Gaussian multiple-access channel with degraded message sets. For a fixed average error probability $\varepsilon \in (0, 1)$ and an arbitrary point on the boundary of the capacity region, we characterize the speed of convergence of rate pairs that converge to that boundary point for codes that have asymptotic error probability no larger than ε. As a stepping stone to this local notion of the second-order asymptotics, we study a global notion, and establish relationships between the two. We provide a numerical example**

given by the set of rate pairs (R_1, R_2) satisfying

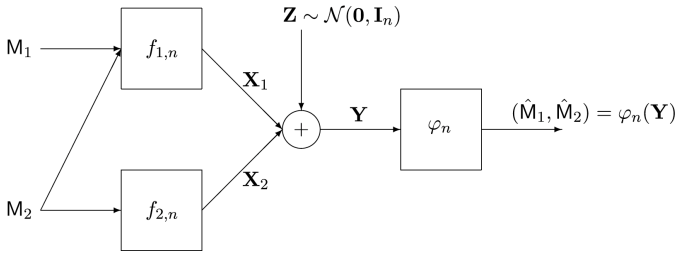
$$R_1 \le \mathsf{C}\big((1-\rho^2)S_1\big) \tag{1}$$

$$R_1 + R_2 \le \mathsf{C}\big(S_1 + S_2 + 2\rho\sqrt{S_1 S_2}\big) \tag{2}$$

for some $\rho \in [0, 1]$, where S_1 and S_2 are the admissible transmit powers, and $\mathsf{C}(x) := \frac{1}{2}\log(1+x)$ is the Gaussian capacity function. The capacity region $\mathcal{C}$ does not depend on

Other Contributions to Second- and Third-Order

Gaussian MAC with degraded message sets with J. Scarlett



Second-Order Asymptotics for the Gaussian MAC With Degraded Message Sets

Jonathan Scarlett, *Member, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

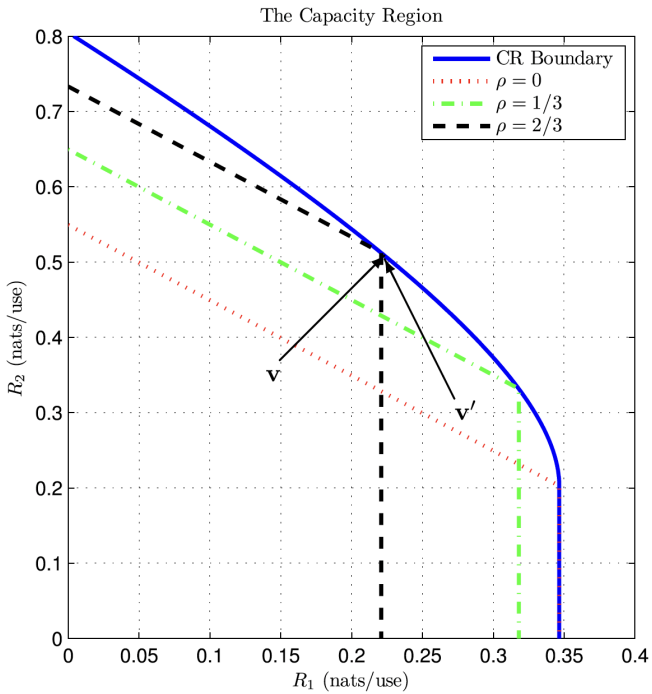
***Abstract*— This paper studies the second-order asymptotics of the Gaussian multiple-access channel with degraded message sets. For a fixed average error probability $\varepsilon \in (0, 1)$ and an arbitrary point on the boundary of the capacity region, we characterize the speed of convergence of rate pairs that converge to that boundary point for codes that have asymptotic error probability no larger than ε. As a stepping stone to this local notion of the second-order asymptotics, we study a global notion, and establish relationships between the two. We provide a numerical example**

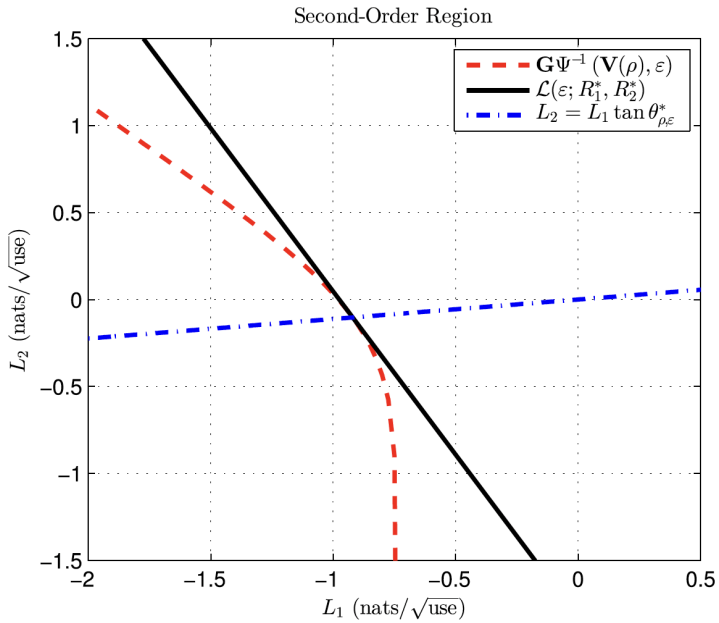
given by the set of rate pairs (R_1, R_2) satisfying

$$R_1 \le \mathsf{C}((1-\rho^2)S_1) \quad (1)$$

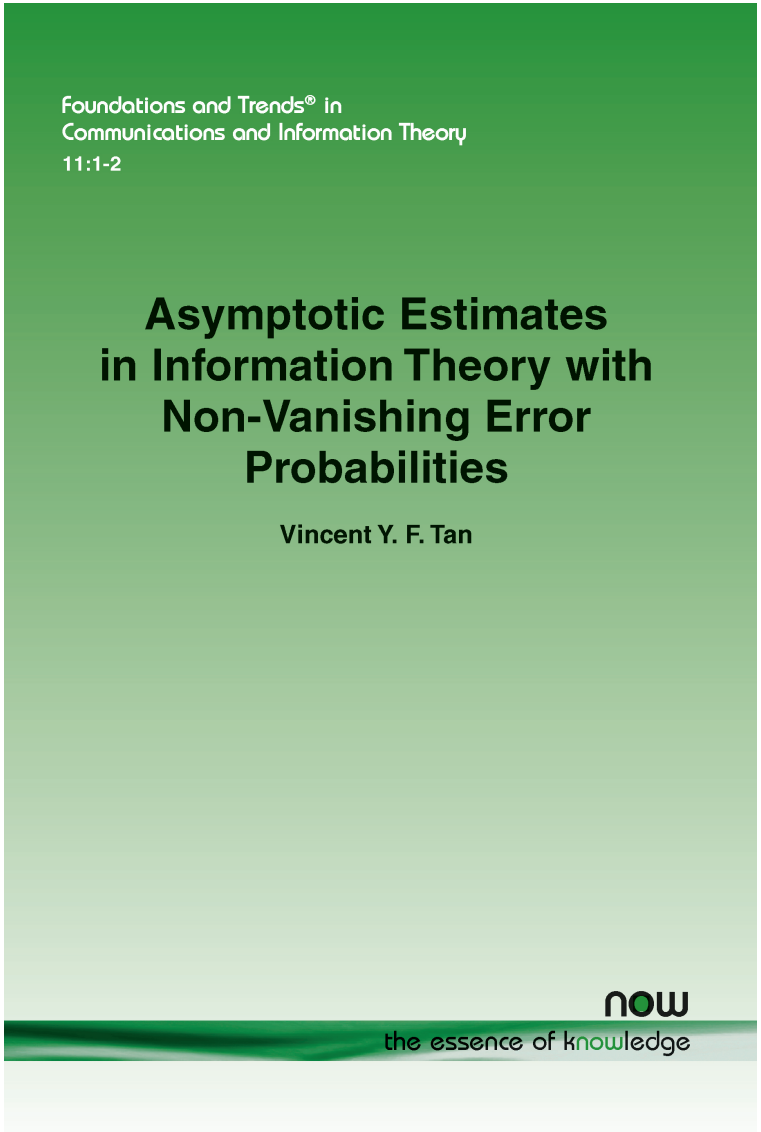
$$R_1 + R_2 \le \mathsf{C}(S_1 + S_2 + 2\rho\sqrt{S_1 S_2}) \quad (2)$$

for some $\rho \in [0, 1]$, where S_1 and S_2 are the admissible transmit powers, and $\mathsf{C}(x) := \frac{1}{2}\log(1+x)$ is the Gaussian capacity function. The capacity region $\mathcal{C}$ does not depend on





Summarized in First Monograph



Outline

Transitioning...

Transitioning...

- (2014 to 2018) Collaborated with M. Hayashi on equivocations for hash functions under Rényi information measures.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 63, NO. 2, FEBRUARY 2017 975

Equivocations, Exponents, and Second-Order Coding Rates Under Various Rényi Information Measures

Masahito Hayashi, *Fellow, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— We evaluate the asymptotics of equivocations, their exponents as well as their second-order coding rates under various Rényi information measures. Specifically, we consider the effect of applying a hash function on a source and we quantify the level of non-uniformity and dependence of the compressed source from another correlated source when the number of copies of the sources is large. Unlike previous works that use Shannon information measures to quantify randomness, information, or uniformity, we define our security measures in terms of a more**

source (A^n, E^n). One of the central tasks in information-theoretic security is to understand the effect of applying a hash function [1] (binning operator) f on A^n. This hash function is used to ensure that the compressed source $f(A^n)$ is almost uniform on its alphabet and also almost independent of another discrete memoryless source E^n. Mathematically, we want to understand the deviation of $f(A^n) \in \{1, \ldots, \lceil e^{nR} \rceil\}$

3734 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 5, MAY 2018

Analysis of Remaining Uncertainties and Exponents Under Various Conditional Rényi Entropies

Vincent Y. F. Tan, *Senior Member, IEEE*, and Masahito Hayashi, *Fellow, IEEE*

***Abstract*— We analyze the asymptotics of the normalized remaining uncertainty of a source when a compressed or hashed version of it and correlated side information is observed. For this system, commonly known as Slepian–Wolf source coding, we establish the optimal (minimum) rate of compression of the source to ensure that the remaining uncertainties vanish. We also study the exponential rate of decay of the remaining uncertainty to zero when the rate is above the optimal rate of compression. In this paper, we consider various classes of random universal hash functions. Instead of measuring remaining uncertainties using traditional Shannon information measures, we do so using two forms of the conditional Rényi entropy.**

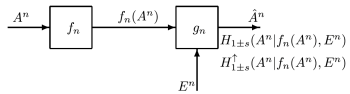


Fig. 1. The Slepian-Wolf [4] source coding problem. We are interesting in quantifying the asymptotic behaviors of the remaining uncertainty of A^n given $(f_n(A^n), E^n)$ measured according to the conditional Rényi entropies $H_{1\pm s}$ and $H^{\uparrow}_{1\pm s}$ defined in (10) and (12).

Transitioning...

- (2014 to 2018) Collaborated with M. Hayashi on equivocations for hash functions under Rényi information measures.

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 63, NO. 2, FEBRUARY 2017 975

Equivocations, Exponents, and Second-Order Coding Rates Under Various Rényi Information Measures

Masahito Hayashi, *Fellow, IEEE*, and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— We evaluate the asymptotics of equivocations, their exponents as well as their second-order coding rates under various Rényi information measures. Specifically, we consider the effect of applying a hash function on a source and we quantify the level of non-uniformity and dependence of the compressed source from another correlated source when the number of copies of the sources is large. Unlike previous works that use Shannon information measures to quantify randomness, information, or uniformity, we define our security measures in terms of a more**

source (A^n, E^n). One of the central tasks in information-theoretic security is to understand the effect of applying a hash function [1] (binning operator) f on A^n. This hash function is used to ensure that the compressed source $f(A^n)$ is almost uniform on its alphabet and also almost independent of another discrete memoryless source E^n. Mathematically, we want to understand the deviation of $f(A^n) \in \{1, \ldots, \lceil e^{nR} \rceil\}$

3734 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 64, NO. 5, MAY 2018

Analysis of Remaining Uncertainties and Exponents Under Various Conditional Rényi Entropies

Vincent Y. F. Tan, *Senior Member, IEEE*, and Masahito Hayashi, *Fellow, IEEE*

***Abstract*— We analyze the asymptotics of the normalized remaining uncertainty of a source when a compressed or hashed version of it and correlated side information is observed. For this system, commonly known as Slepian–Wolf source coding, we establish the optimal (minimum) rate of compression of the source to ensure that the remaining uncertainties vanish. We also study the exponential rate of decay of the remaining uncertainty to zero when the rate is above the optimal rate of compression. In this paper, we consider various classes of random universal hash functions. Instead of measuring remaining uncertainties using traditional Shannon information measures, we do so using two forms of the conditional Rényi entropy.**

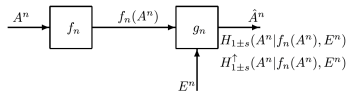


Fig. 1. The Slepian-Wolf [4] source coding problem. We are interesting in quantifying the asymptotic behaviors of the remaining uncertainty of A^n given $(f_n(A^n), E^n)$ measured according to the conditional Rényi entropies $H_{1\pm s}$ and $H^{\uparrow}_{1\pm s}$ defined in (10) and (12).

- (2017 to 2018) Collaborated with Lei Yu on Rényi resolvability.

1862 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 65, NO. 3, MARCH 2019

Rényi Resolvability and Its Applications to the Wiretap Channel

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— The conventional channel resolvability problem refers to the determination of the minimum rate required for an input process so that the output distribution approximates a target distribution in either the total variation distance or the relative entropy. In contrast to previous works, in this paper, we use the (normalized or unnormalized) Rényi divergence (with the Rényi parameter in $[0, 2] \cup \{\infty\}$) to measure the level of approximation. We also provide asymptotic expressions for normalized Rényi divergence when the Rényi parameter is larger than or equal to 1 as well as (lower and upper)**

a target output distribution? This is the so-called *channel resolvability problem*, studied by Han and Verdú [2]. In [2], the total variation (TV) distance and the normalized relative entropy (Kullback-Leibler divergence) were used to measure the level of approximation. The resolvability problem with the *unnormalized* relative entropy was studied by Hayashi [3], [4]. In [2]–[4] it was shown that in the memoryless case the minimum rates of randomness needed for simulating a channel output under the TV, normalized relative entropy, or unnor-

Measures of Information Among Random Variables

- Given two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with joint distribution π_{XY}, how common are they?

