Joint Source and Channel Coding: Fundamental Bounds and Connections to Machine Learning

Deniz Gündüz

Imperial College London

18 April 2019
European School of Information Theory (ESIT)

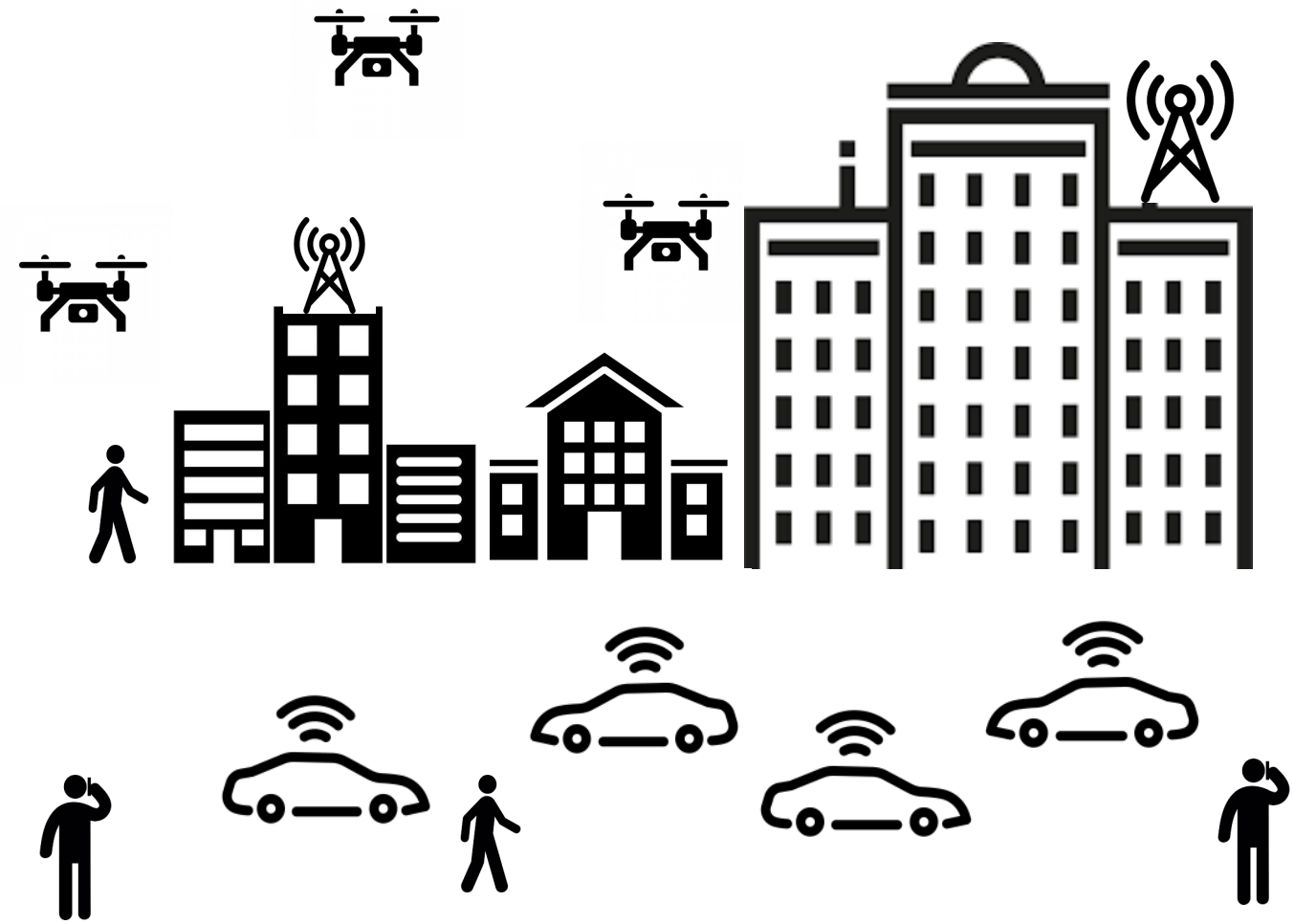
PART I: Information theoretic limits

- Motivation
- Point-to-point oint source-channel coding (JSCC) problem
- Separation Theorem
- JSCC with receiver side information
- JSCC over multi-user networks (broadcast, multi-access and relay channels)

PART II: Practical systems

- Uncoded/ analog transmission
- Compressive sensing for JSCC
- Deep JSCC
- Learning over noisy channels
- Over-the-air stochastic gradient descent

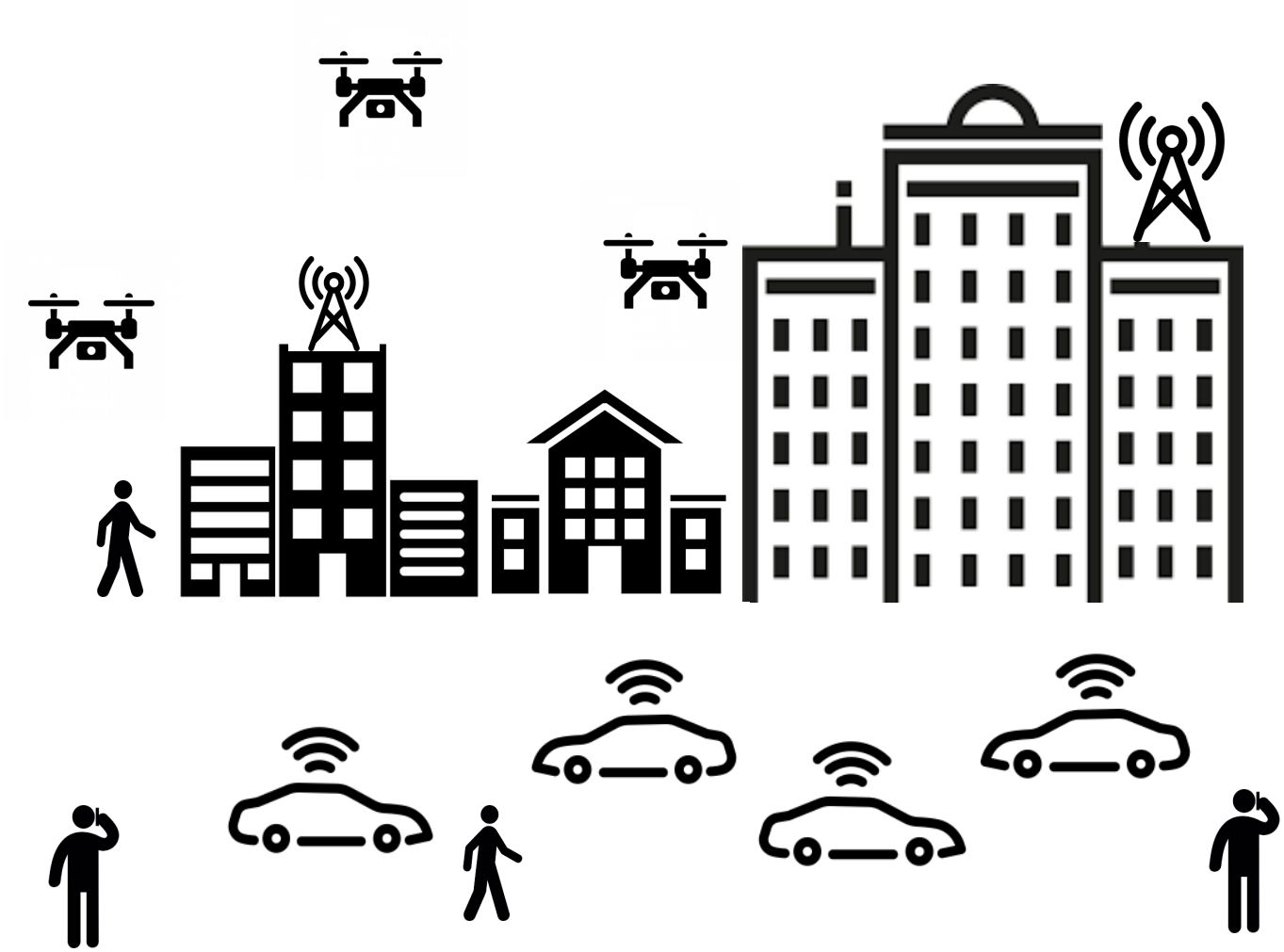
Future Autonomous Systems



Intelligence is the key for future autonomous systems!

and, so is communications ...

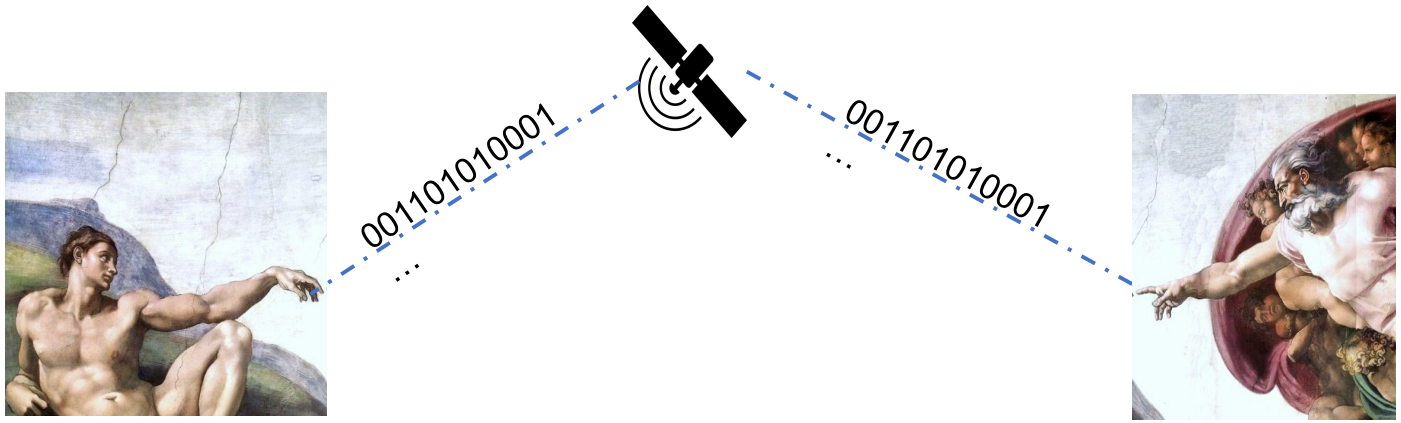
new objectives, new constraints, new problems!



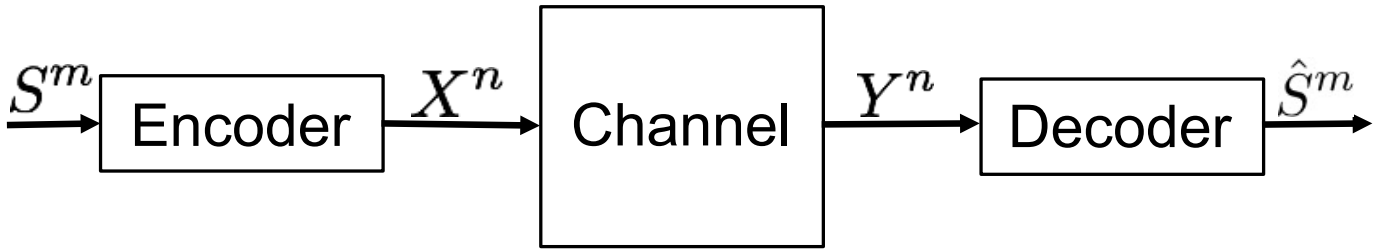
Intelligence is the key for future autonomous systems!
and, so is communications ...
new objectives, new constraints, new problems!

Unkept 5G Promises: Tactile Internet

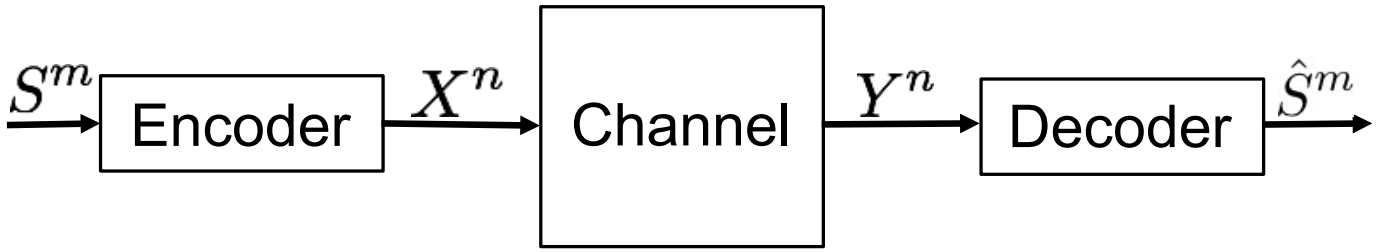
- "Internet network that combines ultra low latency with extremely high availability, reliability and security" (ITU)
- Next generation Internet of Things (IoT): human-machine and machine-machine interaction: haptic interaction with visual feedback
- Augmented reality (AR), virtual reality (VR), automation, robotics, remote education, telepresence, ...
- 1ms round trip delay?



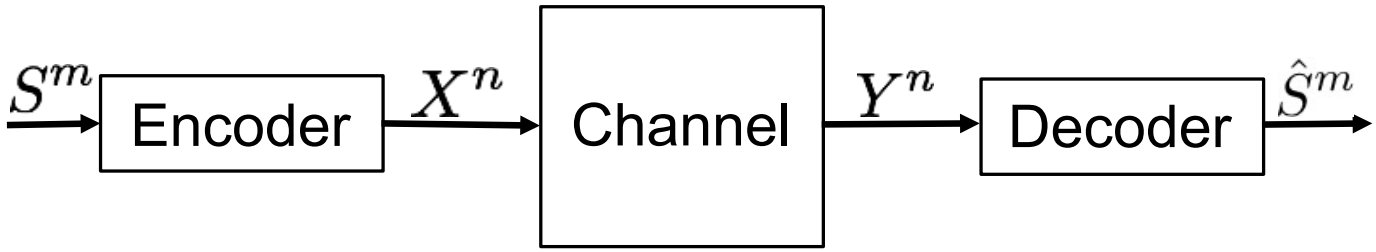
Fundamental Problem of Information Theory



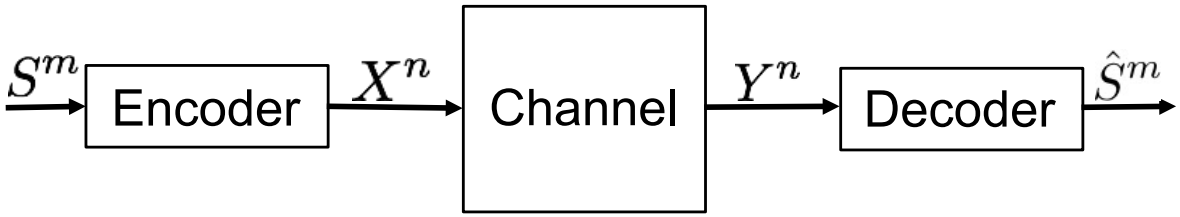
- Transmit source to the destination "reliably"
- Source: i.i.d. samples from p_S
- Channel: memoryless with $p_{Y|X}$
- Encoder: $f^{m,n} : S^m \to X^n$
- Decoder: $g^{m,n} : Y^n \to \hat{S}^m$
- Rate: $\frac{n}{m}$
- Probability of error $P_e^{m,n} = \Pr\{S^m \neq \hat{S}^m\}$
- Rate r is achievable if there exists a sequence of encoders and decoders such that $P_e^{m,n} \to 0$ as $n, m \to \infty$ while $\frac{n}{m} \leq r$.
- Minimum achievable rate is called the *source-channel capacity*



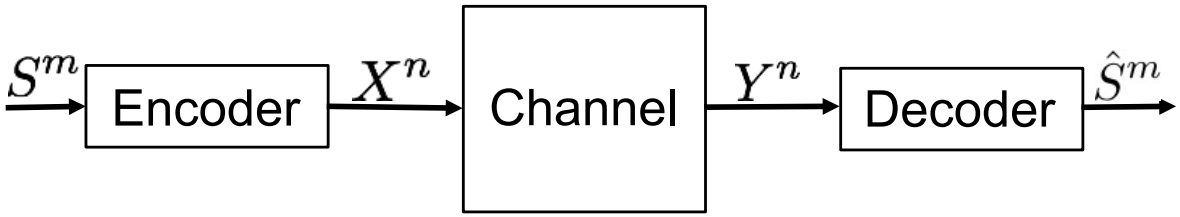
- Transmit source to the destination "reliably"
- Source: i.i.d. samples from p_S
- Channel: memoryless with $p_{Y|X}$
- Encoder: $f^{m,n} : S^m \to X^n$
- Decoder: $g^{m,n} : Y^n \to \hat{S}^m$
- Rate: $\frac{n}{m}$
- Probability of error $P_e^{m,n} = \Pr\{S^m \neq \hat{S}^m\}$
- Rate r is achievable if there exists a sequence of encoders and decoders such that $P_e^{m,n} \to 0$ as $n, m \to \infty$ while $\frac{n}{m} \leq r$.
- Minimum achievable rate is called the *source-channel capacity*



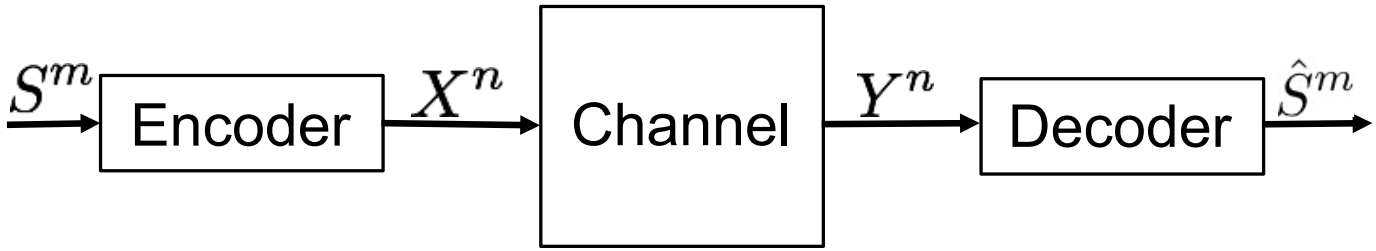
- Transmit source to the destination "reliably"
- Source: i.i.d. samples from p_S
- Channel: memoryless with $p_{Y|X}$
- Encoder: $f^{m,n} : S^m \to X^n$
- Decoder: $g^{m,n} : Y^n \to \hat{S}^m$
- Rate: $\frac{n}{m}$
- Probability of error $P_e^{m,n} = \Pr\{S^m \neq \hat{S}^m\}$
- Rate r is achievable if there exists a sequence of encoders and decoders such that $P_e^{m,n} \to 0$ as $n, m \to \infty$ while $\frac{n}{m} \leq r$.
- Minimum achievable rate is called the *source-channel capacity*



- **Channel Coding**
  - Assume: source is binary 1/2
  - i.e., entropy is $H(S) = 1$ bit per sample
  - Inverse of the minimum source-channel rate is the maximum number of bits per channel use one can transmit reliably over this channel
- **Source Coding**
  - Assume: Channel is error free with capacity 1 bit per channel use
  - Minimum source-channel rate gives us the minimum number of bits per sample we need to compress this source reliably



- **Channel Coding**
  - Assume: source is binary 1/2
  - i.e., entropy is $H(S) = 1$ bit per sample
  - Inverse of the minimum source-channel rate is the maximum number of bits per channel use one can transmit reliably over this channel
- **Source Coding**
  - Assume: Channel is error free with capacity 1 bit per channel use
  - Minimum source-channel rate gives us the minimum number of bits per sample we need to compress this source reliably



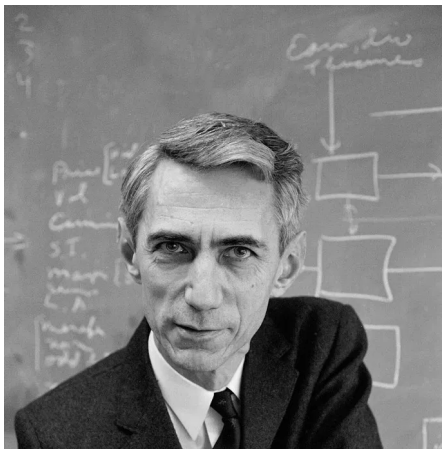
- Additive distortion measure: $d(s, \hat{s})$:

$$d(S^m, \hat{S}^m) = \frac{1}{m} \sum_{i=1}^{m} d(S_i, \hat{S}_i)$$

- A rate- distortion pair (r, D) is achievable if there exists a sequence of encoders and decoders with $\frac{n}{m} \leq r$ and $\lim_{m,n\to\infty} E[d(S^m, \hat{S}^m)] \leq D$.

Shannon's Source-Channel Separation Theorem

- First compress the source
- Match quantized bits to the optimal channel code
- No loss of optimality



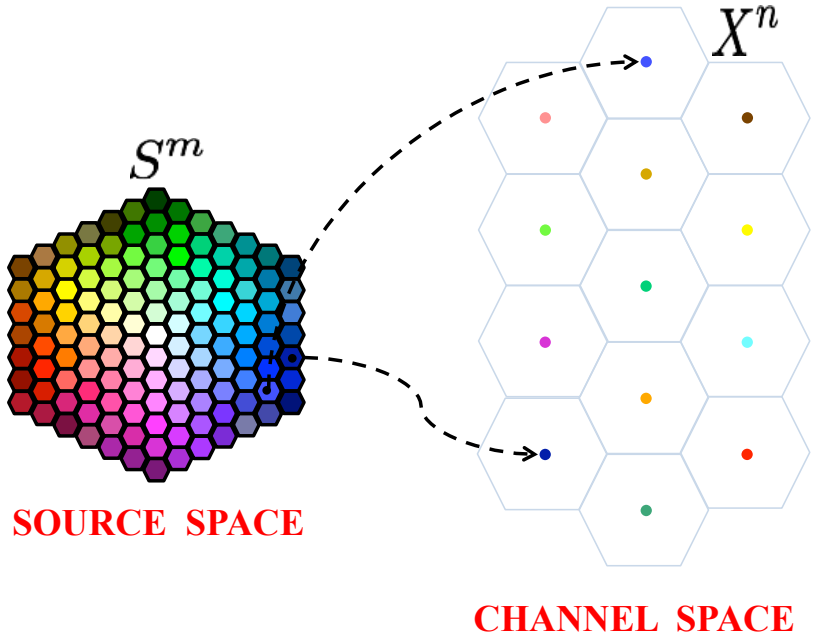
Separation Theorem

(Lossless) Rate r is achievable iff $H(S) \leq rC$
(Lossy) For given rate r and distortion measure $d(\cdot,\cdot)$, the minimum achievable distortion is given by $D(rC)$

where $D(R)$ is the distortion-rate function of the source, and C is the capacity of the channel.

Separate Coding Scheme

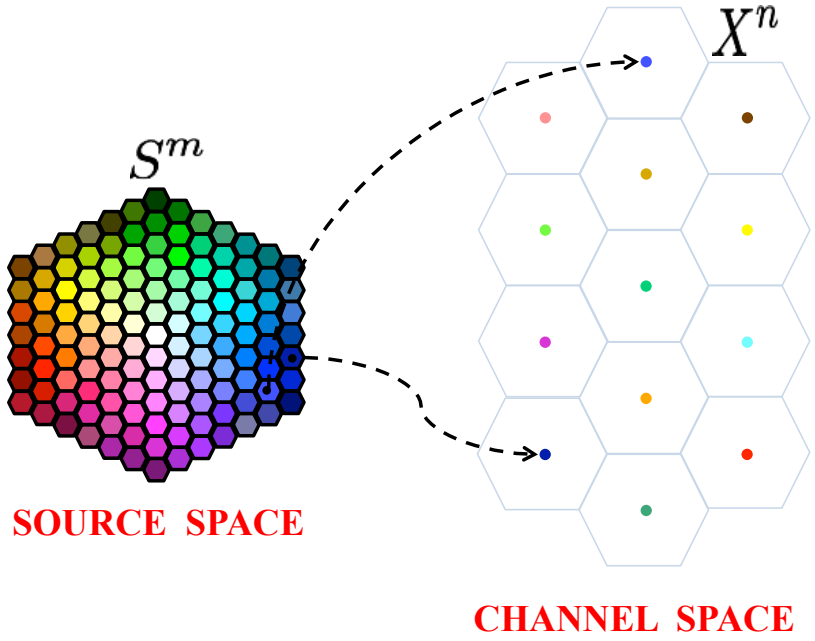
- Optimal source-channel rate is $r = \frac{H(S)}{C}$, where $C = \max p_X I(X;Y)$
- Random coding: generate $2^{mH(S)}$ source codewords of length m with probability p_S
- Also, generate $2^{mH(S)} = 2^{nC}$ length-n channel codewords with capacity achieving input distribution p_X



- First the channel codeword, then the source codeword is decoded with arbitrarily small probability of error
- In practice concatenate near optimal source and channel codes, such as LDGM followed by LDPC etc..

Separate Coding Scheme

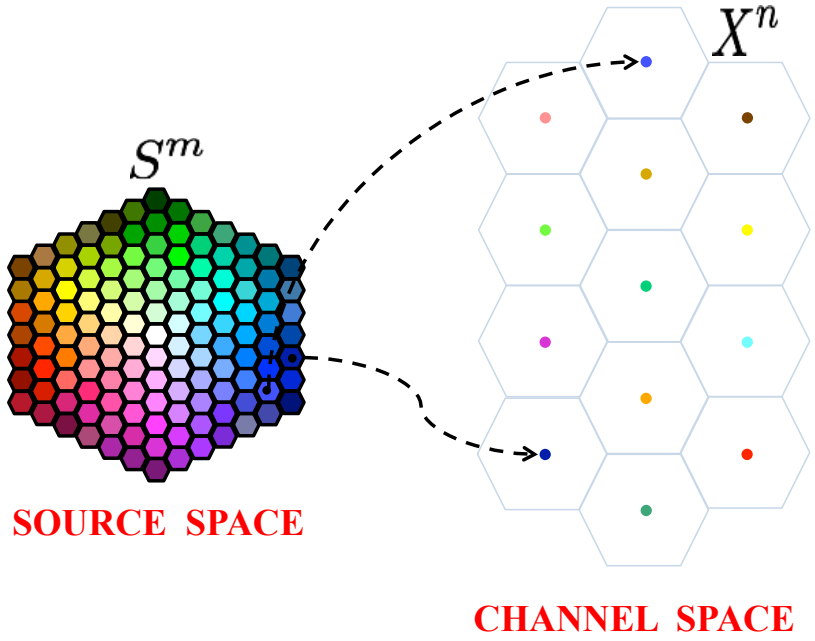
- Optimal source-channel rate is $r = \frac{H(S)}{C}$, where $C = \max p_X I(X;Y)$
- Random coding: generate $2^{mH(S)}$ source codewords of length m with probability p_S
- Also, generate $2^{mH(S)} = 2^{nC}$ length-n channel codewords with capacity achieving input distribution p_X



- First the channel codeword, then the source codeword is decoded with arbitrarily small probability of error
- In practice concatenate near optimal source and channel codes, such as LDGM followed by LDPC etc..

Separate Coding Scheme

- Optimal source-channel rate is $r = \frac{H(S)}{C}$, where $C = \max p_X I(X;Y)$
- Random coding: generate $2^{mH(S)}$ source codewords of length m with probability p_S
- Also, generate $2^{mH(S)} = 2^{nC}$ length-n channel codewords with capacity achieving input distribution p_X



- First the channel codeword, then the source codeword is decoded with arbitrarily small probability of error
- In practice concatenate near optimal source and channel codes, such as LDGM followed by LDPC etc..

