# Learned Task-Aware Compression Methods in Communication Systems

Fabrizio Carpi

Co-Advisors: Prof. Elza Erkip and Prof. Siddharth Garg

NYU TANDON SCHOOL OF ENGINEERING

NYU WIRELESS

Ph.D. Defense – August 15, 2024

# Motivation

Shannon–Weaver identified three levels of communication problems [1]:

1. **Technical problem** $\Longrightarrow$ bits

[1] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949.

# Motivation

Shannon–Weaver identified three levels of communication problems [1]:

1. **Technical problem** $\Longrightarrow$ bits
2. Semantic problem $\Longrightarrow$ bits + source
3. Effectiveness problem $\Longrightarrow$ bits + source + task

[1] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949.

# Motivation

Shannon–Weaver identified three levels of communication problems [1]:

1. **Technical problem** $\Longrightarrow$ bits
2. Semantic problem $\Longrightarrow$ bits + source
3. Effectiveness problem $\Longrightarrow$ bits + source + task

*Semantic/effectiveness* paradigm $\Rightarrow$ specialized network tied to the application [2]
PHY + upper layers

[1] C. E. Shannon and W. Weaver, The Mathematical Theory of Communication. Urbana, IL: University of Illinois Press, 1949.
[2] H. Xie, Z. Qin, G. Y. Li and B. -H. Juang, "Deep Learning Enabled Semantic Communication Systems," in IEEE TSP, 2021

# Task-Aware Compression in Communications Systems

In this thesis:

- Identify communications strategies that can be redefined in a task-aware fashion
  $\Rightarrow$ reusable in general-purpose networks: focus on PHY

# Task-Aware Compression in Communications Systems

In this thesis:

- Identify communications strategies that can be redefined in a task-aware fashion $\Rightarrow$ reusable in general-purpose networks: focus on PHY
- Focus on source coding problems: instead of minimizing distortion, optimize the end-to-end task

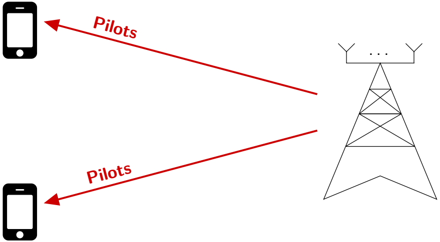
# Task-Aware Compression in Communications Systems

In this thesis:

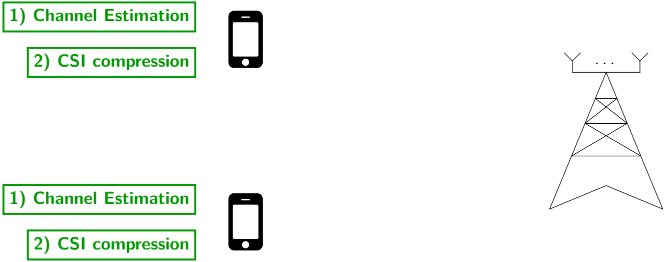
- Identify communications strategies that can be redefined in a task-aware fashion $\Rightarrow$ reusable in general-purpose networks: focus on PHY
- Focus on source coding problems: instead of minimizing distortion, optimize the end-to-end task
- **Task-aware co-design** of the *compressor* (TX) and the *decoder* (RX)
  1. Channel state information (CSI) feedback
  2. Compress-and-Forward (CF) relaying

# Outline

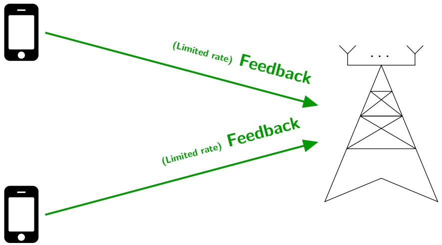
# Conventional CSI Feedback in FDD MIMO



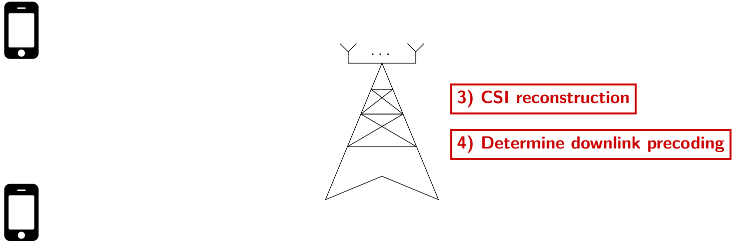
# Conventional CSI Feedback in FDD MIMO



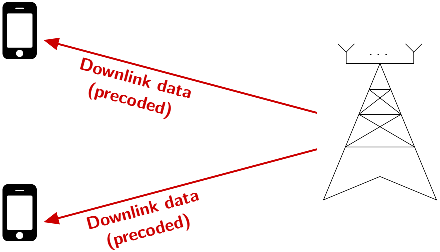
# Conventional CSI Feedback in FDD MIMO



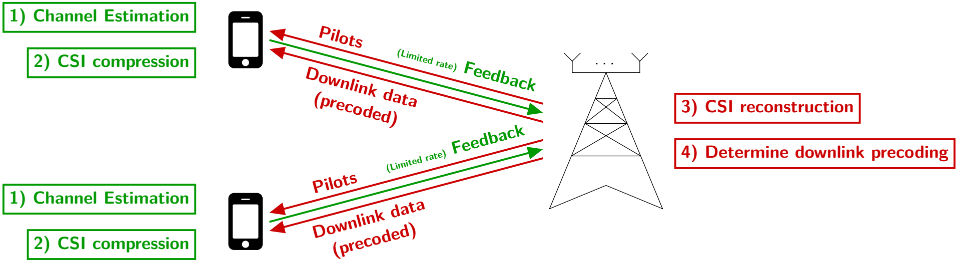
# Conventional CSI Feedback in FDD MIMO



# Conventional CSI Feedback in FDD MIMO



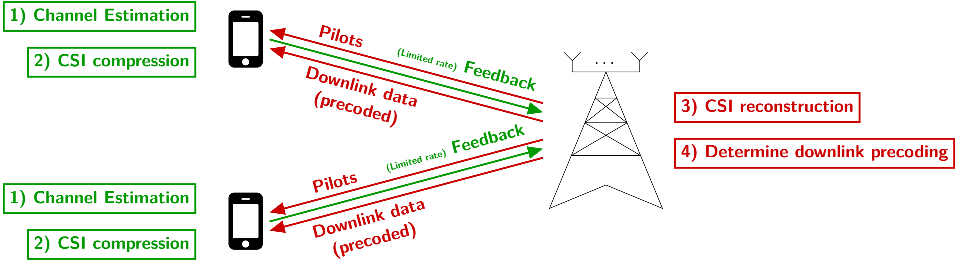
# Conventional CSI Feedback in FDD MIMO




- Conventional CSI compression: vector quantization, compressed sensing [1]
- Deep learning methods: outperform conventional methods, fewer assumptions [2]

[1] D. Love, R. Heath, V. Lau, D. Gesbert, B. Rao, "An overview on limited feedback in wireless communication systems," IEEE JSAC 2008

# Conventional CSI Feedback in FDD MIMO

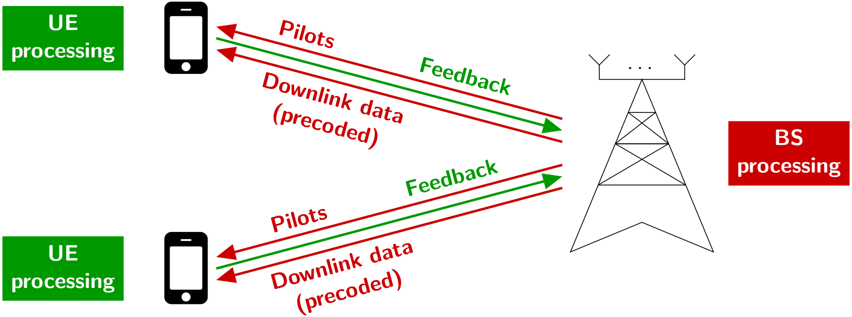



- Conventional CSI compression: vector quantization, compressed sensing [1]
- Deep learning methods: outperform conventional methods, fewer assumptions [2]

**The ultimate metric is spectral efficiency!** $\Longrightarrow$ **Task-oriented CSI compression**

[1] D. Love, R. Heath, V. Lau, D. Gesbert, B. Rao, "An overview on limited feedback in wireless communication systems," IEEE JSAC 2008
[2] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," IEEE TCOM, 2022

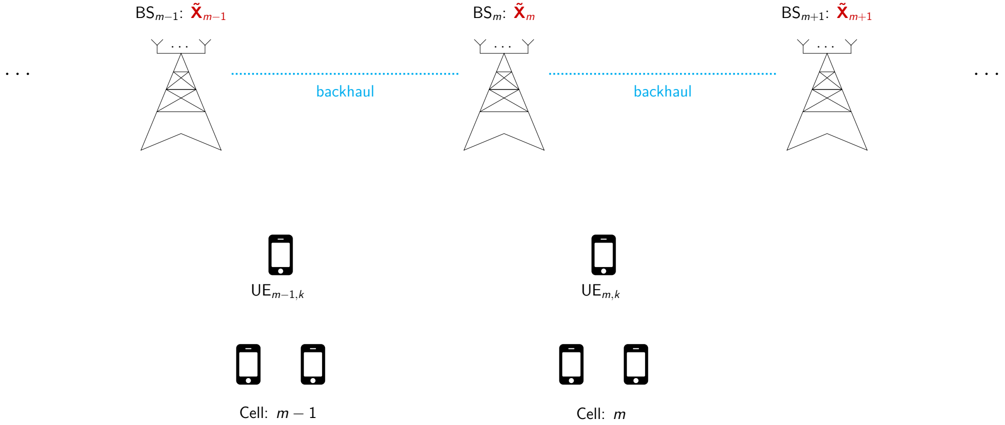
# Task-Aware CSI Feedback: Precoding-Oriented CSI