Measures of Information Among Random Variables

- Given two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with joint distribution π_{XY}, how common are they?
- Pearson correlation coefficient

$$\rho(X;Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1].$$

Measures of Information Among Random Variables

- Given two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with joint distribution π_{XY}, how common are they?
- Pearson correlation coefficient

$$\rho(X;Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1].$$

- Mutual Information

$$I_\pi(X;Y) = \mathbb{E}_{\pi_{XY}}\left[\log \frac{\pi_{XY}(X,Y)}{\pi_X(X)\pi_Y(Y)}\right] = D(\pi_{XY} \,\|\, \pi_X\pi_Y).$$

Measures of Information Among Random Variables

- Given two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with joint distribution π_{XY}, how common are they?
- Pearson correlation coefficient

$$\rho(X;Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1].$$

- Mutual Information

$$I_\pi(X;Y) = \mathbb{E}_{\pi_{XY}}\left[\log \frac{\pi_{XY}(X,Y)}{\pi_X(X)\pi_Y(Y)}\right] = D(\pi_{XY} \,\|\, \pi_X\pi_Y).$$

- As information theorists, we like operational interpretations

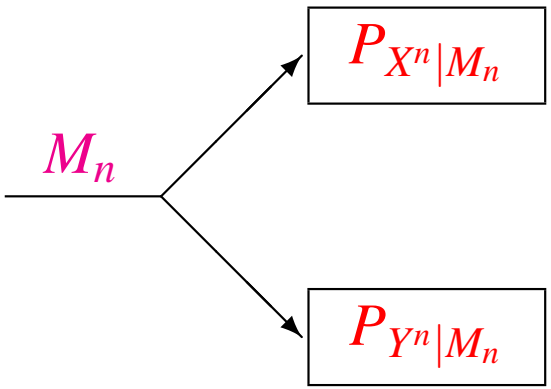
Wyner's Common Information (Wyner, 1975)

Wyner's Common Information (Wyner, 1975)

M_n

- M_n is uniformly distributed over $\mathcal{M}_n := \{1, \ldots, 2^{nR}\}$

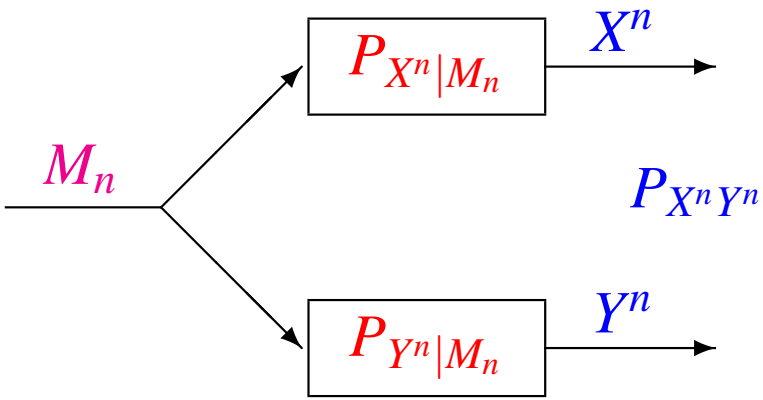
Wyner's Common Information (Wyner, 1975)



- M_n is uniformly distributed over $\mathcal{M}_n := \{1, \ldots, 2^{nR}\}$
- An (n, R)-synthesis code consists of

$$P_{X^n|M_n} : \mathcal{M}_n \to \mathcal{X}^n \quad \text{and} \quad P_{Y^n|M_n} : \mathcal{M}_n \to \mathcal{Y}^n.$$

Wyner's Common Information (Wyner, 1975)



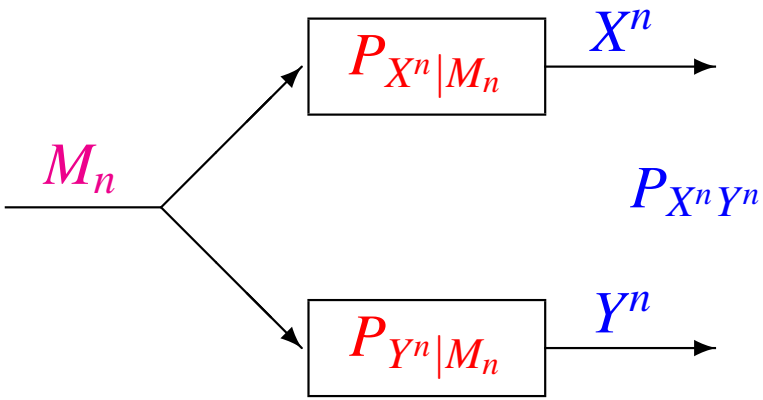
- M_n is uniformly distributed over $\mathcal{M}_n := \{1, \dots, 2^{nR}\}$
- An (n, R)-synthesis code consists of

$$P_{X^n|M_n} : \mathcal{M}_n \to \mathcal{X}^n \quad \text{and} \quad P_{Y^n|M_n} : \mathcal{M}_n \to \mathcal{Y}^n.$$

- The distribution induced by the code $(P_{X^n|M_n}, P_{Y^n|M_n})$ is

$$P_{X^nY^n}(x^n, y^n) := \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} P_{X^n|M_n}(x^n|m) P_{Y^n|M_n}(y^n|m)$$

Wyner's Common Information (Wyner, 1975)



- M_n is uniformly distributed over $\mathcal{M}_n := \{1, \dots, 2^{nR}\}$
- An (n, R)-synthesis code consists of

$$P_{X^n|M_n} : \mathcal{M}_n \to \mathcal{X}^n \quad \text{and} \quad P_{Y^n|M_n} : \mathcal{M}_n \to \mathcal{Y}^n.$$

- The distribution induced by the code $(P_{X^n|M_n}, P_{Y^n|M_n})$ is

$$P_{X^nY^n}(x^n, y^n) := \frac{1}{|\mathcal{M}_n|} \sum_{m \in \mathcal{M}_n} P_{X^n|M_n}(x^n|m) P_{Y^n|M_n}(y^n|m)$$

- Desideratum:

$$P_{X^nY^n} \approx \pi_{XY}^n \quad \text{(target distribution)}$$

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Normalized relative entropy to measure "distance" btw. $P_{X^nY^n}$ and π_{XY}^n

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Normalized relative entropy to measure "distance" btw. $P_{X^nY^n}$ and π_{XY}^n

Theorem (Wyner (1975))

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\}$$

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Normalized relative entropy to measure "distance" btw. $P_{X^nY^n}$ and π_{XY}^n

Theorem (Wyner (1975))

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\}$$
$$= \min_{P_W P_{X|W} P_{Y|W} : P_{XY} = \pi_{XY}} I(XY; W)$$

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Normalized relative entropy to measure "distance" btw. $P_{X^nY^n}$ and π_{XY}^n

Theorem (Wyner (1975))

$$\begin{aligned}
&\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} \\
&= \min_{P_W P_{X|W} P_{Y|W} : P_{XY} = \pi_{XY}} I(XY; W) \\
&=: C_{\mathrm{W}}(\pi_{XY})
\end{aligned}$$

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Normalized relative entropy to measure "distance" btw. $P_{X^nY^n}$ and π_{XY}^n

Theorem (Wyner (1975))

$$\begin{aligned}
&\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} \\
&= \min_{P_W P_{X|W} P_{Y|W} : P_{XY} = \pi_{XY}} I(XY; W) \\
&=: C_{\mathrm{W}}(\pi_{XY})
\end{aligned}$$

Wyner's Common Information (Wyner, 1975)

IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. IT-21, NO. 2, MARCH 1975

The Common Information of Two Dependent Random Variables

AARON D. WYNER, SENIOR MEMBER, IEEE

Normalized relative entropy to measure "distance" btw. $P_{X^nY^n}$ and π_{XY}^n

Theorem (Wyner (1975))

$$\begin{aligned}&\inf\left\{R:\lim_{n\to\infty}\frac{D(P_{X^nY^n}\,\|\,\pi_{XY}^n)}{n}=0\right\}\\&=\min_{P_WP_{X|W}P_{Y|W}:\,P_{XY}=\pi_{XY}} I(XY;W)\\&=:C_{\mathrm{W}}(\pi_{XY})\end{aligned}$$

where $C_{\mathrm{W}}(\pi_{XY})$ is named Wyner's Common Information.

Example: Doubly Symmetric Binary Source (DSBS)

- DSBS $(X, Y) \in \{0, 1\}^2$

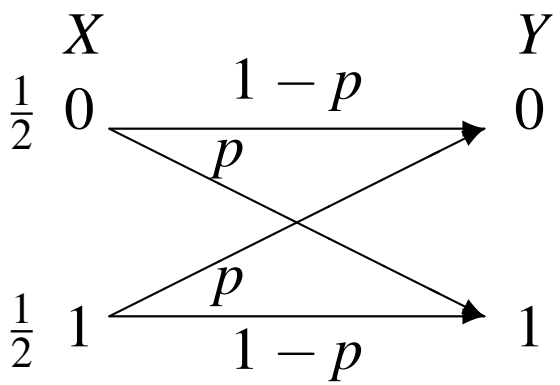
$$\pi_{XY} = \begin{bmatrix} (1-p)/2 & p/2 \\ p/2 & (1-p)/2 \end{bmatrix}$$

Example: Doubly Symmetric Binary Source (DSBS)

- DSBS $(X, Y) \in \{0, 1\}^2$

$$\pi_{XY} = \begin{bmatrix} (1-p)/2 & p/2 \\ p/2 & (1-p)/2 \end{bmatrix}$$

- Interpretation in terms of $X - W - Y$

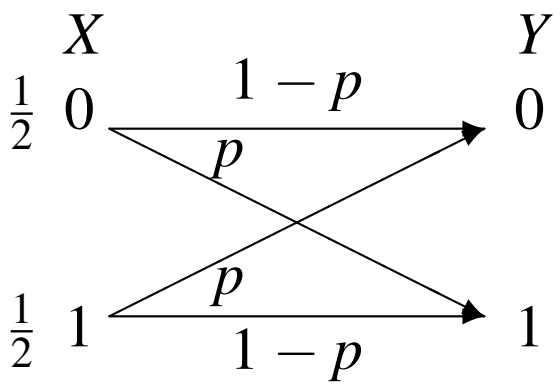


Example: Doubly Symmetric Binary Source (DSBS)

- DSBS $(X, Y) \in \{0, 1\}^2$

$$\pi_{XY} = \begin{bmatrix} (1-p)/2 & p/2 \\ p/2 & (1-p)/2 \end{bmatrix}$$

- Interpretation in terms of $X - W - Y$

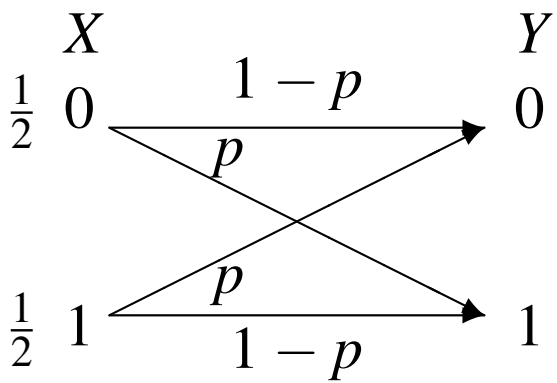


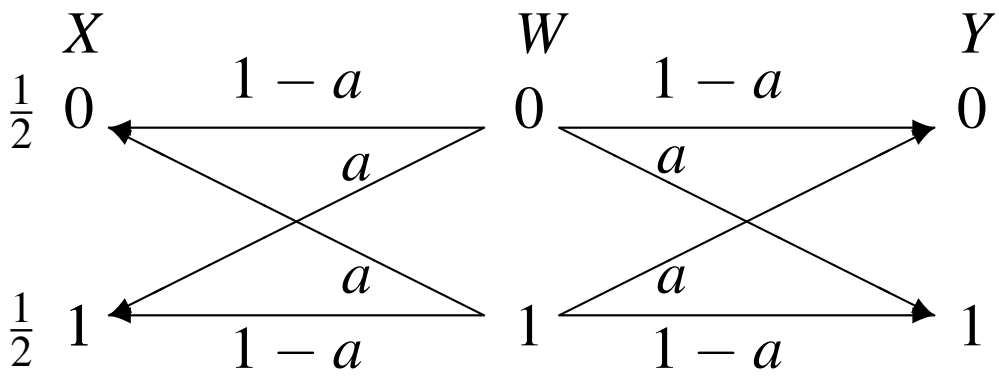
Example: Doubly Symmetric Binary Source (DSBS)

- DSBS $(X, Y) \in \{0, 1\}^2$

$$\pi_{XY} = \begin{bmatrix} (1-p)/2 & p/2 \\ p/2 & (1-p)/2 \end{bmatrix}$$

- Interpretation in terms of $X - W - Y$



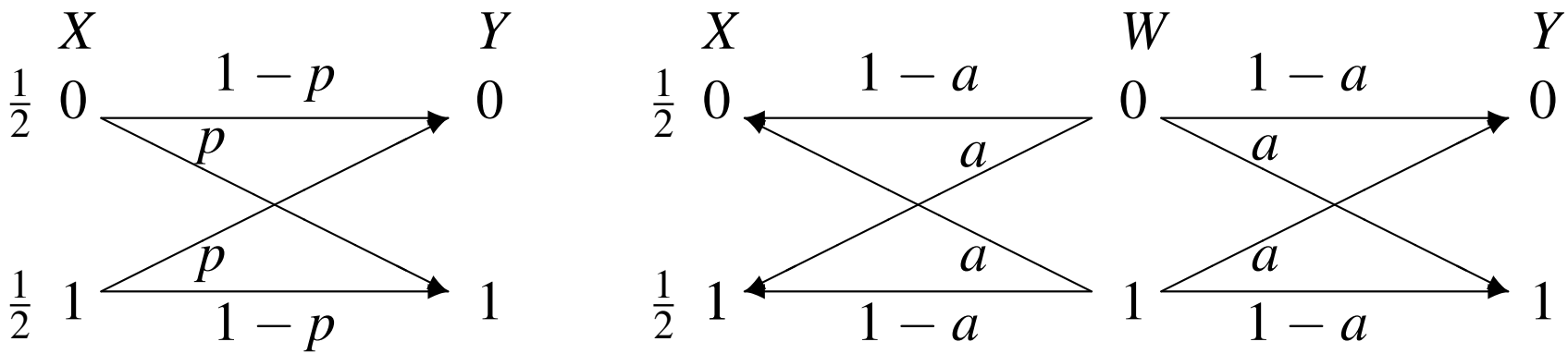


Example: Doubly Symmetric Binary Source (DSBS)

- DSBS $(X, Y) \in \{0, 1\}^2$

$$\pi_{XY} = \begin{bmatrix} (1-p)/2 & p/2 \\ p/2 & (1-p)/2 \end{bmatrix}$$

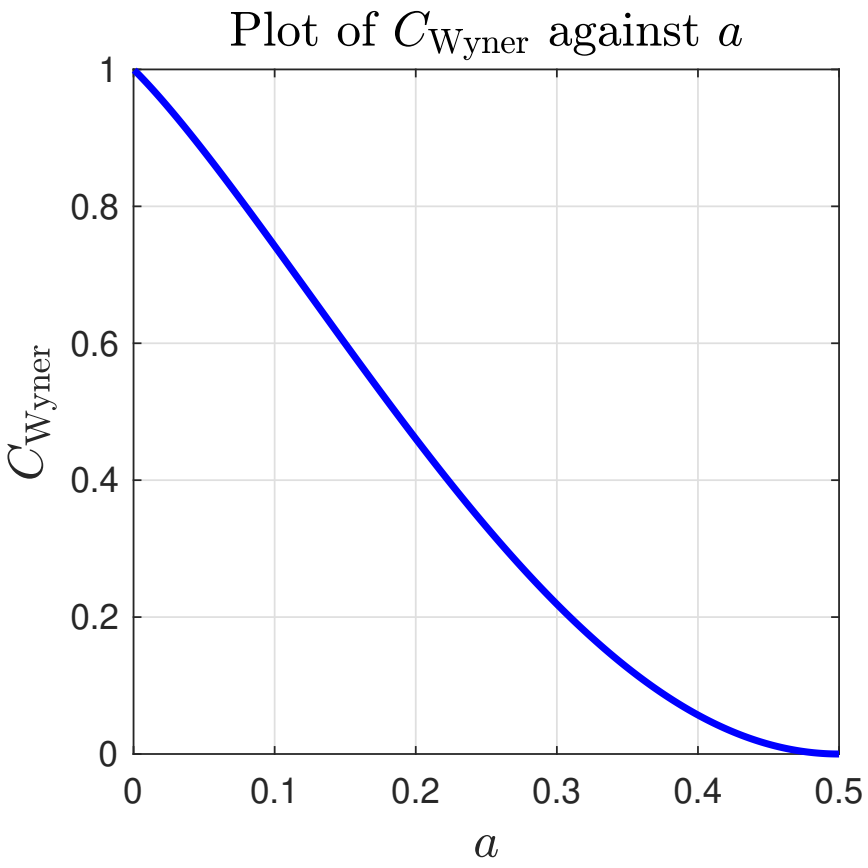
- Interpretation in terms of $X - W - Y$