- If $P_e \to 0$, then $H(S) < rC$, for any sequence of encoder-decoder pairs with $n \leq r \cdot m$.
- From Fano's inequality:

$$H(S^m|\hat{S}^m) \leq 1 + P_e^{m,n} \log |\mathcal{S}^m| = 1 + P_e^{m,n} m \log |\mathcal{S}|$$

Hence,

$$\begin{aligned} H(S) &= \frac{1}{m} H(S^m|\hat{S}^m) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Chain rule)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Fano's inequality)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(X^n; Y^n) \\ &\quad \text{(Data processing inequality, } S^m - X^n - Y^n - \hat{S}^m) \\ &\leq \frac{1}{m} + P_e^{m,n} \log |\mathcal{S}| + rC \quad \text{(Capacity theorem)} \end{aligned}$$

Letting $m, n \to \infty$, if $P_e^{m,n} \to 0$, we get $H(S) \leq rC$.

Optimality of separation continues to hold in the presence of feedback!

- If $P_e \to 0$, then $H(S) < rC$, for any sequence of encoder-decoder pairs with $n \leq r \cdot m$.
- From Fano's inequality:

$$H(S^m|\hat{S}^m) \leq 1 + P_e^{m,n} \log |\mathcal{S}^m| = 1 + P_e^{m,n} m \log |\mathcal{S}|$$

Hence,

$$\begin{aligned} H(S) &= \frac{1}{m} H(S^m|\hat{S}^m) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Chain rule)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Fano's inequality)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(X^n; Y^n) \\ &\quad \text{(Data processing inequality, } S^m - X^n - Y^n - \hat{S}^m) \\ &\leq \frac{1}{m} + P_e^{m,n} \log |\mathcal{S}| + rC \quad \text{(Capacity theorem)} \end{aligned}$$

Letting $m, n \to \infty$, if $P_e^{m,n} \to 0$, we get $H(S) \leq rC$.

Optimality of separation continues to hold in the presence of feedback!

- If $P_e \to 0$, then $H(S) < rC$, for any sequence of encoder-decoder pairs with $n \leq r \cdot m$.
- From Fano's inequality:

$$H(S^m|\hat{S}^m) \leq 1 + P_e^{m,n} \log |\mathcal{S}^m| = 1 + P_e^{m,n} m \log |\mathcal{S}|$$

Hence,

$$\begin{aligned} H(S) &= \frac{1}{m} H(S^m|\hat{S}^m) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Chain rule)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Fano's inequality)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(X^n; Y^n) \\ &\quad \text{(Data processing inequality, } S^m - X^n - Y^n - \hat{S}^m\text{)} \\ &\leq \frac{1}{m} + P_e^{m,n} \log |\mathcal{S}| + rC \quad \text{(Capacity theorem)} \end{aligned}$$

Letting $m, n \to \infty$, if $P_e^{m,n} \to 0$, we get $H(S) \leq rC$.

Optimality of separation continues to hold in the presence of feedback!

- If $P_e \to 0$, then $H(S) < rC$, for any sequence of encoder-decoder pairs with $n \leq r \cdot m$.
- From Fano's inequality:

$$H(S^m|\hat{S}^m) \leq 1 + P_e^{m,n} \log |\mathcal{S}^m| = 1 + P_e^{m,n} m \log |\mathcal{S}|$$

Hence,

$$\begin{aligned} H(S) &= \frac{1}{m} H(S^m|\hat{S}^m) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Chain rule)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Fano's inequality)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(X^n; Y^n) \\ &\quad \text{(Data processing inequality, } S^m - X^n - Y^n - \hat{S}^m\text{)} \\ &\leq \frac{1}{m} + P_e^{m,n} \log |\mathcal{S}| + rC \quad \text{(Capacity theorem)} \end{aligned}$$

Letting $m, n \to \infty$, if $P_e^{m,n} \to 0$, we get $H(S) \leq rC$.

Optimality of separation continues to hold in the presence of feedback!

- If $P_e \to 0$, then $H(S) < rC$, for any sequence of encoder-decoder pairs with $n \leq r \cdot m$.
- From Fano's inequality:

$$H(S^m|\hat{S}^m) \leq 1 + P_e^{m,n} \log |\mathcal{S}^m| = 1 + P_e^{m,n} m \log |\mathcal{S}|$$

Hence,

$$\begin{aligned} H(S) &= \frac{1}{m} H(S^m|\hat{S}^m) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Chain rule)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(S^m; \hat{S}^m) \quad \text{(Fano's inequality)} \\ &\leq \frac{1}{m} (1 + P_e^{m,n} m \log |\mathcal{S}|) + \frac{1}{m} I(X^n; Y^n) \\ &\quad \text{(Data processing inequality, } S^m - X^n - Y^n - \hat{S}^m\text{)} \\ &\leq \frac{1}{m} + P_e^{m,n} \log |\mathcal{S}| + rC \quad \text{(Capacity theorem)} \end{aligned}$$

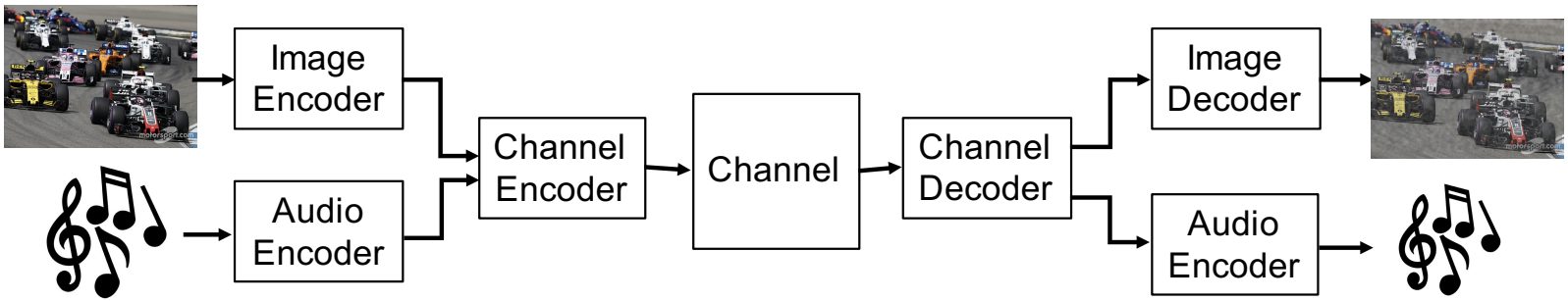
Letting $m, n \to \infty$, if $P_e^{m,n} \to 0$, we get $H(S) \leq rC$.

Optimality of separation continues to hold in the presence of feedback!

Benefits and Limitations of Separation

Separation is good, because ...

- brings modularity,
- we can benefit from existing source and channel coding techniques



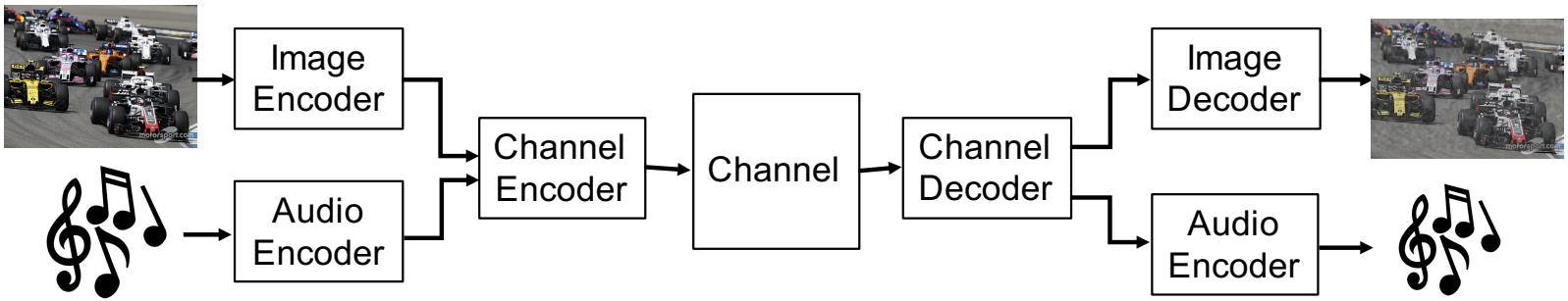
but ...

- infinite delay and complexity,
- ergodic source and channel assumption
- and no separation theorem for multi-user networks

Benefits and Limitations of Separation

Separation is good, because ...

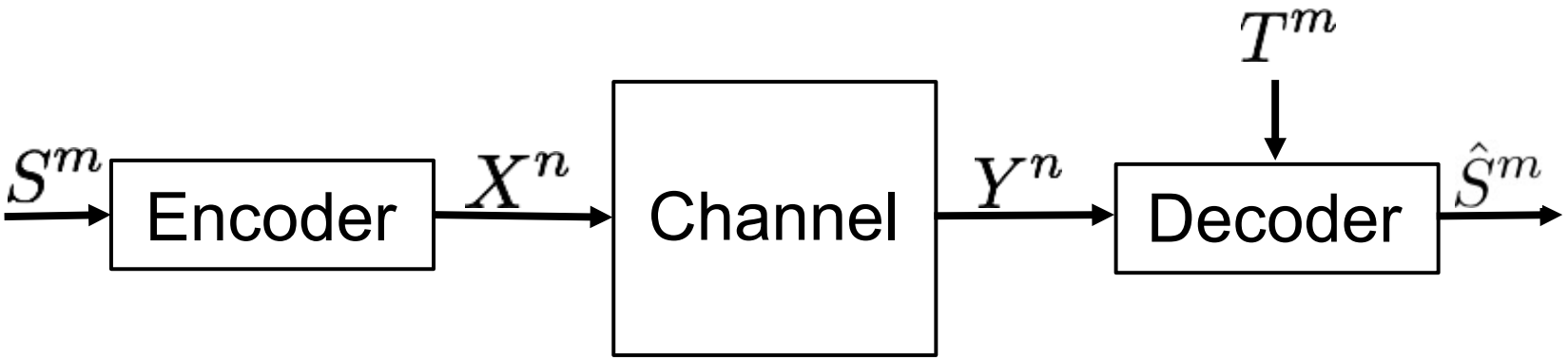
- brings modularity,
- we can benefit from existing source and channel coding techniques



but ...

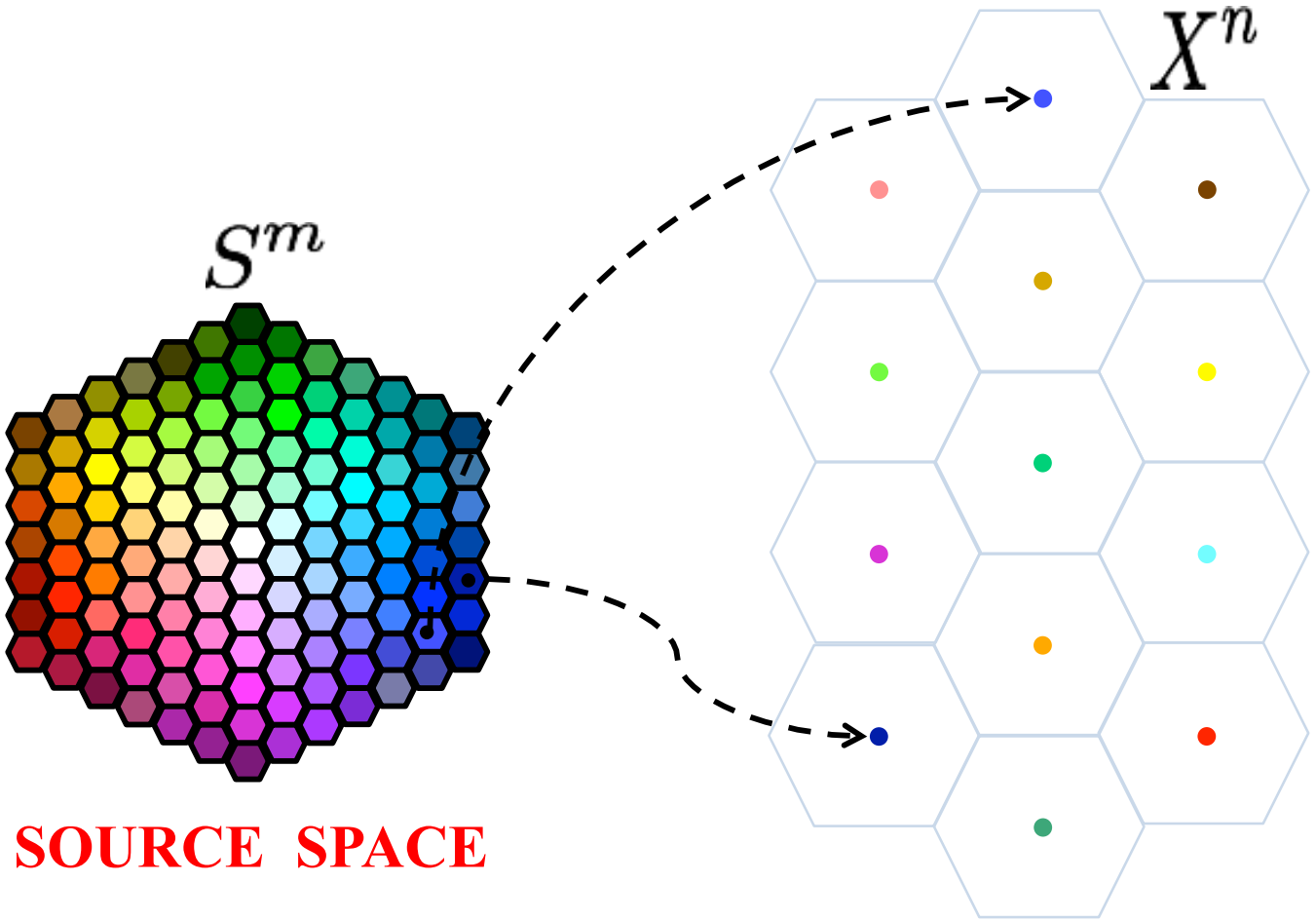
- infinite delay and complexity,
- ergodic source and channel assumption
- and no separation theorem for multi-user networks

Receiver Side Information

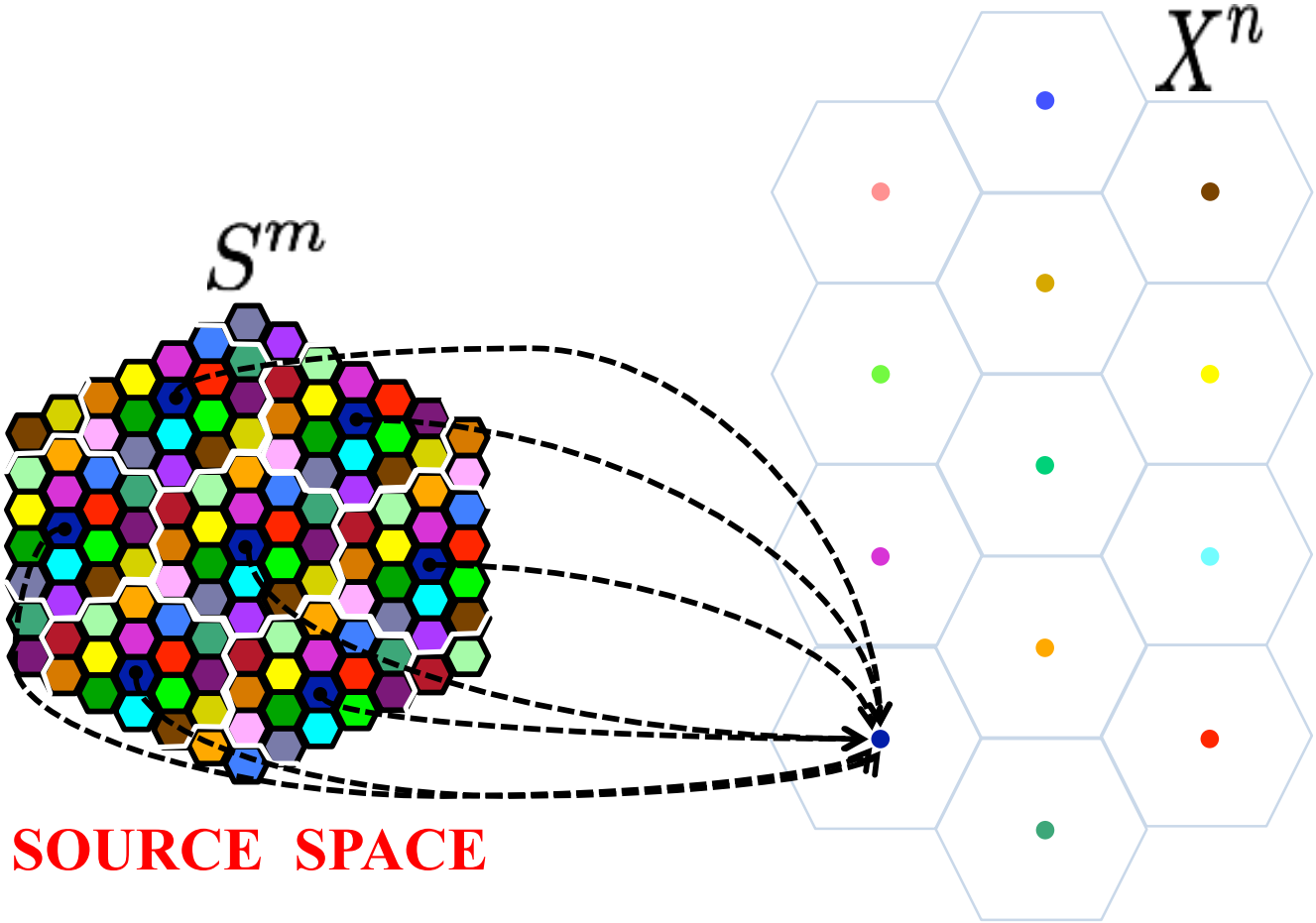


- Receiver has correlated side information: sensor network
- Separation optimal (Shamai, Verdu, '95): Optimal source-channel rate $r = \frac{H(S|T)}{C}$
- Lossy transmission: minimum distortion $\mathrm{D}^{\mathrm{WZ}}(rC)$, where D^{WZ} is the Wyner-Ziv rate-distortion function

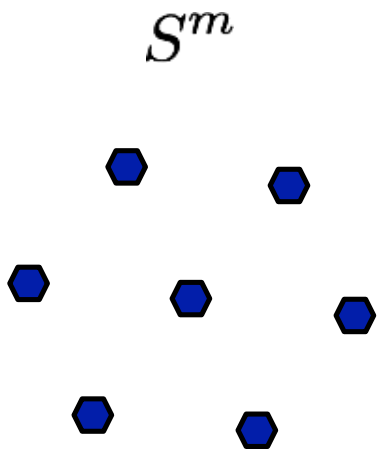
When there is no side information, no need for binning.



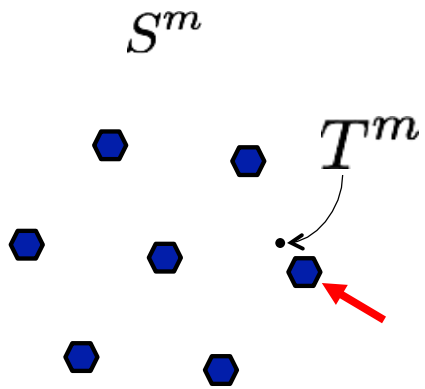
When there is side information at the receiver, we map multiple source codewords to the same channel codeword:



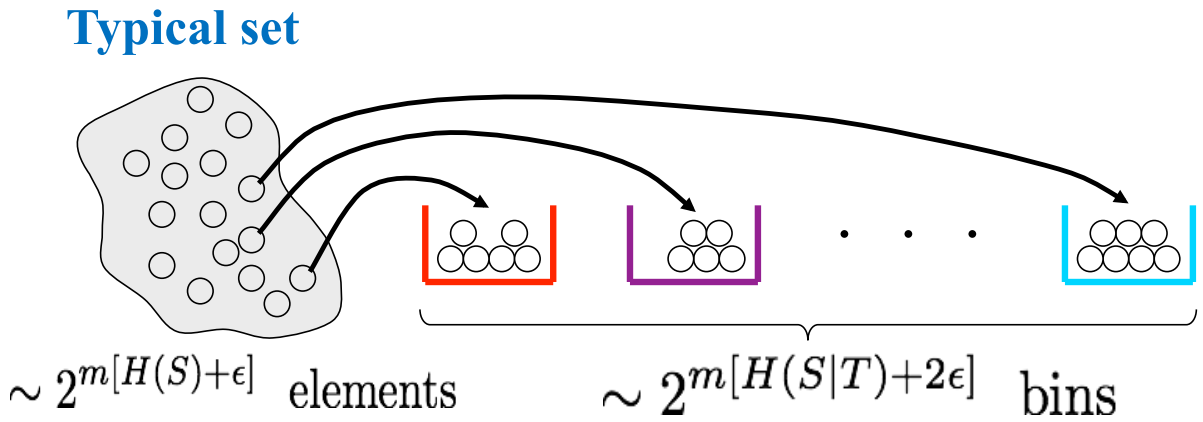
First decode channel codeword. There are multiple candidates for source codeword from the same bin:



Correlated side information T^m: Choose source codeword in the bin jointly typical with T^m:



Random Binning (Slepian-Wolf Coding)



- Randomly assign source vectors to bins such that there are $\sim 2^{m[I(S;T)-\epsilon]}$ elements in each bin.
- Sufficiently few elements in each bin to decode S^m using typicality.
- Even if the sender knew T^m, source coding rate could not be lower than $H(S|T)$.

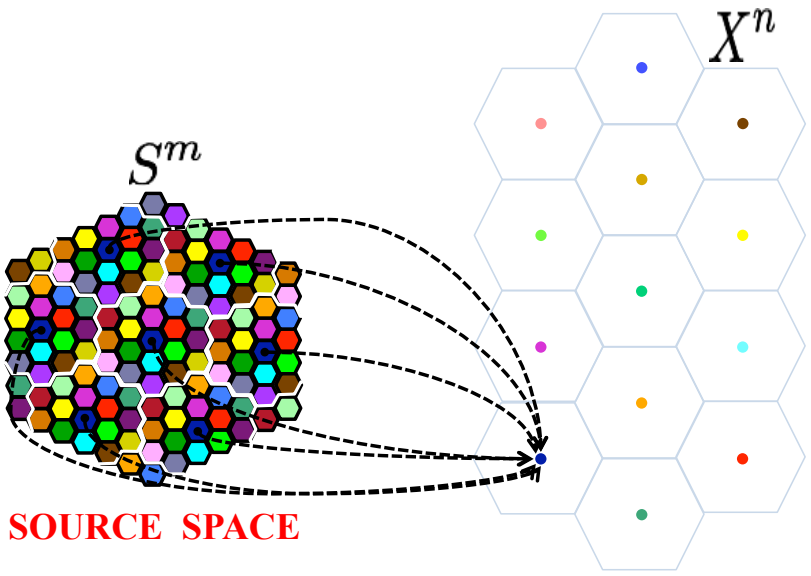
In lossy transmission, we first quantize, then bin:

- Fix $P_{W|S}$. Create a codebook of m-length codewords W^m of size $\sim 2^{m[I(S;W)+\epsilon]}$.
- Randomly assign these codewords into bins such that there are $\sim 2^{n[I(T;W)-\epsilon]}$ elements in each bin.
- Sufficiently few elements in each bin to decode W^m using typicality.
- Since $T-S-W$, correct W^m satisfies typicality (conditional typicality lemma)
- Once W^m is decoded, use it with side information T^m through a single-letter function $\hat{S}_i = \phi(T_i, W_i)$.

Minimum source coding rate within distortion D :

$$\begin{aligned} R^{WZ}(D) &= \min_{W,\phi: T-S-W, E[d(S,\phi(T,W))]\leq D} I(S;T) - I(W;T) \\ &= \min_{W,\phi: T-S-W, E[d(S,\phi(T,W))]\leq D} I(S;W|T) \end{aligned}$$

Generalized Coding Scheme



- Generate $M = 2^{mR}$ bins with

$$H(S|T) \leq R \leq H(S)$$

- Randomly allocate source sequences to bins.
  $\mathcal{B}(i)$: sequences in ith bin

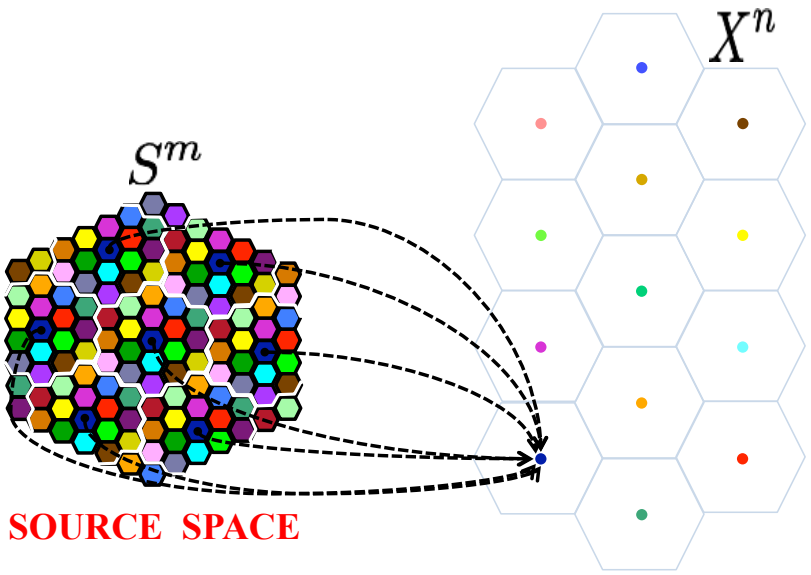
- Joint decoding: Find bin index s.t.
  1. corresponding channel input $x^n(i)$ is typical with channel output Y^n,
  2. there exist exactly one codeword in the bin jointly typical with side information T^m

- Prob of error: Prob. of having another bin satisfying above conditions:

$$2^{mR} 2^{-n(I(X;Y)-3\epsilon)} \left|\mathcal{B}(i) \cap \mathcal{A}_\epsilon^m(S)\right| 2^{-m(I(S;T)-3\epsilon)}$$
$$\leq 2^{-n(I(X;Y)-3\epsilon)} 2^{-m(H(S|T)-2\epsilon)}$$

goes to zero if $m(H(S|T)) \leq nI(X;Y)$.

Generalized Coding Scheme



- Generate $M = 2^{mR}$ bins with

$$H(S|T) \leq R \leq H(S)$$

- Randomly allocate source sequences to bins.
  $\mathcal{B}(i)$: sequences in ith bin

- Joint decoding: Find bin index s.t.
  1. corresponding channel input $x^n(i)$ is typical with channel output Y^n,
  2. there exist exactly one codeword in the bin jointly typical with side information T^m

- Prob of error: Prob. of having another bin satisfying above conditions:

$$2^{mR}2^{-n(I(X;Y)-3\epsilon)}\left|\mathcal{B}(i)\cap\mathcal{A}_\epsilon^m(S)\right|2^{-m(I(S;T)-3\epsilon)}$$
$$\leq 2^{-n(I(X;Y)-3\epsilon)}2^{-m(H(S|T)-2\epsilon)}$$

goes to zero if $m(H(S|T)) \leq nI(X;Y)$.