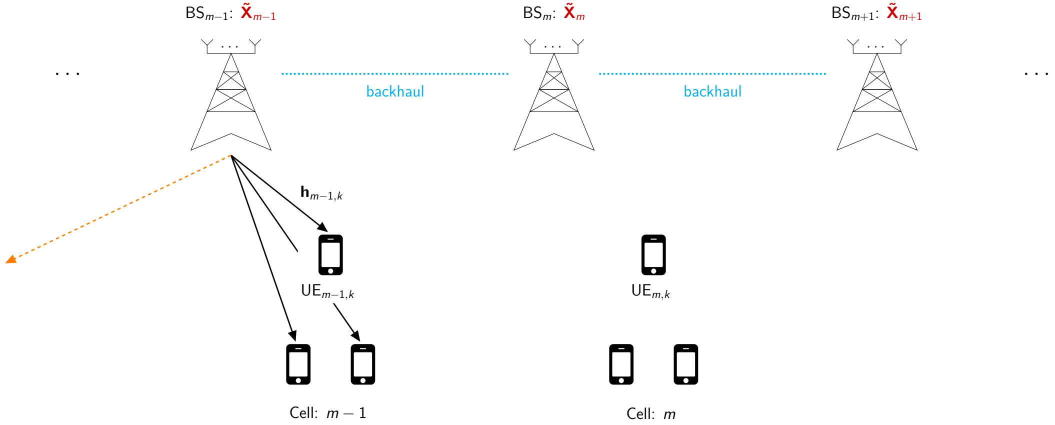

**Task**: MIMO precoding

**Goal**: max achievable rate with limited feedback overhead

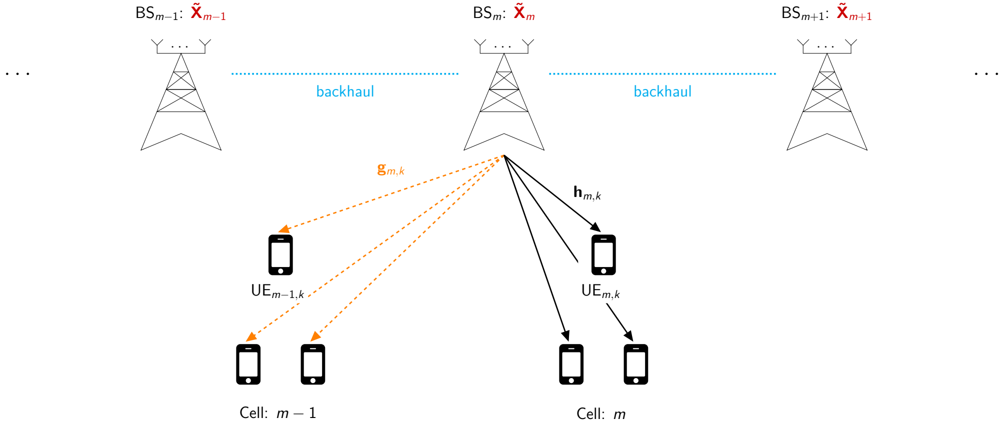
# Multi-Cell Downlink Channel Estimation with Pilots



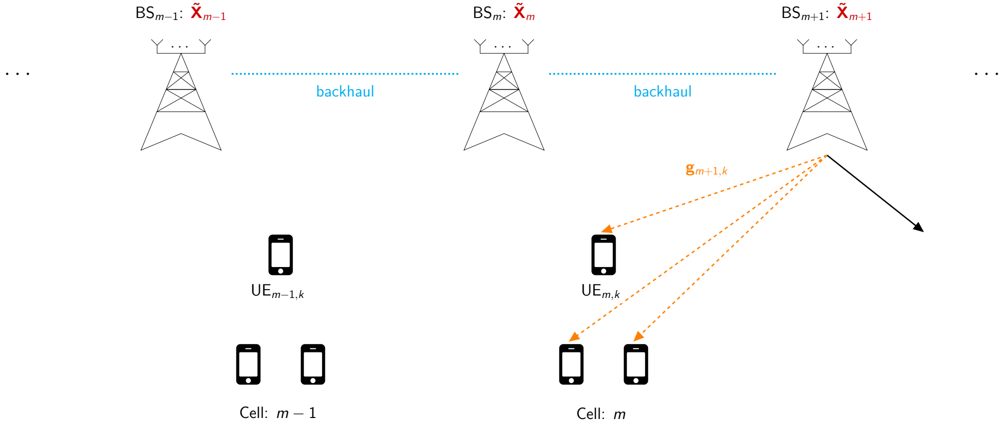
# Multi-Cell Downlink Channel Estimation with Pilots



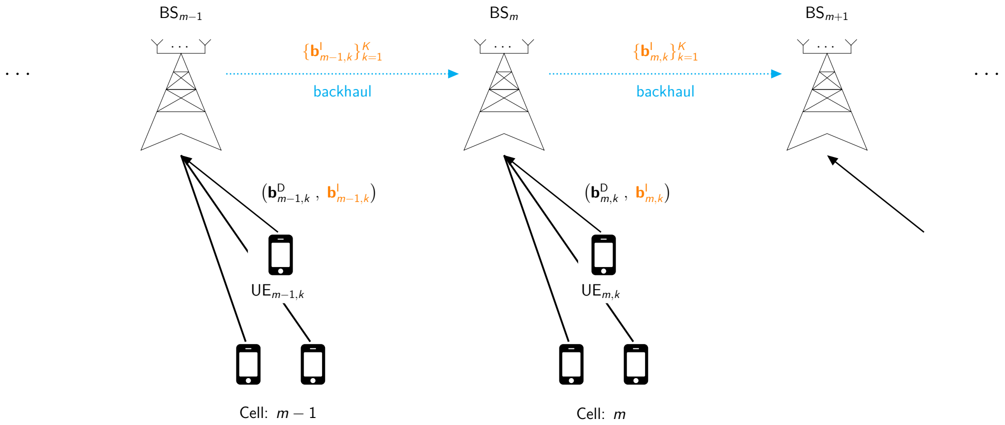
# Multi-Cell Downlink Channel Estimation with Pilots



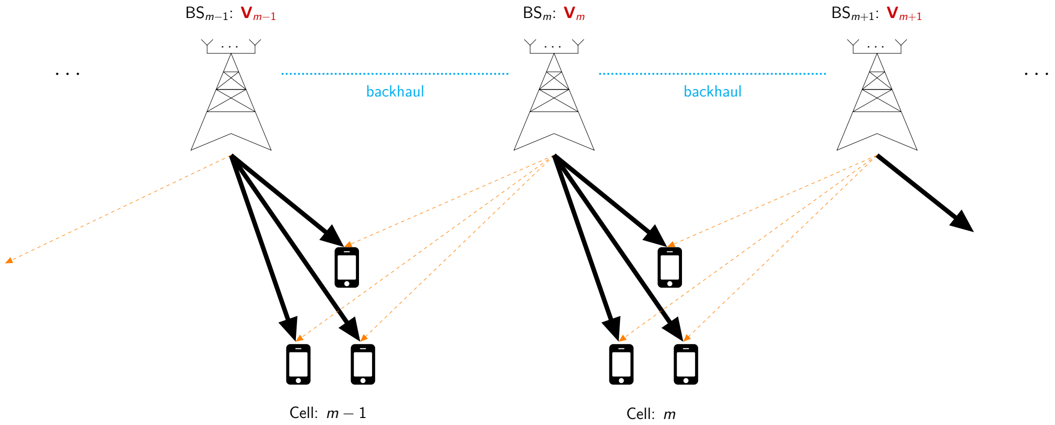
# Multi-Cell Downlink Channel Estimation with Pilots



# Multi-Cell Uplink Feedback



# Multi-Cell Downlink Precoding



# System Model

$m$-th BS downlink signal $\mathbf{x}_m = \sum_k \mathbf{v}_{m,k} s_{m,k} = \mathbf{V}_m \mathbf{s}_m$

# System Model

$m$-th BS downlink signal $\mathbf{x}_m = \sum_k \mathbf{v}_{m,k} s_{m,k} = \mathbf{V}_m \mathbf{s}_m$
The received signal at the $k$-th user in the $m$-th cell is

$$y_{m,k} = \sqrt{\gamma_{m,k}}\, \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} s_{m,k} + \underbrace{\sqrt{\gamma_{m,k}} \sum_{j \neq k} \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} s_{m,j}}_{\text{intra-cell interference}} + \underbrace{\sqrt{\eta_{m,k}} \sum_{i} \mathbf{g}_{m+1,i}^H \mathbf{v}_{m+1,k} s_{m+1,i}}_{\text{inter-cell interference}} + z_{m,k}$$

# System Model

$m$-th BS downlink signal $\mathbf{x}_m = \sum_k \mathbf{v}_{m,k} s_{m,k} = \mathbf{V}_m \mathbf{s}_m$
The received signal at the $k$-th user in the $m$-th cell is

$$y_{m,k} = \sqrt{\gamma_{m,k}}\, \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} s_{m,k} + \underbrace{\sqrt{\gamma_{m,k}} \sum_{j \neq k} \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} s_{m,j}}_{\text{intra-cell interference}} + \underbrace{\sqrt{\eta_{m,k}} \sum_{i} \mathbf{g}_{m+1,i}^H \mathbf{v}_{m+1,k} s_{m+1,i}}_{\text{inter-cell interference}} + z_{m,k}$$

$$\mathrm{SINR}_{m,k} = \frac{|\mathbf{h}_{m,k}^H \mathbf{v}_{m,k}|^2}{\sum_{j \neq k} |\mathbf{h}_{m,k}^H \mathbf{v}_{m,k}|^2 + \alpha_{m,k} \sum_i |\mathbf{g}_{m+1,k}^H \mathbf{v}_{m+1,i}|^2 + 1/\rho_{m,k}}$$

Interference ratio: $\alpha_{m,k} = \eta_{m,k}/\gamma_{m,k} \in [0, 1]$, SNR: $\rho_{m,k} = \gamma_{m,k}/\sigma_{m,k}^2$

# System Model

$m$-th BS downlink signal $\mathbf{x}_m = \sum_k \mathbf{v}_{m,k} s_{m,k} = \mathbf{V}_m \mathbf{s}_m$
The received signal at the $k$-th user in the $m$-th cell is