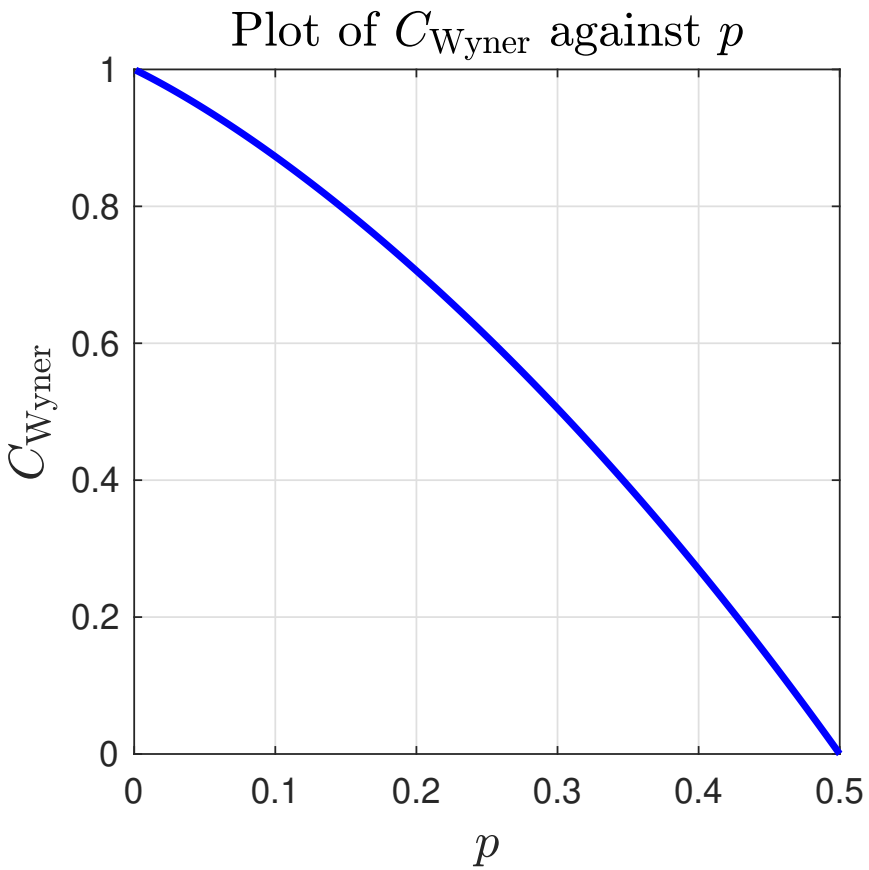


- Here, $a * a = p$ and

$$a = \frac{1 - \sqrt{1-2p}}{2} \in (0, 1/2).$$

Example: DSBS





Plots of Wyner's common information for the DSBS in terms of p and a

Motivation for Alternative Measures

Motivation for Alternative Measures

- Wyner used the normalized relative entropy, i.e.,

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} = C_{\mathrm{W}}(\pi_{XY}) = \min_{X-W-Y} I(W;XY).$$

Motivation for Alternative Measures

- Wyner used the normalized relative entropy, i.e.,

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} = C_{\mathrm{W}}(\pi_{XY}) = \min_{X-W-Y} I(W;XY).$$

- What if we do not normalize?

$$\tilde{T}(\pi_{XY}) := \inf\left\{R : \lim_{n\to\infty} D(P_{X^nY^n} \,\|\, \pi_{XY}^n) = 0\right\} \geq C_{\mathrm{W}}(\pi_{XY}).$$

We get a stronger measure of dependence.

Motivation for Alternative Measures

- Wyner used the normalized relative entropy, i.e.,

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} = C_{\mathrm{W}}(\pi_{XY}) = \min_{X-W-Y} I(W;XY).$$

- What if we do not normalize?

$$\tilde{T}(\pi_{XY}) := \inf\left\{R : \lim_{n\to\infty} D(P_{X^nY^n} \,\|\, \pi_{XY}^n) = 0\right\} \geq C_{\mathrm{W}}(\pi_{XY}).$$

We get a stronger measure of dependence.

Motivation for Alternative Measures

- Wyner used the normalized relative entropy, i.e.,

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} = C_{\mathrm{W}}(\pi_{XY}) = \min_{X-W-Y} I(W;XY).$$

- What if we do not normalize?

$$\tilde{T}(\pi_{XY}) := \inf\left\{R : \lim_{n\to\infty} D(P_{X^nY^n} \,\|\, \pi_{XY}^n) = 0\right\} \geq C_{\mathrm{W}}(\pi_{XY}).$$

We get a stronger measure of dependence. Even stronger?

Motivation for Alternative Measures

- Wyner used the normalized relative entropy, i.e.,

$$\inf\left\{R : \lim_{n\to\infty} \frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\} = C_{\mathrm{W}}(\pi_{XY}) = \min_{X-W-Y} I(W;XY).$$

- What if we do not normalize?

$$\tilde{T}(\pi_{XY}) := \inf\left\{R : \lim_{n\to\infty} D(P_{X^nY^n} \,\|\, \pi_{XY}^n) = 0\right\} \geq C_{\mathrm{W}}(\pi_{XY}).$$

We get a stronger measure of dependence. Even stronger?

- Rényi common information for orders ≥ 1 (Yu and Tan, 2018)!

$$T_{1+s}(\pi_{XY}) := \inf\left\{R : \lim_{n\to\infty} \frac{D_{1+s}(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} = 0\right\}$$

$$\tilde{T}_{1+s}(\pi_{XY}) := \inf\left\{R : \lim_{n\to\infty} D_{1+s}(P_{X^nY^n} \,\|\, \pi_{XY}^n) = 0\right\}$$

Rényi Common Information

Rényi Common Information

- Rényi divergence

$$D_{1+s}(P\,\|\,Q) := \frac{1}{s}\log \sum_{x\in\mathrm{supp}(P)} P(x)^{1+s}Q(x)^{-s} \quad s\in[-1,\infty)$$

$$D_{\infty}(P\,\|\,Q) := \lim_{s\to\infty} D_{1+s}(P\,\|\,Q) = \log \max_{x\in\mathrm{supp}(P)} \frac{P(x)}{Q(x)}.$$

Rényi Common Information

- Rényi divergence

$$D_{1+s}(P\,\|\,Q) := \frac{1}{s}\log \sum_{x\in\mathrm{supp}(P)} P(x)^{1+s}Q(x)^{-s} \quad s\in[-1,\infty)$$

$$D_{\infty}(P\,\|\,Q) := \lim_{s\to\infty} D_{1+s}(P\,\|\,Q) = \log \max_{x\in\mathrm{supp}(P)} \frac{P(x)}{Q(x)}.$$

- The Rényi divergence if monotonically non-decreasing, i.e.,

$$D_{1+s}(P\,\|\,Q) \le D_{1+t}(P\,\|\,Q) \qquad s\le t.$$

Rényi Common Information

- Rényi divergence

$$D_{1+s}(P\,\|\,Q) := \frac{1}{s}\log \sum_{x\in\mathrm{supp}(P)} P(x)^{1+s}Q(x)^{-s} \quad s\in[-1,\infty)$$

$$D_{\infty}(P\,\|\,Q) := \lim_{s\to\infty} D_{1+s}(P\,\|\,Q) = \log \max_{x\in\mathrm{supp}(P)} \frac{P(x)}{Q(x)}.$$

- The Rényi divergence if monotonically non-decreasing, i.e.,

$$D_{1+s}(P\,\|\,Q) \le D_{1+t}(P\,\|\,Q) \qquad s\le t.$$

- Hence, the Rényi common information is also non-decreasing, i.e.,

$$\text{(normalized)} \qquad T_{1+s}(\pi_{XY}) \le T_{1+t}(\pi_{XY}) \qquad s\le t$$

$$\text{(unnormalized)} \qquad \tilde{T}_{1+s}(\pi_{XY}) \le \tilde{T}_{1+t}(\pi_{XY}) \qquad s\le t.$$

Rényi Common Information

- Rényi divergence

$$D_{1+s}(P\,\|\,Q) := \frac{1}{s}\log \sum_{x\in\mathrm{supp}(P)} P(x)^{1+s}Q(x)^{-s} \quad s\in[-1,\infty)$$

$$D_{\infty}(P\,\|\,Q) := \lim_{s\to\infty} D_{1+s}(P\,\|\,Q) = \log \max_{x\in\mathrm{supp}(P)} \frac{P(x)}{Q(x)}.$$

- The Rényi divergence if monotonically non-decreasing, i.e.,

$$D_{1+s}(P\,\|\,Q) \le D_{1+t}(P\,\|\,Q) \qquad s\le t.$$

- Hence, the Rényi common information is also non-decreasing, i.e.,

$$\text{(normalized)} \qquad T_{1+s}(\pi_{XY}) \le T_{1+t}(\pi_{XY}) \qquad s\le t$$

$$\text{(unnormalized)} \qquad \tilde{T}_{1+s}(\pi_{XY}) \le \tilde{T}_{1+t}(\pi_{XY}) \qquad s\le t.$$

- And for a fixed order $1+s\in[0,\infty]$,

$$\text{(normalized)} \qquad T_{1+s}(\pi_{XY}) \le \tilde{T}_{1+s}(\pi_{XY}) \qquad \text{(unnormalized)}.$$

Rényi CI: The Case $s > 0$

- For $s > 0$,

$$C_{\mathrm{W}}(\pi_{XY}) = T_1(\pi_{XY})$$

Rényi CI: The Case $s > 0$

- For $s > 0$,

$$C_{\mathrm{W}}(\pi_{XY}) = T_1(\pi_{XY}) \leq T_{1+s}(\pi_{XY})$$

Rényi CI: The Case $s > 0$

- For $s > 0$,

$$C_{\mathrm{W}}(\pi_{XY}) = T_1(\pi_{XY}) \leq T_{1+s}(\pi_{XY}) \leq \underbrace{T_\infty(\pi_{XY})}_{\text{normalized}}$$

Rényi CI: The Case $s > 0$

- For $s > 0$,

$$C_{\mathrm{W}}(\pi_{XY}) = T_1(\pi_{XY}) \leq T_{1+s}(\pi_{XY}) \leq \underbrace{T_\infty(\pi_{XY})}_{\text{normalized}} \leq \underbrace{\tilde{T}_\infty(\pi_{XY})}_{\text{unnormalized}} .$$

Rényi CI: The Case $s > 0$

- For $s > 0$,

$$C_{\mathrm{W}}(\pi_{XY}) = T_1(\pi_{XY}) \leq T_{1+s}(\pi_{XY}) \leq \underbrace{T_\infty(\pi_{XY})}_{\text{normalized}} \leq \underbrace{\tilde{T}_\infty(\pi_{XY})}_{\text{unnormalized}} .$$

- Only concerned with $s = \infty$.

Rényi CI: The Case $s > 0$

- For $s > 0$,

$$C_{\mathrm{W}}(\pi_{XY}) = T_1(\pi_{XY}) \le T_{1+s}(\pi_{XY}) \le \underbrace{T_\infty(\pi_{XY})}_{\text{normalized}} \le \underbrace{\tilde{T}_\infty(\pi_{XY})}_{\text{unnormalized}} .$$

- Only concerned with $s = \infty$.

Definition

Maximal cross entropy w.r.t. $(X, Y) \sim \pi_{XY}$ over couplings of (P_X, P_Y) is

$$\mathsf{H}_\infty(P_X, P_Y \,\|\, \pi_{XY}) := \max_{Q_{XY} \in \mathcal{C}(P_X, P_Y)} \sum_{x,y} Q_{XY}(x,y) \log \frac{1}{\pi_{XY}(x,y)},$$

where $\mathcal{C}(P_X, P_Y) := \{Q_{XY} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : Q_X = P_X, Q_Y = P_Y\}$.

Upper and Lower Pseudo Common Informations

Definition

The upper pseudo-common information is

$$\overline{\Gamma}_\infty(\pi_{XY}) := \min_{\substack{P_W P_{X|W} P_{Y|W}: \\ P_{XY}=\pi_{XY}}} -H(XY|W) + \mathbb{E}_{P_W}\left[\mathsf{H}_\infty(P_{X|W}, P_{Y|W} \,\|\, \pi_{XY})\right].$$

The lower pseudo-common information is

$$\begin{aligned}\underline{\Gamma}_\infty(\pi_{XY}) := &\min_{\substack{P_W P_{X|W} P_{Y|W}: \\ P_{XY}=\pi_{XY}}} -H(XY|W) \\ &+ \inf_{Q_{WW'} \in \mathcal{C}(P_W, P_W)} \mathbb{E}_{Q_{WW'}}\left[\mathsf{H}_\infty(P_{X|W}, P_{Y|W'} \,\|\, \pi_{XY})\right].\end{aligned}$$

Upper and Lower Pseudo Common Informations

Definition

The upper pseudo-common information is

$$\overline{\Gamma}_\infty(\pi_{XY}) := \min_{\substack{P_W P_{X|W} P_{Y|W}: \\ P_{XY}=\pi_{XY}}} -H(XY|W) + \mathbb{E}_{P_W}\left[\mathsf{H}_\infty(P_{X|W}, P_{Y|W} \,\|\, \pi_{XY})\right].$$

The lower pseudo-common information is

$$\begin{aligned}\underline{\Gamma}_\infty(\pi_{XY}) := &\min_{\substack{P_W P_{X|W} P_{Y|W}: \\ P_{XY}=\pi_{XY}}} -H(XY|W) \\ &+ \inf_{Q_{WW'} \in \mathcal{C}(P_W, P_W)} \mathbb{E}_{Q_{WW'}}\left[\mathsf{H}_\infty(P_{X|W}, P_{Y|W'} \,\|\, \pi_{XY})\right].\end{aligned}$$

cf. $$C_{\mathrm{W}}(\pi_{XY}) = \min_{P_W P_{X|W} P_{Y|W}: P_{XY}=\pi_{XY}} -H(XY|W) + H(XY).$$

Rényi Common Information of order ∞

Theorem (Yu and Tan (2020))

$$\max\left\{\underline{\Gamma}_\infty(\pi_{XY}), C_{\mathrm{W}}(\pi_{XY})\right\} \le T_\infty(\pi_{XY}) \le \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

Rényi Common Information of order ∞

Theorem (Yu and Tan (2020))

$$\max\left\{\underline{\Gamma}_\infty(\pi_{XY}), C_{\mathrm{W}}(\pi_{XY})\right\} \le T_\infty(\pi_{XY}) \le \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

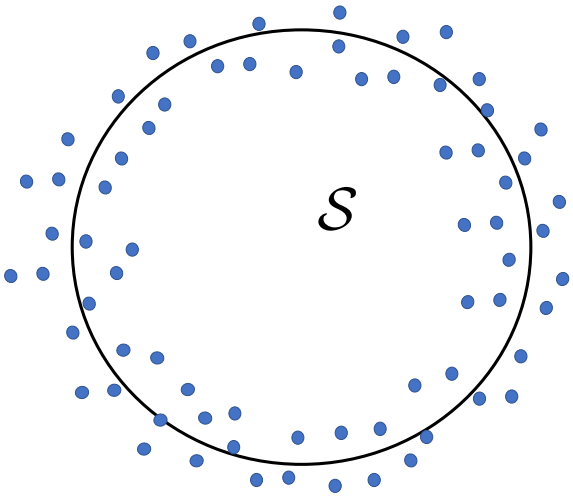
Achievability: Rényi soft-covering (Yu and Tan, 2019) and truncated product distributions.

Rényi Common Information of order ∞

Theorem (Yu and Tan (2020))

$$\max\left\{\underline{\Gamma}_\infty(\pi_{XY}), C_{\mathrm{W}}(\pi_{XY})\right\} \le T_\infty(\pi_{XY}) \le \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

Achievability: Rényi soft-covering (Yu and Tan, 2019) and truncated product distributions.



Product distribution

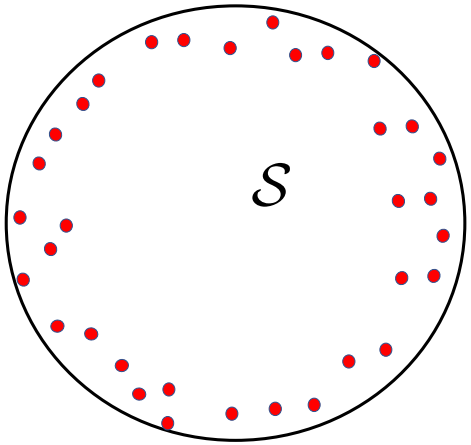
$$P_W^n(w^n) = \prod_{i=1}^n P_W(w_i)$$

Rényi Common Information of order ∞

Theorem (Yu and Tan (2020))

$$\max\left\{\underline{\Gamma}_\infty(\pi_{XY}), C_{\mathrm{W}}(\pi_{XY})\right\} \le T_\infty(\pi_{XY}) \le \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

Achievability: Rényi soft-covering (Yu and Tan, 2019) and truncated product distributions.