Generalized Coding Scheme

- Separate decoding: List indices i s.t. $x^n(i)$ and Y^n are jointly typical. Source decoder finds the bin with a jointly typical sequence with T^m
- Separate source and channel coding is a special case for $R = H(S|T)$: single element in list
- Works without any binning at all: generate an iid channel codeword for each source outcome, i.e., $R = \log |\mathcal{S}_0|$
- Decoder outputs only typical sequences: no point having $\geq 2^{m(H(S)+\epsilon)}$ bins. $R = H(S)$ equivalent to *no-binning*
- Transfer complexity of binning from encoder to decoder

Generalized Coding Scheme

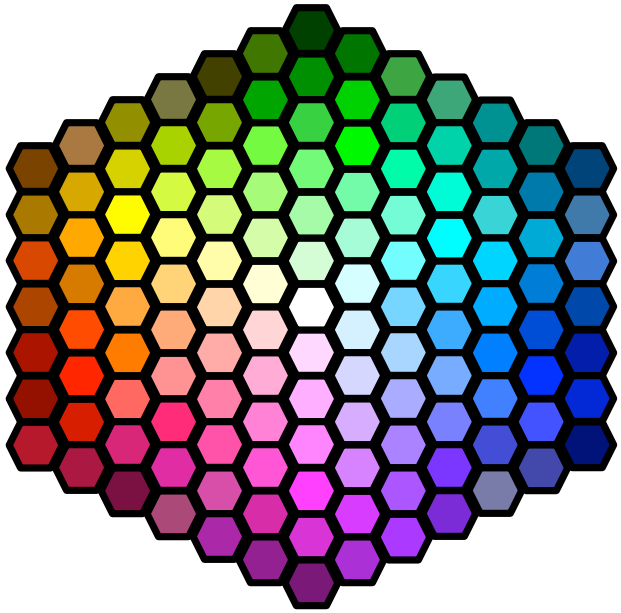
- Separate decoding: List indices i s.t. $x^n(i)$ and Y^n are jointly typical. Source decoder finds the bin with a jointly typical sequence with T^m
- Separate source and channel coding is a special case for $R = H(S|T)$: single element in list
- Works without any binning at all: generate an iid channel codeword for each source outcome, i.e., $R = \log |\mathcal{S}_0|$
- Decoder outputs only typical sequences: no point having $\geq 2^{m(H(S)+\epsilon)}$ bins. $R = H(S)$ equivalent to *no-binning*
- Transfer complexity of binning from encoder to decoder

Generalized Coding Scheme

- Separate decoding: List indices i s.t. $x^n(i)$ and Y^n are jointly typical. Source decoder finds the bin with a jointly typical sequence with T^m
- Separate source and channel coding is a special case for $R = H(S|T)$: single element in list
- Works without any binning at all: generate an iid channel codeword for each source outcome, i.e., $R = \log |\mathcal{S}_0|$
- Decoder outputs only typical sequences: no point having $\geq 2^{m(H(S)+\epsilon)}$ bins. $R = H(S)$ equivalent to *no-binning*
- Transfer complexity of binning from encoder to decoder

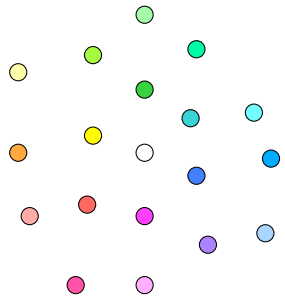
Channel is virtually binning the channel codewords; equivalently the source codewords (or, outcomes)

S^m



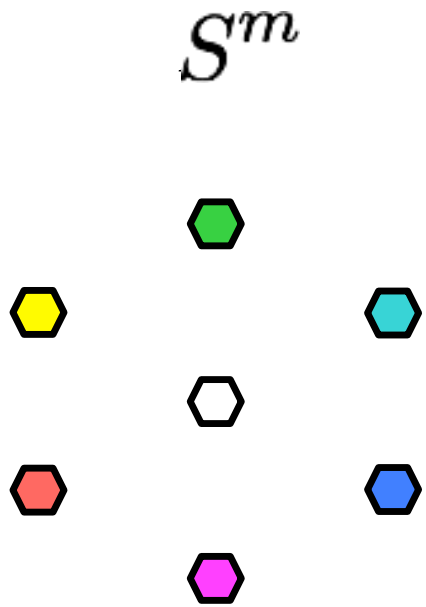
SOURCE SPACE

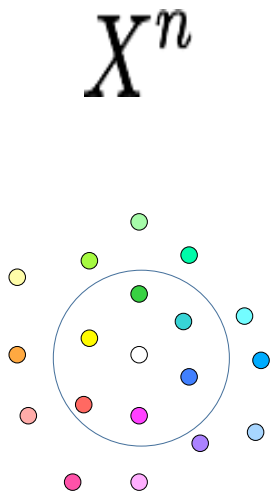
X^n



CHANNEL SPACE

When the channel is good, there will be fewer candidates in the list



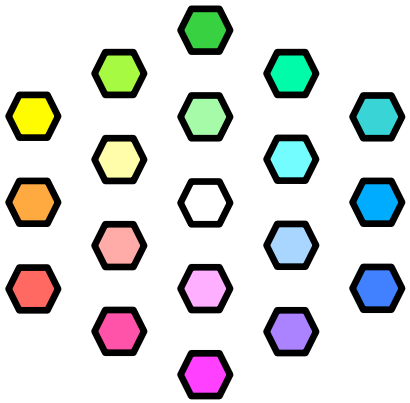


SOURCE SPACE

CHANNEL SPACE

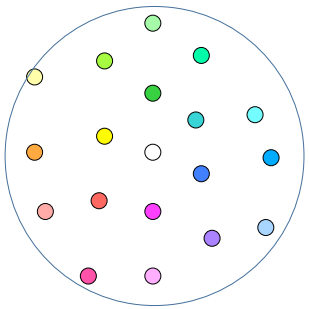
When the channel is weak, there will be more candidates

S^m



SOURCE SPACE

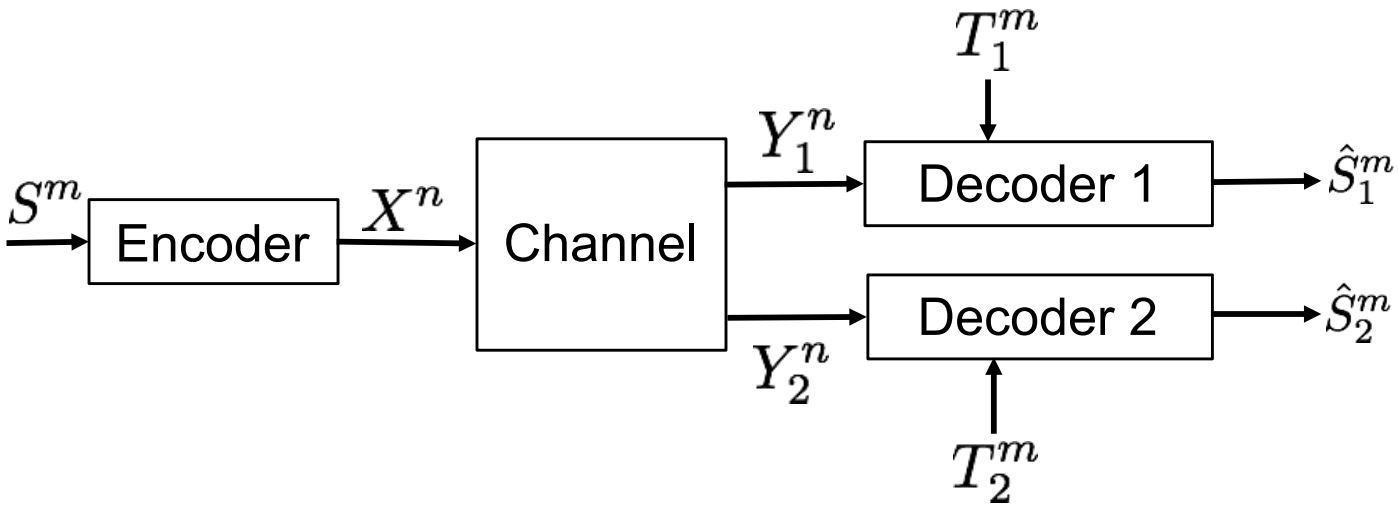
X^n



CHANNEL SPACE

When Does It Help?

- Multiple receivers with different side information.
- Strict separation suboptimal.



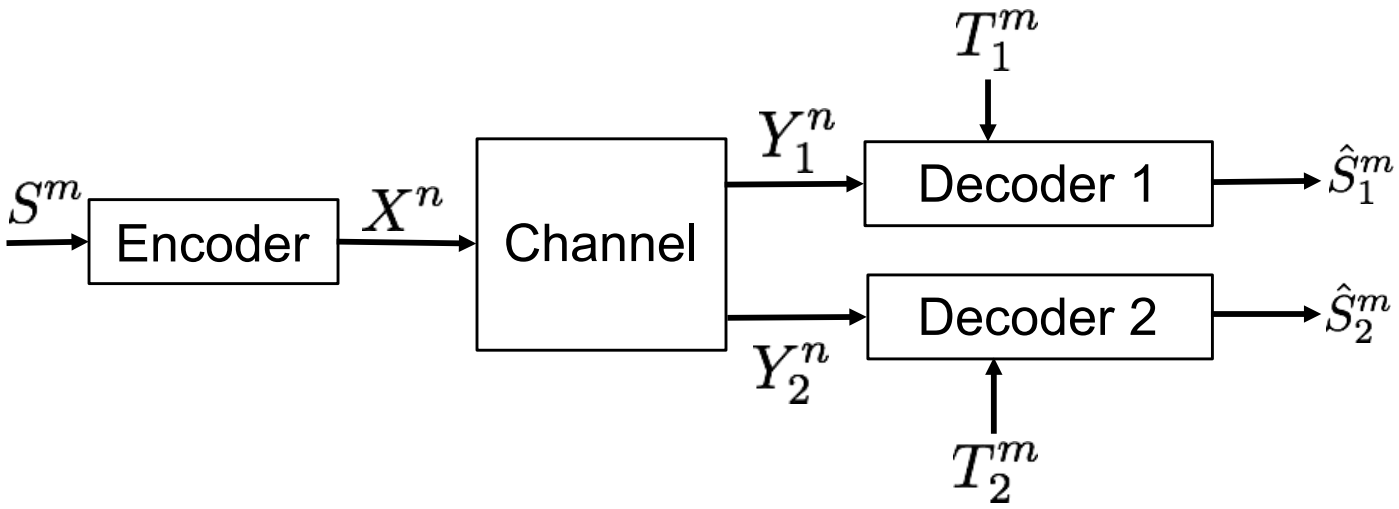
- Source-channel capacity:

$$\max_{p(x)} \min_{i=1,2} \frac{I(X;Y_i)}{H(S|T_i)}$$

- If $p(x)$ maximizes both $I(X;Y_1)$ and $I(X;Y_2)$, then we can use the channel at full capacity for each user.

E. Tuncel, **Slepian–Wolf coding over broadcast channels**, *IEEE Trans. Information Theory*, Apr. 2006.

- Multiple receivers with different side information.
- Strict separation suboptimal.



- Source-channel capacity:

$$\max_{p(x)} \min_{i=1,2} \frac{I(X;Y_i)}{H(S|T_i)}$$

- If $p(x)$ maximizes both $I(X;Y_1)$ and $I(X;Y_2)$, then we can use the channel at full capacity for each user.

E. Tuncel, **Slepian–Wolf coding over broadcast channels**, *IEEE Trans. Information Theory*, Apr. 2006.

- Randomly partition all source outputs into
  - $M_1 = 2^{nH(S|T_1)}$ bins for Receiver 1
  - $M_2 = 2^{nH(S|T_2)}$ bins for Receiver 2
- Fix $p(x)$. Generate
  - $M_1 M_2$ length-n codewords with $\prod_{i=1}^{n} p(x_i)$: $x^n(w_1, w_2)$, $w_i \in [1 : M_i]$.

	$\mathbf{1}$	$\cdots$	$\mathbf{M_2}$
$\mathbf{1}$	$x^n(1,1)$		$x^n(1,M_2)$
$\vdots$			
$\mathbf{M_1}$	$x^n(M_1,1)$		$x^n(M_1,M_2)$

D. Gunduz, E. Erkip, A. Goldsmith and H. V. Poor, **Reliable joint source-channel cooperative transmission over relay networks**, *IEEE Trans. Information Theory*, Apr. 2013.

Backward decoding

- Send Bm samples over $(B+1)n$ channel uses with $n/m = r$.
- $w_{1,i} \in [1 : M_1]$: bin index for receiver 1, $i = 1, \ldots, B$
- $w_{2,i} \in [1 : M_2]$: bin index for receiver 2, $i = 1, \ldots, B$

Block 1	Block 2	$\cdots$	Block i	$\cdots$	Block $B+1$
$x^n(w_{1,1}, 1)$	$x^n(w_{1,2}, w_{2,1})$	$\cdots$	$x^n(w_{1,i}, w_{2,i-1})$	$\cdots$	$x^n(1, w_{2,B})$

- Receiver 1 decodes reliably if

$$H(S|T_1) \leq r \cdot I(X; Y_1)$$

- Receiver 2 decodes reliably if

$$H(S|T_2) \leq r \cdot I(X; Y_2)$$

Backward decoding

- Send Bm samples over $(B+1)n$ channel uses with $n/m = r$.
- $w_{1,i} \in [1 : M_1]$: bin index for receiver 1, $i = 1, \ldots, B$
- $w_{2,i} \in [1 : M_2]$: bin index for receiver 2, $i = 1, \ldots, B$

Block 1	Block 2	$\cdots$	Block i	$\cdots$	Block $B+1$
$x^n(w_{1,1}, 1)$	$x^n(w_{1,2}, w_{2,1})$	$\cdots$	$x^n(w_{1,i}, w_{2,i-1})$	$\cdots$	$x^n(1, w_{2,B})$

- Receiver 1 decodes reliably if

$$H(S|T_1) \leq r \cdot I(X; Y_1)$$

- Receiver 2 decodes reliably if

$$H(S|T_2) \leq r \cdot I(X; Y_2)$$

Backward decoding

- Send Bm samples over $(B+1)n$ channel uses with $n/m = r$.
- $w_{1,i} \in [1 : M_1]$: bin index for receiver 1, $i = 1, \ldots, B$
- $w_{2,i} \in [1 : M_2]$: bin index for receiver 2, $i = 1, \ldots, B$

Block 1	Block 2	$\cdots$	Block i	$\cdots$	Block $B+1$
$x^n(w_{1,1}, 1)$	$x^n(w_{1,2}, w_{2,1})$	$\cdots$	$x^n(w_{1,i}, w_{2,i-1})$	$\cdots$	$x^n(1, w_{2,B})$

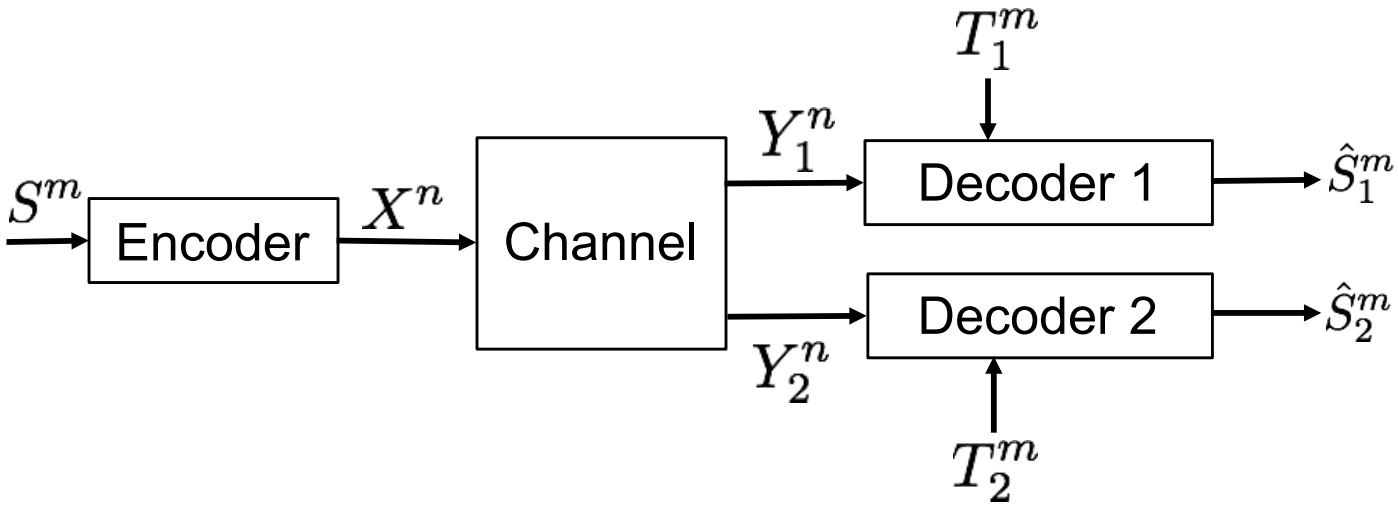
- Receiver 1 decodes reliably if

$$H(S|T_1) \leq r \cdot I(X; Y_1)$$

- Receiver 2 decodes reliably if

$$H(S|T_2) \leq r \cdot I(X; Y_2)$$

- First quantize, then broadcat quantized codeword



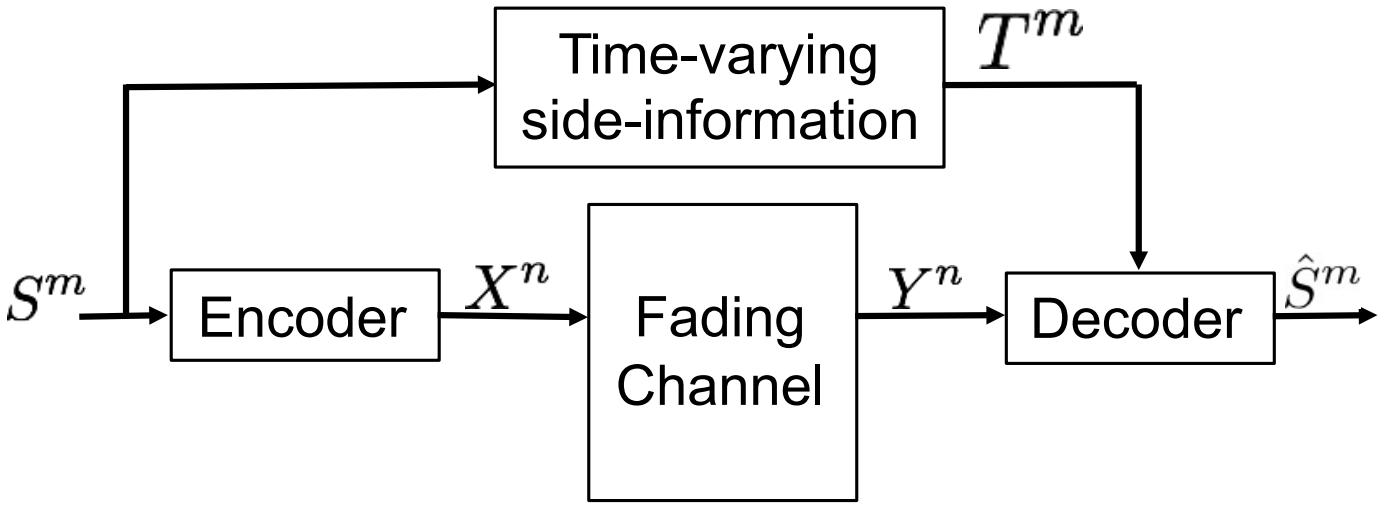
(D_1, D_2) is achievable at rate r if there exist W satisfying $W - S - (T_1, T_2)$, input distribution $p_X(x)$ and reconstruction functions ϕ_1, ϕ_2 such that

$$I(S; W|T_i) \leq rI(X; Y_i),$$
$$\mathrm{E}[d_k(S, \phi_i(W, T_i))] \leq D_i$$

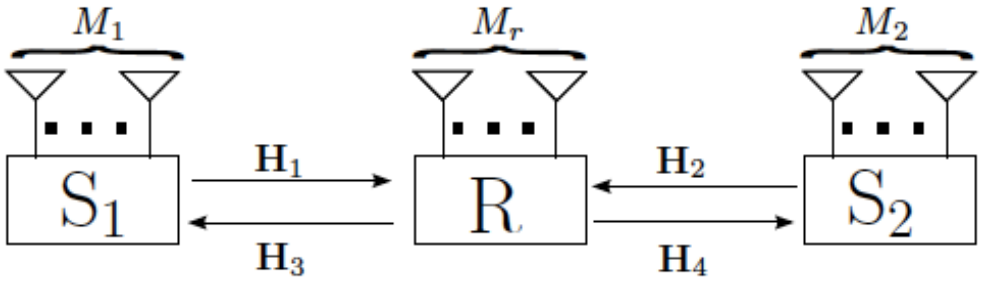
for $i = 1, 2$.

J. Nayak, E. Tuncel, D. Gunduz, **Wyner-Ziv coding over broadcast channels: Digital schemes**, *IEEE Trans. Information Theory*, Apr. 2010.

Time-varying Channel and Side Information



I. E. Aguerri and D. Gunduz, Joint source-channel coding with time-varying channel and side-information, IEEE Trans. Information Theory, vol. 62, no. 2, pp. 736 - 753, Feb. 2016.

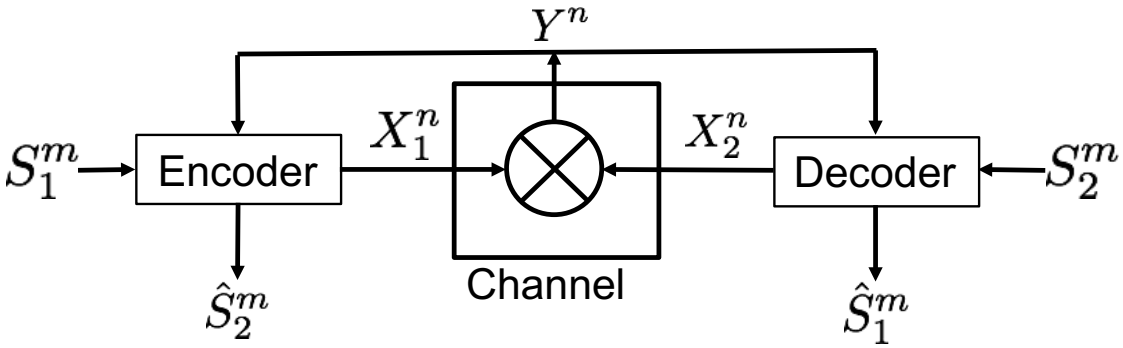


- Compress-and-forward at the relay
- Lossy broadcasting with side information
- Achieves optimal diversity-multiplexing trade-off

D. Gunduz, A. Goldsmith, and H. V. Poor, **MIMO two-way relay channel: Diversity-multiplexing trade-off analysis**, *Asilomar Conference*, Oct. 2008.

D. Gunduz, E. Tuncel, and J. Nayak, Rate regions for the separated two-way relay channel, Allerton Conf. on Comm., Control, and Computing, Sep. 2008.

- Separation does not hold for multi-user channels



- Binary two-way multiplying channel: $X_i \in \{0,1\}$, $i = 1, 2$

$$Y = X_1 \cdot X_2$$

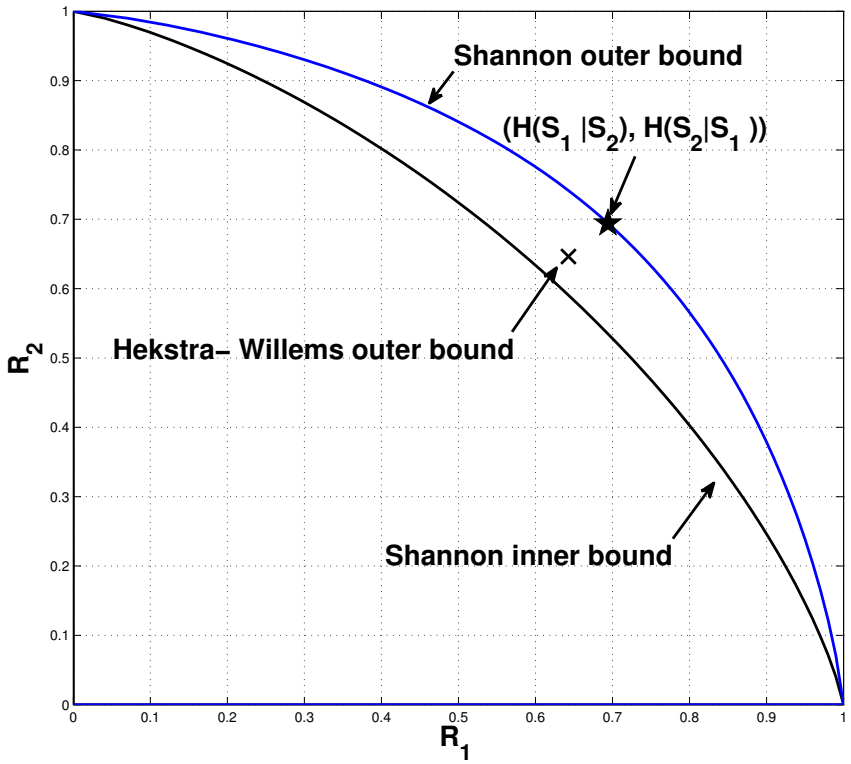
- Capacity still open: Shannon provided inner/ outer bounds
- Consider correlated signals S_1 and S_2:

	0	**1**
0	0	0.275
1	0.275	0.45

- With separation, they need to exchange rates

$$H(S_1|S_2) = H(S_2|S_1) = 0.6942 \text{ bpss}$$

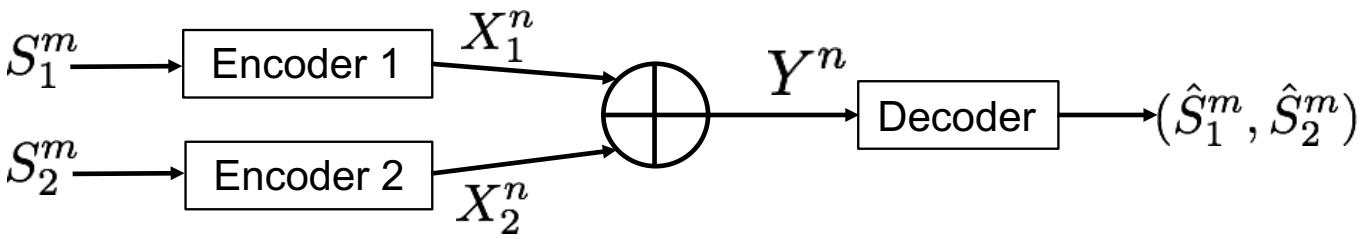
C. E. Shannon, **Two-way communication channels**, *in Proc. 4th Berkeley Symp. Math. Satist. Probability,* vol. 1, 1961, pp. 611-644.



- Symmetric transmission rate with independent channel inputs bounded by 0.64628 bpcu (Hekstra and Willems)
- Uncoded transmission allows reliable decoding!