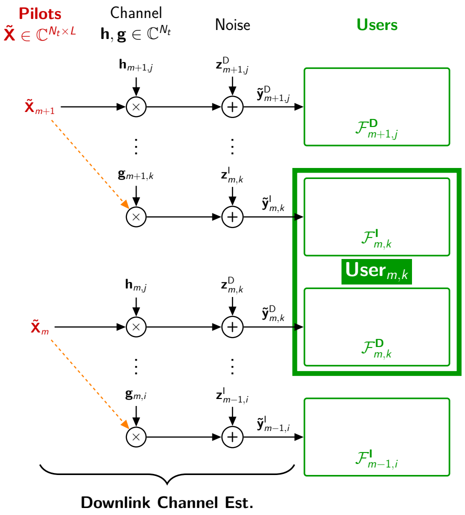
$$y_{m,k} = \sqrt{\gamma_{m,k}}\, \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} s_{m,k} + \underbrace{\sqrt{\gamma_{m,k}} \sum_{j \neq k} \mathbf{h}_{m,k}^H \mathbf{v}_{m,k} s_{m,j}}_{\text{intra-cell interference}} + \underbrace{\sqrt{\eta_{m,k}} \sum_{i} \mathbf{g}_{m+1,i}^H \mathbf{v}_{m+1,k} s_{m+1,i}}_{\text{inter-cell interference}} + z_{m,k}$$

$$\mathrm{SINR}_{m,k} = \frac{|\mathbf{h}_{m,k}^H \mathbf{v}_{m,k}|^2}{\sum_{j \neq k} |\mathbf{h}_{m,k}^H \mathbf{v}_{m,k}|^2 + \alpha_{m,k} \sum_i |\mathbf{g}_{m+1,k}^H \mathbf{v}_{m+1,i}|^2 + 1/\rho_{m,k}}$$

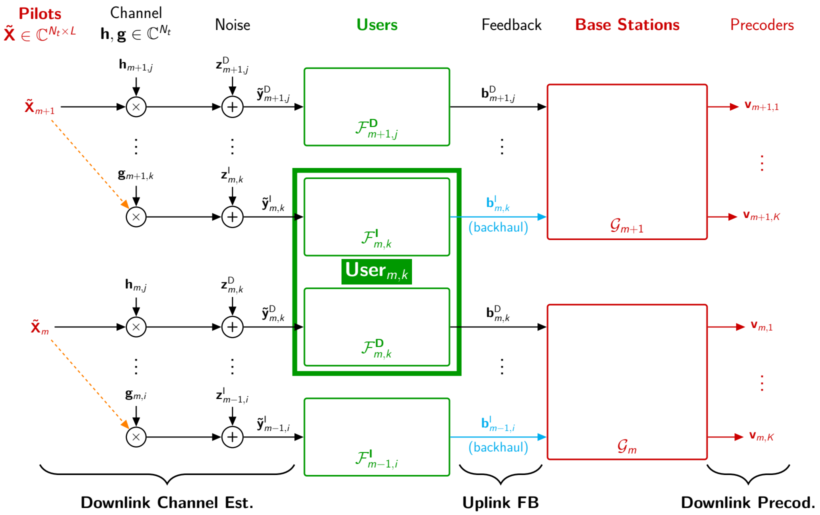
Interference ratio: $\alpha_{m,k} = \eta_{m,k}/\gamma_{m,k} \in [0,1]$, SNR: $\rho_{m,k} = \gamma_{m,k}/\sigma_{m,k}^2$

**Metric**: network sum rate $R = \sum_m \sum_k \log_2 (1 + \mathrm{SINR}_{m,k})$

# System Model: Block Diagram



# System Model: Block Diagram



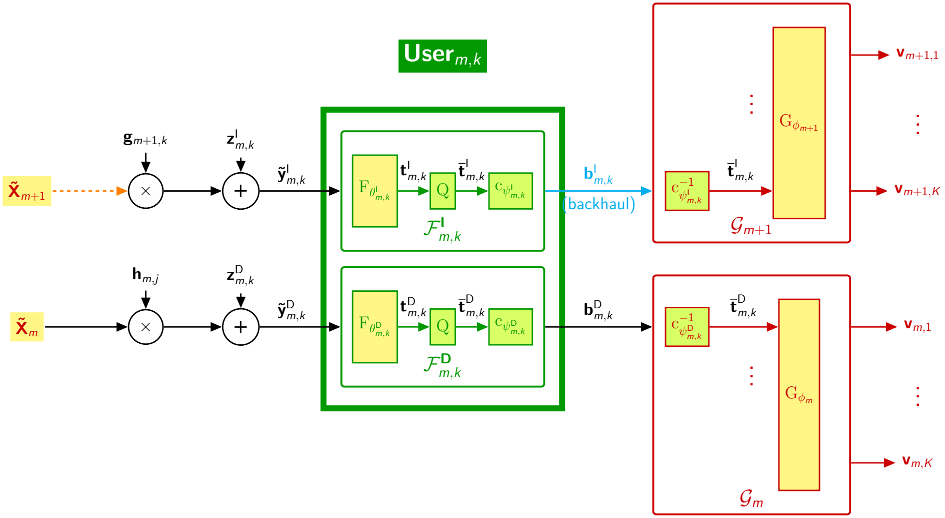
# Previous Work

| Ref | Task-oriented? | Feedback optimization? | Multi-User? | Multi-Cell? |
|---|---|---|---|---|
| [1] | Yes, precoding | No, fixed | Yes | No |
| [2] | No, ch. reconstr. | Yes [4] | Yes | No |
| [3] | Yes, precoding | No, fixed | No | Yes |
| Ours | Yes, precoding | Yes [4] | Yes | Yes |

- [4]: image compression with neural networks (autoencoder), the loss function includes a tradeoff between feedback overhead (rate) and image reconstruction performance

[1] F. Sohrabi, K.Attiah, W.Yu,"Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO,"IEEE TWC 2021
[2] M. B. Mashhadi, Q. Yang, D. Gunduz, "Distributed deep convolutional compression for massive MIMO CSI feedback," IEEE TWC 2021
[3] J. Guo, C.Wen,S.Jin,"Deep Learning-Based CSI Feedback for Beamforming in Single- and Multi-Cell Massive MIMO Systems," IEEE JSAC '21
[4] J. Ballé, V. Laparra, E. P. Simoncelli, "End-to-end optimized image compression," ICLR 2017

# Learned Neural Compression

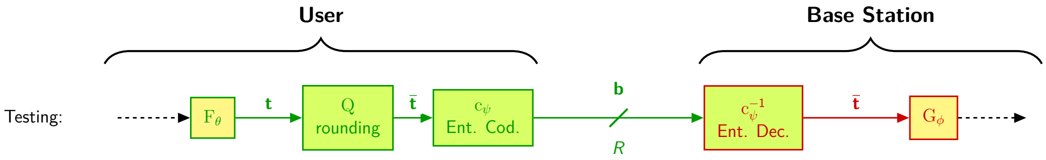


# Feedback Overhead Optimization

Feedback quantization: Neural network output to bitstream — from [Ballé et al., 2017]

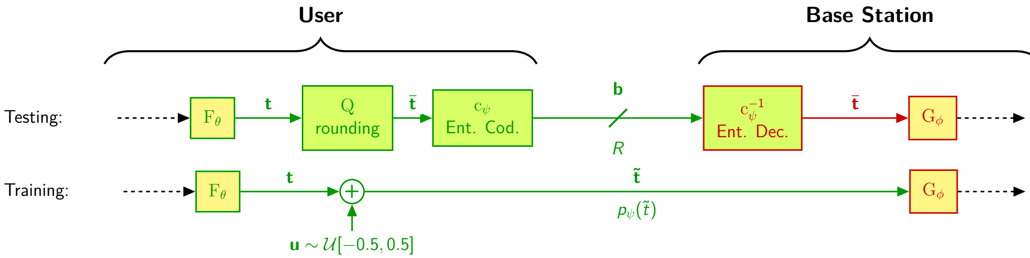
# Feedback Overhead Optimization

Feedback quantization: Neural network output to bitstream — from [Ballé et al., 2017]



# Feedback Overhead Optimization

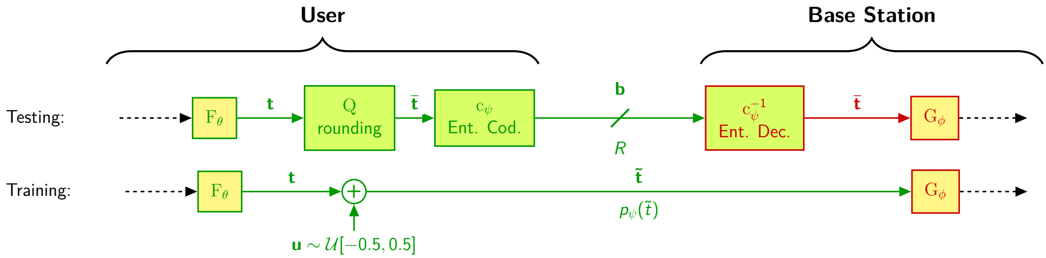
Feedback quantization: Neural network output to bitstream — from [Ballé et al., 2017]




- During training: pseudo-quantized features $\tilde{\mathbf{t}} = \mathbf{t} + \mathbf{u}$

# Feedback Overhead Optimization

Feedback quantization: Neural network output to bitstream — from [Ballé et al., 2017]




- During training: pseudo-quantized features $\tilde{\mathbf{t}} = \mathbf{t} + \mathbf{u}$
- Note: the probability distribution of $\tilde{\mathbf{t}}$ is a continuous relaxation of the one of $\bar{\mathbf{t}}$
- $\psi$: parameters learned during training $\Longrightarrow p_\psi(\tilde{\mathbf{t}}) \sim$ probability distribution of $\tilde{\mathbf{t}}$
- Entropy of $\tilde{\mathbf{t}}$ as an estimate for $R$: $\mathbb{E}\left[-\log_2 p_\psi(\tilde{\mathbf{t}})\right]$

# Loss Function

$$\mathcal{L}(\Theta, \Phi, \Psi) = \mathcal{O} - \lambda \mathcal{R}$$

- Feedback overhead $\mathcal{O}$: entropy (rate) of the pseudo-quantized features $\tilde{\mathbf{t}}_{m,k}$
- Performance $\mathcal{R}$: network sum rate achieved with precoding $\mathbf{V}_m$