Truncated product distribution

$$P_{W^n}(w^n) \propto \left(\prod_{i=1}^{n} P_W(w_i)\right) \mathbb{1}\{w^n \in \mathcal{S}\}$$

Exact Common Information

- In Wyner's distributed source simulation problem,

$$\frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} \to 0.$$

Exact Common Information

- In Wyner's distributed source simulation problem,

$$\frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} \to 0.$$

- What if we require

$$P_{X^nY^n} = \pi_{XY}^n \quad \text{for some} \quad n \in \mathbb{N}?$$

Exact Common Information

- In Wyner's distributed source simulation problem,

$$\frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} \to 0.$$

- What if we require

$$P_{X^nY^n} = \pi_{XY}^n \quad \text{for some} \quad n \in \mathbb{N}?$$

- Fixed-length block codes $\implies$ Rate $\leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$.

Exact Common Information

- In Wyner's distributed source simulation problem,

$$\frac{D(P_{X^nY^n} \,\|\, \pi_{XY}^n)}{n} \to 0.$$

- What if we require

$$P_{X^nY^n} = \pi_{XY}^n \quad \text{for some} \quad n \in \mathbb{N}?$$

- Fixed-length block codes $\implies$ Rate $\leq \min\{\log|\mathcal{X}|, \log|\mathcal{Y}|\}$.
- Kumar, Li, and El Gamal (2014) introduced

2014 IEEE International Symposium on Information Theory

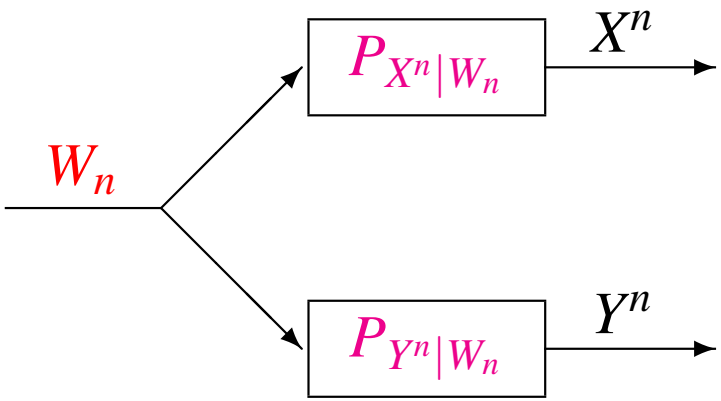
Exact Common Information

Gowtham Ramani Kumar
Electrical Engineering
Stanford University
Email: gowthamr@stanford.edu

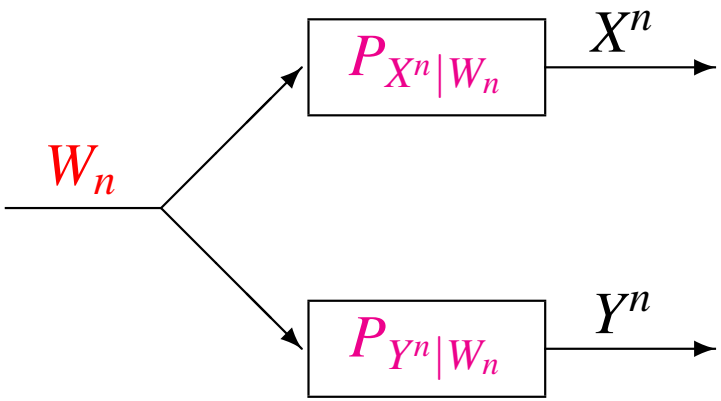
Cheuk Ting Li
Electrical Engineering
Stanford University
Email: ctli@stanford.edu

Abbas El Gamal
Electrical Engineering
Stanford University
Email: abbas@stanford.edu

Exact Common Information

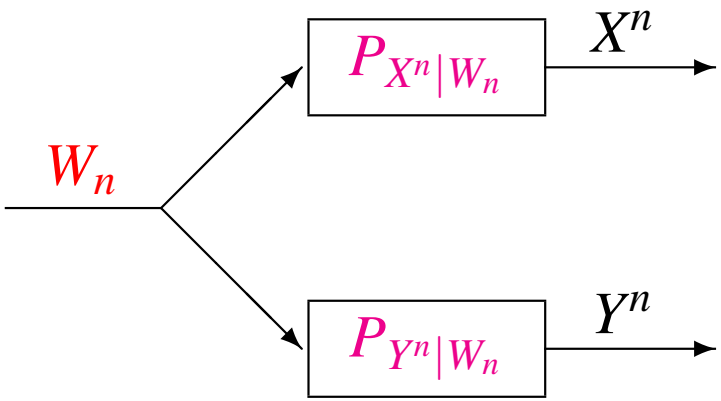


Exact Common Information



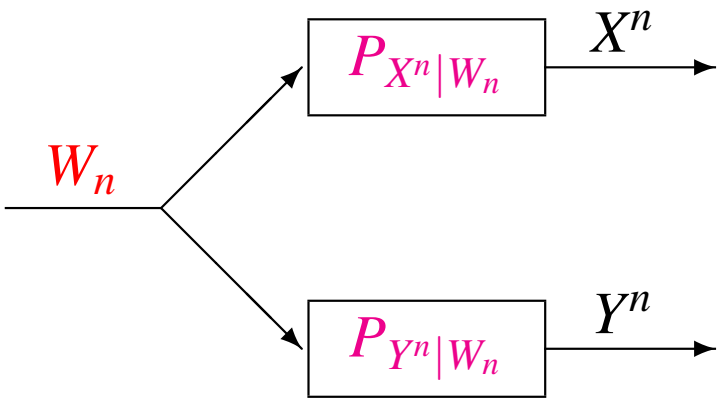
- A synthesis code $(P_{W_n}, P_{X^n|W_n}, P_{Y^n|W_n})$.

Exact Common Information



- A synthesis code $(P_{W_n}, P_{X^n|W_n}, P_{Y^n|W_n})$.
- W_n can be any not necessarily uniform discrete random variable.

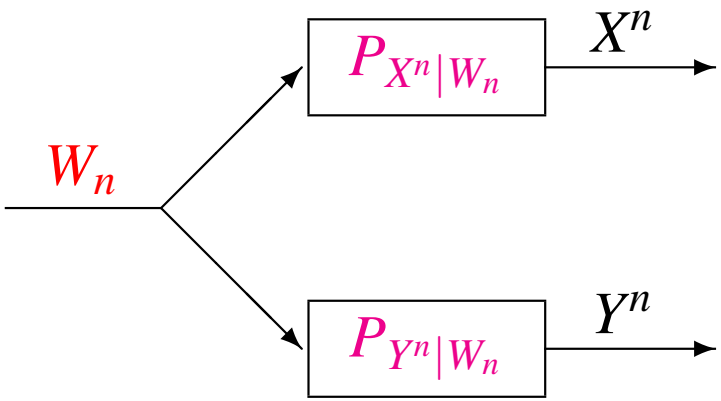
Exact Common Information



- A synthesis code $(P_{W_n}, P_{X^n|W_n}, P_{Y^n|W_n})$.
- W_n can be any not necessarily uniform discrete random variable.
- Distribution induced by the code is

$$P_{X^nY^n}(x^n, y^n) := \sum_w P_{W_n}(w) P_{X^n|W_n}(x^n|w) P_{Y^n|W_n}(y^n|w).$$

Exact Common Information



- A synthesis code $(P_{W_n}, P_{X^n|W_n}, P_{Y^n|W_n})$.
- W_n can be any not necessarily uniform discrete random variable.
- Distribution induced by the code is

$$P_{X^nY^n}(x^n, y^n) := \sum_w P_{W_n}(w) P_{X^n|W_n}(x^n|w) P_{Y^n|W_n}(y^n|w).$$

- Require

$$P_{X^nY^n} = \pi_{XY}^n \quad \text{for some} \quad n \in \mathbb{N}.$$

Exact Common Information

Asymptotic rate induced by the code is

$$\lim_{n\to\infty} \frac{H(W_n)}{n}$$

Exact Common Information

Asymptotic rate induced by the code is

$$\lim_{n\to\infty} \frac{H(W_n)}{n}$$

- Compress W_n by a prefix-free, zero-error variable-length code.

Exact Common Information

Asymptotic rate induced by the code is

$$\lim_{n\to\infty} \frac{H(W_n)}{n}$$

- Compress W_n by a prefix-free, zero-error variable-length code.
- Let the length of W_n be $\ell(W_n)$.

Exact Common Information

Asymptotic rate induced by the code is

$$\lim_{n\to\infty} \frac{H(W_n)}{n}$$

- Compress W_n by a prefix-free, zero-error variable-length code.
- Let the length of W_n be $\ell(W_n)$.
- Optimal expected codeword length $L(W_n) = \mathbb{E}[\ell(W_n)]$ satisfies

$$H(W_n) \le L(W_n) < H(W_n) + 1$$

which implies that

$$\lim_{n\to\infty} \frac{L(W_n)}{n} = \lim_{n\to\infty} \frac{H(W_n)}{n}.$$

Exact Common Information

Definition

Exact common information between $(X, Y) \sim \pi_{XY}$

$$T_{\mathrm{Ex}}(\pi_{XY}) := \inf \left\{ \lim_{n\to\infty} \frac{L(W_n)}{n} : P_{X^nY^n} = \pi_{XY}^n \text{ for some } n \geq 1 \right\}$$

Exact Common Information

Definition

Exact common information between $(X, Y) \sim \pi_{XY}$

$$T_{\mathrm{Ex}}(\pi_{XY}) := \inf \left\{ \lim_{n\to\infty} \frac{L(W_n)}{n} : P_{X^nY^n} = \pi_{XY}^n \text{ for some } n \geq 1 \right\}$$

Theorem (Kumar, Li, and El Gamal (2014))

$$T_{\mathrm{Ex}}(\pi_{XY}) = \lim_{n\to\infty} \min_{X^n - W_n - Y^n : P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}.$$

Exact Common Information

Definition

Exact common information between $(X, Y) \sim \pi_{XY}$

$$T_{\mathrm{Ex}}(\pi_{XY}) := \inf \left\{ \lim_{n\to\infty} \frac{L(W_n)}{n} : P_{X^nY^n} = \pi_{XY}^n \text{ for some } n \geq 1 \right\}$$

Theorem (Kumar, Li, and El Gamal (2014))

$$T_{\mathrm{Ex}}(\pi_{XY}) = \lim_{n\to\infty} \min_{X^n - W_n - Y^n : P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}.$$

- Multi-letter characterization!

Exact Common Information

Definition

Exact common information between $(X, Y) \sim \pi_{XY}$

$$T_{\mathrm{Ex}}(\pi_{XY}) := \inf \left\{ \lim_{n\to\infty} \frac{L(W_n)}{n} : P_{X^nY^n} = \pi_{XY}^n \text{ for some } n \geq 1 \right\}$$

Theorem (Kumar, Li, and El Gamal (2014))

$$T_{\mathrm{Ex}}(\pi_{XY}) = \lim_{n\to\infty} \min_{X^n - W_n - Y^n : P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}.$$

- Multi-letter characterization!
- Exact CI $\geq$ Wyner's CI

Exact Common Information

Definition

Exact common information between $(X, Y) \sim \pi_{XY}$

$$T_{\mathrm{Ex}}(\pi_{XY}) := \inf \left\{ \lim_{n\to\infty} \frac{L(W_n)}{n} : P_{X^nY^n} = \pi_{XY}^n \text{ for some } n \geq 1 \right\}$$

Theorem (Kumar, Li, and El Gamal (2014))

$$T_{\mathrm{Ex}}(\pi_{XY}) = \lim_{n\to\infty} \min_{X^n - W_n - Y^n : P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}.$$

- Multi-letter characterization!
- Exact CI $\geq$ Wyner's CI
- Exact CI $>$ Wyner's CI?
- Was an open problem

Exact Common Information

Definition

Exact common information between $(X, Y) \sim \pi_{XY}$

$$T_{\mathrm{Ex}}(\pi_{XY}) := \inf \left\{ \lim_{n\to\infty} \frac{L(W_n)}{n} : P_{X^nY^n} = \pi_{XY}^n \text{ for some } n \geq 1 \right\}$$

Theorem (Kumar, Li, and El Gamal (2014))

$$T_{\mathrm{Ex}}(\pi_{XY}) = \lim_{n\to\infty} \min_{X^n - W_n - Y^n \,:\, P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}.$$

- Multi-letter characterization!
- Exact CI $\geq$ Wyner's CI
- Exact CI $>$ Wyner's CI?
- Was an open problem

As expected the exact common information rate is greater than or equal to the Wyner common information.

Proposition 3.

$$\overline{G}(X;Y) \geq J(X;Y).$$

In the following section, we show that they are equal for the SBES in Example 1. We do not know if this is the case in general, however.

Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

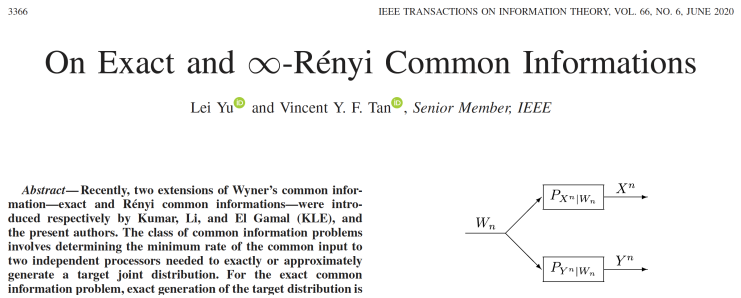
$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$

Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$



On Exact and ∞-Rényi Common Informations

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

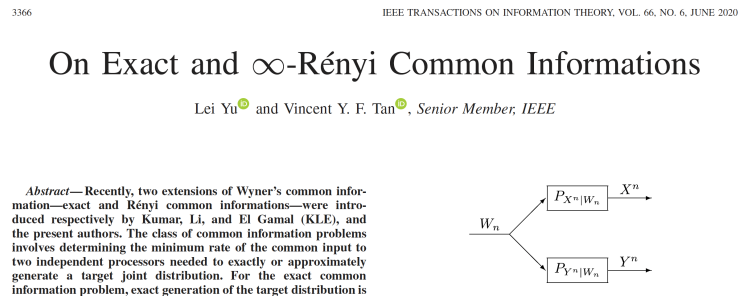
***Abstract*— Recently, two extensions of Wyner's common information—exact and Rényi common informations—were introduced respectively by Kumar, Li, and El Gamal (KLE), and the present authors. The class of common information problems involves determining the minimum rate of the common input to two independent processors needed to exactly or approximately generate a target joint distribution. For the exact common information problem, exact generation of the target distribution is**

Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$

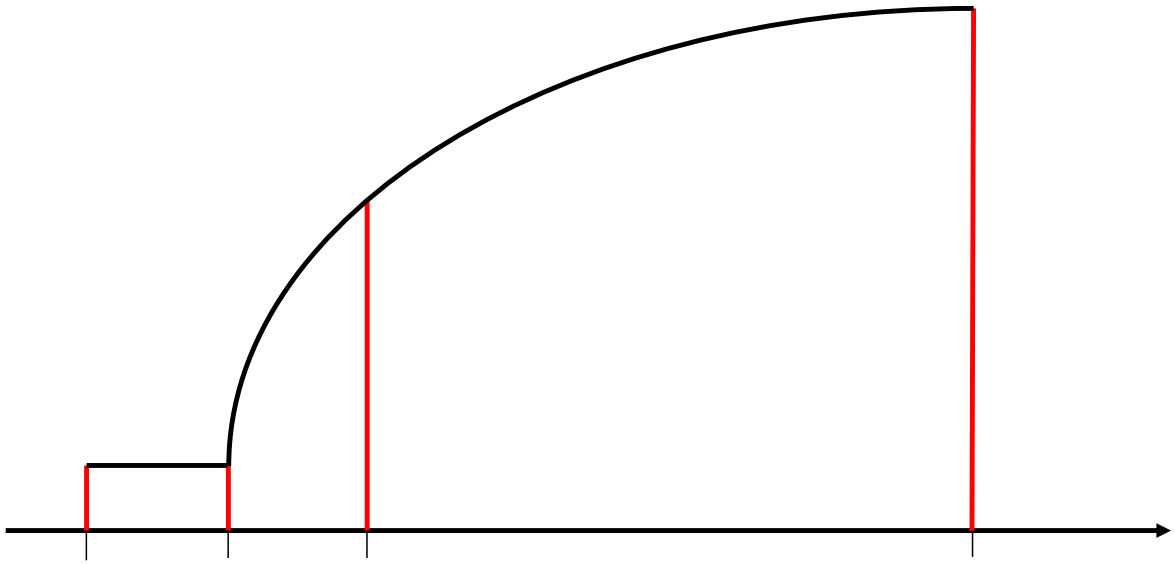


3366 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 6, JUNE 2020

On Exact and ∞-Rényi Common Informations

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— Recently, two extensions of Wyner's common information—exact and Rényi common informations—were introduced respectively by Kumar, Li, and El Gamal (KLE), and the present authors. The class of common information problems involves determining the minimum rate of the common input to two independent processors needed to exactly or approximately generate a target joint distribution. For the exact common information problem, exact generation of the target distribution is**