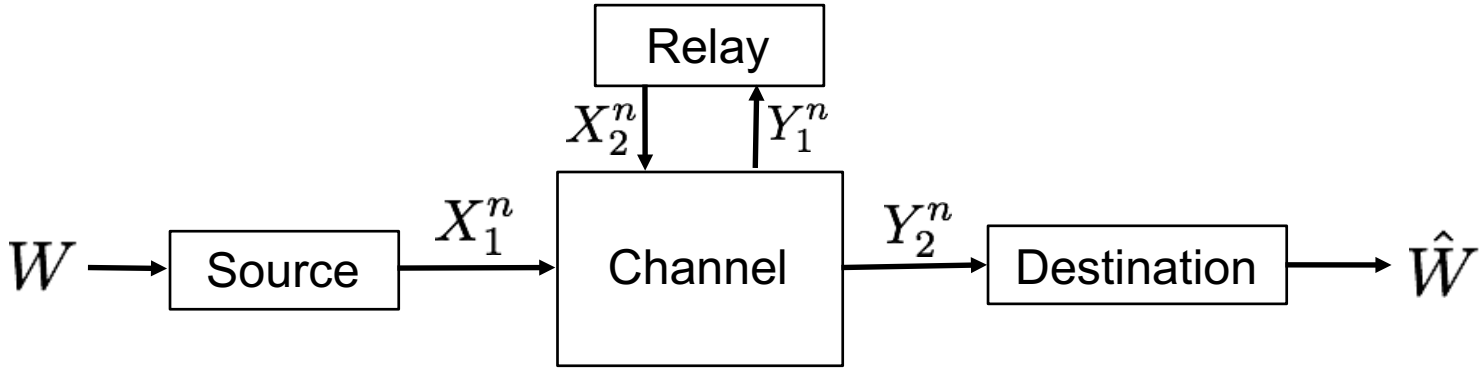
A. P. Hekstra and F. M. W. Willems, **Dependence balance bounds for single-output two-way channels**, *IEEE Trans. Inform. Theory*, Jan. 1989.

Multiple Access Channel (MAC) with Correlated Sources



- Binary input adder channel: $X_i \in \{0,1\}$, $Y = X_1 + X_2$
- $p(s_1, s_2)$: $p(0,0) = p(1,0) = p(0,1) = 1/3$
- $H(S_1, S_2) = \log 3 = 1.58$ bits/sample
- Max. sum rate with independent inputs: 1.5 bits/channel use
- Separation fails, while uncoded transmission is optimal

T. M. Cover, A. El Gamal and M. Salehi, **Multiple access channels with arbitrarily correlated sources**, *IEEE Trans. Information Theory*, Nov. 1980.



- Introduced by van der Meulen
- Characterized by $p(y_1, y_2|x_1, x_2)$
- Capacity of relay channel not known
- Multi-letter capacity given by van der Meulen:

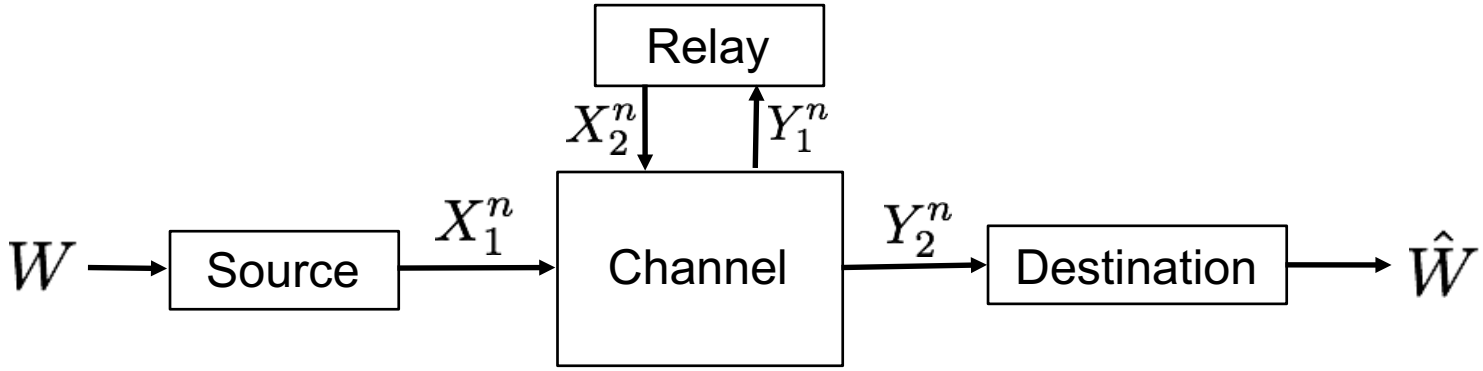
$$C = \sup_k C^k = \lim_{k\to\infty} C^k$$

where

$$C^k \triangleq \max_{p(x_1^k),\{x_{2i}(y_1^{i-1})\}_{i=1}^k} \frac{1}{k} I(X_1^k; Y_2^k)$$

- Various achievable schemes: amplify-and-forward, decode-and-forward, compress-and-forward

T. M. Cover and A. E. Gamal, **Capacity theorems for the relay channel**, *IEEE Trans. Inf. Theory*, Sep. 1979.



- Introduced by van der Meulen
- Characterized by $p(y_1, y_2|x_1, x_2)$
- Capacity of relay channel not known
- Multi-letter capacity given by van der Meulen:

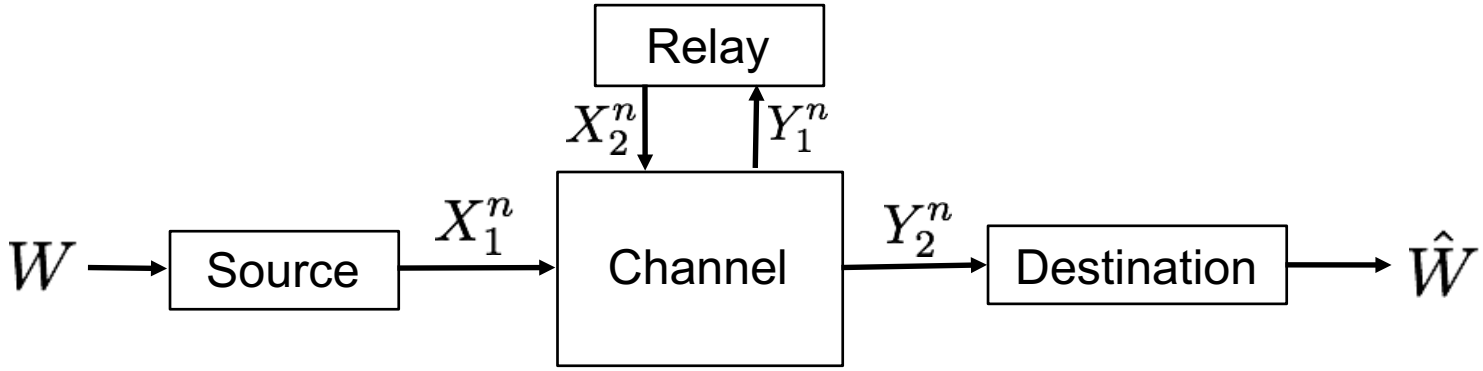
$$C = \sup_k C^k = \lim_{k\to\infty} C^k$$

where

$$C^k \triangleq \max_{p(x_1^k),\{x_{2i}(y_1^{i-1})\}_{i=1}^k} \frac{1}{k} I(X_1^k; Y_2^k)$$

- Various achievable schemes: amplify-and-forward, decode-and-forward, compress-and-forward

T. M. Cover and A. E. Gamal, **Capacity theorems for the relay channel**, *IEEE Trans. Inf. Theory*, Sep. 1979.



- Introduced by van der Meulen
- Characterized by $p(y_1, y_2|x_1, x_2)$
- Capacity of relay channel not known
- Multi-letter capacity given by van der Meulen:

$$C = \sup_k C^k = \lim_{k\to\infty} C^k$$

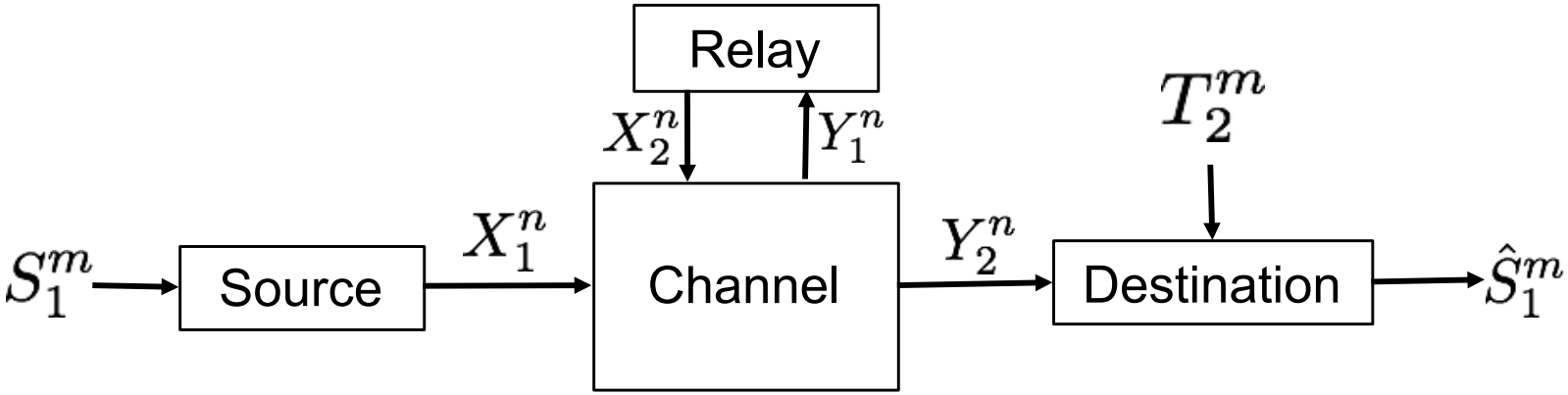
where

$$C^k \triangleq \max_{p(x_1^k),\{x_{2i}(y_1^{i-1})\}_{i=1}^k} \frac{1}{k} I(X_1^k; Y_2^k)$$

- Various achievable schemes: amplify-and-forward, decode-and-forward, compress-and-forward

T. M. Cover and A. E. Gamal, **Capacity theorems for the relay channel**, *IEEE Trans. Inf. Theory*, Sep. 1979.

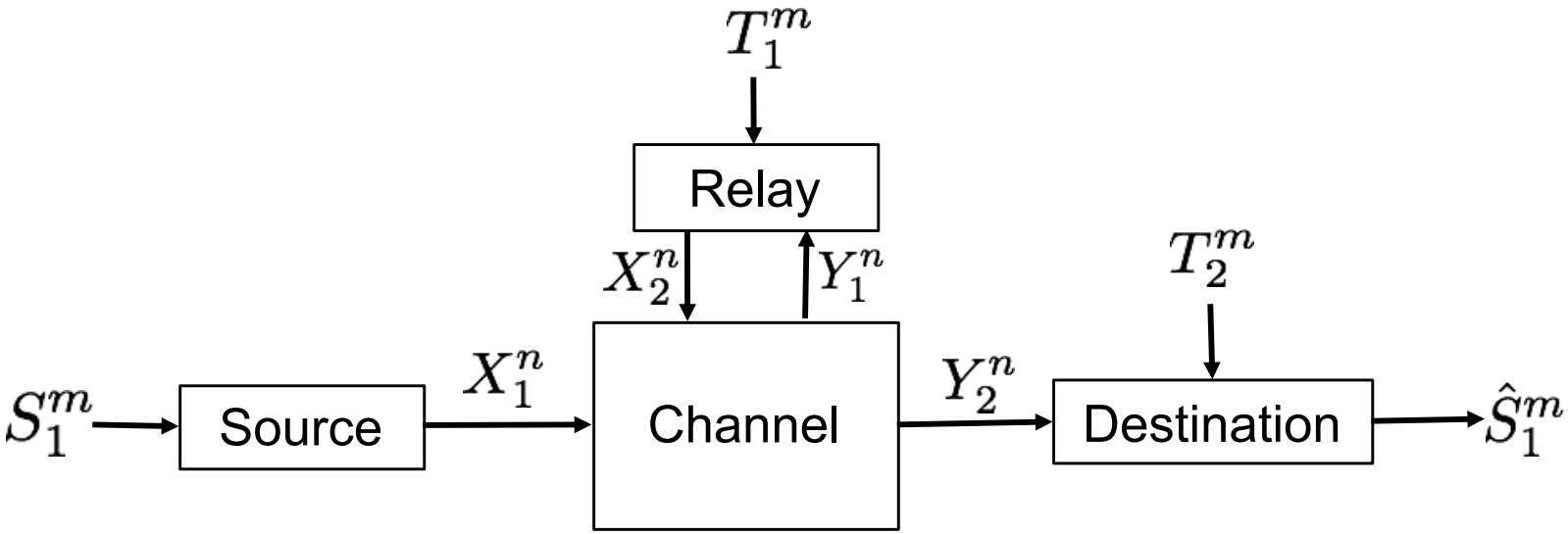
Relay Channel with Destination Side Information



- Separation still optimal
- Proof of separation in a network whose capacity is not known!

D. Gunduz, E. Erkip, A. Goldsmith and H. Poor, **Reliable joint source-channel cooperative transmission over relay networks**, *IEEE Trans. Inform. Theory*, Apr. 2013.

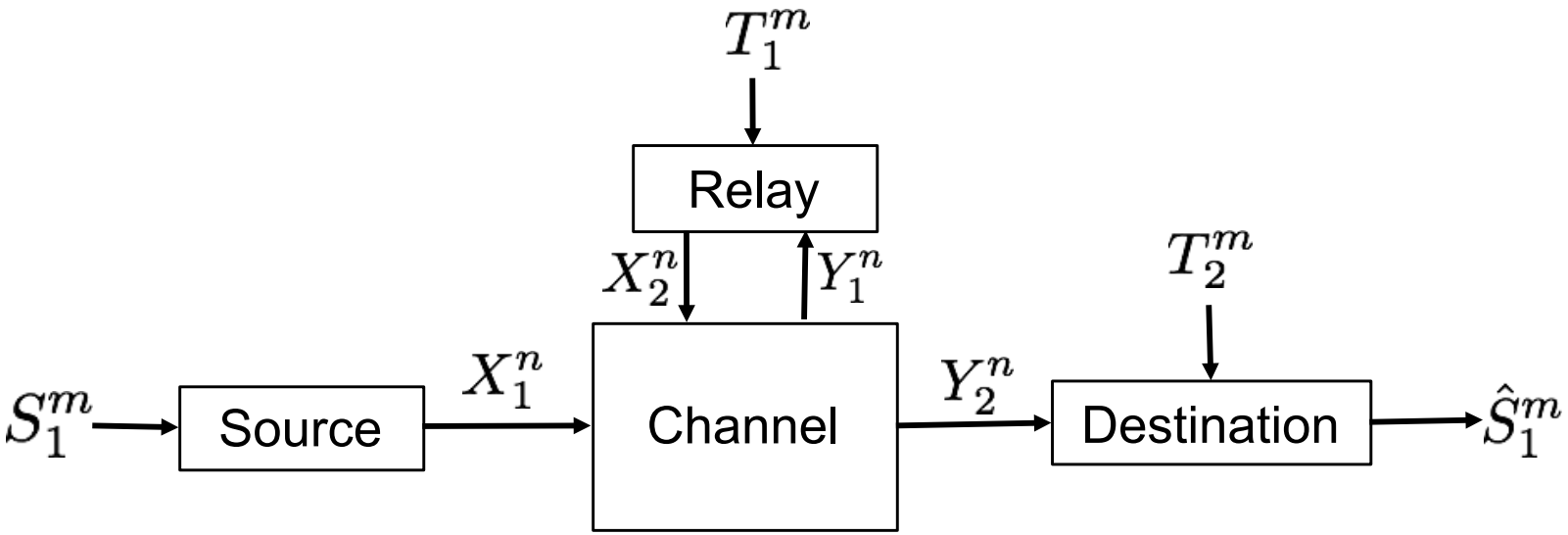
Relay and Destination Side Information



- Source-channel rate r is achievable if,

$$r \cdot H(S|T_1) \leq I(X_1; Y_1|X_2)$$
$$r \cdot H(S|T_2) \leq I(X_1, X_2; Y_2)$$

  for some $p(x_1, x_2)$.
- Decode-and-forward transmission
- Optimal for physically degraded relay channel $(X_1 - (X_2, Y_1) - Y_2)$ with degraded side information $(S_1 - T_1 - T_2)$



- Source-channel rate r is achievable if,

$$r \cdot H(S|T_1) \leq I(X_1; Y_1|X_2)$$
$$r \cdot H(S|T_2) \leq I(X_1, X_2; Y_2)$$

  for some $p(x_1, x_2)$.
- Decode-and-forward transmission
- Optimal for physically degraded relay channel $(X_1 - (X_2, Y_1) - Y_2)$ with degraded side information $(S_1 - T_1 - T_2)$

Achievability

- Block Markov encoding
- Regular encoding and joint source-channel sliding window decoding
  - More complicated decoder
  - Less delay
- Regular encoding and separate source-channel backward decoding
  - Simpler decoder
  - More delay

Backward decoding

- Randomly partition all source outputs into
  - $M_1 = 2^{nH(S|T_1)}$ bins: Relay bins
  - $M_2 = 2^{nH(S|T_2)}$ bins: Destination bins
- Fix $p(x_1, x_2)$. Generate
  - M_1 codewords of length n with $\prod_{i=1}^{n} p(x_{2,i})$. Enumerate as $x_2^n(w_2)$.
  - For each $x_2^n(w_2)$, generate M_1 codewords of length n with $\prod_{i=1}^{n} p(x_{1,i}|x_{2,i}^n)$. Enumerate as $x_1^n(w_1, w_2)$

Backward decoding

- Send Bm samples over $(B+1)n$ channel uses with $n/m = r$.
- $w_{1,i} \in [1, M_1]$: relay bin index of source block $i = 1, \ldots, B$
- $w_{2,i} \in [1, M_2]$: destination bin index of block $i = 1, \ldots, B$

Block 1	Block 2	$\cdots$	Block i	$\cdots$	Block $B+1$
$x_1^n(w_{1,1}, 1)$	$x_1^n(w_{1,2}, w_{2,1})$	$\cdots$	$x_1^n(w_{1,i}, w_{2,i-1})$	$\cdots$	$x_1^n(1, w_{2,B})$
$x_2^n(1)$	$x_2^n(w'_{2,1})$	$\cdots$	$x_2^n(w'_{2,i-1})$	$\cdots$	$x_2^n(w'_{2,B})$

- Relay decodes reliably if

$$H(S|T_1) \leq r \cdot I(X_1; Y_1|X_2)$$

- Destination decodes reliably if

$$H(S|T_2) \leq r \cdot I(X_1, X_2; Y_1)$$

Backward decoding

- Send Bm samples over $(B+1)n$ channel uses with $n/m = r$.
- $w_{1,i} \in [1, M_1]$: relay bin index of source block $i = 1, \ldots, B$
- $w_{2,i} \in [1, M_2]$: destination bin index of block $i = 1, \ldots, B$

Block 1	Block 2	$\cdots$	Block i	$\cdots$	Block $B+1$
$x_1^n(w_{1,1}, 1)$	$x_1^n(w_{1,2}, w_{2,1})$	$\cdots$	$x_1^n(w_{1,i}, w_{2,i-1})$	$\cdots$	$x_1^n(1, w_{2,B})$
$x_2^n(1)$	$x_2^n(w'_{2,1})$	$\cdots$	$x_2^n(w'_{2,i-1})$	$\cdots$	$x_2^n(w'_{2,B})$

- Relay decodes reliably if

$$H(S|T_1) \leq r \cdot I(X_1; Y_1|X_2)$$

- Destination decodes reliably if

$$H(S|T_2) \leq r \cdot I(X_1, X_2; Y_1)$$

Backward decoding

- Send Bm samples over $(B+1)n$ channel uses with $n/m = r$.
- $w_{1,i} \in [1, M_1]$: relay bin index of source block $i = 1, \ldots, B$
- $w_{2,i} \in [1, M_2]$: destination bin index of block $i = 1, \ldots, B$

Block 1	Block 2	$\cdots$	Block i	$\cdots$	Block $B+1$
$x_1^n(w_{1,1}, 1)$	$x_1^n(w_{1,2}, w_{2,1})$	$\cdots$	$x_1^n(w_{1,i}, w_{2,i-1})$	$\cdots$	$x_1^n(1, w_{2,B})$
$x_2^n(1)$	$x_2^n(w'_{2,1})$	$\cdots$	$x_2^n(w'_{2,i-1})$	$\cdots$	$x_2^n(w'_{2,B})$

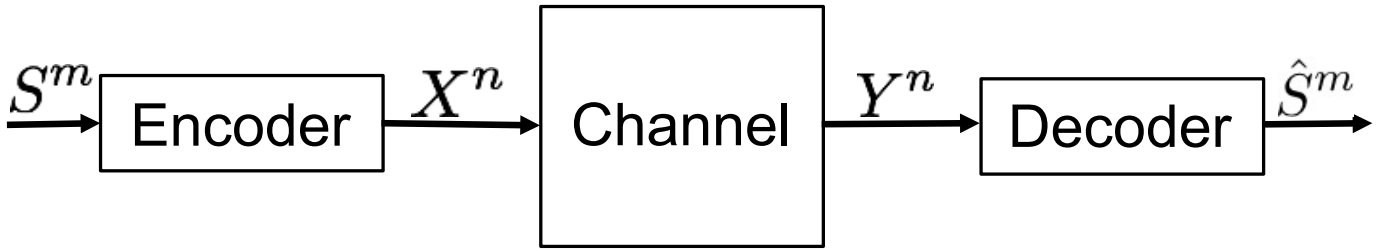
- Relay decodes reliably if

$$H(S|T_1) \leq r \cdot I(X_1; Y_1|X_2)$$

- Destination decodes reliably if

$$H(S|T_2) \leq r \cdot I(X_1, X_2; Y_1)$$

Do we need coding?



- Let $S_i \sim \mathcal{N}(0,1)$ i.i.d. Gaussian
- Memoryless Gaussian channel:

$$Y_i = X_i + Z_i, \quad Z_i \mathcal{N}(0,N), \qquad \frac{1}{m}\mathbb{E}[X^m (X^m)^T] \leq P$$

- Capacity: $\frac{1}{2}\log\left(1+\frac{P}{N}\right)$
- Distortion-rate function: $D(R) = 2^{-2R}$
-

$$D_{min} = \left(1+\frac{P}{N}\right)^{-1}$$

- What about uncoded/ analog transmision?

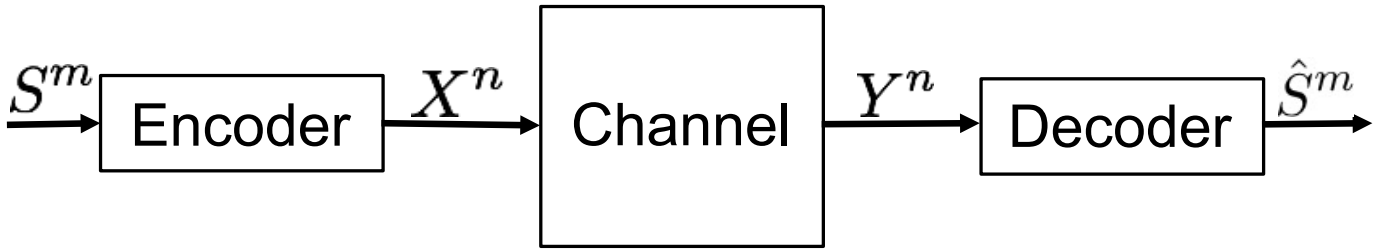
$$X_i = \sqrt{P} S_i$$

MMSE at the receiver

- Uncoded symbol-by-symbol transmission is optimal!

T. J. Goblick, **Theoretical limitations on the transmission of data from analog sources**, *IEEE Trans. Inf. Theory*, vol. 11, pp. 558- 567, Oct. 1965.

Do we need coding?



- Let $S_i \sim \mathcal{N}(0,1)$ i.i.d. Gaussian
- Memoryless Gaussian channel:

$$Y_i = X_i + Z_i, \quad Z_i \mathcal{N}(0,N), \qquad \frac{1}{m}\mathbb{E}[X^m (X^m)^T] \leq P$$

- Capacity: $\frac{1}{2}\log\left(1+\frac{P}{N}\right)$
- Distortion-rate function: $D(R) = 2^{-2R}$
- $$D_{min} = \left(1+\frac{P}{N}\right)^{-1}$$
- What about uncoded/ analog transmision?

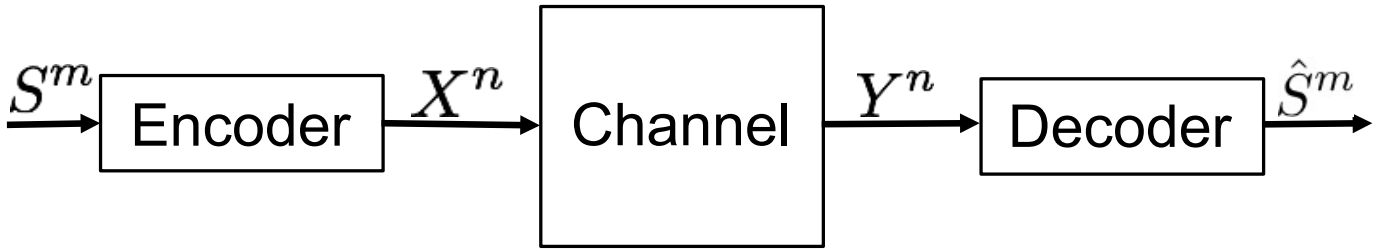
$$X_i = \sqrt{P} S_i$$

MMSE at the receiver

- Uncoded symbol-by-symbol transmission is optimal!

T. J. Goblick, **Theoretical limitations on the transmission of data from analog sources**, *IEEE Trans. Inf. Theory*, vol. 11, pp. 558- 567, Oct. 1965.

Do we need coding?



- Let $S_i \sim \mathcal{N}(0,1)$ i.i.d. Gaussian
- Memoryless Gaussian channel:

$$Y_i = X_i + Z_i, \quad Z_i \mathcal{N}(0,N), \qquad \frac{1}{m}\mathbb{E}[X^m (X^m)^T] \leq P$$

- Capacity: $\frac{1}{2}\log\left(1+\frac{P}{N}\right)$
- Distortion-rate function: $D(R) = 2^{-2R}$
-

$$D_{min} = \left(1+\frac{P}{N}\right)^{-1}$$

- What about uncoded/ analog transmision?

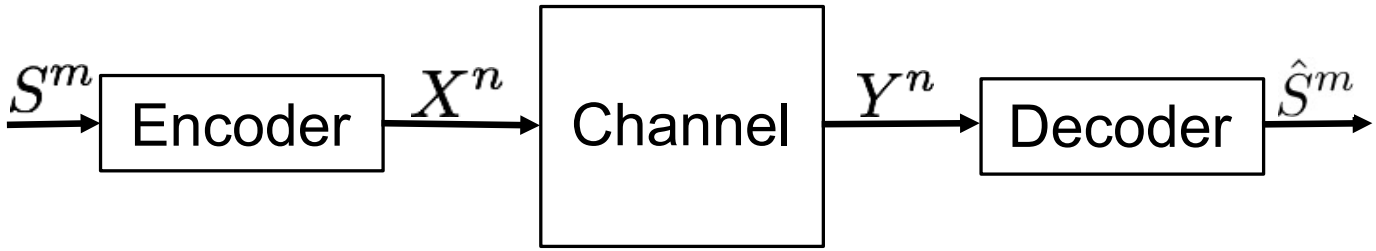
$$X_i = \sqrt{P} S_i$$

MMSE at the receiver

- Uncoded symbol-by-symbol transmission is optimal!

T. J. Goblick, **Theoretical limitations on the transmission of data from analog sources**, *IEEE Trans. Inf. Theory*, vol. 11, pp. 558- 567, Oct. 1965.

Do we need coding?