# Loss Function

$$\mathcal{L}(\Theta, \Phi, \Psi) = \mathcal{O} - \lambda \mathcal{R}$$

- Feedback overhead $\mathcal{O}$: entropy (rate) of the pseudo-quantized features $\tilde{\mathbf{t}}_{m,k}$

$$\mathcal{O}(\Theta, \Psi) = \sum_{m=1}^{M} \sum_{k=1}^{K} \mathcal{O}_{m,k}(\Theta, \Psi)$$

$$\mathcal{O}_{m,k}(\Theta, \Psi) = \mathbb{E}_{\mathbf{h},\mathbf{g},\mathbf{u},\mathbf{z}} \left[ -\log_2 p_{\psi^{\mathsf{D}}}(\tilde{\mathbf{t}}^{\mathsf{D}}_{m,k}) - \log_2 p_{\psi^{\mathsf{I}}}(\tilde{\mathbf{t}}^{\mathsf{I}}_{m,k}) \right]$$

- Performance $\mathcal{R}$: network sum rate achieved with precoding $\mathbf{V}_m$

$$\mathcal{R}(\Theta, \Psi, \Phi) = \sum_{m=1}^{M} \sum_{k=1}^{K} \mathcal{R}_{m,k}(\Theta, \Psi, \Phi)$$

$$\mathcal{R}_{m,k}(\Theta, \Psi, \Phi) = \mathbb{E}_{\mathbf{h},\mathbf{g},\mathbf{u},\mathbf{z}} \log_2 \left( 1 + \mathsf{SINR}_{m,k} \right)$$

# Simulation Scenario

Channel model: multipath, BS with uniform linear array.

$$\mathbf{h}_{m,k} = \frac{1}{\sqrt{L_p^{\mathrm{D}}}} \sum_{\ell=1}^{L_p^{\mathrm{D}}} \alpha_{m,k,\ell}^{\mathrm{D}} \mathbf{a}_t(\beta_{m,k,\ell}^{\mathrm{D}}), \qquad \mathbf{g}_{m,k} = \frac{1}{\sqrt{L_p^{\mathrm{I}}}} \sum_{\ell=1}^{L_p^{\mathrm{I}}} \alpha_{m,k,\ell}^{\mathrm{I}} \mathbf{a}_t(\beta_{m,k,\ell}^{\mathrm{I}})$$

where $\alpha_{m,k,\ell}$ is the complex gain, $\mathbf{a}_t(\beta_{m,k,\ell})$ is the ULA response for AoD $\beta_{m,k,\ell}$.

# Simulation Scenario

Channel model: multipath, BS with uniform linear array.

$$\mathbf{h}_{m,k} = \frac{1}{\sqrt{L_p^{\mathrm{D}}}} \sum_{\ell=1}^{L_p^{\mathrm{D}}} \alpha_{m,k,\ell}^{\mathrm{D}} \mathbf{a}_t(\beta_{m,k,\ell}^{\mathrm{D}}), \qquad \mathbf{g}_{m,k} = \frac{1}{\sqrt{L_p^{\mathrm{I}}}} \sum_{\ell=1}^{L_p^{\mathrm{I}}} \alpha_{m,k,\ell}^{\mathrm{I}} \mathbf{a}_t(\beta_{m,k,\ell}^{\mathrm{I}})$$

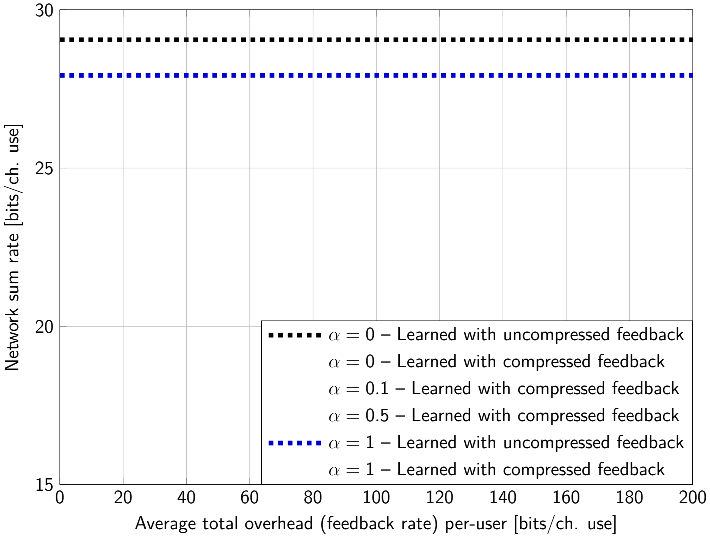
where $\alpha_{m,k,\ell}$ is the complex gain, $\mathbf{a}_t(\beta_{m,k,\ell})$ is the ULA response for AoD $\beta_{m,k,\ell}$.

Scenarios:

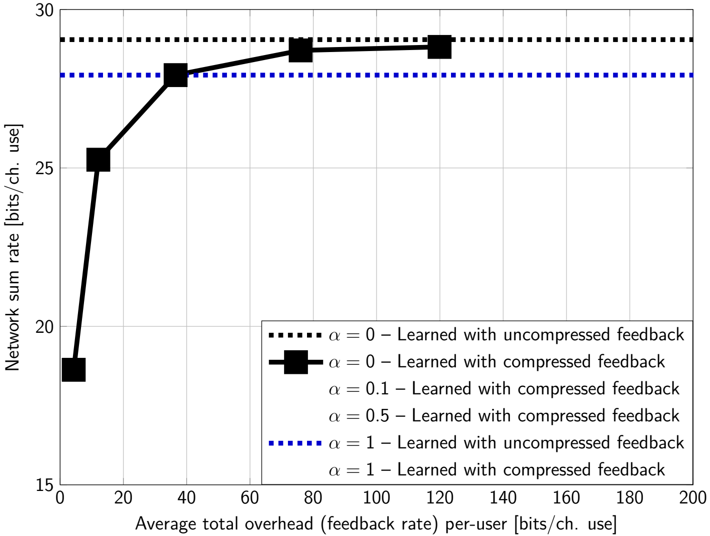
- Single-cell
  - Single-user
  - Multi-user (intra-cell interference)
- Multi-cell
  - Single-user (inter-cell interference)
  - Multi-user (intra- and inter-cell interference)

Results for $N_t = 64$ TX antennas, $L_p = 2$ paths channel, $L = 8$ pilots.

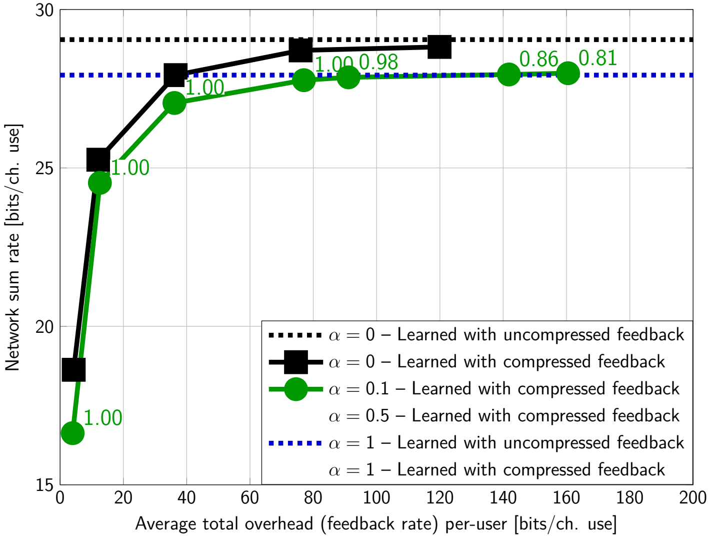
# 2 cells, 2 users each (4 users total), SNR $\rho = 10$ dB



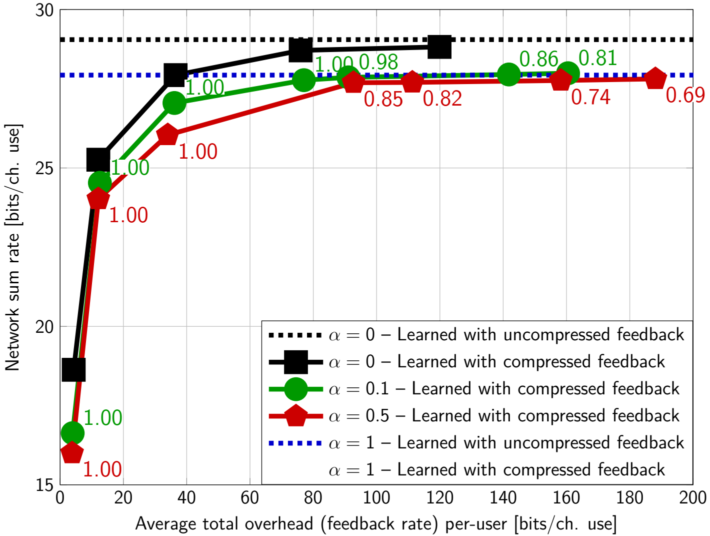
# 2 cells, 2 users each (4 users total), SNR $\rho = 10$ dB



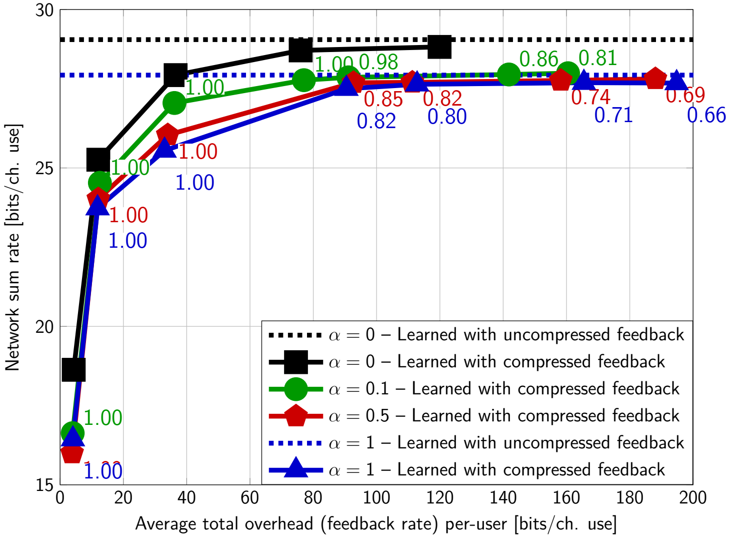
# 2 cells, 2 users each (4 users total), SNR $\rho = 10$ dB