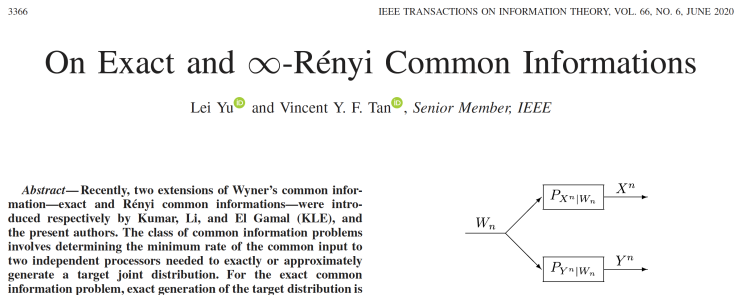


Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$

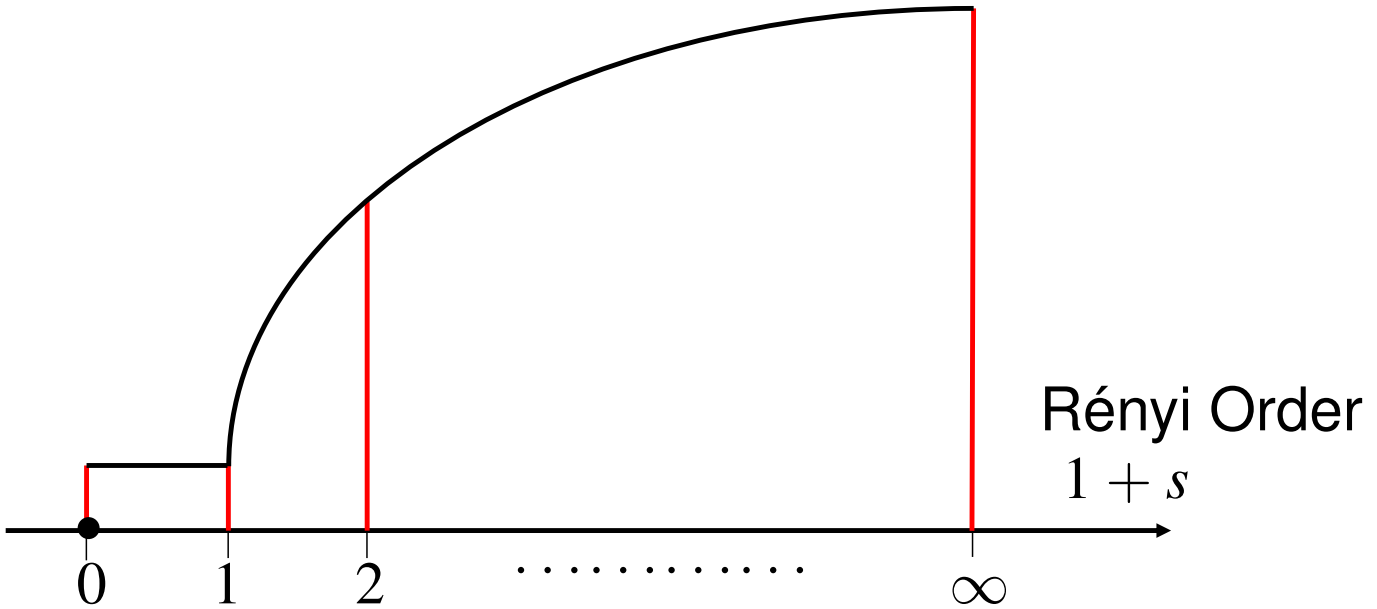


3366 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 6, JUNE 2020

On Exact and ∞-Rényi Common Informations

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— Recently, two extensions of Wyner's common information—exact and Rényi common informations—were introduced respectively by Kumar, Li, and El Gamal (KLE), and the present authors. The class of common information problems involves determining the minimum rate of the common input to two independent processors needed to exactly or approximately generate a target joint distribution. For the exact common information problem, exact generation of the target distribution is**

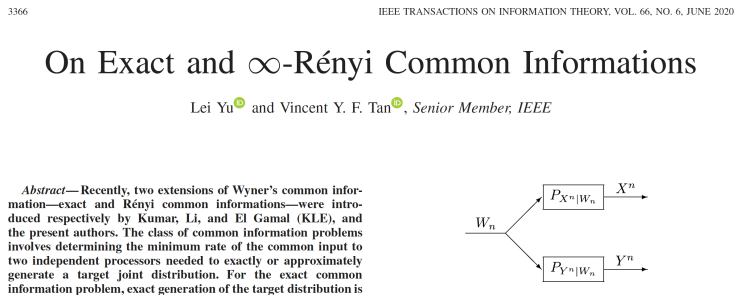


Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$

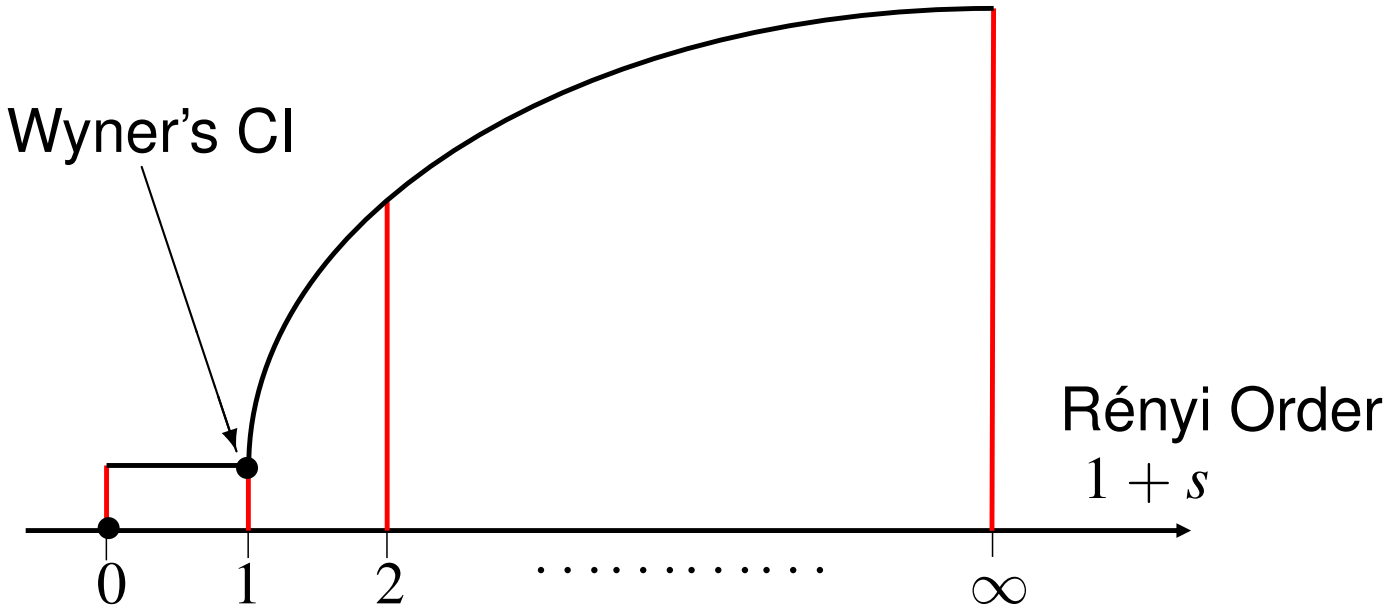


3366 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 6, JUNE 2020

On Exact and ∞-Rényi Common Informations

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— Recently, two extensions of Wyner's common information—exact and Rényi common informations—were introduced respectively by Kumar, Li, and El Gamal (KLE), and the present authors. The class of common information problems involves determining the minimum rate of the common input to two independent processors needed to exactly or approximately generate a target joint distribution. For the exact common information problem, exact generation of the target distribution is**

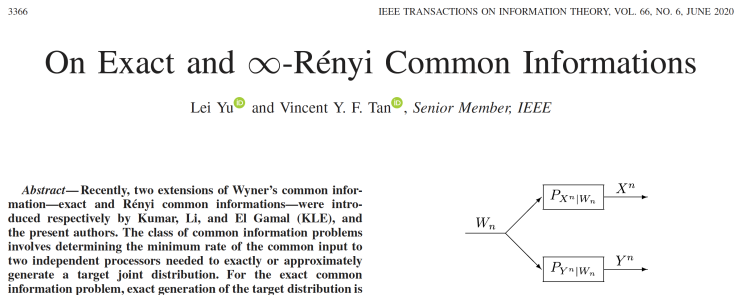


Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$

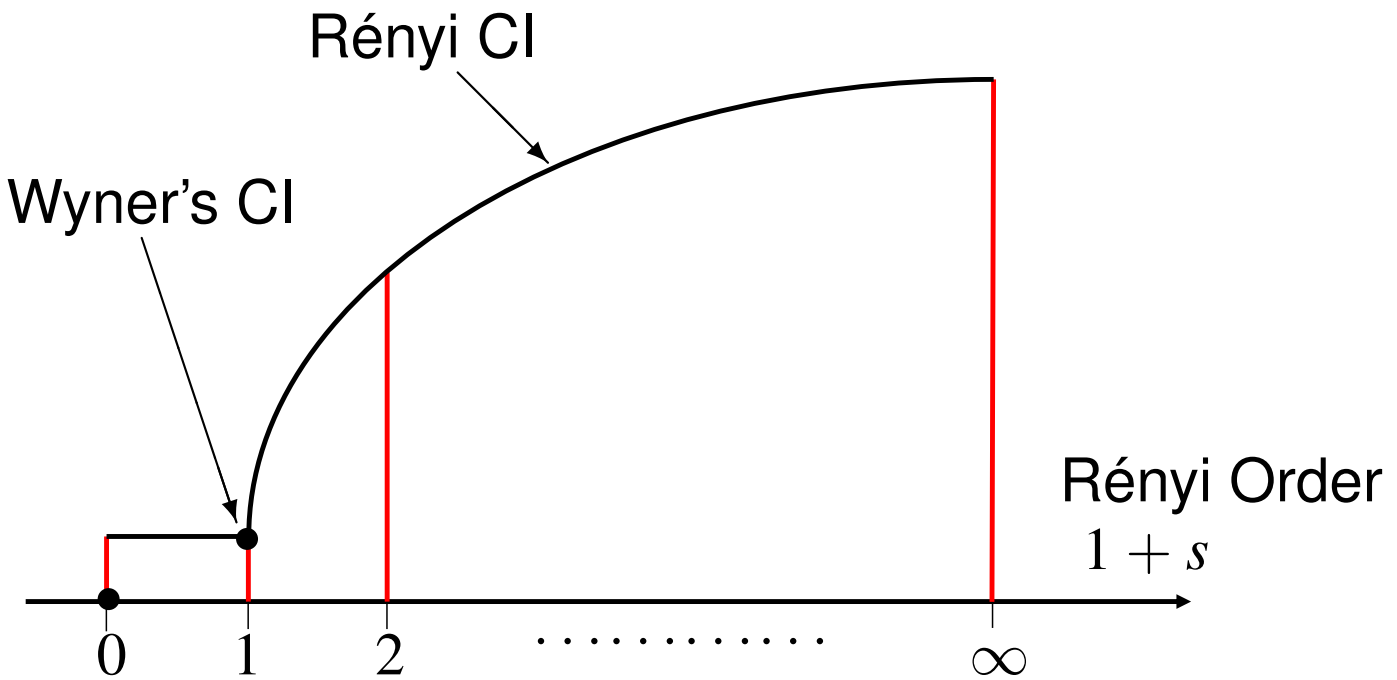


3366 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 6, JUNE 2020

On Exact and ∞-Rényi Common Informations

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— Recently, two extensions of Wyner's common information—exact and Rényi common informations—were introduced respectively by Kumar, Li, and El Gamal (KLE), and the present authors. The class of common information problems involves determining the minimum rate of the common input to two independent processors needed to exactly or approximately generate a target joint distribution. For the exact common information problem, exact generation of the target distribution is**

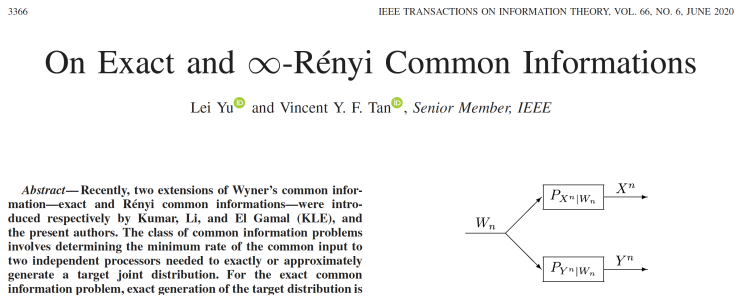


Surprising Equivalence: ∞-Rényi CI and Exact CI

Theorem (Yu and Tan (2020))

For a source π_{XY} on a finite alphabet,

$$T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_{\infty}(\pi_{XY}).$$

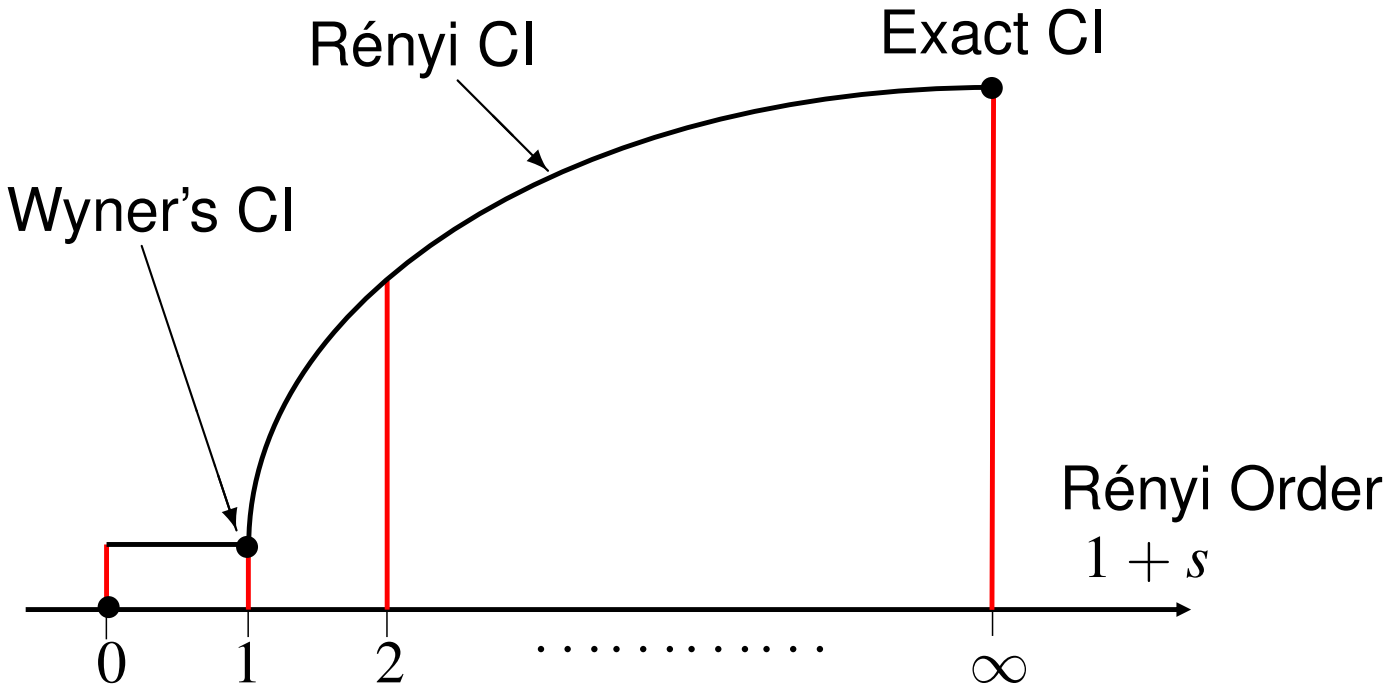


3366 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 66, NO. 6, JUNE 2020

On Exact and ∞-Rényi Common Informations

Lei Yu and Vincent Y. F. Tan, *Senior Member, IEEE*

***Abstract*— Recently, two extensions of Wyner's common information—exact and Rényi common informations—were introduced respectively by Kumar, Li, and El Gamal (KLE), and the present authors. The class of common information problems involves determining the minimum rate of the common input to two independent processors needed to exactly or approximately generate a target joint distribution. For the exact common information problem, exact generation of the target distribution is**