- Let $S_i \sim \mathcal{N}(0,1)$ i.i.d. Gaussian
- Memoryless Gaussian channel:

$$Y_i = X_i + Z_i, \quad Z_i \mathcal{N}(0,N), \qquad \frac{1}{m}\mathbb{E}[X^m (X^m)^T] \leq P$$

- Capacity: $\frac{1}{2}\log\left(1+\frac{P}{N}\right)$
- Distortion-rate function: $D(R) = 2^{-2R}$
-

$$D_{min} = \left(1+\frac{P}{N}\right)^{-1}$$

- What about uncoded/ analog transmision?

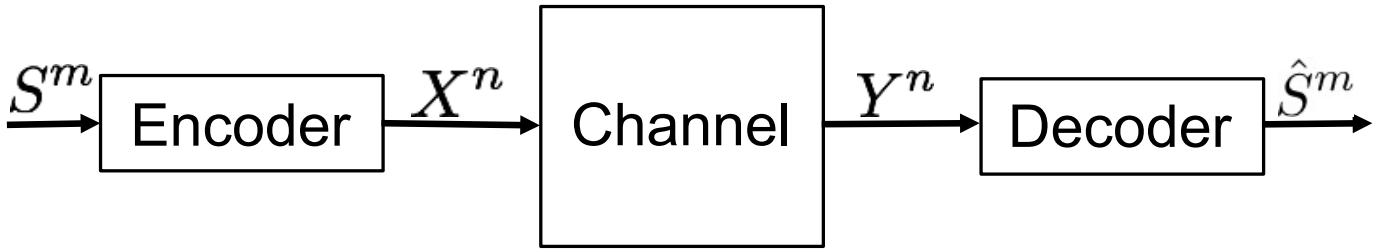
$$X_i = \sqrt{P} S_i$$

MMSE at the receiver

- Uncoded symbol-by-symbol transmission is optimal!

T. J. Goblick, **Theoretical limitations on the transmission of data from analog sources**, *IEEE Trans. Inf. Theory*, vol. 11, pp. 558- 567, Oct. 1965.

Do we need coding?



- Let $S_i \sim \mathcal{N}(0,1)$ i.i.d. Gaussian
- Memoryless Gaussian channel:

$$Y_i = X_i + Z_i, \quad Z_i \mathcal{N}(0,N), \qquad \frac{1}{m}\mathbb{E}[X^m (X^m)^T] \leq P$$

- Capacity: $\frac{1}{2}\log\left(1+\frac{P}{N}\right)$
- Distortion-rate function: $D(R) = 2^{-2R}$
-

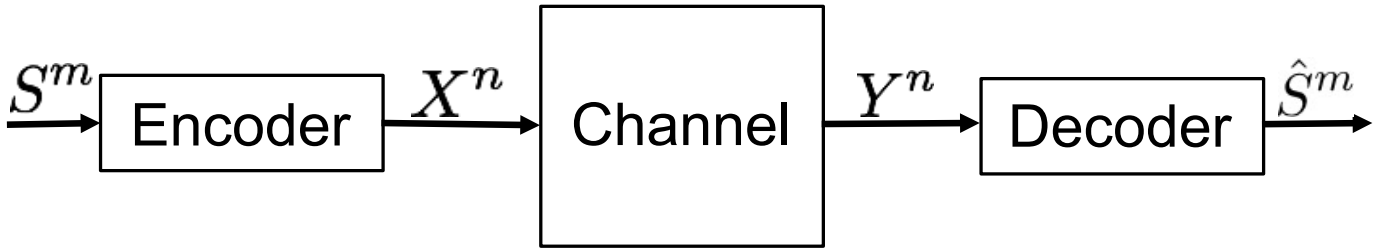
$$D_{min} = \left(1+\frac{P}{N}\right)^{-1}$$

- What about uncoded/ analog transmision?

$$X_i = \sqrt{P} S_i$$

  MMSE at the receiver
- Uncoded symbol-by-symbol transmission is optimal!

T. J. Goblick, **Theoretical limitations on the transmission of data from analog sources**, *IEEE Trans. Inf. Theory*, vol. 11, pp. 558- 567, Oct. 1965.

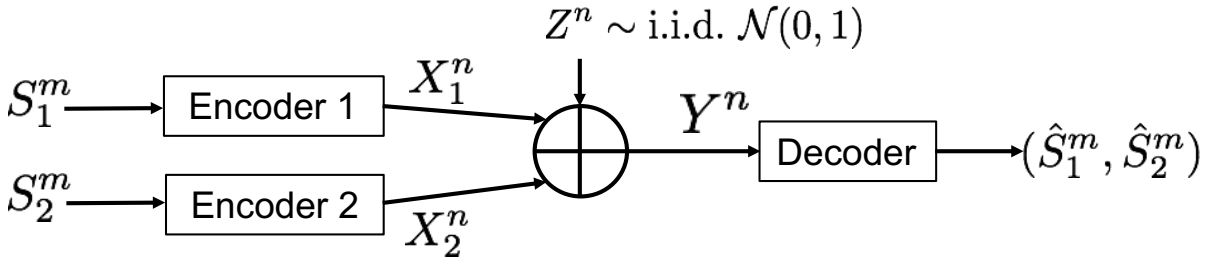


S can be communicated over channel $p(y|x)$ uncoded if

- $X \sim p_S(x)$ attains the capacity $C = \max_{p(x)} I(X;Y)$
- test channel $p_{Y|X}(\hat{s}|s)$ attains the rate-distortion function $R(D) = \min_{p(\hat{s}|s):\mathrm{E}[d(S,\hat{S})\leq D]} I(S;\hat{S})$

Then, we have $C = R(D)$.

M. Gastpar, B. Rimoldi, and M. Vetterli, **To code, or not to code: Lossy source-channel communication revisited**, *IEEE Trans. Inf. Theory*, May 2003.



- Correlated Gausssian sources: $\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$
- Memoryless Gaussian MAC:

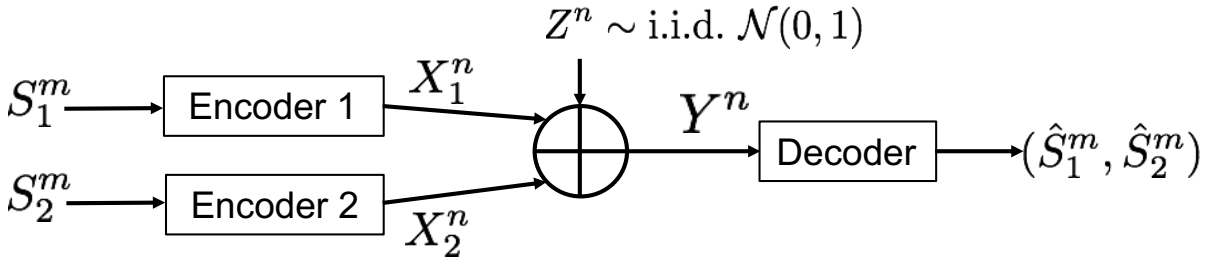
$$Y_j = X_{1,j} + X_{2,j} + Z_j, \quad Z_j \sim \mathcal{N}(0,1), \qquad \frac{1}{m}\mathbb{E}[X_i^m (X_i^m)^T] \leq P$$

- Mean squared-error distortion measure:
  $D_i = \mathrm{E}\left[\frac{1}{m}\sum_{j=1}^{m} |S_{i,j} - \hat{S}_{i,j}|^2\right], \; i = 1, 2.$
- Necessary conditions: $R_{S_1,S_2}(D_1, D_2) \leq \frac{1}{2}\log(1 + 2P(1+\rho))$

Corollary

Uncoded transmission is optimal in the low SNR regime, i.e., if $P \leq \frac{\rho}{1-\rho^2}$.

A. Lapidoth and S. Tinguely, **Sending a bivariate Gaussian over a Gaussian MAC**, *IEEE Transactions on Information Theory*, Jun. 2010.



- Correlated Gausssian sources: $\begin{pmatrix} S_1 \\ S_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$
- Memoryless Gaussian MAC:

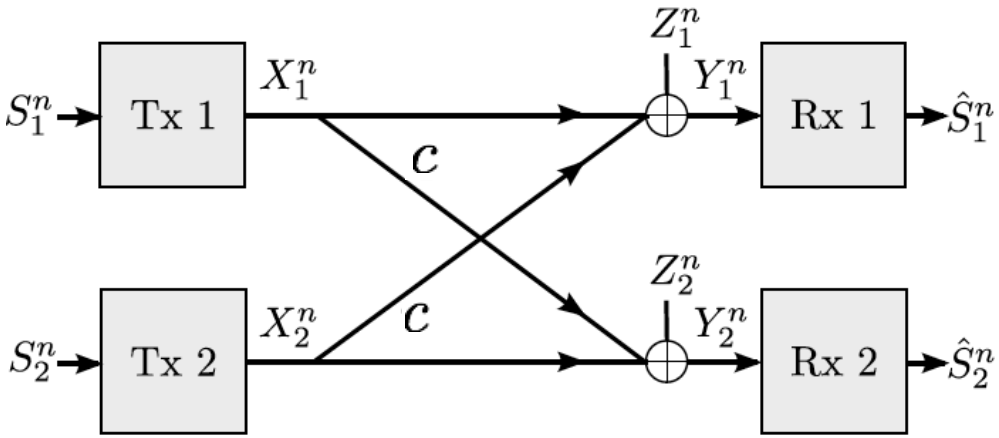
$$Y_j = X_{1,j} + X_{2,j} + Z_j, \quad Z_j \sim \mathcal{N}(0,1), \qquad \frac{1}{m}\mathbb{E}[X_i^m (X_i^m)^T] \leq P$$

- Mean squared-error distortion measure:
  $D_i = \mathrm{E}\left[\frac{1}{m}\sum_{j=1}^{m} |S_{i,j} - \hat{S}_{i,j}|^2\right], \; i = 1, 2.$
- Necessary conditions: $R_{S_1,S_2}(D_1, D_2) \leq \frac{1}{2}\log(1 + 2P(1+\rho))$

Corollary

Uncoded transmission is optimal in the low SNR regime, i.e., if $P \leq \frac{\rho}{1-\rho^2}$.

A. Lapidoth and S. Tinguely, **Sending a bivariate Gaussian over a Gaussian MAC**, *IEEE Transactions on Information Theory*, Jun. 2010.



- Correlated Gausssian sources with correlation coefficient ρ
- Memoryless Gaussian weak interference channel ($c \leq 1$):

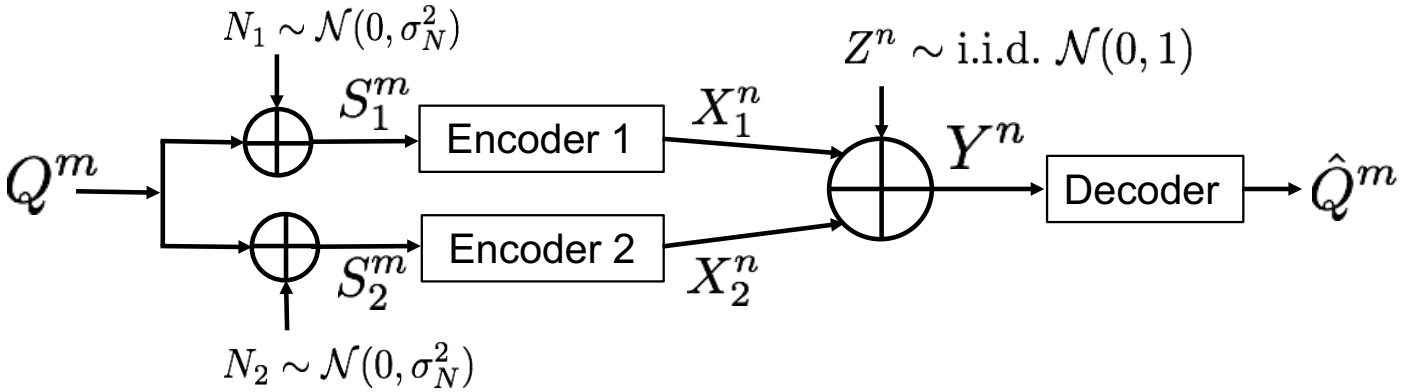
$$Y_{1,j} = X_{1,j} + cX_{2,j} + Z_{1,j},$$
$$Y_{2,j} = cX_{1,j} + X_{2,j} + Z_{2,j},$$

with $\frac{1}{m}\mathbb{E}[X_i^m (X_i^m)^T] \leq P$

Corollary

Uncoded transmission is optimal in the low SNR regime, i.e., if $cP \leq \frac{\rho}{1-\rho^2}$.

I. E. Aguerri and D. Gunduz, **Correlated Gaussian sources over Gaussian weak interference channels**, *IEEE Inform. Theory Workshop (ITW)*, Oct. 2015.



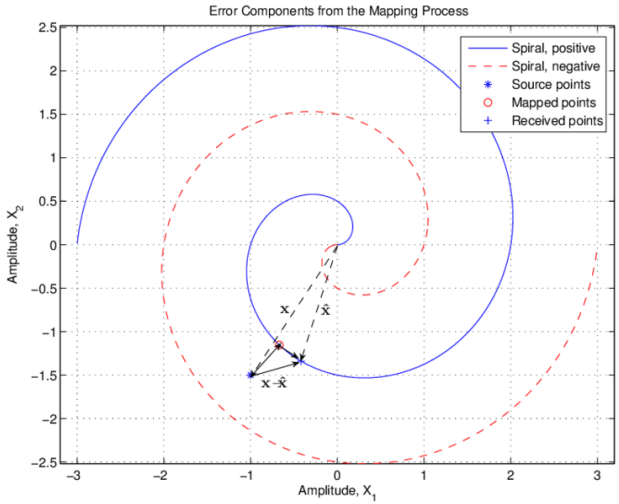
Memoryless Gaussian MAC:

$$Y_i = X_{1,j} + X_{2,j} + Z_i, \quad Z_i \sim \mathcal{N}(0,1), \qquad \frac{1}{m}\mathbb{E}[X_i^m (X_i^m)^T] \leq P$$

Uncoded transmission is always optimal!

M. Gastpar, **Uncoded transmission is exactly optimal for a simple Gaussian sensor network,** *IEEE Trans. Inf. Theory*, Nov. 2008.

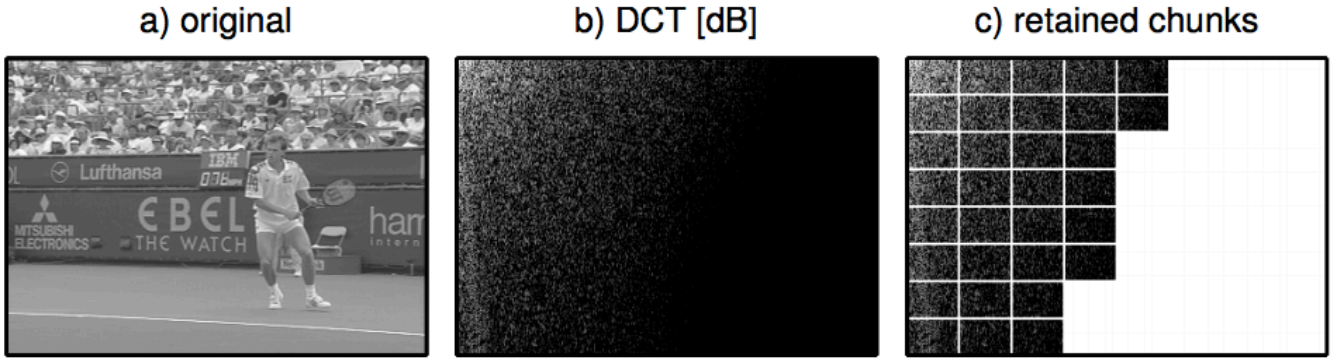
- How do we map 2 Gaussian sample into 1 channel use? or, 1 sample to 2 channel uses?
- Optimal mappings (encoder and decoder) are either noth linear or both nonlinear.
- Can be optimized numerically.
- What about 1 sample and unlimited bandwidth?



E Akyol, KB Viswanatha, K Rose, TA Ramstad, **On zero-delay source-channel coding**, *IEEE Transactions on Information Theory,* Dec. 2012.

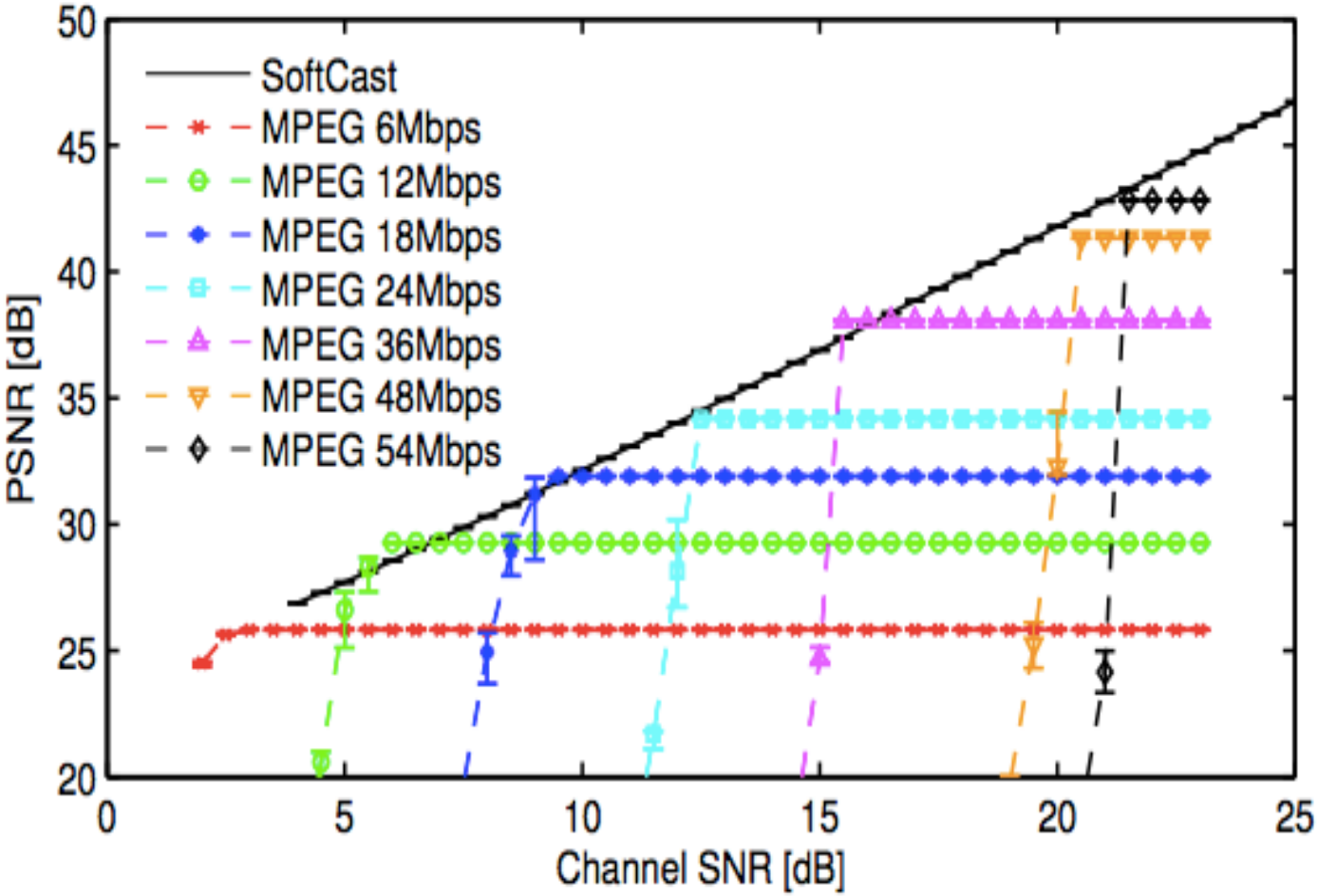
E. Koken, E. Tuncel, and D. Gunduz, **Energy-distortion exponents in lossy transmission of Gaussian sources over Gaussian channels**, *IEEE Trans. Information Theory*, Feb. 2017.

SoftCast: Uncoded image/video transmission

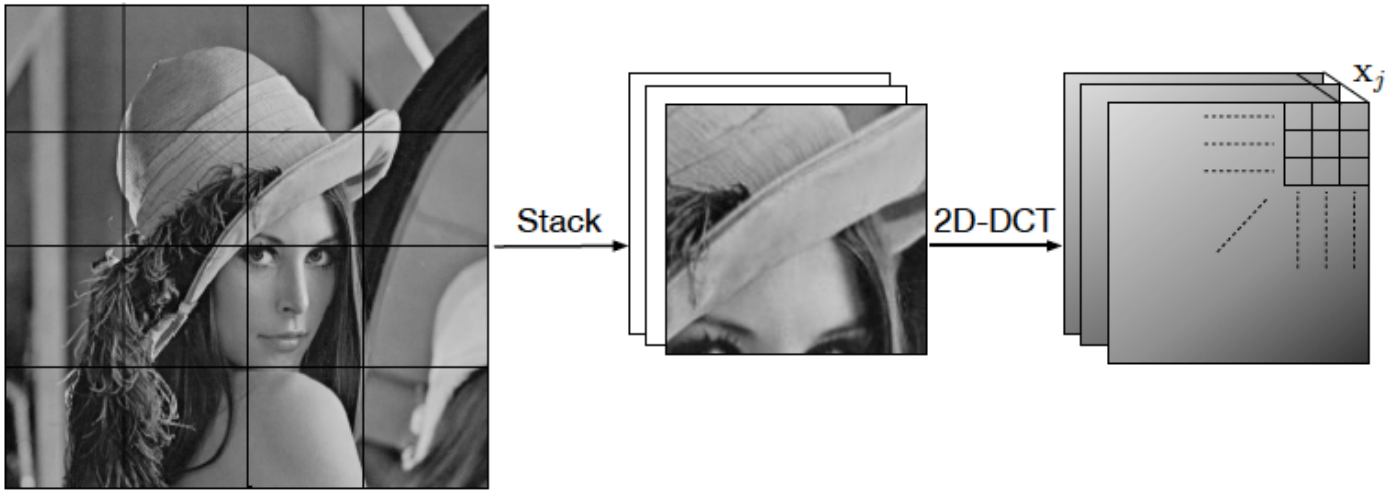


- Divide DCT coefficients into blocks
- Find empirical variance ("energy") of each block
- Compression: Remove blocks with low energy
- Remaining blocks transmitted uncoded
- Power allocation according to block energies

S. Jakubczak and D. Katabi, **Softcast: One-size-fits-all wireless video,** *in Proc. ACM SIGCOMM*, New York, NY, Aug. 2010, pp. 449–450.



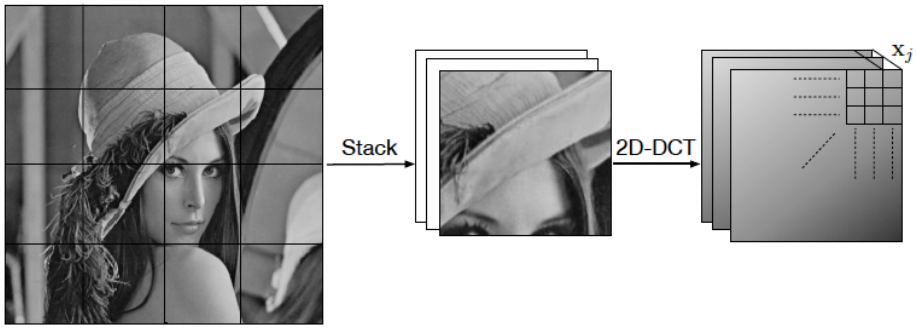
S. Jakubczak and D. Katabi, **Softcast: One-size-fits-all wireless video,** *in Proc. ACM SIGCOMM*, New York, NY, Aug. 2010, pp. 449–450.

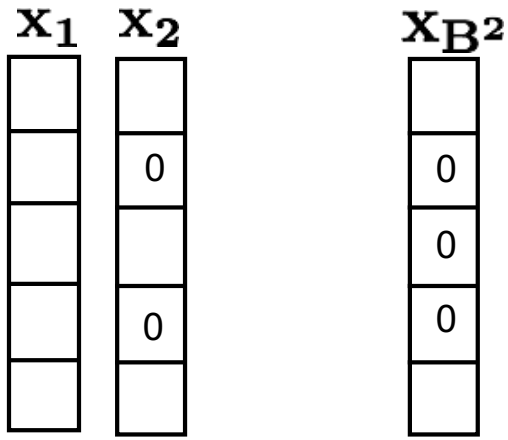


- SparseCast: Hybrid digital-analog image transmission
- Block-based DCT transform
- One vector for each frequency component
- Thresholding for compression (remove small components)
- Compressive sensing for transmission

Tung and Gunduz, **SparseCast: Hybrid Digital-Analog Wireless Image Transmission Exploiting Frequency Domain Sparsity**, *IEEE Comm. Letters*, 2018.

Exploit Sparsity for Bandwidth Efficiency



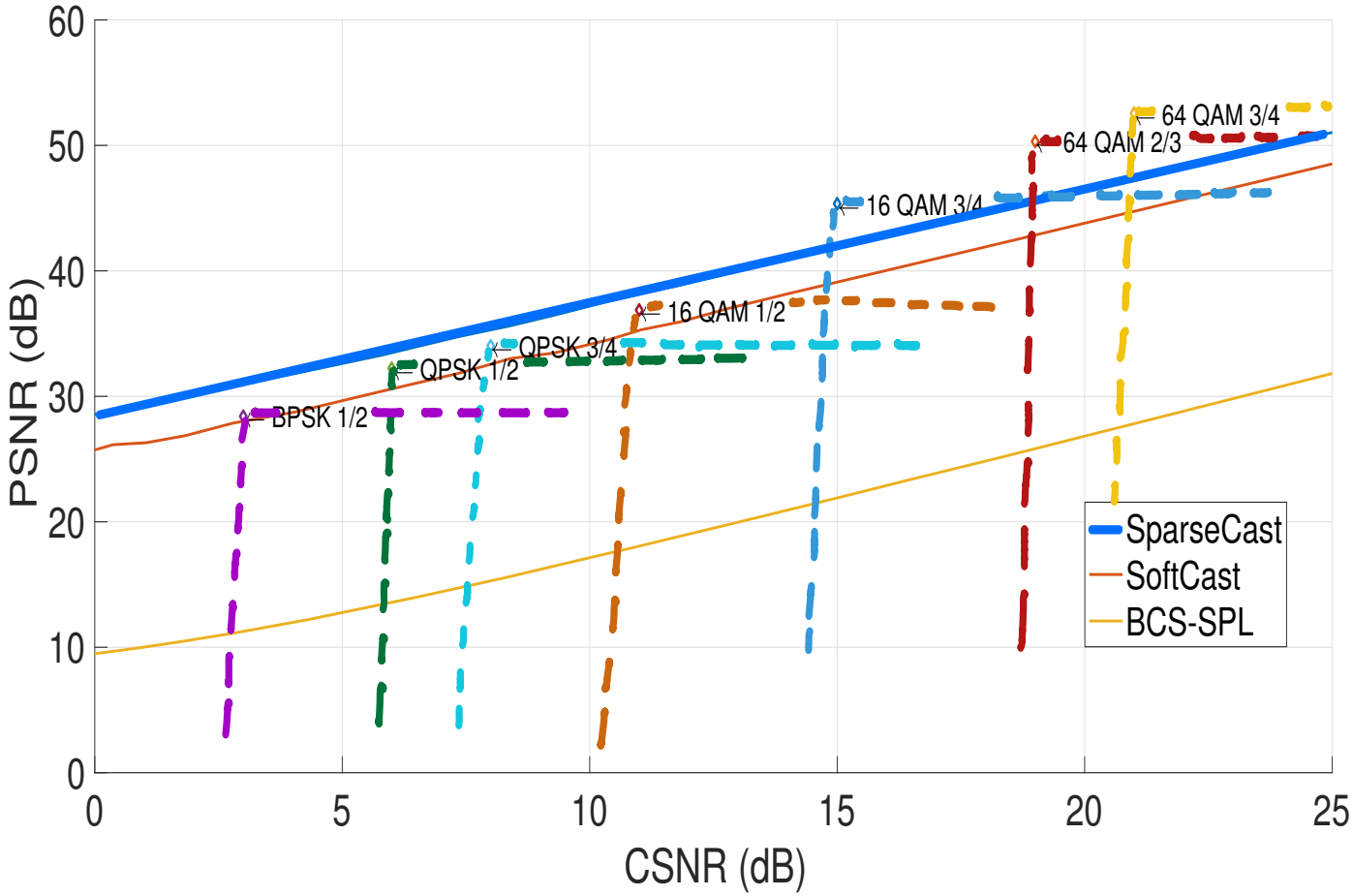


$$\mathbf{Y_k} = \mathbf{A_k x_k} + \mathbf{Z_k}$$

- $N \times N$ grayscale image
- $B \times B$ block DCT transform
- B^2 vectors (of length N^2/B^2 each)
- Thresholding for compression
- Compressive transmission: measurement matrix $\mathbf{A_k}$
  - dimension chosen according to sparsity of $\mathbf{x_k}$
  - finite set of sparsity levels
  - variance according to power allocation
- Approximate message passing (AMP) receiver

Tung and Gunduz, **SparseCast: Hybrid Digital-Analog Wireless Image Transmission Exploiting Frequency Domain Sparsity**, *IEEE Comm. Letters*, 2018.