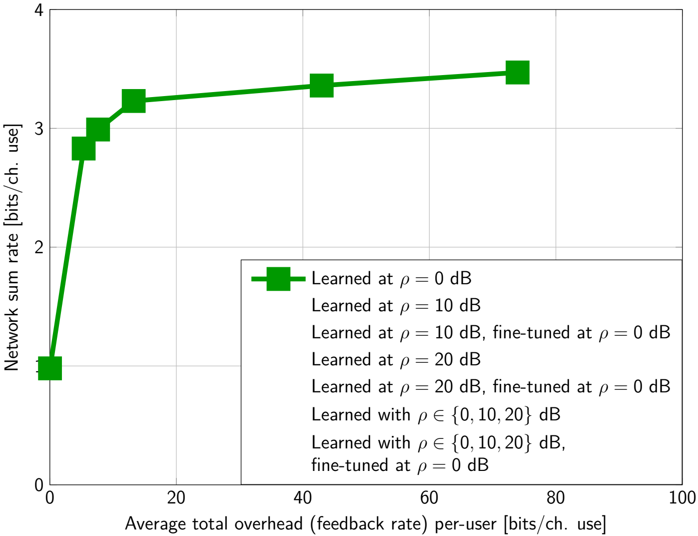
# 2 cells, 2 users each (4 users total), SNR $\rho = 10$ dB



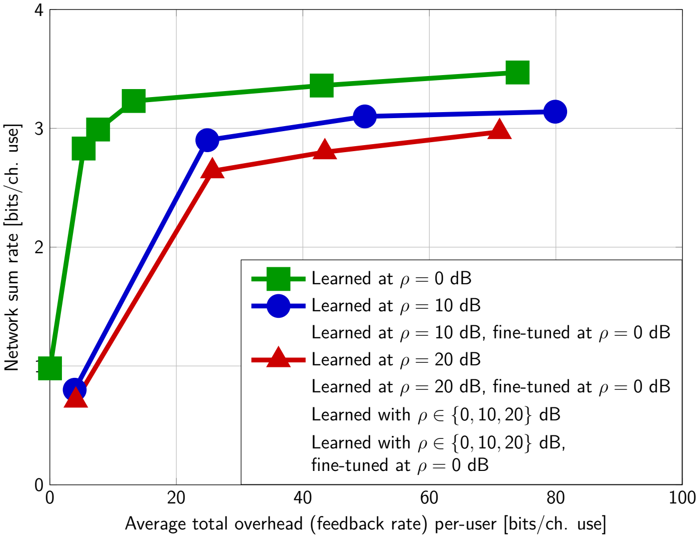
# 2 cells, 2 users each (4 users total), SNR $\rho = 10$ dB



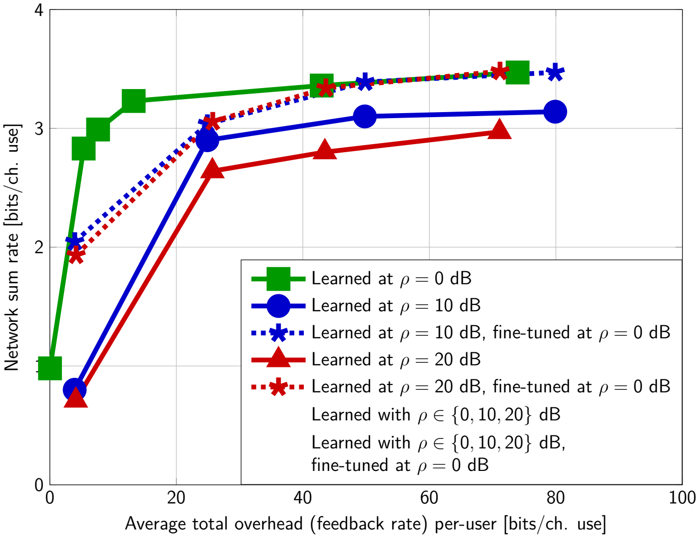
# 1 cell, 2 users, test SNR $\rho = 0$ dB, train SNR $\in \{0, 10, 20\}$ dB



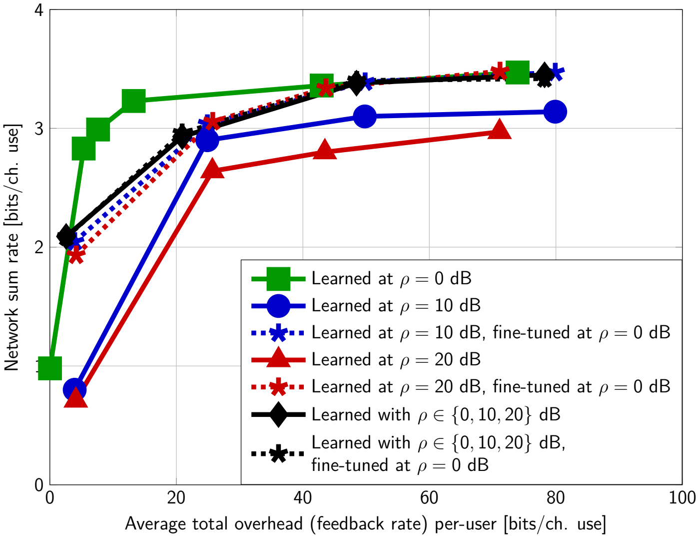
# 1 cell, 2 users, test SNR $\rho = 0$ dB, train SNR $\in \{0, 10, 20\}$ dB



# 1 cell, 2 users, test SNR $\rho = 0$ dB, train SNR $\in \{0, 10, 20\}$ dB



# 1 cell, 2 users, test SNR $\rho = 0$ dB, train SNR $\in \{0, 10, 20\}$ dB



# Summary on the Precoding-oriented CSI Feedback

- Analysis of the tradeoff between feedback overhead and system performance for multi-cell multi-user MIMO systems
- The fine-tuned BS provides robustness, since it can compensate for the use of mismatched user models
- The unstructured loss function learns an allocation strategy that recalls the water-filling policy

# Summary on the Precoding-oriented CSI Feedback

- Analysis of the tradeoff between feedback overhead and system performance for multi-cell multi-user MIMO systems
- The fine-tuned BS provides robustness, since it can compensate for the use of mismatched user models
- The unstructured loss function learns an allocation strategy that recalls the water-filling policy

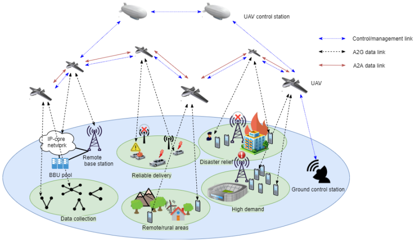
**Future directions:**

- Further investigation of asymmetric settings (e.g., different user SNRs)
- Run on different channel models
- Extension to MIMO-OFDM systems
- Include link- and system-level simulations

# Outline

# Introduction

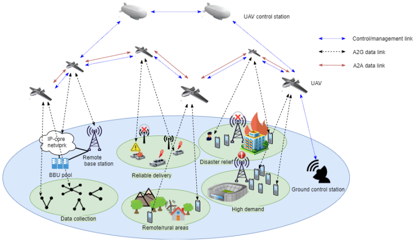
- Relay channel: fundamental building block of cooperative communications.
  - Applications: relays to improve throughput/coverage, e.g., RIS, drones.




Gholami et. al. "Joint Mobility-Aware UAV Placement and Routing in Multi-Hop UAV Relaying Systems"

# Introduction

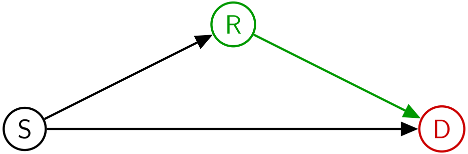
- Relay channel: fundamental building block of cooperative communications.
  - Applications: relays to improve throughput/coverage, e.g., RIS, drones.




Gholami et. al. "Joint Mobility-Aware UAV Placement and Routing in Multi-Hop UAV Relaying Systems"

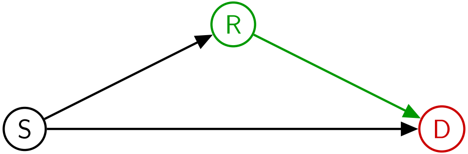
- Capacity for the general relay channel is unknown, but several relaying strategies have been proposed.
  - Amplify-and-forward, decode-and-forward...
  - **Compress-and-forward (CF)**: the relay sends a quantized version of its rx signal.

# Motivation




- Relay and destination signals are correlated: **distributed compression** techniques like Wyner-Ziv (WZ) coding can be used
- ... but practical distributed compressors have not been fully developed

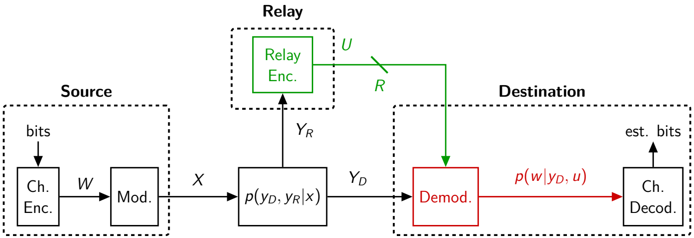
# Motivation



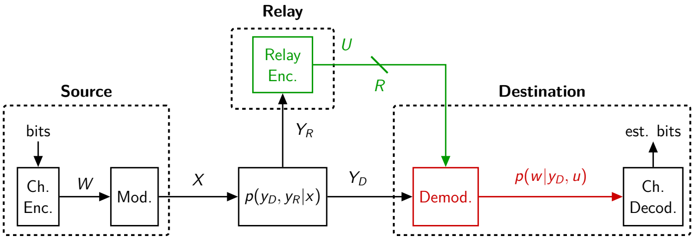

- Relay and destination signals are correlated: **distributed compression** techniques like Wyner-Ziv (WZ) coding can be used
- … but practical distributed compressors have not been fully developed
- We model **relays as learned WZ compressors** [1] in a simple communication system $\Rightarrow$ learned CF strategy

[1] E. Ozyilkan, J. Ballé, and E. Erkip, "Learned Wyner–Ziv compressors recover binning," in IEEE ISIT 2023