Proof of $\Longrightarrow$ Part of Equivalence Theorem

Lemma (Kumar, Li, and El Gamal '14 & Vellambi and Kliewer '16)

$\exists$ *rate-*R ∞*-Rényi CI code* $\Longrightarrow$ $\exists$ *rate-*R *Exact CI code*

Proof of $\Longrightarrow$ Part of Equivalence Theorem

Lemma (Kumar, Li, and El Gamal '14 & Vellambi and Kliewer '16)

$\exists$ *rate-R ∞-Rényi CI code* $\Longrightarrow$ $\exists$ *rate-R Exact CI code*

- $\exists$ rate-R ∞-Rényi CI code

$$D_\infty(P_{X^nY^n} \,\|\, \pi_{XY}^n) < \epsilon \quad \Longrightarrow \quad P_{X^nY^n}(x^n, y^n) < 2^\epsilon \pi_{XY}^n(x^n, y^n)$$

Proof of $\Longrightarrow$ Part of Equivalence Theorem

Lemma (Kumar, Li, and El Gamal '14 & Vellambi and Kliewer '16)

$\exists$ *rate-R ∞-Rényi CI code* $\Longrightarrow$ $\exists$ *rate-R Exact CI code*

- $\exists$ rate-R ∞-Rényi CI code

$$D_\infty(P_{X^nY^n} \,\|\, \pi^n_{XY}) < \epsilon \quad \Longrightarrow \quad P_{X^nY^n}(x^n, y^n) < 2^\epsilon \pi^n_{XY}(x^n, y^n)$$

- Define

$$\widehat{P}_{X^nY^n}(x^n, y^n) := \frac{2^\epsilon \pi^n_{XY}(x^n, y^n) - P_{X^nY^n}(x^n, y^n)}{2^\epsilon - 1},$$

then obviously, $\widehat{P}_{X^nY^n}(x^n, y^n)$ is a valid distribution.

Proof of $\Longrightarrow$ Part of Equivalence Theorem

Lemma (Kumar, Li, and El Gamal '14 & Vellambi and Kliewer '16)

$\exists$ *rate-R ∞-Rényi CI code* $\Longrightarrow$ $\exists$ *rate-R Exact CI code*

- $\exists$ rate-R ∞-Rényi CI code

$$D_\infty(P_{X^nY^n} \,\|\, \pi^n_{XY}) < \epsilon \quad \Longrightarrow \quad P_{X^nY^n}(x^n, y^n) < 2^\epsilon \pi^n_{XY}(x^n, y^n)$$

- Define

$$\widehat{P}_{X^nY^n}(x^n, y^n) := \frac{2^\epsilon \pi^n_{XY}(x^n, y^n) - P_{X^nY^n}(x^n, y^n)}{2^\epsilon - 1},$$

  then obviously, $\widehat{P}_{X^nY^n}(x^n, y^n)$ is a valid distribution.

- Hence π^n_{XY} can be written as a mixture distribution

$$\pi^n_{XY}(x^n, y^n) = 2^{-\epsilon} P_{X^nY^n}(x^n, y^n) + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}(x^n, y^n)$$

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^n Y^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^n Y^n}$$

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^{n} = 2^{-\epsilon} P_{X^n Y^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^n Y^n}$$

- A time-sharing variable-length scheme:

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

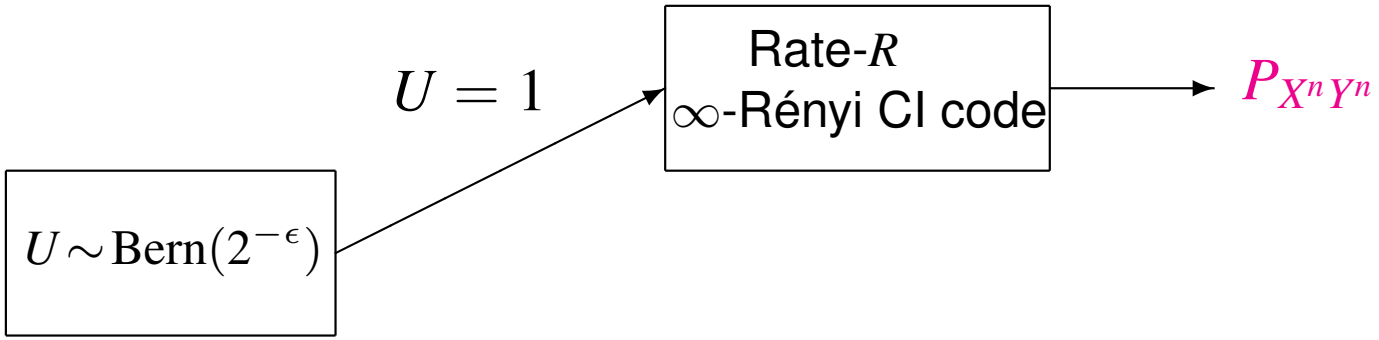
- A time-sharing variable-length scheme:

$$\boxed{U \sim \text{Bern}(2^{-\epsilon})}$$

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^{n} = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

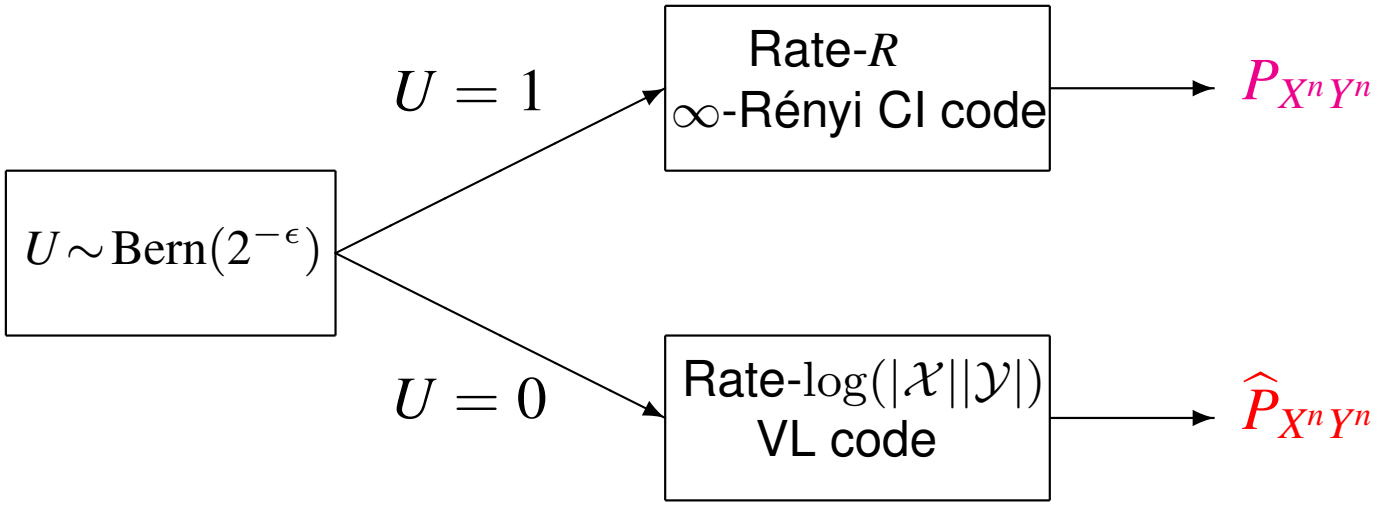
- A time-sharing variable-length scheme:



Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^{n} = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

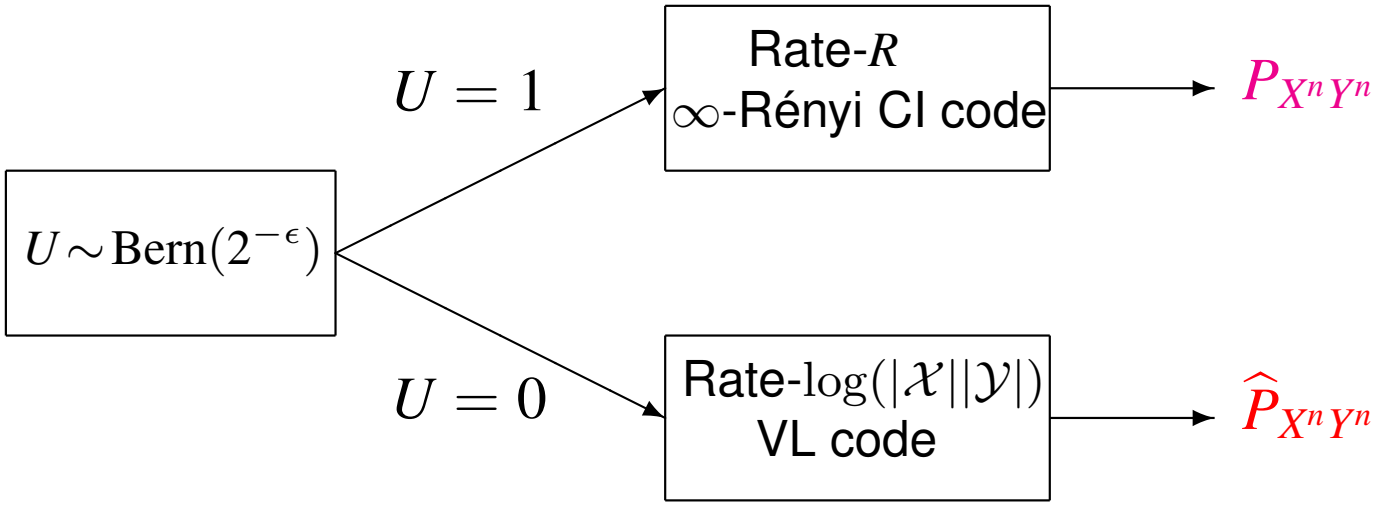
- A time-sharing variable-length scheme:



Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

- A time-sharing variable-length scheme:

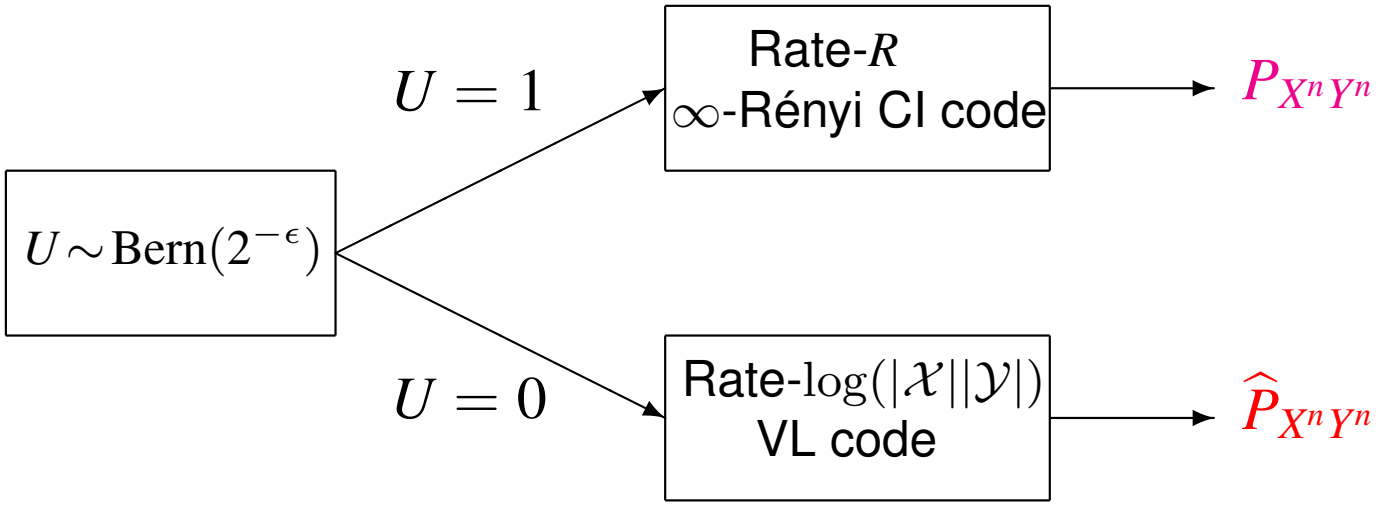


- Induced distribution is exactly π_{XY}^n.

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

- A time-sharing variable-length scheme:



- Induced distribution is exactly π_{XY}^n.
- Expected rate/normalized length of code

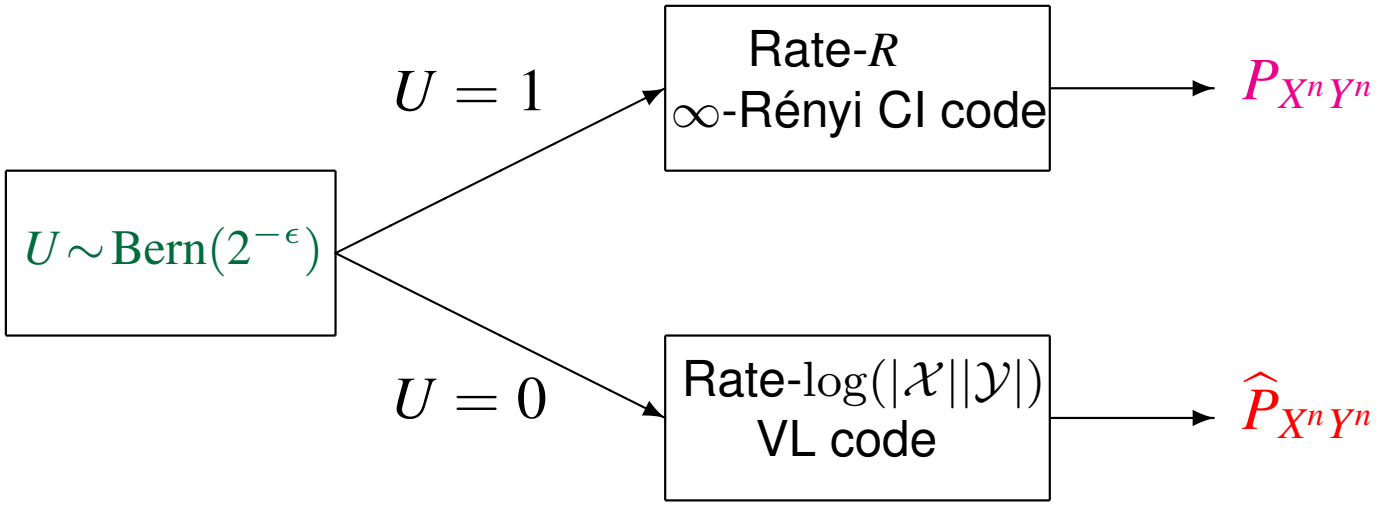
$$\leq \frac{1}{n} + 2^{-\epsilon} R + \left(1 - 2^{-\epsilon}\right) \log(|\mathcal{X}||\mathcal{Y}|) \ \to R$$

as $n \to \infty$ then as $\epsilon \to 0$.