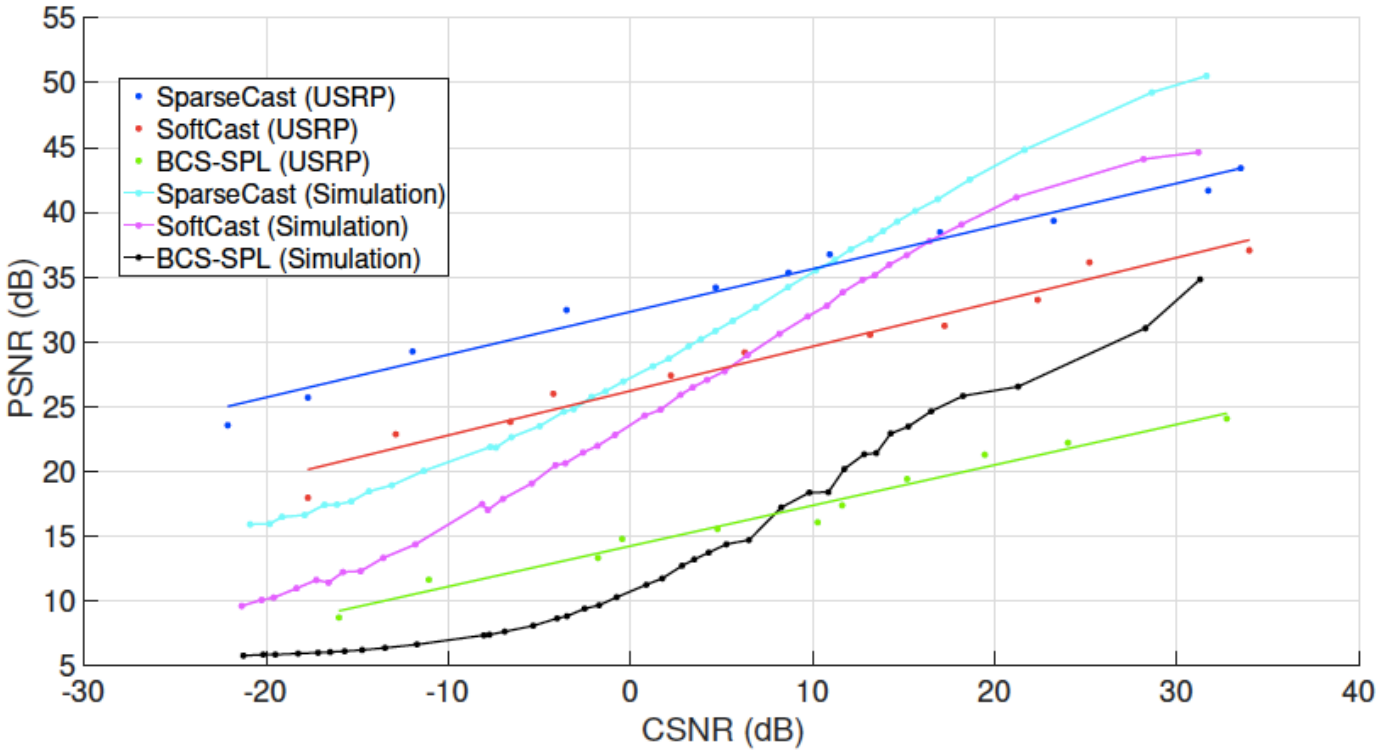
$131K$ channel symbols transmitted



Metadata size: SoftCast: 17 Kbits, SoftCast $10-16$ Kbits (depending on block threshold)

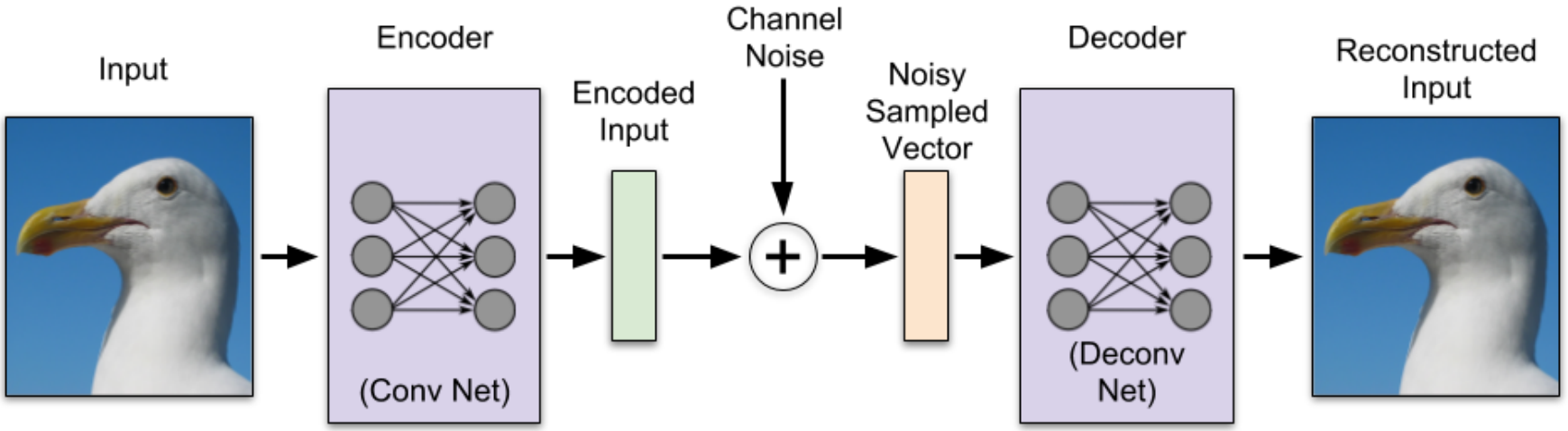
Tung and Gunduz, **SparseCast: Hybrid Digital-Analog Wireless Image Transmission Exploiting Frequency Domain Sparsity**, *IEEE Comm. Letters*, 2018.

$75K$ channel symbols transmitted

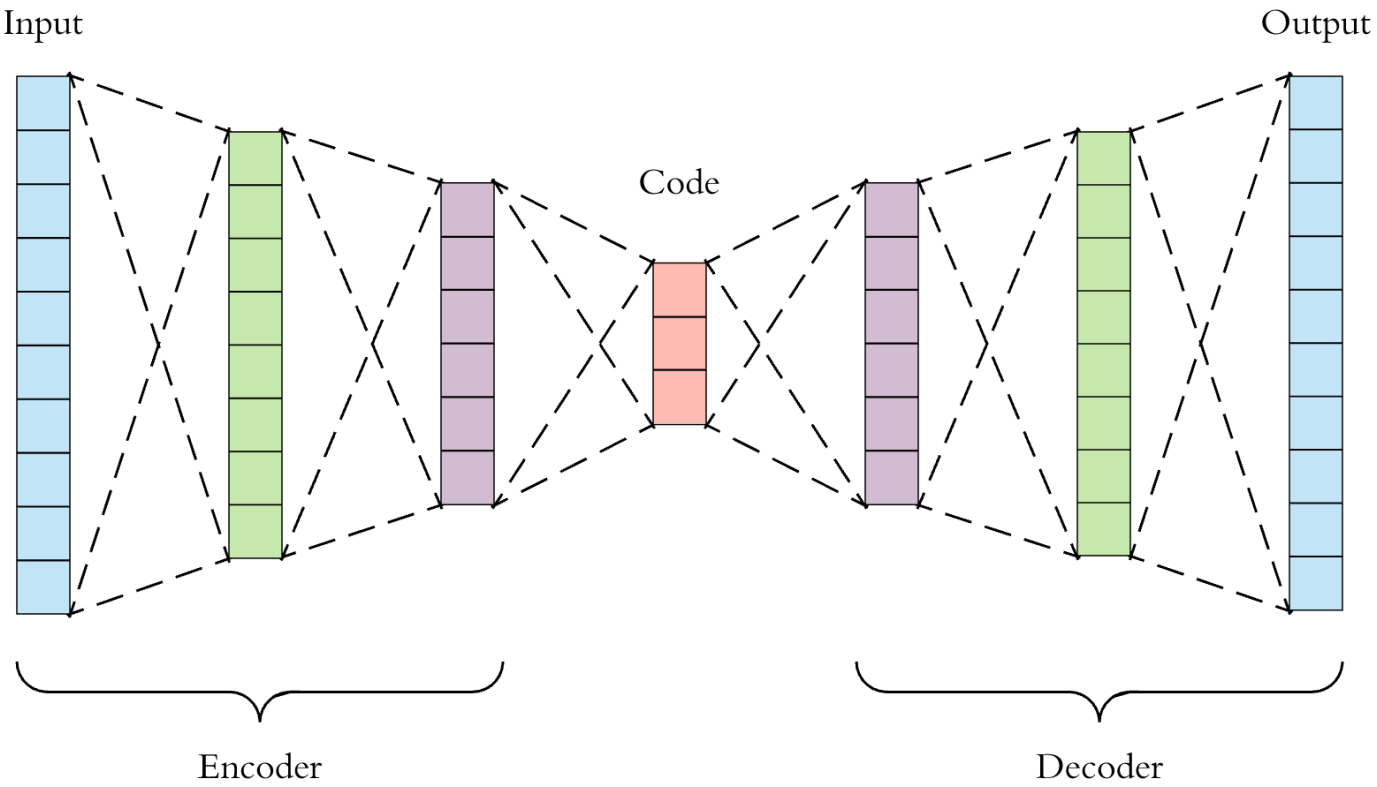


Tung and Gunduz, **SparseCast: Hybrid Digital-Analog Wireless Image Transmission Exploiting Frequency Domain Sparsity**, *IEEE Comm. Letters*, 2018.

- Forget about compression, channel coding, modulation, channel estimation, equalization, etc.
- Deep neural networks for code design

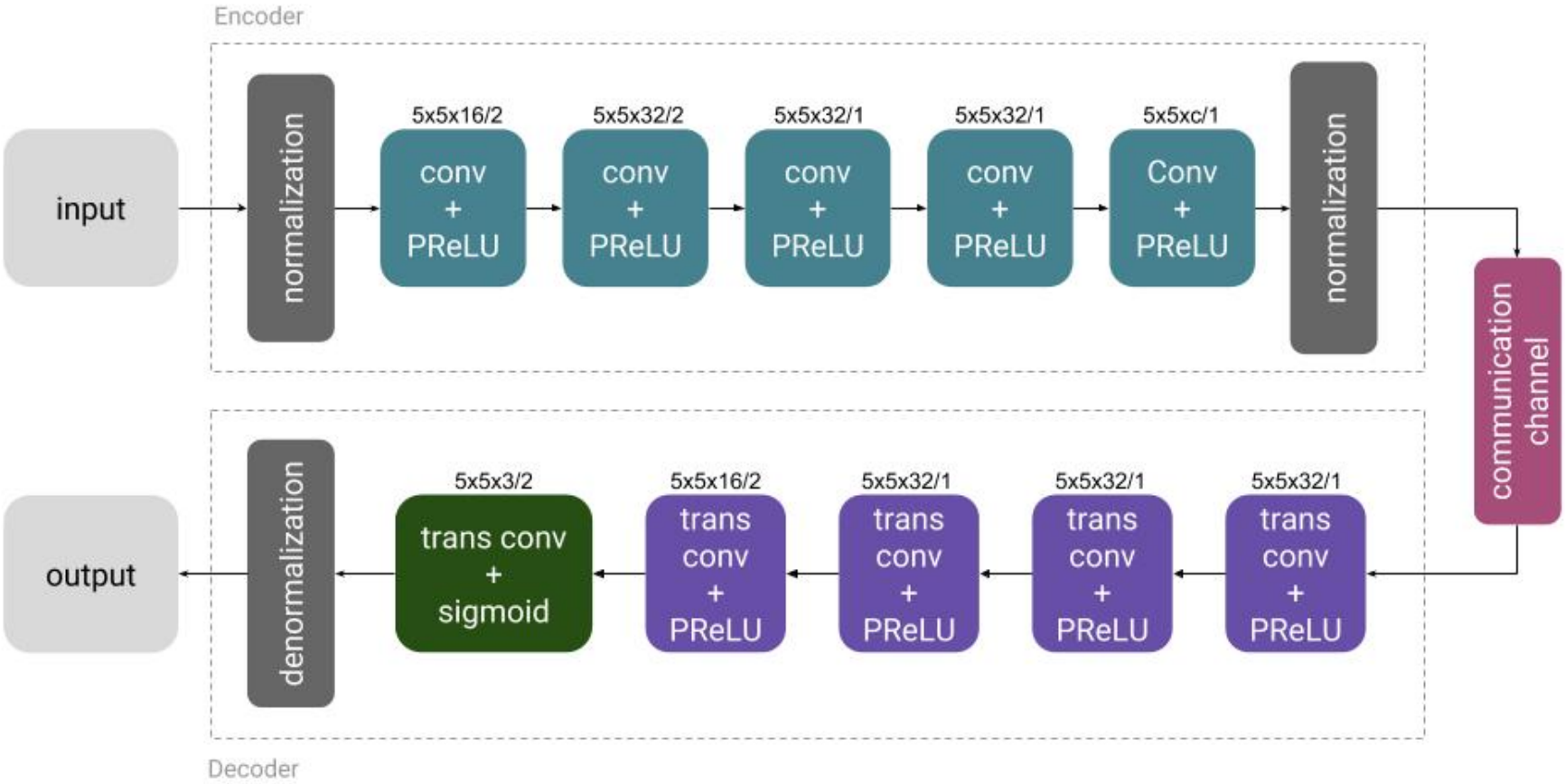


Autoencoder: Dimensionality Reduction with Neural Networks (NNs)



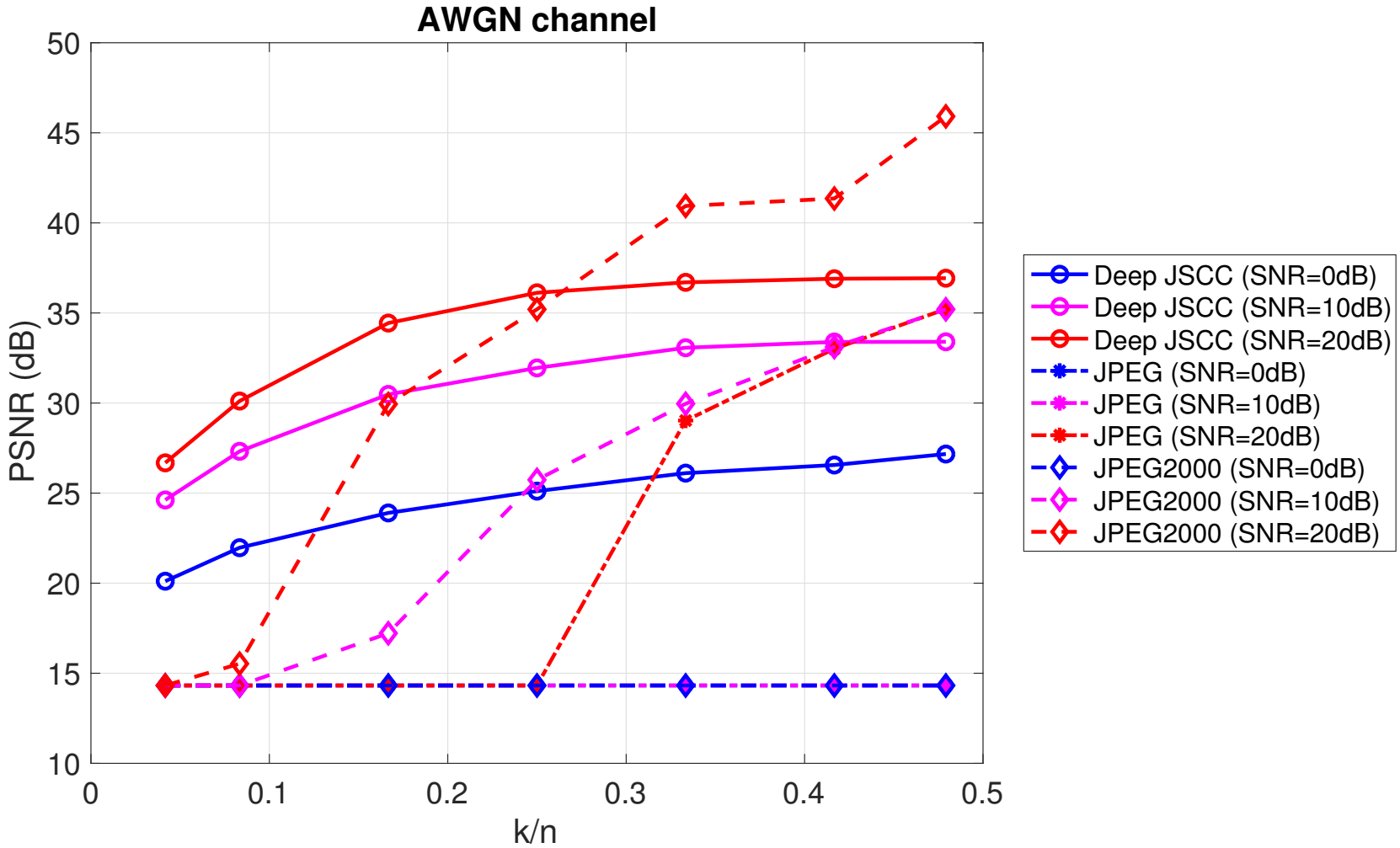
- Example of unsupervised learning
- Two NNs trained together: Goal is to reconstruct the original input with highest fidelity

Deep JSCC Architecture



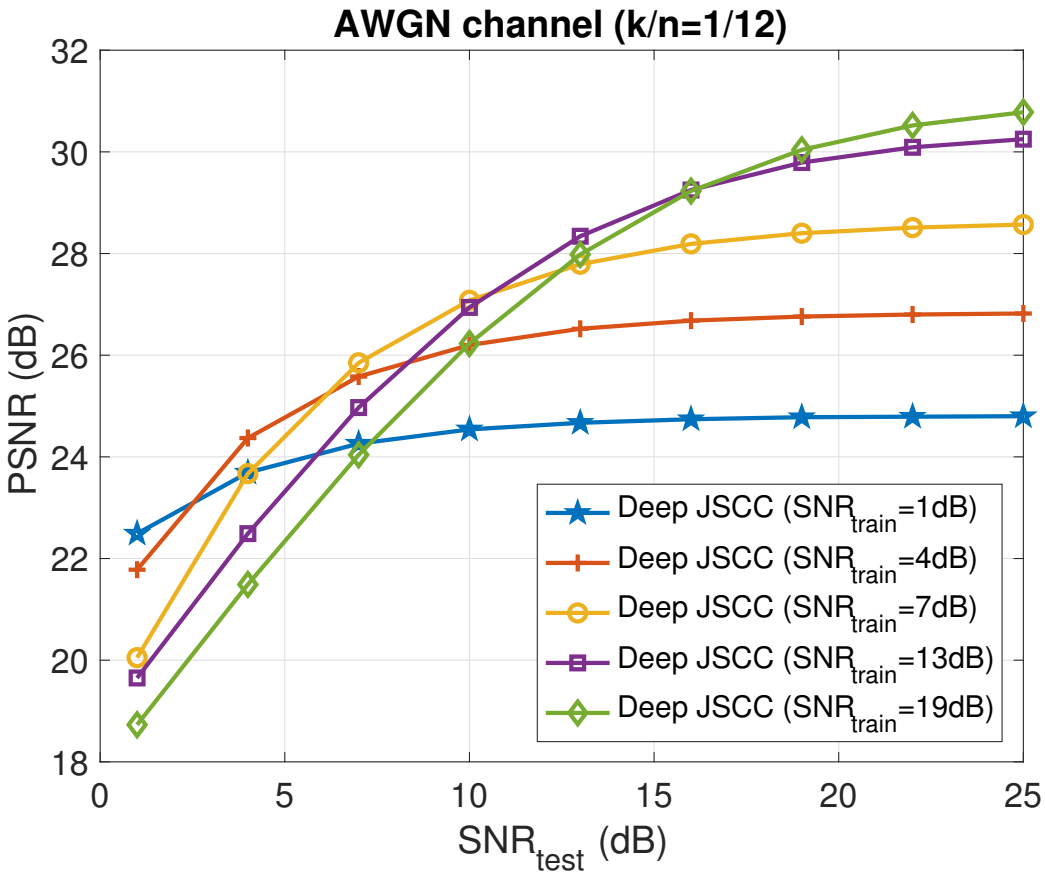
E. Bourtsoulatze, D. Burth Kurka and D. Gunduz, **Deep joint source-channel coding for wireless image transmission-journal**, submitted, IEEE TCCN, Sep. 2018.

Deep JSCC - PSNR vs. Channel Bandwidth



E. Bourtsoulatze, D. Burth Kurka and D. Gunduz, **Deep joint source-channel coding for wireless image transmission-journal**, submitted, IEEE TCCN, Sep. 2018.

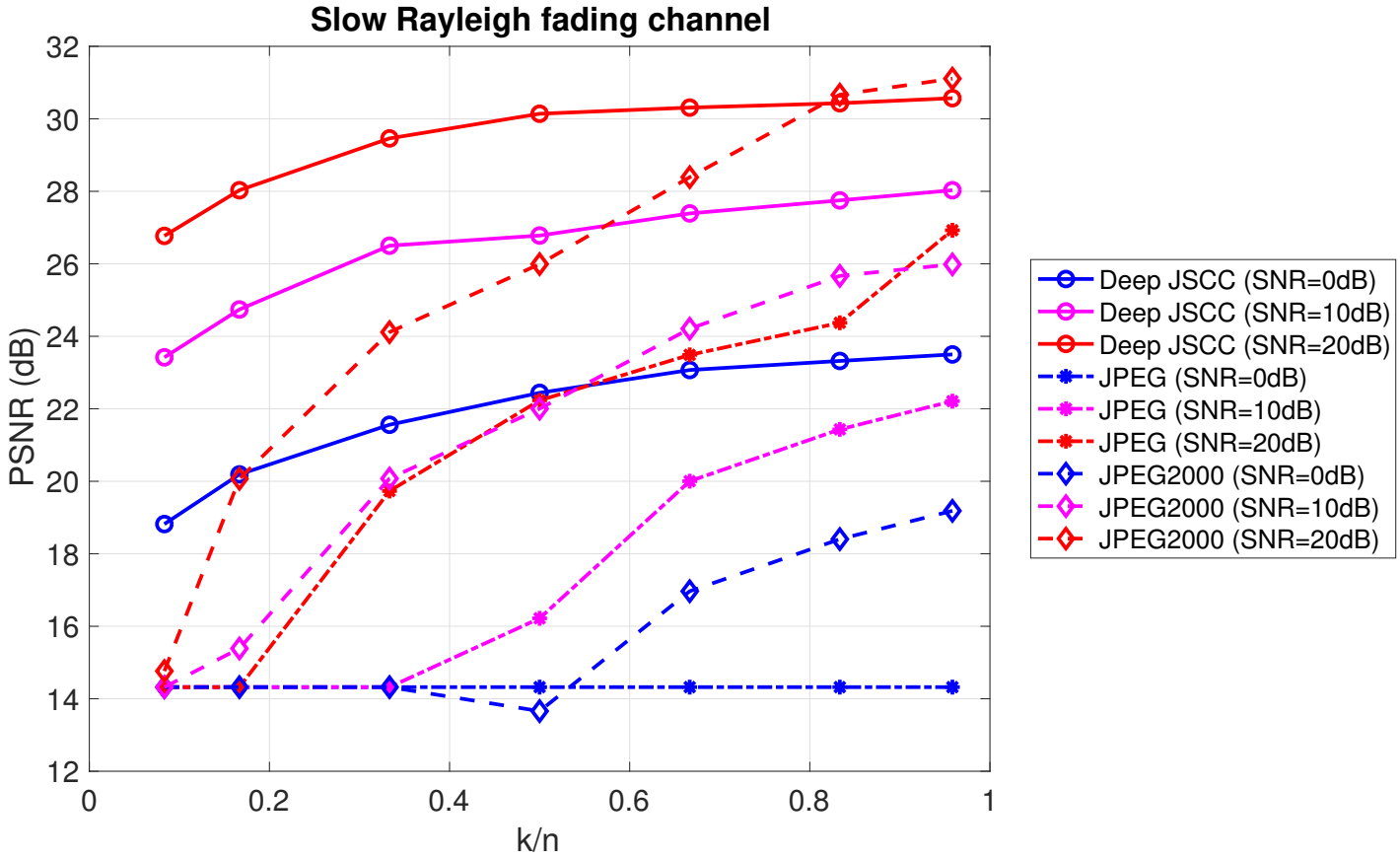
Deep JSCC - PSNR vs. Test SNR



- Provides graceful degradation with channel SNR!
- More like analog communications than digital.