# Primitive Relay Channel (PRC) – *out-of-band* relay

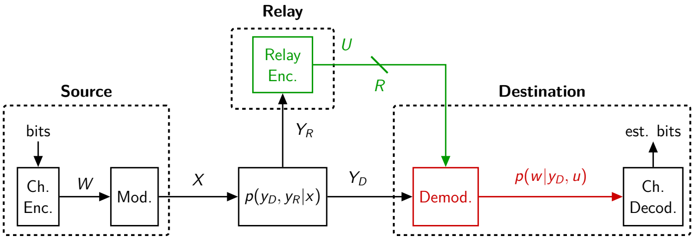


# Primitive Relay Channel (PRC) – *out-of-band* relay




- **Relay's POV**: compress $Y_R$ to help the destination decode $W$

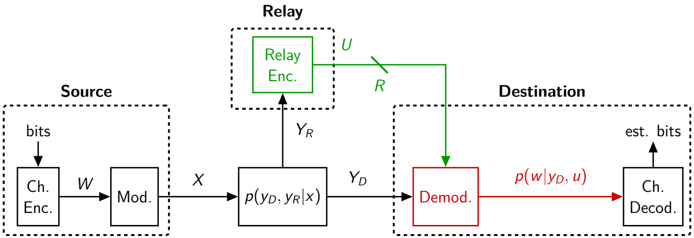
# Primitive Relay Channel (PRC) – *out-of-band* relay




- **Relay's POV**: compress $Y_R$ to help the destination decode $W$
- CF is optimal for *oblivious* relaying [1]

[1] O. Simeone, E. Erkip, S. Shamai, "On codebook information for interference relay channels with out-of-band relaying," IEEE TIT 2011

# Primitive Relay Channel (PRC) – *out-of-band* relay

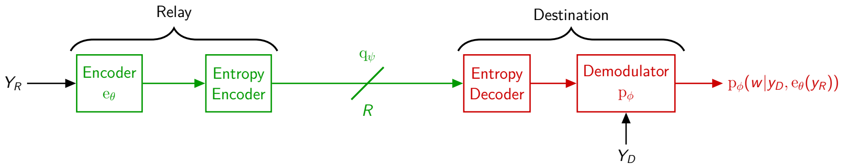



- **Relay's POV**: compress $Y_R$ to help the destination decode $W$
- CF is optimal for *oblivious* relaying [1]
- **Task-aware design**: detection-oriented relays
  - Task: symbol detection (demodulation)
  - Goal: maximize communication rate $\mathrm{I}(X; Y_D, U)$ subject to rate constraint $R$

[1] O. Simeone, E. Erkip, S. Shamai, "On codebook information for interference relay channels with out-of-band relaying," IEEE TIT 2011

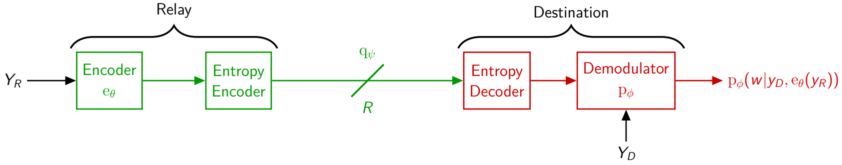
# Learned CF $\Rightarrow$ Neural WZ compressors [Ozyilkan et al, 2023]

- Marginal formulation:

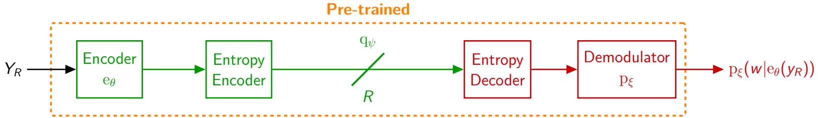


# Learned CF $\Rightarrow$ Neural WZ compressors [Ozyilkan et al, 2023]

- Marginal formulation:

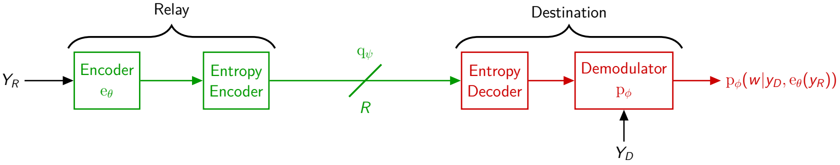



- Relay compression without side information (point-to-point):

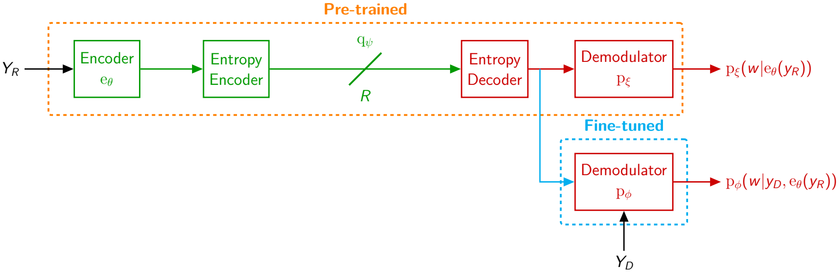


# Learned CF $\Rightarrow$ Neural WZ compressors [Ozyilkan et al, 2023]

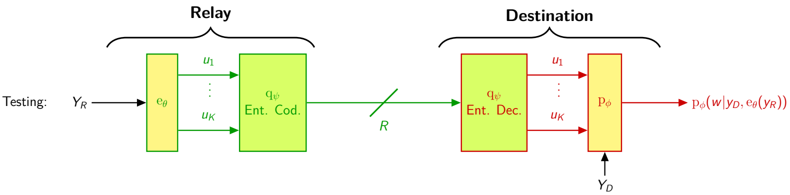
- Marginal formulation:




- Relay compression without side information (point-to-point):

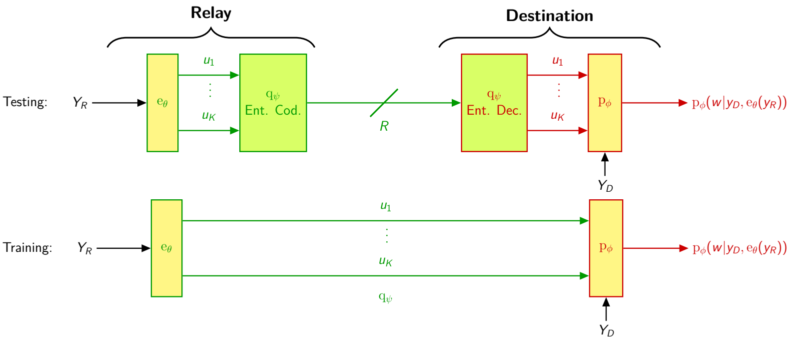


# Relay rate $R$ optimization




- Encoder's output $[u_1, \ldots, u_K] \sim$ *one-hot* — possible messages $1, \ldots, K$
- Testing: one-hot vector, probability distribution $q_\psi(u)$

# Relay rate $R$ optimization




- Encoder's output $[u_1, \ldots, u_K] \sim$ *one-hot* — possible messages $1, \ldots, K$
- Testing: one-hot vector, probability distribution $q_\psi(u)$
- Training: Gumbel-softmax trick with decreasing temperature parameter
- $\psi$: parameters learned during training $\Rightarrow \sim$ probability distribution of messages $u$
- Entropy of $q_\psi(u)$ as an estimate of $R$: $\mathbb{E}\left[-\log_2 q_\psi(e_\theta(y_R))\right]$

# Loss Function for Learned Compress-and-Forward

**Objective** [Simeone et al, 2011]: $C = \max \, \mathrm{I}(X; Y_D, U) \quad \text{s.t.} \quad R \geq \mathrm{I}(Y_R; U \mid Y_D)$

# Loss Function for Learned Compress-and-Forward

**Objective** [Simeone et al, 2011]: $C = \max \mathrm{I}(X; Y_D, U) \quad \text{s.t.} \quad R \geq \mathrm{I}(Y_R; U \mid Y_D)$

**Upper bound on relay rate**

$$\mathrm{I}(Y_R; U \mid Y_D) \leq H(U \mid Y_D) \leq \mathbb{E}\left[-\log_2 \mathrm{q}_\psi(\mathrm{e}_\theta(y_R))\right]$$

# Loss Function for Learned Compress-and-Forward

**Objective** [Simeone et al, 2011]: $C = \max \mathrm{I}(X; Y_D, U) \quad \text{s.t.} \quad R \geq \mathrm{I}(Y_R; U \mid Y_D)$

**Upper bound on relay rate**

$$\mathrm{I}(Y_R; U \mid Y_D) \leq H(U \mid Y_D) \leq \mathbb{E}\left[-\log_2 \mathrm{q}_\psi(\mathrm{e}_\theta(y_R))\right]$$

**Lower bound on communication rate**

$$I(X; Y_D, U) = H(W) - H(W \mid Y_D, U) \geq H(W) - \mathbb{E}\left[-\log_2(\mathrm{p}_\phi(w \mid y_D, \mathrm{e}_\theta(y_R)))\right]$$

# Loss Function for Learned Compress-and-Forward

**Objective** [Simeone et al, 2011]: $C = \max \mathrm{I}(X; Y_D, U) \quad \text{s.t.} \quad R \geq \mathrm{I}(Y_R; U \mid Y_D)$

**Upper bound on relay rate**

$$\mathrm{I}(Y_R; U \mid Y_D) \leq H(U \mid Y_D) \leq \mathbb{E}\left[-\log_2 \mathrm{q}_\psi(\mathrm{e}_\theta(y_R))\right]$$

**Lower bound on communication rate**

$$I(X; Y_D, U) = H(W) - H(W \mid Y_D, U) \geq H(W) - \mathbb{E}\left[-\log_2(\mathrm{p}_\phi(w \mid y_D, \mathrm{e}_\theta(y_R)))\right]$$

**Loss function**

$$\mathcal{L}(\mathrm{e}_\theta, \mathrm{q}_\psi, \mathrm{p}_\phi) = \underbrace{\mathbb{E}\left[-\log_2 \mathrm{q}_\psi(\mathrm{e}_\theta(y_R))\right]}_{\text{compression rate } \tilde{R}} + \lambda \underbrace{\mathbb{E}\left[-\log_2(\mathrm{p}_\phi(w \mid y_D, \mathrm{e}_\theta(y_R)))\right]}_{\text{cross-entropy } \tilde{D}}$$

# Loss Function for Learned Compress-and-Forward

**Objective** [Simeone et al, 2011]: $C = \max \mathrm{I}(X; Y_D, U) \quad \text{s.t.} \quad R \geq \mathrm{I}(Y_R; U \mid Y_D)$