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

- A time-sharing variable-length scheme:



- Induced distribution is exactly π_{XY}^n.
- Expected rate/normalized length of code

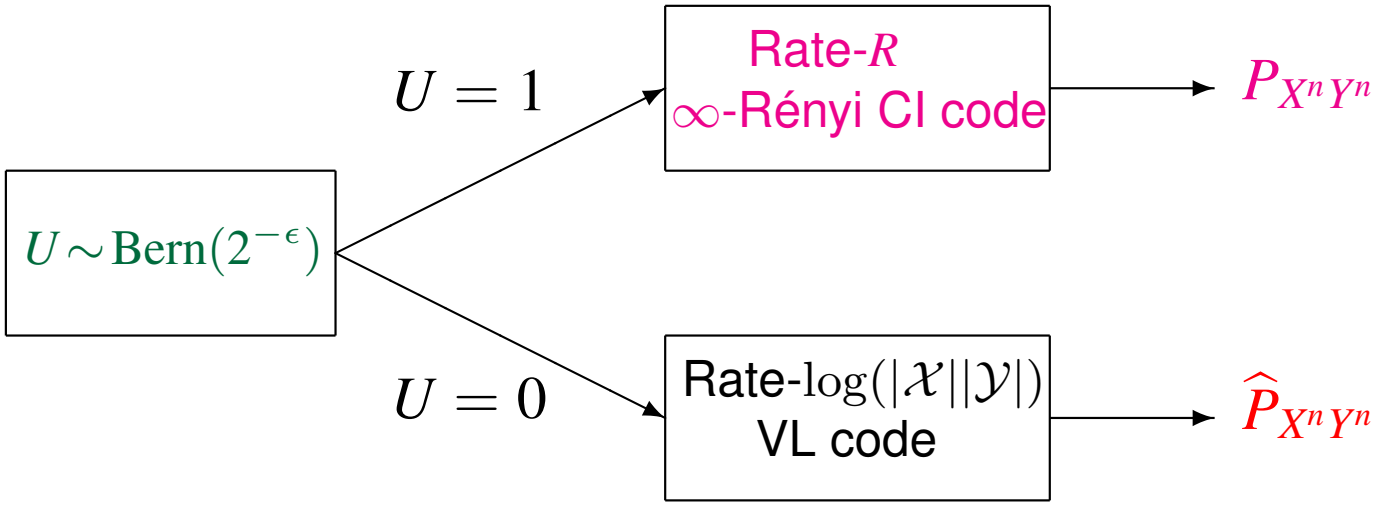
$$\leq \frac{1}{n} + 2^{-\epsilon} R + \left(1 - 2^{-\epsilon}\right) \log(|\mathcal{X}||\mathcal{Y}|) \ \to R$$

as $n \to \infty$ then as $\epsilon \to 0$.

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

- A time-sharing variable-length scheme:



- Induced distribution is exactly π_{XY}^n.
- Expected rate/normalized length of code

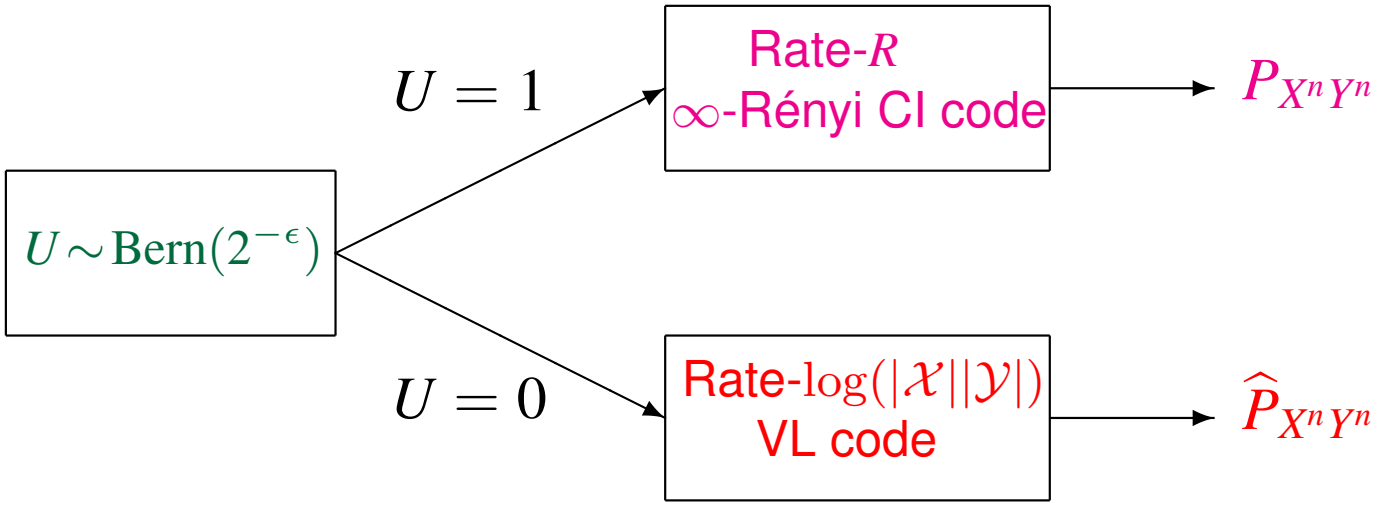
$$\leq \frac{1}{n} + 2^{-\epsilon} R + \left(1 - 2^{-\epsilon}\right) \log(|\mathcal{X}||\mathcal{Y}|) \ \to R$$

as $n \to \infty$ then as $\epsilon \to 0$.

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

- A time-sharing variable-length scheme:



- Induced distribution is exactly π_{XY}^n.
- Expected rate/normalized length of code

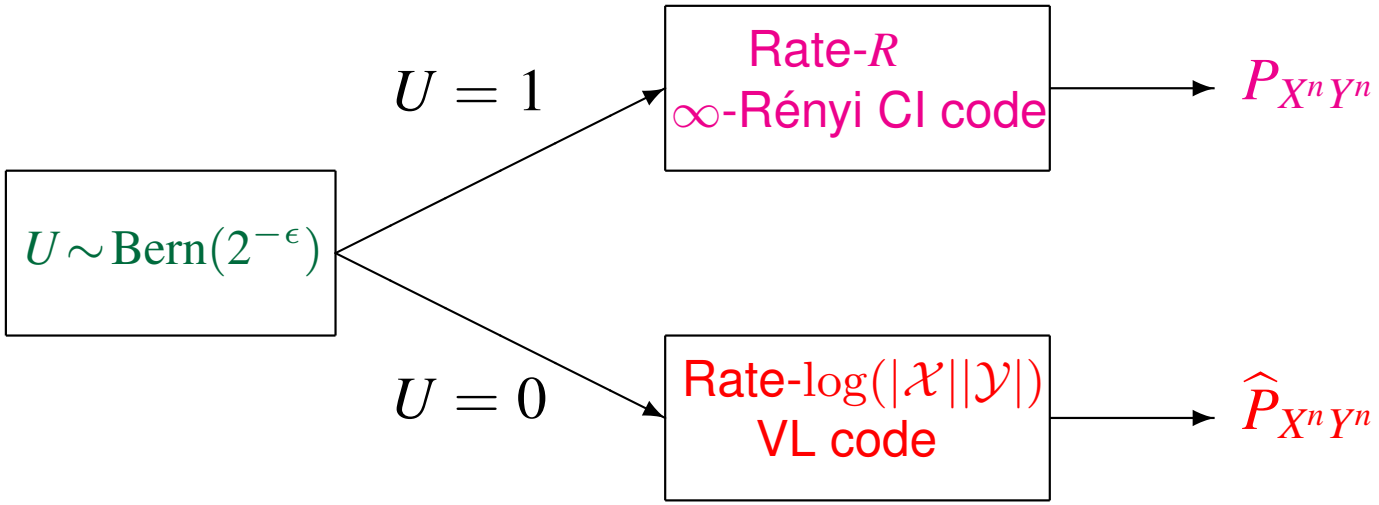
$$\leq \frac{1}{n} + 2^{-\epsilon} R + \left(1 - 2^{-\epsilon}\right) \log(|\mathcal{X}||\mathcal{Y}|) \ \to R$$

as $n \to \infty$ then as $\epsilon \to 0$.

Proof of $\Longrightarrow$ Part of Equivalence Theorem

$$\pi_{XY}^n = 2^{-\epsilon} P_{X^nY^n} + \left(1 - 2^{-\epsilon}\right) \widehat{P}_{X^nY^n}$$

- A time-sharing variable-length scheme:



- Induced distribution is exactly π_{XY}^n.
- Expected rate/normalized length of code

$$\leq \frac{1}{n} + 2^{-\epsilon} R + \left(1 - 2^{-\epsilon}\right) \log(|\mathcal{X}||\mathcal{Y}|) \ \to R$$

as $n \to \infty$ then as $\epsilon \to 0$.

Combining with Single-Letter Bounds from Rényi CI

Theorem (Yu and Tan (2020))

For $(X, Y) \sim \pi_{XY}$ on a finite alphabet,

$$\underline{\Gamma}_\infty(\pi_{XY}) \le T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

Combining with Single-Letter Bounds from Rényi CI

Theorem (Yu and Tan (2020))

For $(X, Y) \sim \pi_{XY}$ on a finite alphabet,

$$\underline{\Gamma}_\infty(\pi_{XY}) \le T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

- Multi-letter expression by Kumar, Li, and El Gamal (2014)

$$\lim_{n\to\infty} \min_{X^n - W_n - Y^n : P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}$$

to single-letter bounds.

Combining with Single-Letter Bounds from Rényi CI

Theorem (Yu and Tan (2020))

For $(X, Y) \sim \pi_{XY}$ on a finite alphabet,

$$\underline{\Gamma}_\infty(\pi_{XY}) \le T_{\mathrm{Ex}}(\pi_{XY}) = \tilde{T}_\infty(\pi_{XY}) \le \overline{\Gamma}_\infty(\pi_{XY}).$$

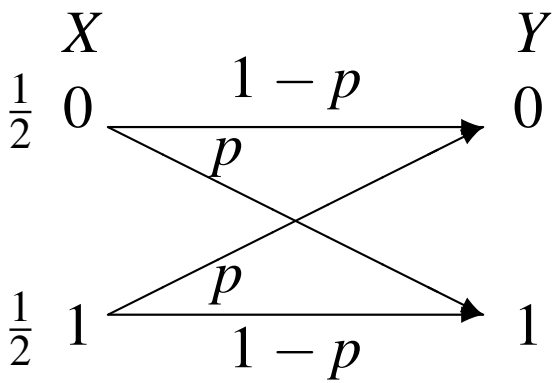
- Multi-letter expression by Kumar, Li, and El Gamal (2014)

$$\lim_{n\to\infty} \min_{X^n - W_n - Y^n : P_{X^nY^n} = \pi_{XY}^n} \frac{H(W_n)}{n}$$

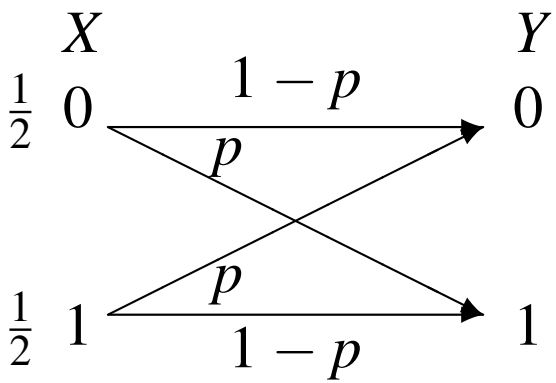
  to single-letter bounds.

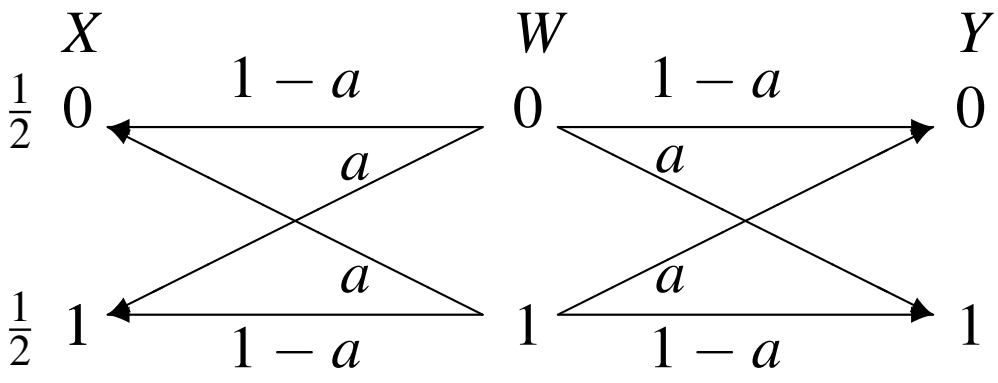
- Bounds are more amenable to numerical evaluation.

Revisiting the DBSS

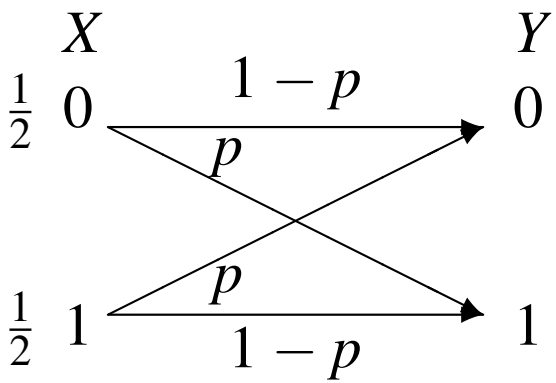


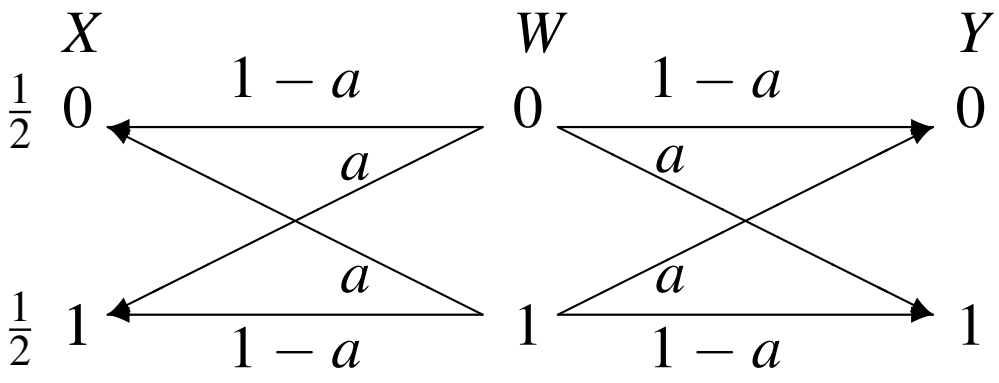
Revisiting the DBSS





Revisiting the DBSS





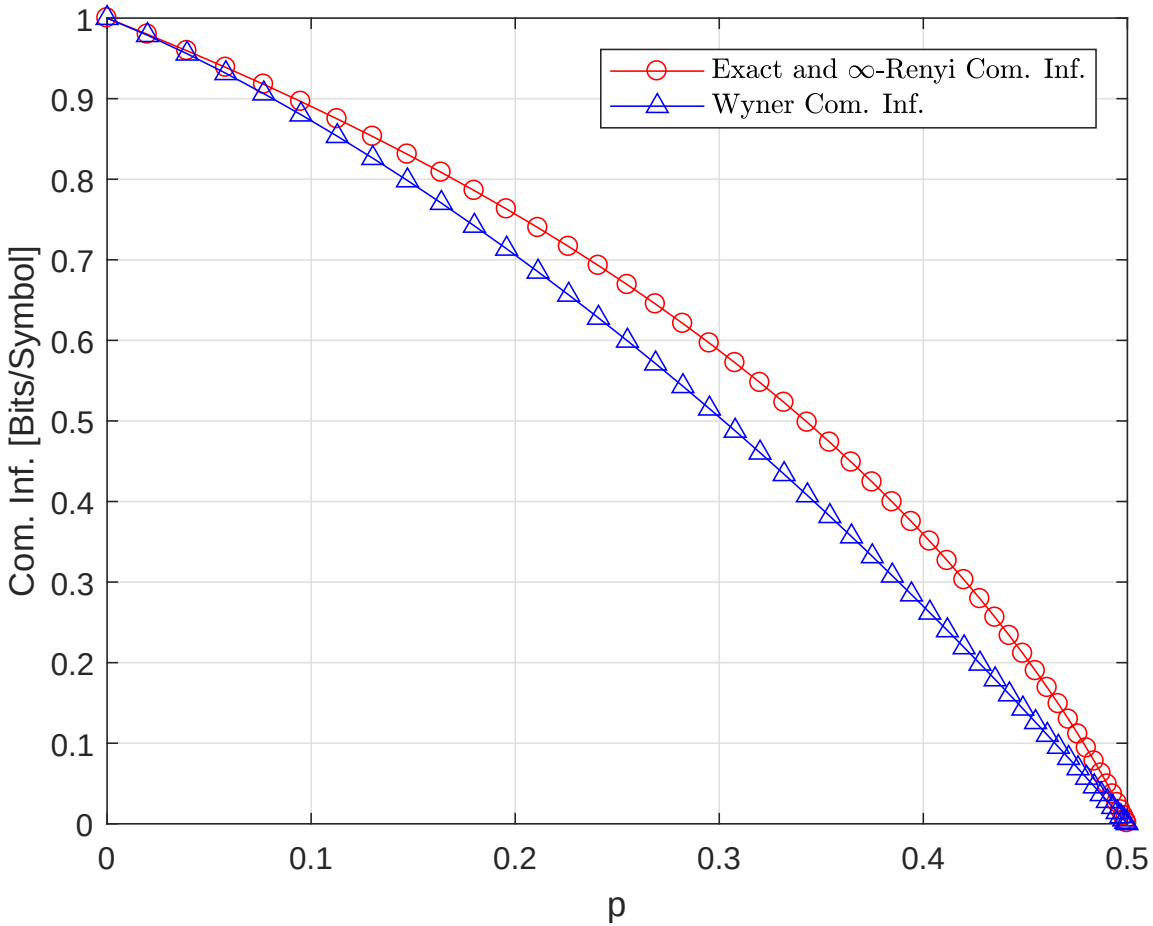
Theorem (Evaluation of Upper and Lower Bounds for DSBS(p))