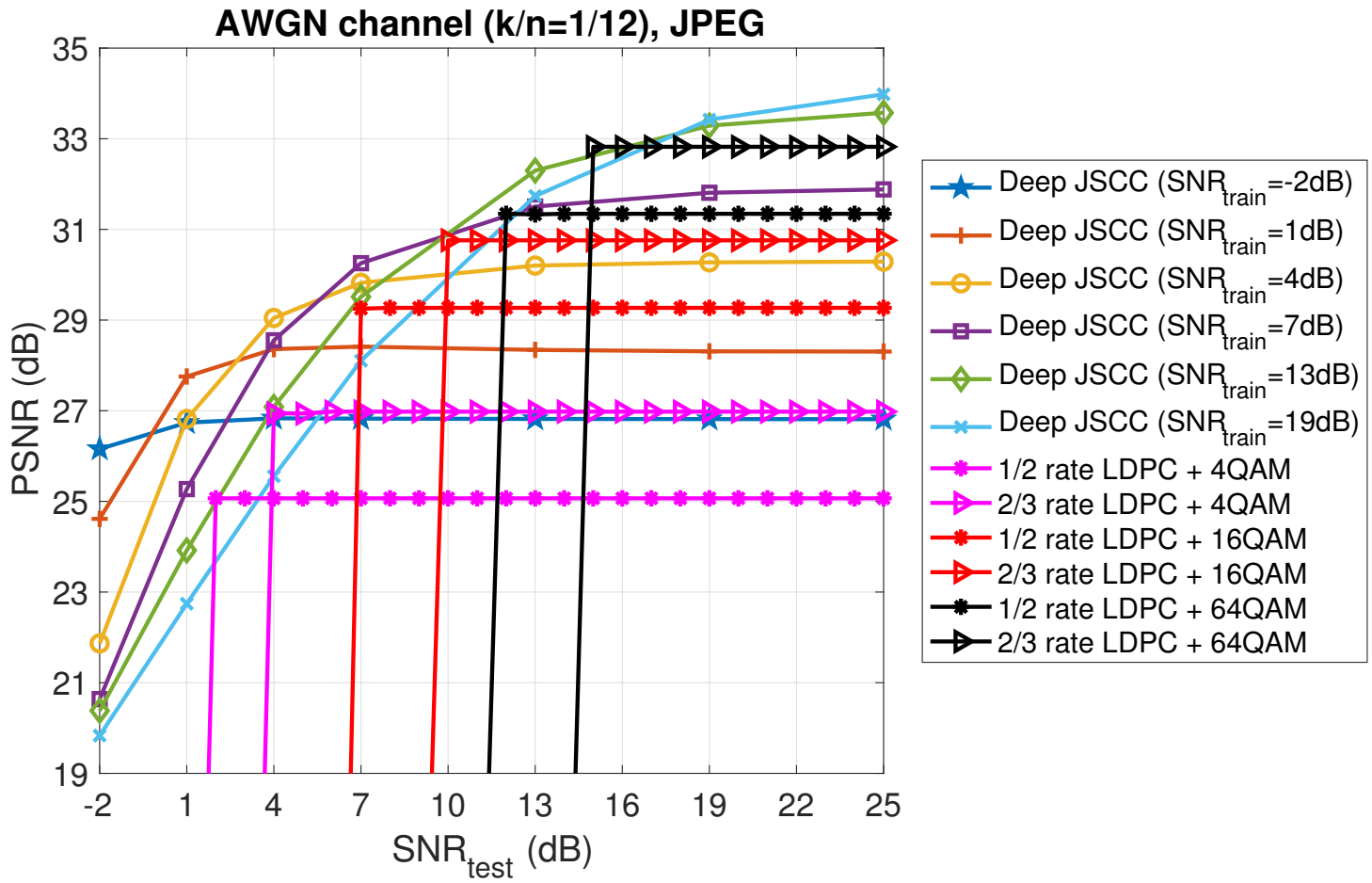
E. Bourtsoulatze, D. Burth Kurka and D. Gunduz, **Deep joint source-channel coding for wireless image transmission-journal**, submitted, IEEE TCCN, Sep. 2018.

Deep JSCC over a Rayleigh Fading Channel



No pilot signal or explicit channel estimation is needed!

E. Bourtsoulatze, D. Burth Kurka and D. Gunduz, **Deep joint source-channel coding for wireless image transmission-journal**, submitted, IEEE TCCN, Sep. 2018.



Train on ImageNet, test with Kodak dataset (24 images of size 768 x 512)

E. Bourtsoulatze, D. Burth Kurka and D. Gunduz, **Deep joint source-channel coding for wireless image transmission-journal**, submitted, IEEE TCCN, Sep. 2018.

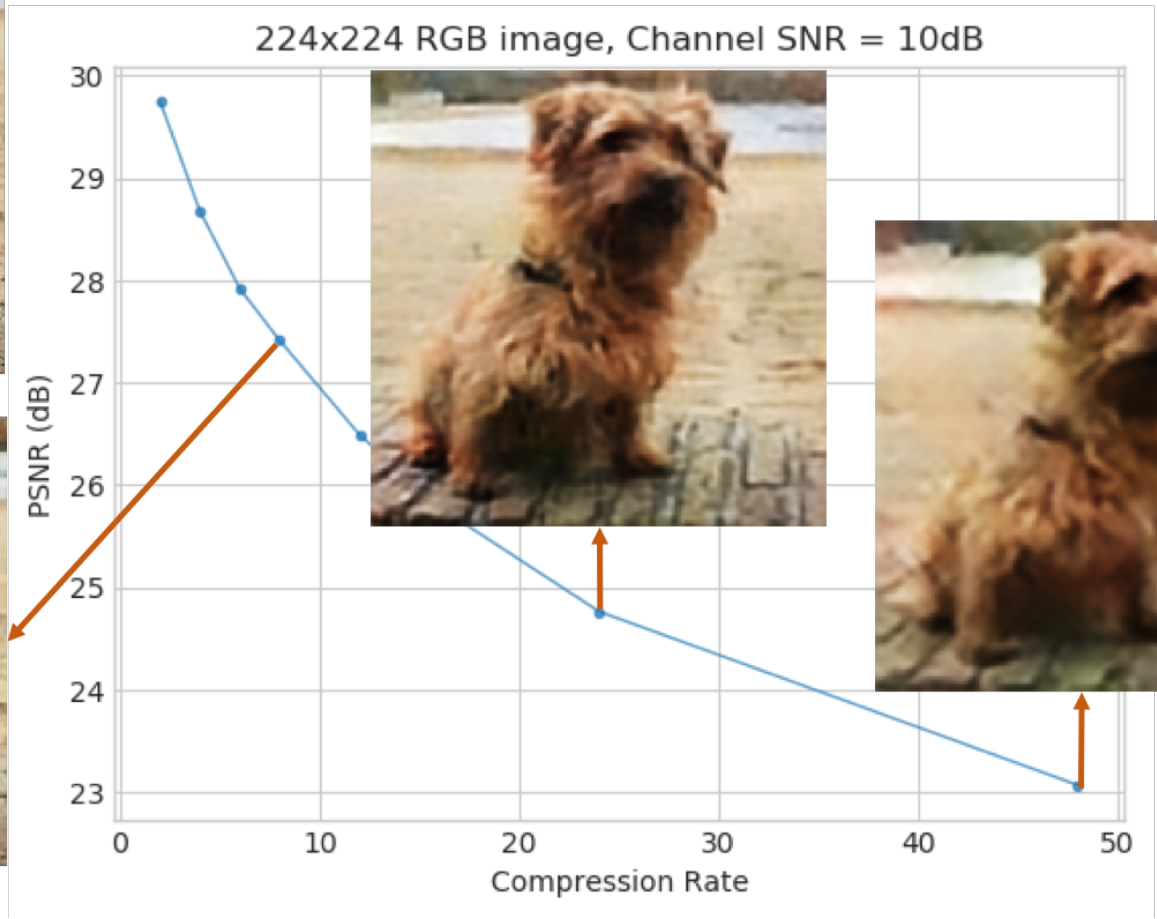
Larger Images

Original	Deep JSCC	JPEG	JPEG2000
			N/A
	30.9dB	22.68dB	
			
	31.92dB	31.65dB	36.40dB
			
	32.90dB	34.36dB	38.46dB
			
	35.34dB	36.45dB	40.5dB

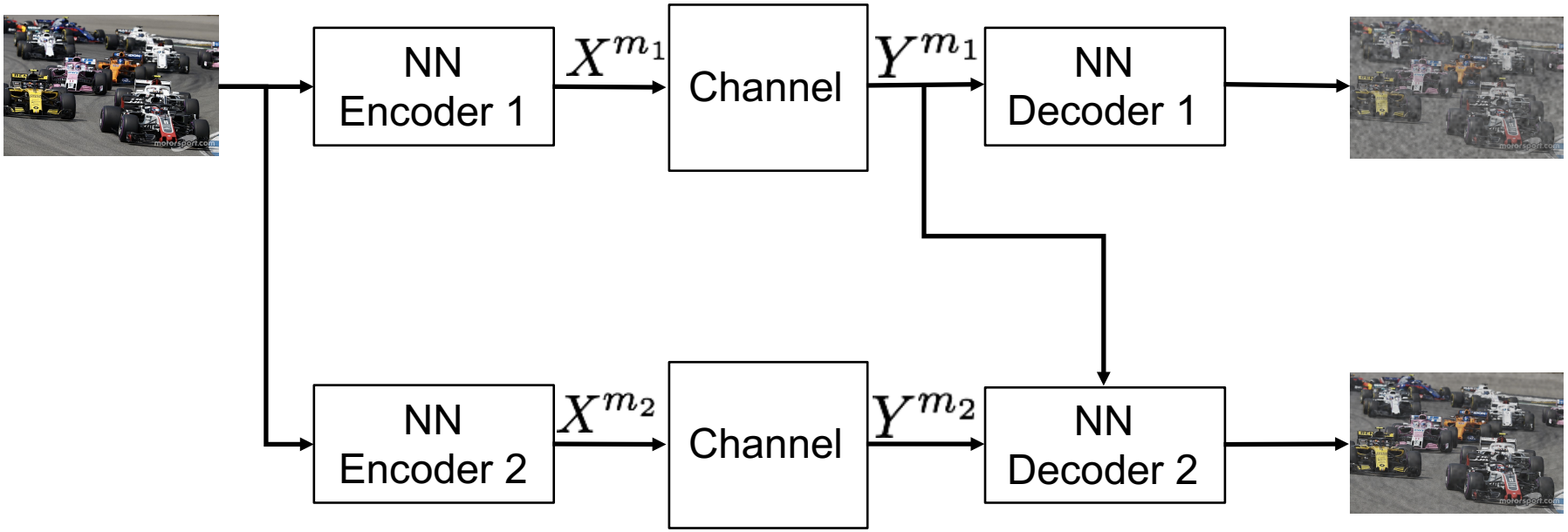
Larger Images

Original	Deep JSCC	JPEG	JPEG2000
			
	25.07dB	20.63dB	24.11dB
			
	26.86dB	24.78dB	27.5dB
			
	28.45dB	27.14dB	30.15dB
			
	31.46dB	29.81dB	33.03dB

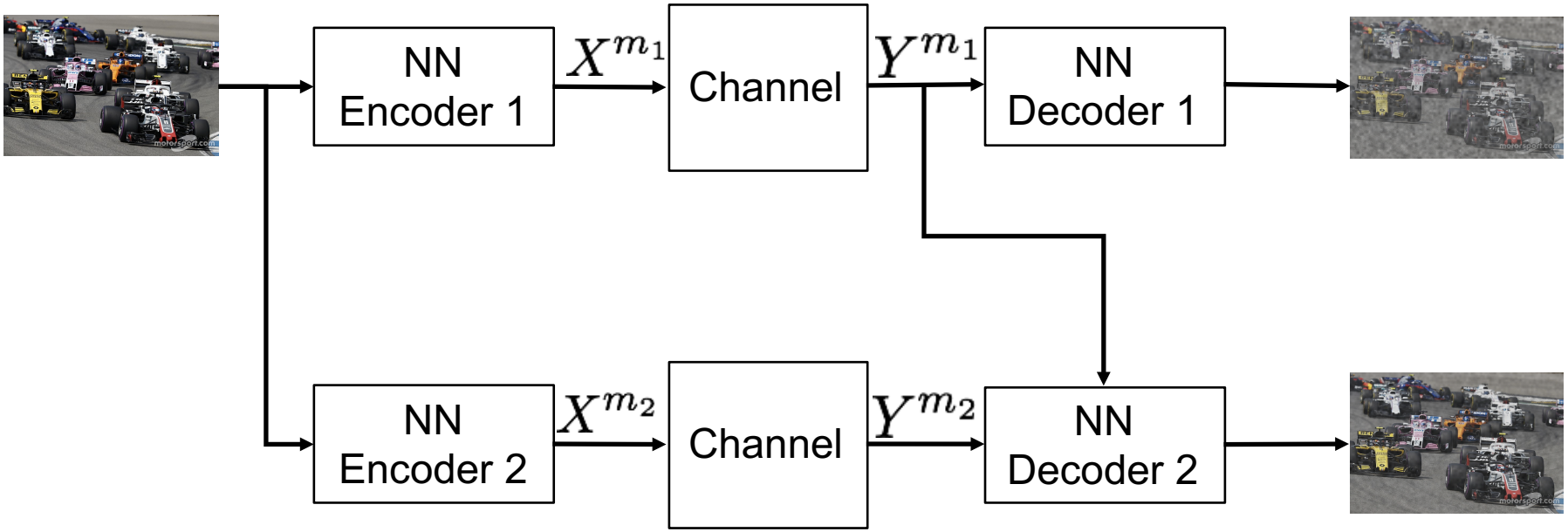
Quality vs. Compression Rate



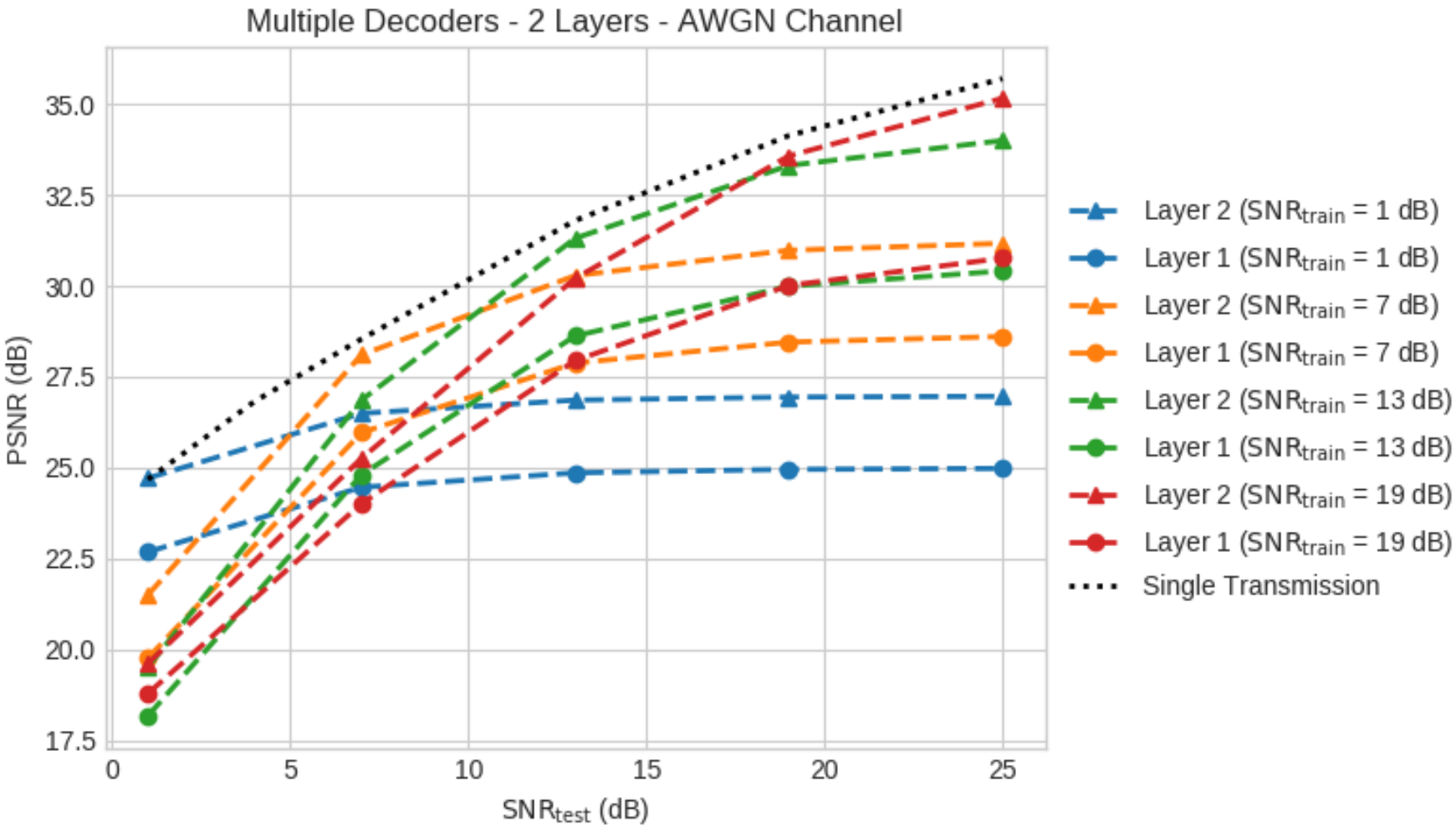
Deep Wireless Successive Refinement



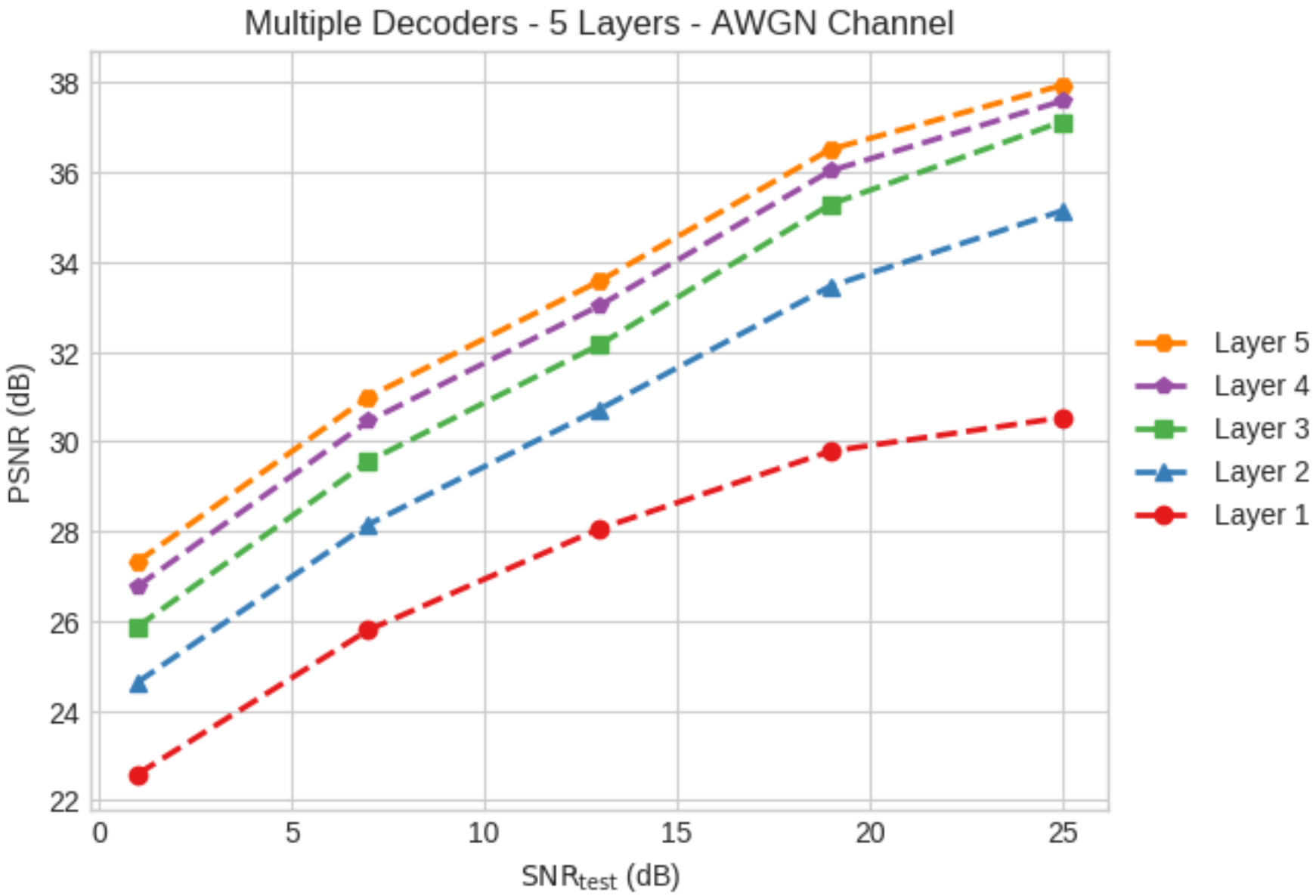
Deep Wireless Successive Refinement



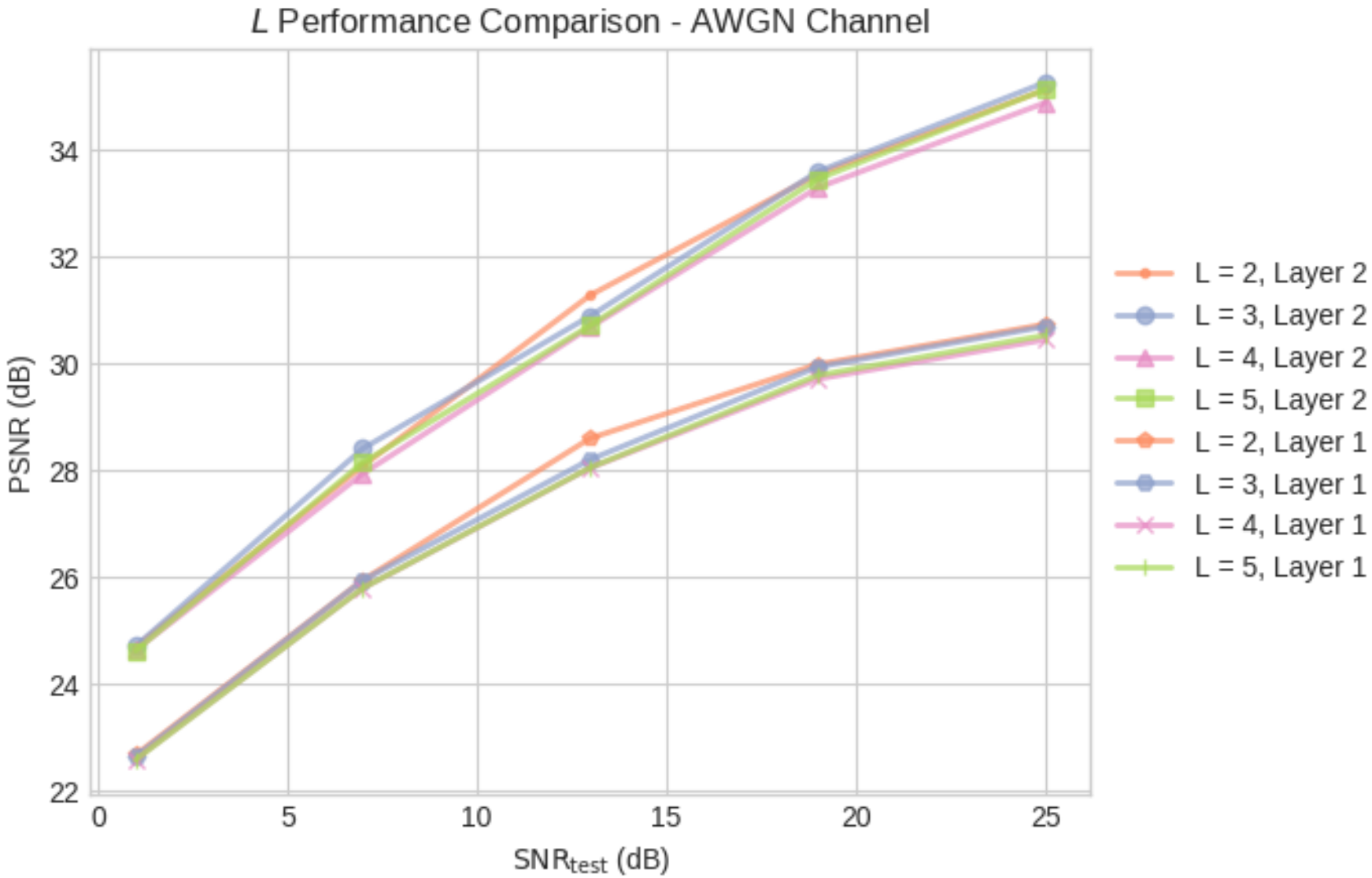
Two-layer Successive Refinement



Five-layer Successive Refinement



First Two Layer Comparison



Null hypothesis $H_0 : U^k \sim \prod_{i=1}^{k} P_U$, Alternate hypothesis $H_1 : U^k \sim \prod_{i=1}^{k} Q_U$.

Acceptance region for H_0: $\mathcal{A}^{(n)} \subseteq \mathcal{Y}^n$

Definition

Type-2 error exponent κ is (τ, ϵ) **achievable** if there exist k, n, such that $n \leq \tau \cdot k$, and

$$\liminf_{k,n\to\infty} -\frac{1}{k} \log\left(Q_{Y^n}(\mathcal{A}^{(n)})\right) \geq \kappa$$

$$\limsup_{k,n\to\infty} -\frac{1}{k} \log\left(1 - P_{Y^n}(\mathcal{A}^{(n)})\right) \leq \epsilon$$

$$\kappa(\tau, \epsilon) \triangleq \sup\{\kappa' : \kappa' \text{ is achievable}\}$$

Null hypothesis $H_0 : U^k \sim \prod_{i=1}^{k} P_U$, Alternate hypothesis $H_1 : U^k \sim \prod_{i=1}^{k} Q_U$.

$$E_c \triangleq \max_{(x,x')\in\mathcal{X}\times\mathcal{X}} D(P_{Y|X=x}||P_{Y|X=x'})$$

$$\kappa(\tau,\epsilon) = \min\left(D(P_U||Q_U), \tau E_c\right)$$

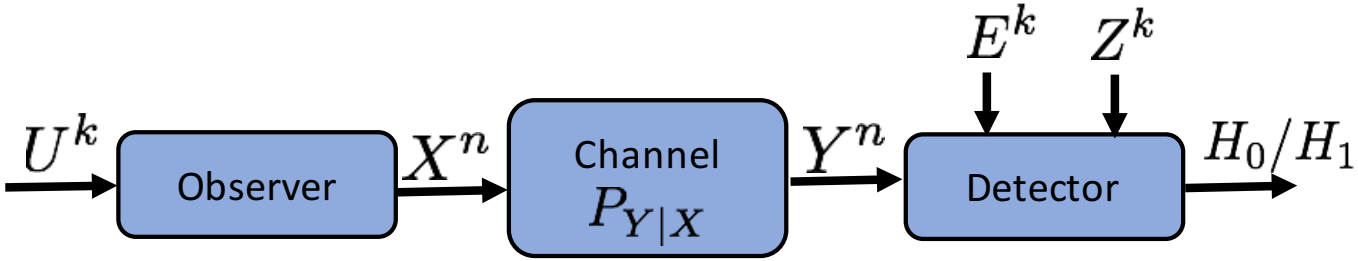
Making decisions locally at the observer, and communicating it to the detector is optimal.