**Upper bound on relay rate**

$$\mathrm{I}(Y_R; U \mid Y_D) \leq H(U \mid Y_D) \leq \mathbb{E}\left[-\log_2 \mathrm{q}_\psi(\mathrm{e}_\theta(y_R))\right]$$

**Lower bound on communication rate**

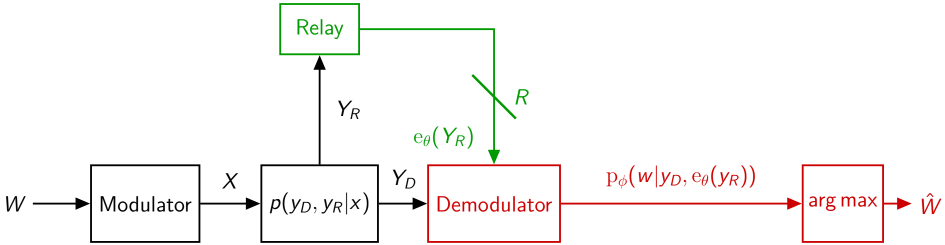
$$I(X; Y_D, U) = H(W) - H(W \mid Y_D, U) \geq H(W) - \mathbb{E}\left[-\log_2(\mathrm{p}_\phi(w \mid y_D, \mathrm{e}_\theta(y_R)))\right]$$

**Loss function**

$$\mathcal{L}(\mathrm{e}_\theta, \mathrm{q}_\psi, \mathrm{p}_\phi) = \underbrace{\mathbb{E}\left[-\log_2 \mathrm{q}_\psi(\mathrm{e}_\theta(y_R))\right]}_{\text{compression rate } \tilde{R}} + \lambda \underbrace{\mathbb{E}\left[-\log_2(\mathrm{p}_\phi(w \mid y_D, \mathrm{e}_\theta(y_R)))\right]}_{\text{cross-entropy } \tilde{D}}$$

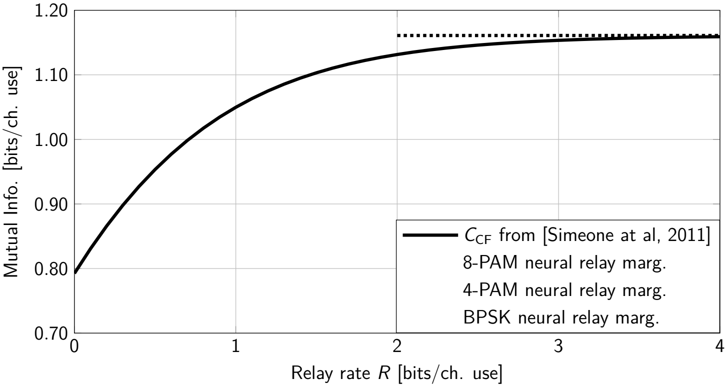
Cross-entropy is also a proxy for $P(W \neq \hat{W})$, $\hat{W} = \arg\max_w \mathrm{p}_\phi(w|Y_D, \mathrm{e}_\theta(Y_R))$

# Simulation Scenario



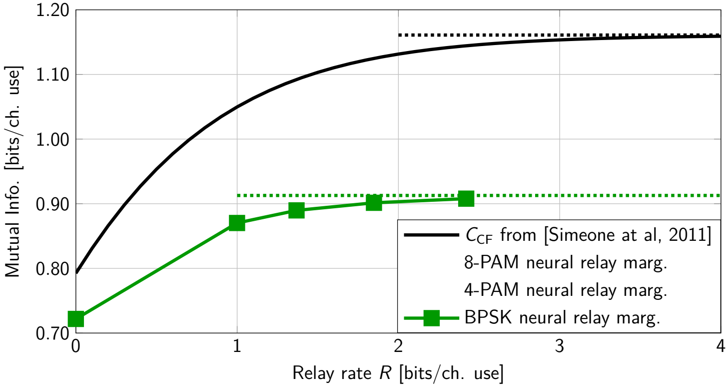

- Source: equally likely symbols, power constraint $P = \mathbb{E}[|X|^2]$
  - Real channel: BPSK, 4-PAM, 8-PAM
  - Complex channel: QAM, 16-QAM
- Channel: $Y_D = X + N_D$ and $Y_R = X + N_R$, with $N_D \perp N_R$
  - $(N_D, N_R)$ (complex) Gaussian noise with variance $(\sigma_D^2, \sigma_R^2)$
- SNR: $\gamma_D = P/\sigma_D^2$, $\gamma_R = P/\sigma_R^2$.

# Mutual Information for marginal model at $\gamma_D = \gamma_R = 3$ dB



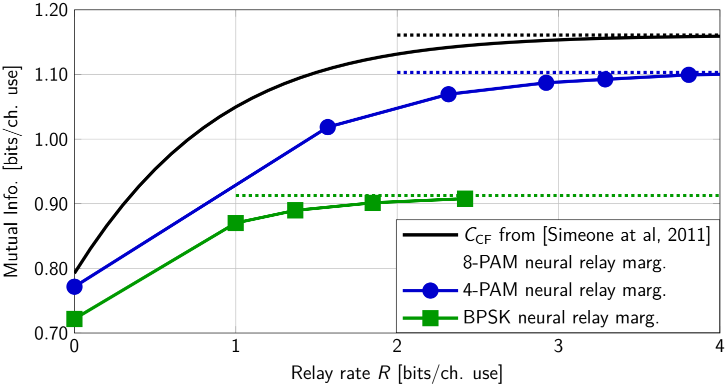

CF achievable rate [Simeone et al, 2011]: $C_{\mathrm{CF}} = \frac{1}{2}\log_2\left(1 + \gamma_D + \frac{\gamma_R}{1 + \frac{1+\gamma_D+\gamma_R}{(2^{2R}-1)(\gamma_D+1)}}\right)$

# Mutual Information for marginal model at $\gamma_D = \gamma_R = 3$ dB



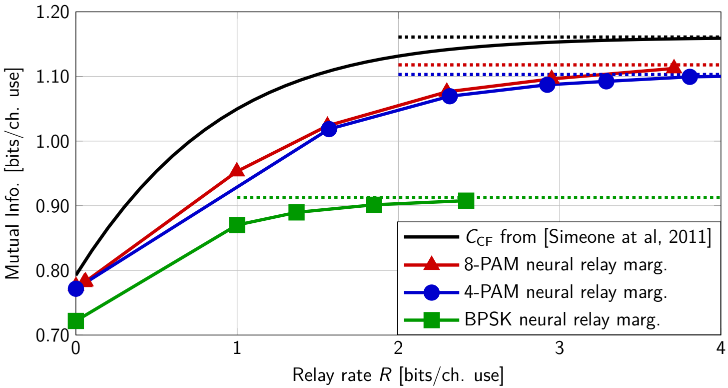

CF achievable rate [Simeone et al, 2011]: $$C_{\mathrm{CF}} = \frac{1}{2}\log_2\left(1+\gamma_D+\frac{\gamma_R}{1+\frac{1+\gamma_D+\gamma_R}{(2^{2R}-1)(\gamma_D+1)}}\right)$$

# Mutual Information for marginal model at $\gamma_D = \gamma_R = 3$ dB



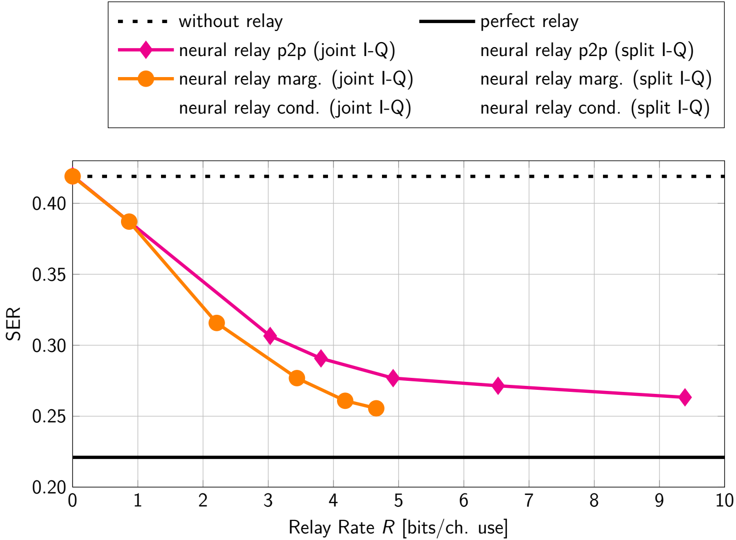

CF achievable rate [Simeone et al, 2011]: $$C_{CF} = \frac{1}{2}\log_2\left(1 + \gamma_D + \frac{\gamma_R}{1 + \frac{1+\gamma_D+\gamma_R}{(2^{2R}-1)(\gamma_D+1)}}\right)$$

# Mutual Information for marginal model at $\gamma_D = \gamma_R = 3$ dB

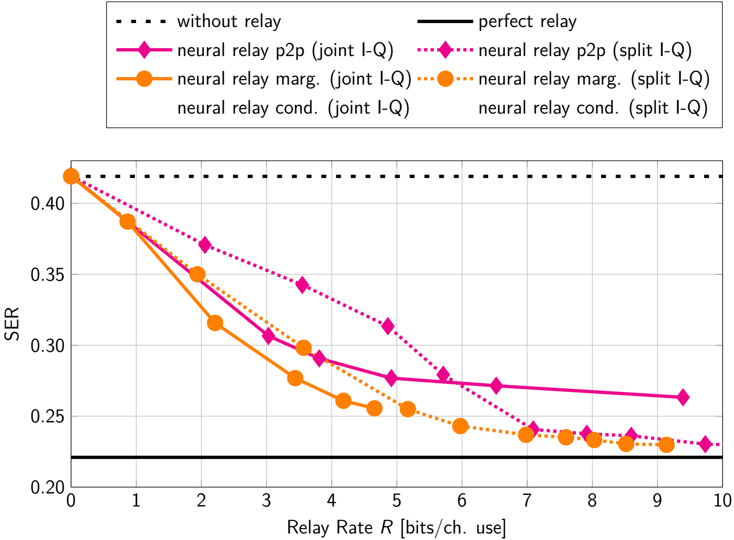



CF achievable rate [Simeone et al, 2011]: $C_{\mathrm{CF}} = \frac{1}{2}\log_2\left(1+\gamma_D+\frac{\gamma_R}{1+\frac{1+\gamma_D+\gamma_R}{(2^{2R}-1)(\gamma_D+1)}}\right)$