For a DSBS $(X, Y) \sim \mathrm{DSBS}(p)$ *with crossover probability* $p \in (0, 1/2)$*,*

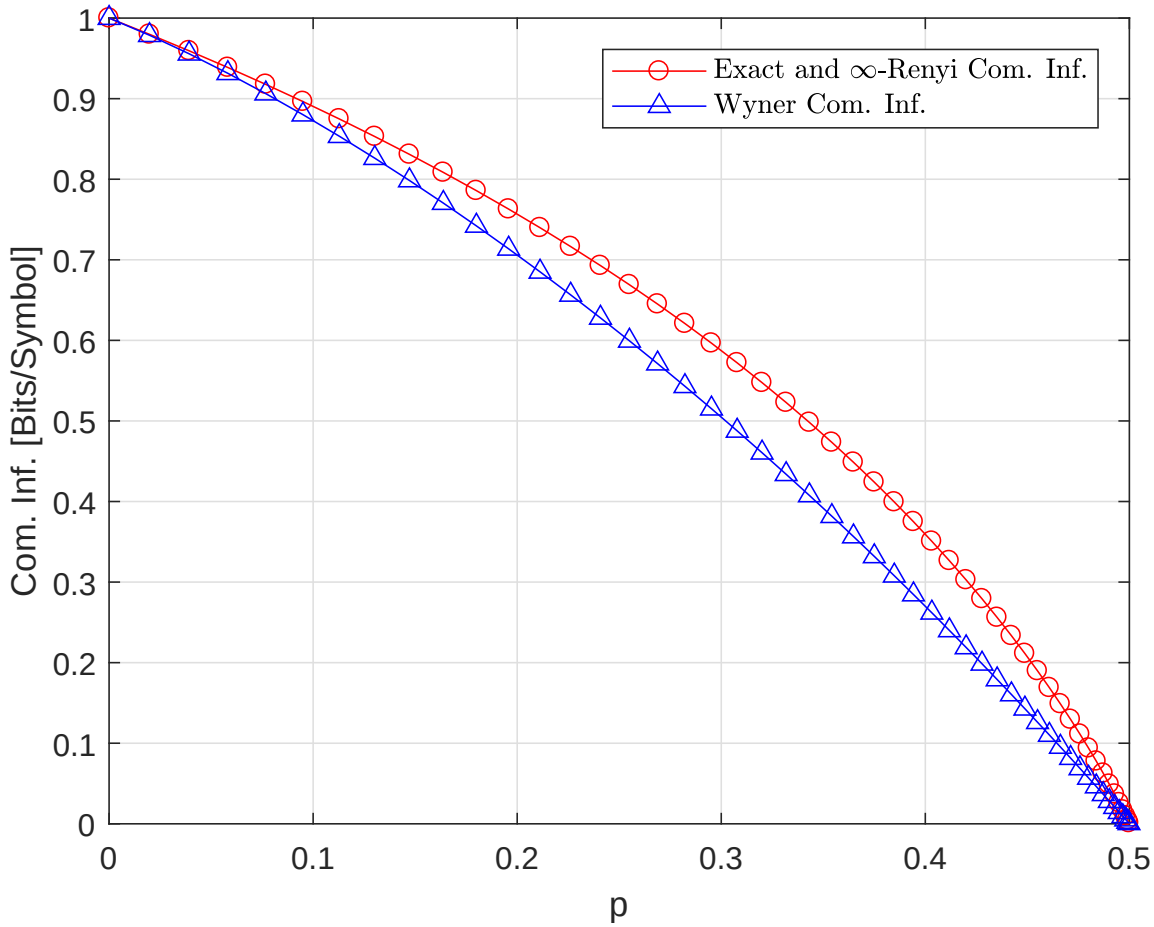
$$\begin{aligned}\tilde{T}_\infty(\pi_{XY}) &= T_{\mathrm{Ex}}(\pi_{XY}) \\ &= -2h(a) - (1-2a)\log\left[\frac{1}{2}\left(a^2 + (1-a)^2\right)\right] - 2a\log\left[a(1-a)\right],\end{aligned}$$

where $a := \frac{1-\sqrt{1-2p}}{2} \in (0, \frac{1}{2})$ *and* $h(a) := -a\log a - (1-a)\log(1-a)$*.*

Numerical Results — DSBS



Numerical Results — DSBS



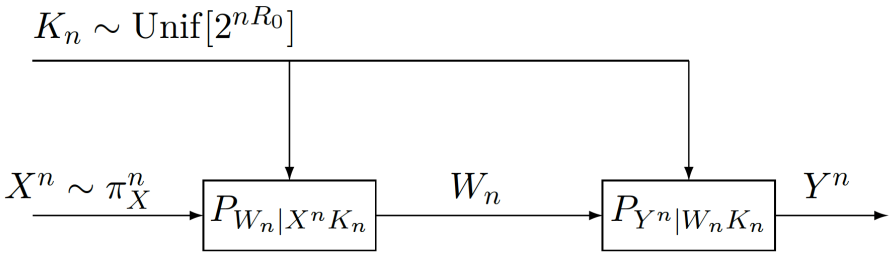
$$T_{\mathrm{Ex}}(\mathrm{DSBS}(p)) > C_{\mathrm{W}}(\mathrm{DSBS}(p)) \quad \forall\, p \in (0, 1/2).$$

Answers the open question in Kumar, Li, and El Gamal (2014).

Other Extensions of Common Information (with Lei Yu)

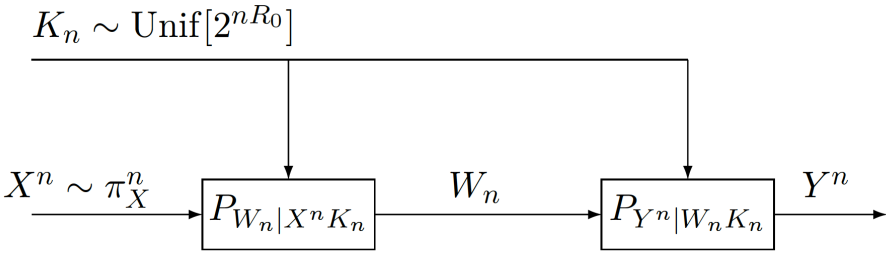
Other Extensions of Common Information (with Lei Yu)

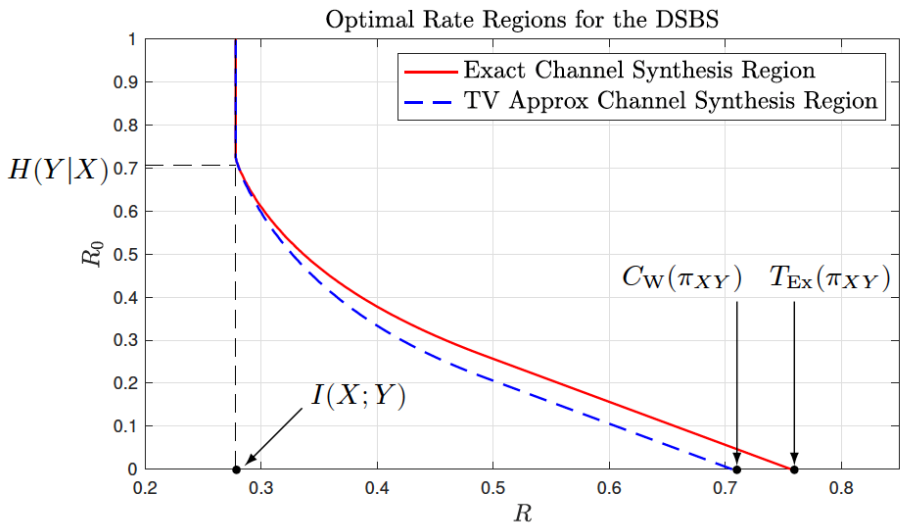
- Exact channel synthesis



Other Extensions of Common Information (with Lei Yu)

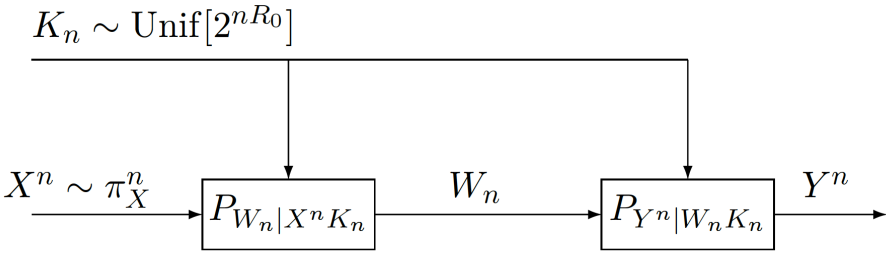
- Exact channel synthesis

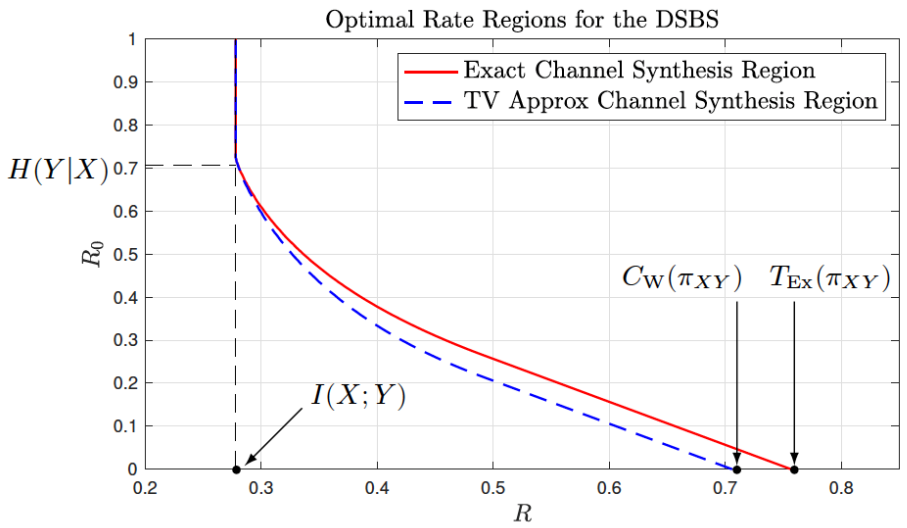




Other Extensions of Common Information (with Lei Yu)

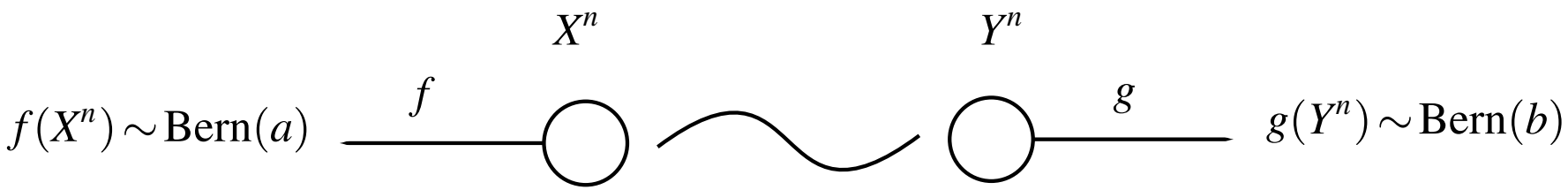
- Exact channel synthesis





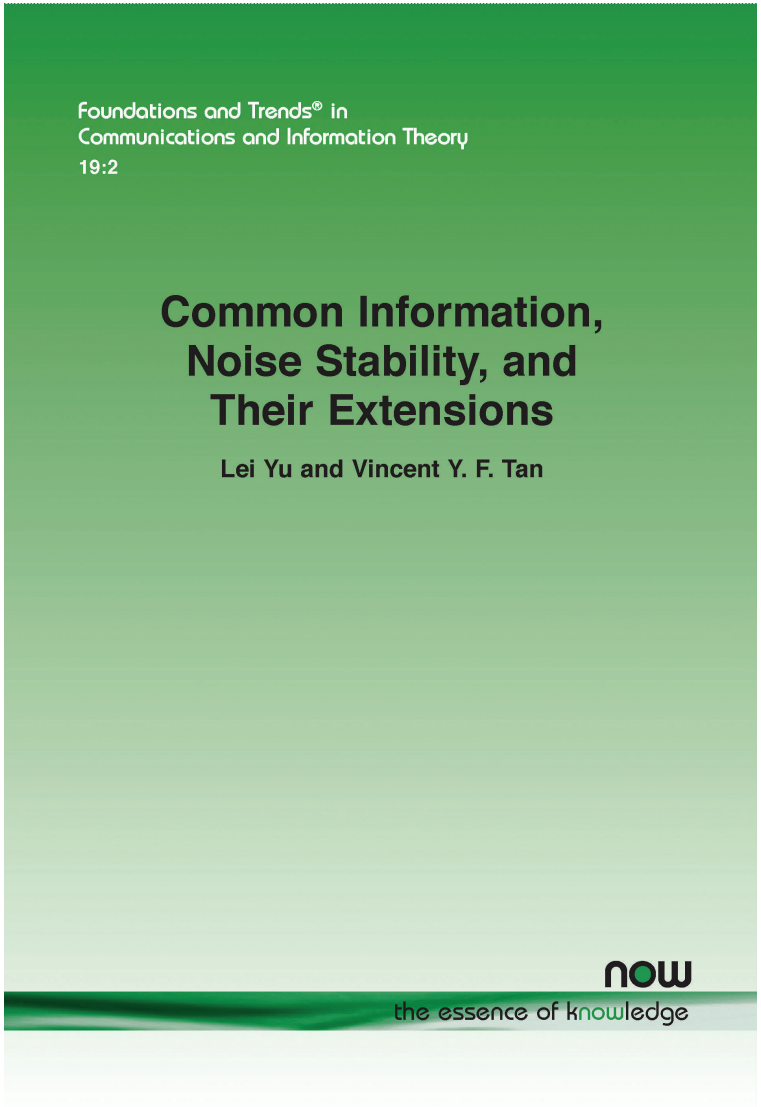
- Non-interactive correlation distillation (cf. Gács–Körner–Witsenhausen CI)

DSBS with correlation coefficient ρ



$$\max / \min \Pr\left(f(X^n) = g(Y^n)\right)$$

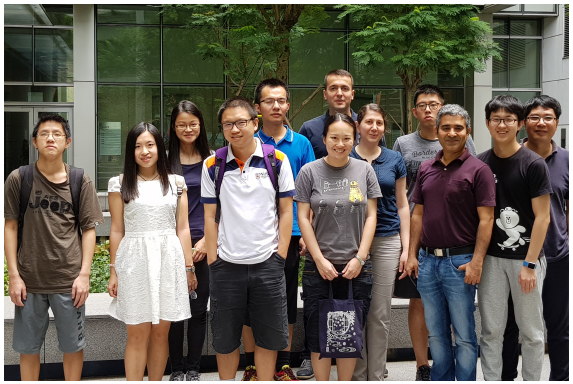
Summarized in Second Monograph (with Lei Yu)



Acknowledgements

Collaborators, students and postdocs!

2017



2018

2021

2022

2023

2024