Null hypothesis $H_0 : U^k \sim \prod_{i=1}^{k} P_U$, Alternate hypothesis $H_1 : U^k \sim \prod_{i=1}^{k} Q_U$.

$$E_c \triangleq \max_{(x,x')\in\mathcal{X}\times\mathcal{X}} D(P_{Y|X=x}||P_{Y|X=x'})$$

$$\kappa(\tau,\epsilon) = \min\left(D(P_U||Q_U), \tau E_c\right)$$

Making decisions locally at the observer, and communicating it to the detector is optimal.



$$H_0 : (U^k, E^K, Z^K) \sim \prod_{i=1}^{k} P_{UEZ}, \qquad H_1 : (U^k, E^K, Z^K) \sim \prod_{i=1}^{k} Q_{UEZ}.$$

- Problem open for general Q
- Let $\kappa(\tau) = \lim_{\epsilon \to 0} \kappa(\tau, \epsilon)$

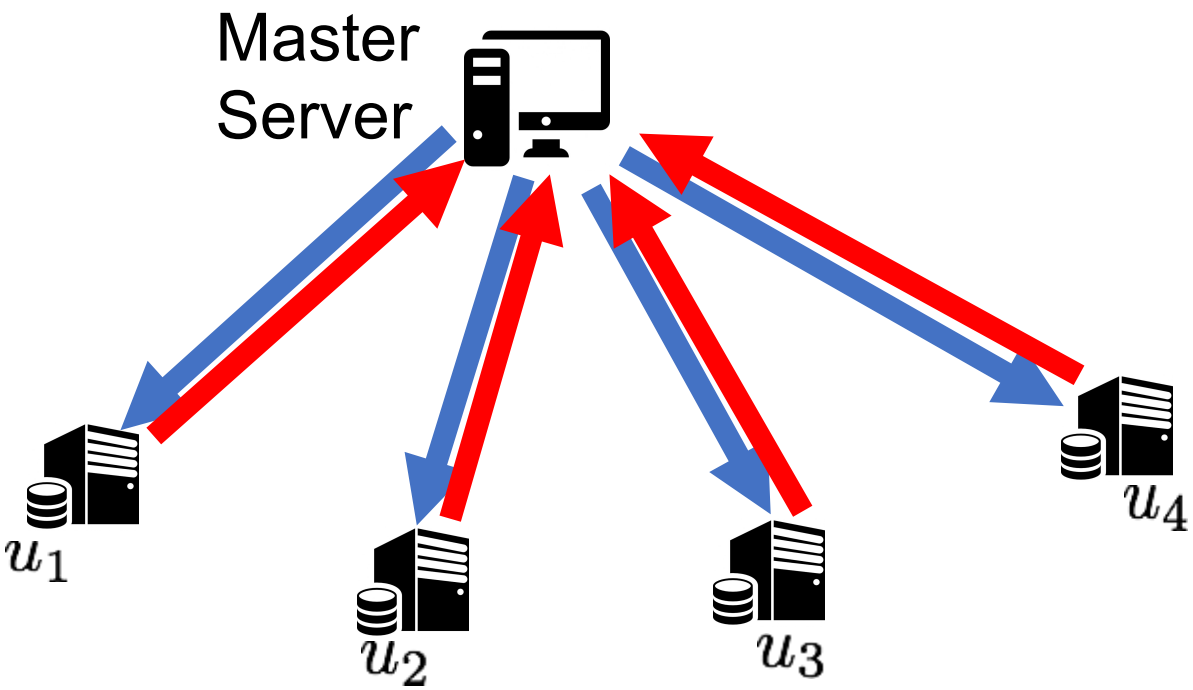
Testing Against Conditional Independence: $Q_{UEZ} = P_{UE} P_{E|Z}$

$$\kappa(\tau) = \sup \left\{ \begin{array}{c} I(E;W|Z) : \exists\ W \text{ s.t. } I(U;W|Z) \le \tau C(P_{Y|X}), \\ (Z,E) - U - W,\ |\mathcal{W}| \le |\mathcal{U}| + 1. \end{array} \right\}, \ \tau \ge 0.$$

Optimal performance achieved by a separation-based scheme.

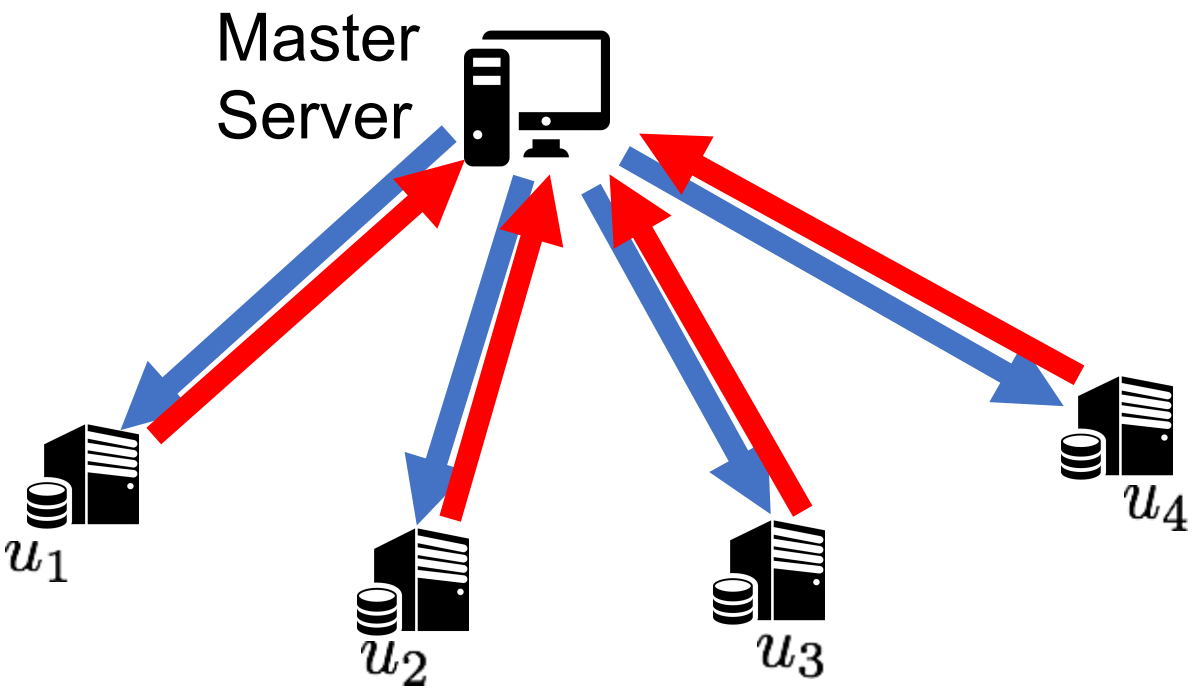
Machine Learning (ML) at the Edge

- Significant amount of data will be collected by IoT devices at network edge
- Standard approach: Powerful centralized ML algorithms to make sense of data
- Requires sending data to the cloud
  - Costy in terms of bandwidth/ energy
  - May conflict with privacy requirements
- Alternative: distributed/ federated learning



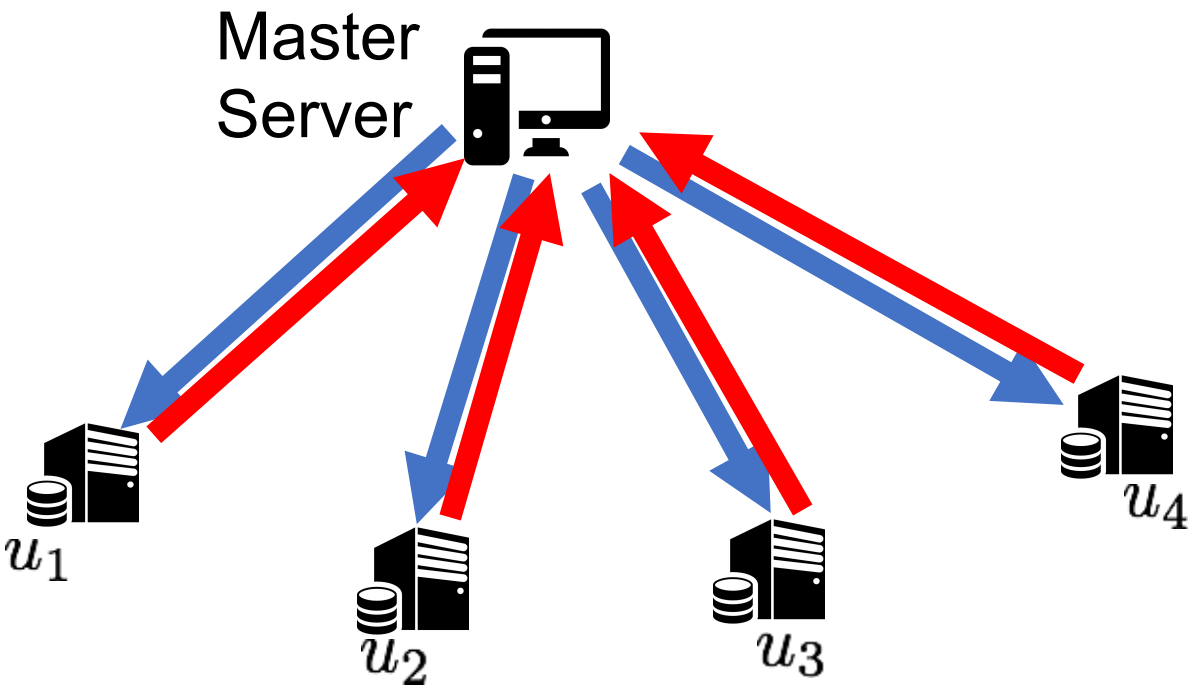
Machine Learning (ML) at the Edge

- Significant amount of data will be collected by IoT devices at network edge
- Standard approach: Powerful centralized ML algorithms to make sense of data
- Requires sending data to the cloud
  - Costy in terms of bandwidth/ energy
  - May conflict with privacy requirements
- Alternative: distributed/ federated learning



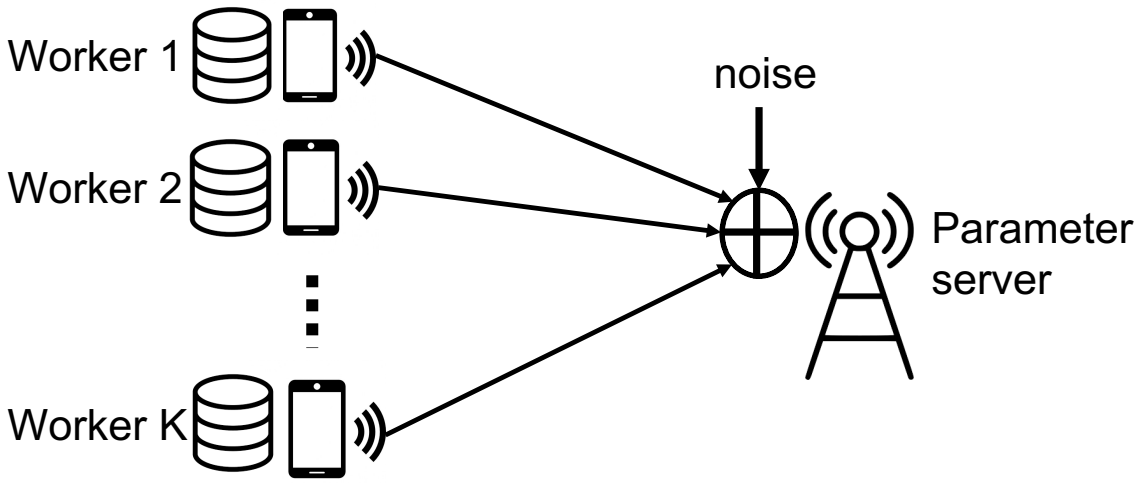
Data set: $(\mathbf{u_1}, y_1), \ldots, (\mathbf{u_N}, y_N)$

$$F(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} f(\boldsymbol{\theta}, \mathbf{u_n})$$



$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \frac{1}{N} \sum_{n=1}^{N} \nabla f(\boldsymbol{\theta}_t, \mathbf{u_n})$$

- Communication is bottleneck in distributed learning
- ML literature focuses on reducing the number and size of gradient informaton transmitted from each worker
- Underlying channel ignored
- In edge learning, wireless channel is limited in bandwidth and may suffer from interference



- Workers operate on capacity boundary of underlying MAC
  - Choose equal-rate point
  - Allow power allocation across iterations
  - For s channel uses

$$R_t = \frac{s}{2M} \log_2 \left(1 + \frac{MP_t}{s\sigma^2}\right),$$

- Each worker has a bit budget to convey its gradient estimate
- Gradient quantization
  - Set all but highest q and lowest q entries of gradient estimate to 0
  - Find mean values of all positive and all negative entries
  - Find the one with the larger magnitude, and set the others to zero
  - Send the larger value, and positions of corresponding entries
- Employ error accumulation

F. Sattler et al. **Sparse binary compression: Towards distributed deep learning with minimal communication**, arXiv:1805.08768v1 [cs.LG], May 2018.

F. Seide et al. **1-bit stochastic gradientdescent and its application to data-parallel distributed training of speech DNNs**, in INTERSPEECH, Singapore, Sep. 2014.

Analog Distributed Gradient Descent

- A distributed joint source-channel coding problem
- Goal: Compute the average of the sources
- Simultaneously transmit gradients in an uncoded fashion: over-the-air computation
- Challenge:
  - Gradient dimension can be very large: VGG Net ∼140 million, ResNet ∼26 million parameters
  - Introduces significant delay
- Proposed scheme:
  - Apply thresholding to sparsify gradient estimates
  - CS-based JSCC: Project onto a lower dimensional space (same projection matrix at all edge devices)

M. Mohammadi Amiri and D. Gunduz, Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air, submitted, Jan. 2019.

Analog Distributed Gradient Descent

- A distributed joint source-channel coding problem
- Goal: Compute the average of the sources
- Simultaneously transmit gradients in an uncoded fashion: over-the-air computation
- Challenge:
  - Gradient dimension can be very large: VGG Net ∼140 million, ResNet ∼26 million parameters
  - Introduces significant delay
- Proposed scheme:
  - Apply thresholding to sparsify gradient estimates
  - CS-based JSCC: Project onto a lower dimensional space (same projection matrix at all edge devices)

M. Mohammadi Amiri and D. Gunduz, Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air, submitted, Jan. 2019.

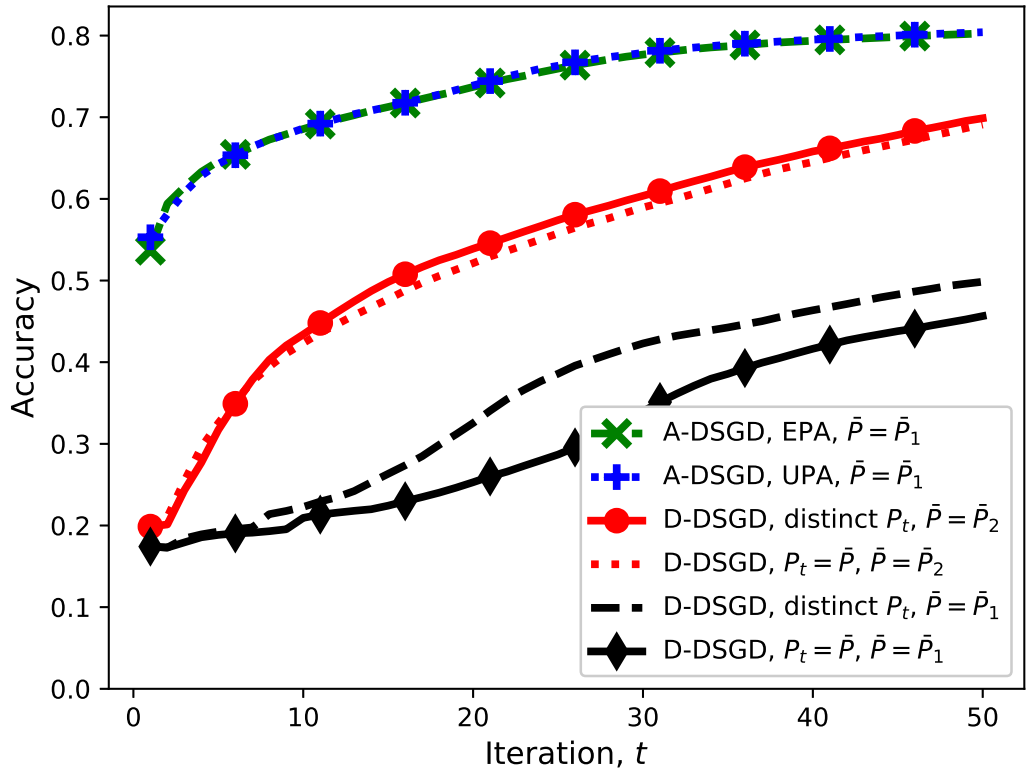
Analog Distributed Gradient Descent

- A distributed joint source-channel coding problem
- Goal: Compute the average of the sources
- Simultaneously transmit gradients in an uncoded fashion: over-the-air computation
- Challenge:
  - Gradient dimension can be very large: VGG Net ∼140 million, ResNet ∼26 million parameters
  - Introduces significant delay
- Proposed scheme:
  - Apply thresholding to sparsify gradient estimates
  - CS-based JSCC: Project onto a lower dimensional space (same projection matrix at all edge devices)

M. Mohammadi Amiri and D. Gunduz, Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air, submitted, Jan. 2019.

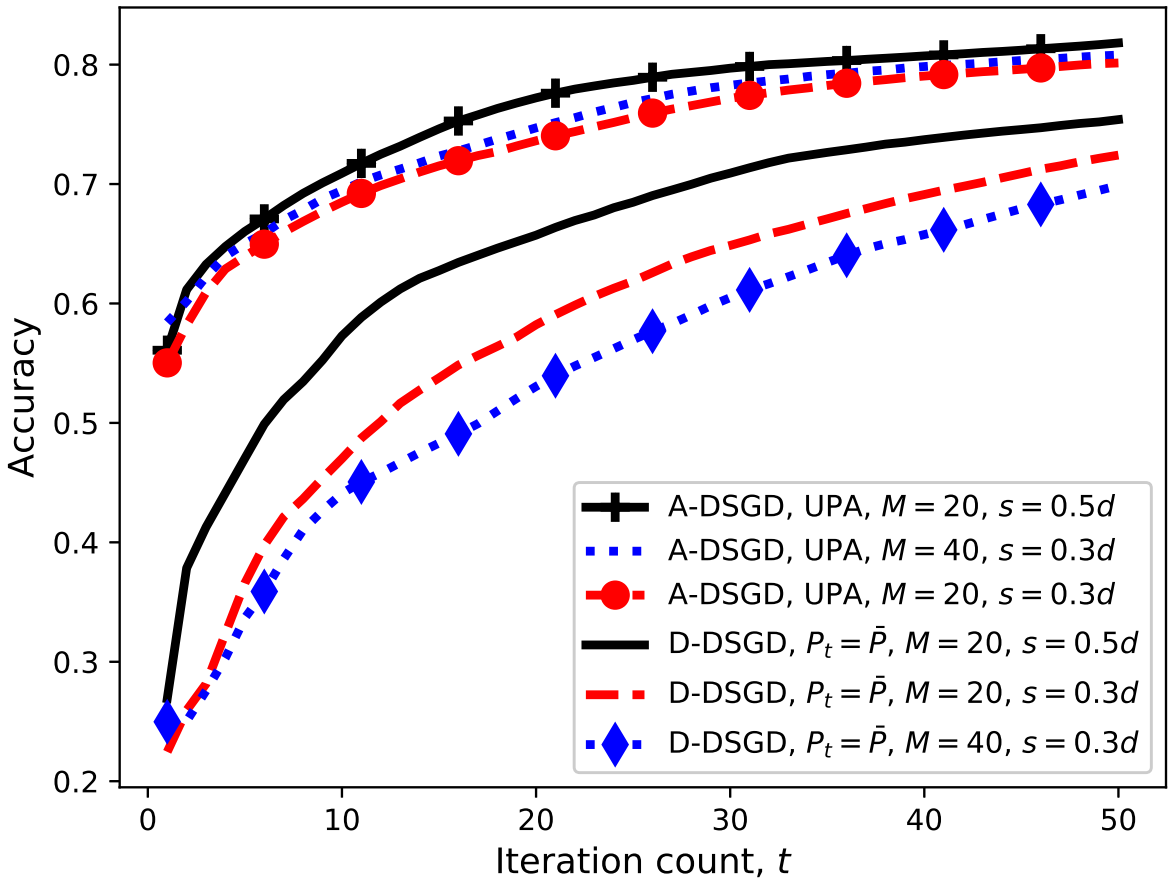
Experiments: Digital vs. Analog Gradient Descent

- Distributed MNIST classification (single layer with 10 neurons, ADAM optimizer)
- Parameter vector size $d = 28 \times 28 \times 10 + 10 = 7850$
- $P_1 = 127$, $P_2 = 422$

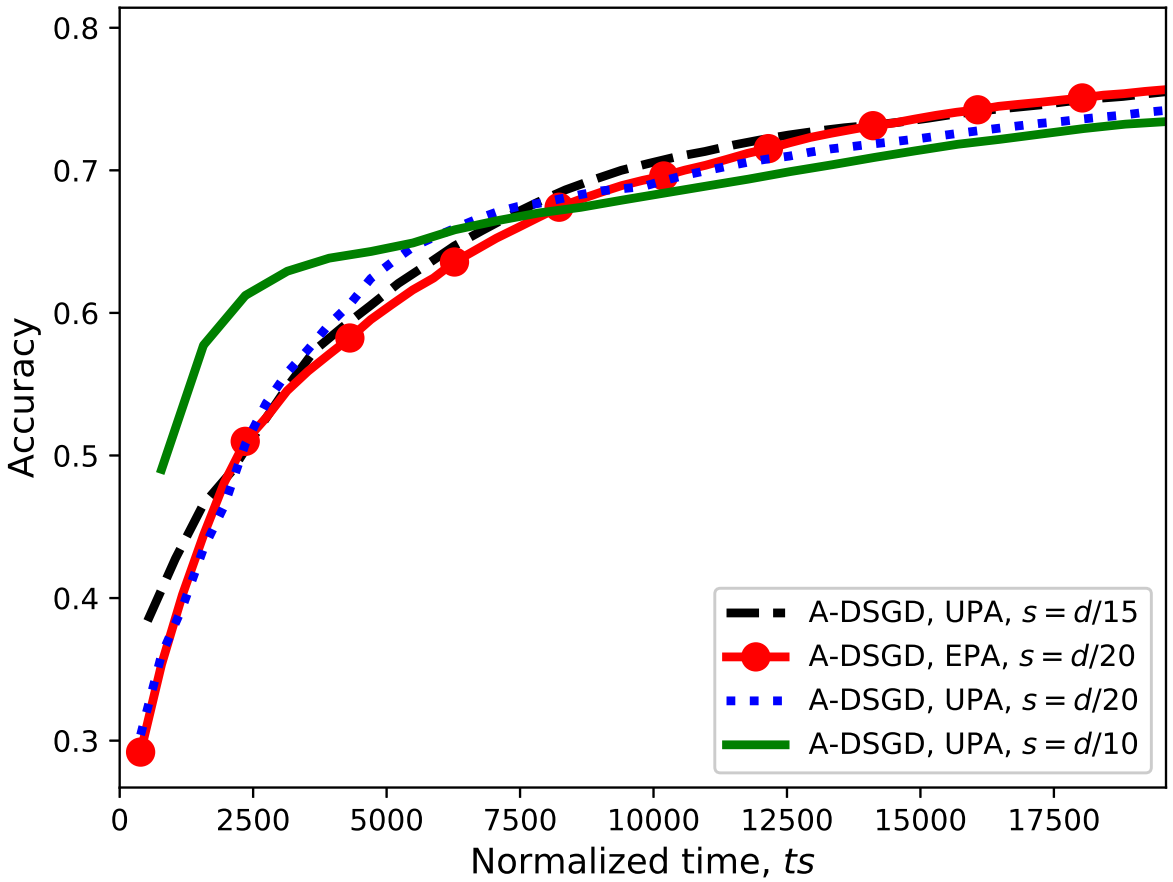


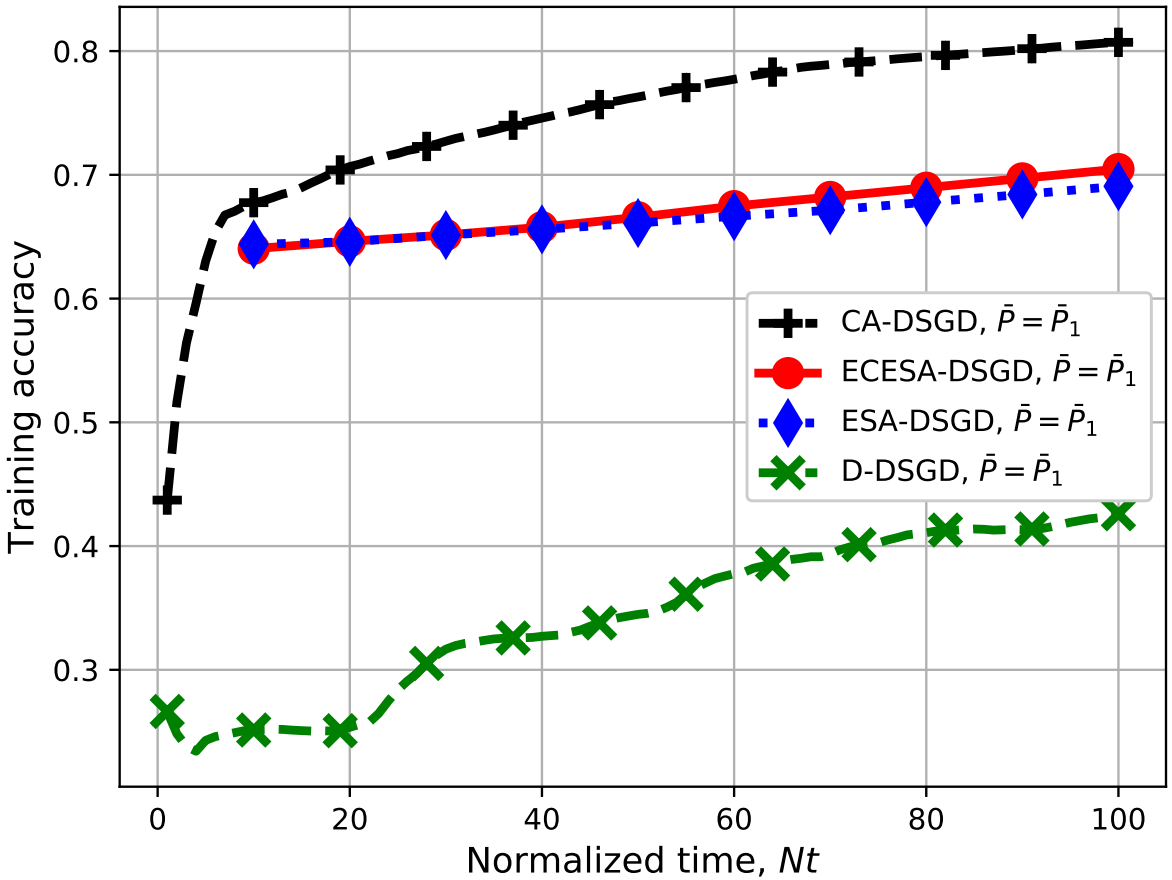
Experiments: Number of Devices

- d: dimension of parameter vector
- s: symbols per iteration
- M: number of devices



- d: dimension of parameter vector
- s: symbols per iteration




0.8
0.7
0.6
0.5
0.4
0.3
Training accuracy
0
20
40
60
80
100
Normalized time, Nt
CA-DSGD, $\bar{P} = \bar{P}_1$
ECESA-DSGD, $\bar{P} = \bar{P}_1$
ESA-DSGD, $\bar{P} = \bar{P}_1$
D-DSGD, $\bar{P} = \bar{P}_1$

Conclusions

- JSCC is a fundamental problem in information theory with many applications
- Becoming essential for modern communication systems with extremely low latency and low power requirements
- Machine learning tools can help us design practical joint source-channel codes that can beat state-of-the-art
- Distributed wireless learning can benefit from JSCC for over-the-air computation