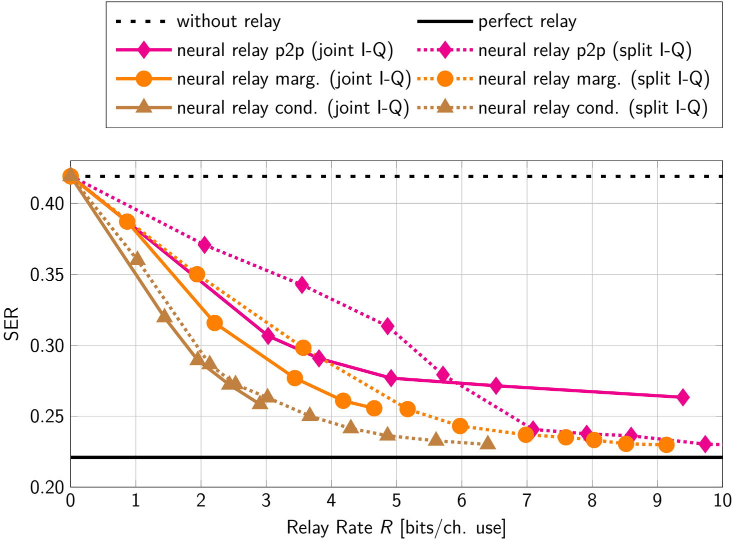
# Symbol Error Rate for 16-QAM at $\gamma_D = \gamma_R = 7$ dB



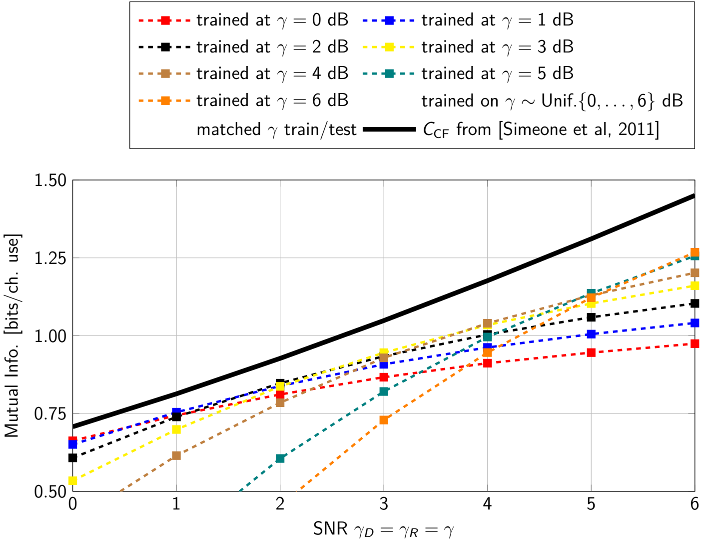
# Symbol Error Rate for 16-QAM at $\gamma_D = \gamma_R = 7$ dB



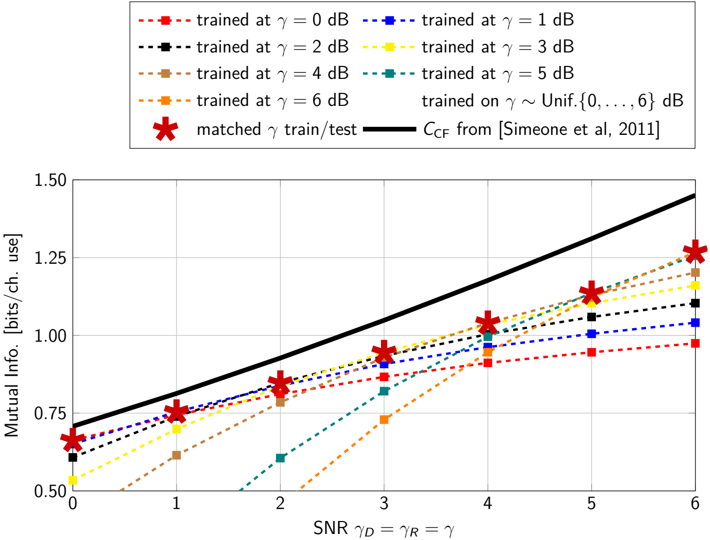
# Symbol Error Rate for 16-QAM at $\gamma_D = \gamma_R = 7$ dB



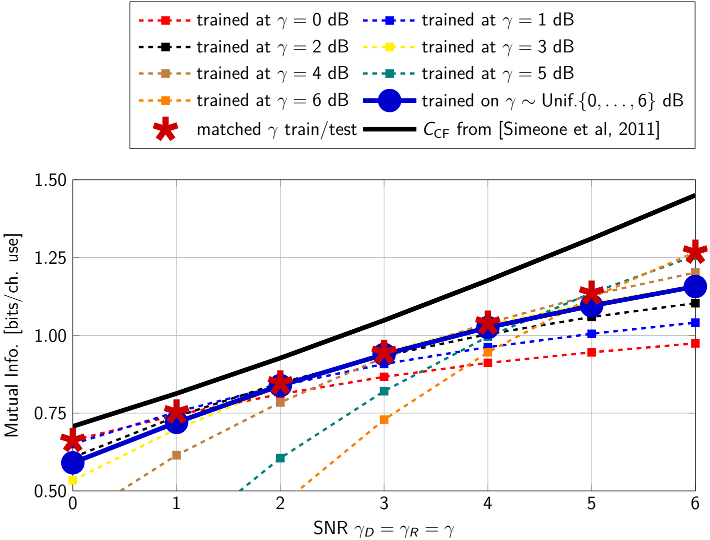
# Robustness for 4-PAM, $\gamma_D = \gamma_R = \gamma$ dB, $R \approx 1$



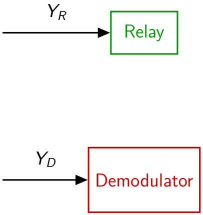
# Robustness for 4-PAM, $\gamma_D = \gamma_R = \gamma$ dB, $R \approx 1$

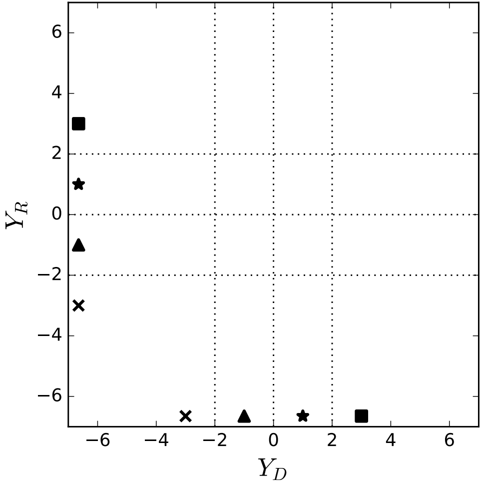


# Robustness for 4-PAM, $\gamma_D = \gamma_R = \gamma$ dB, $R \approx 1$

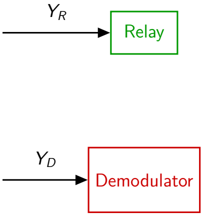


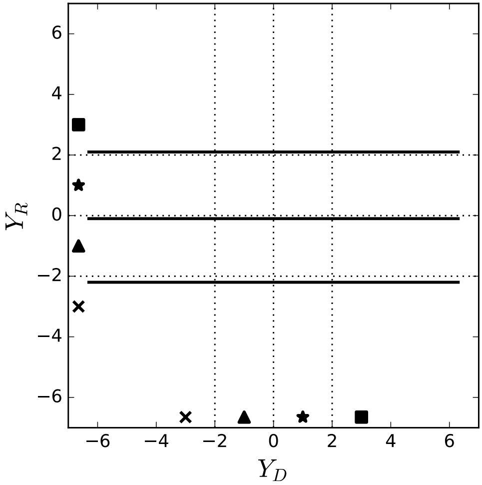
# Quantization and decisions for 4-PAM, $\gamma_D = \gamma_R = 13$ dB, $R \approx 1$



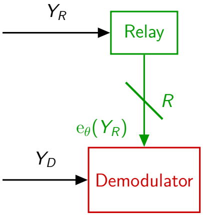



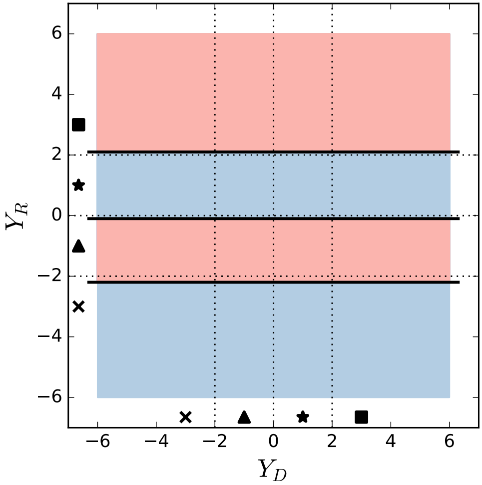
# Quantization and decisions for 4-PAM, $\gamma_D = \gamma_R = 13$ dB, $R \approx 1$



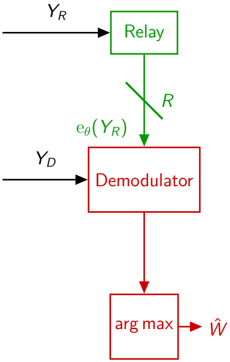



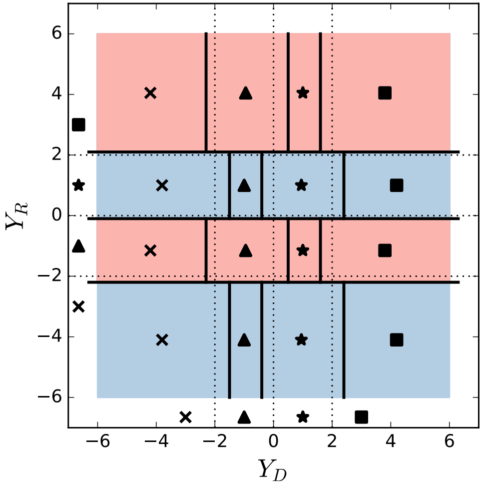
# Quantization and decisions for 4-PAM, $\gamma_D = \gamma_R = 13$ dB, $R \approx 1$



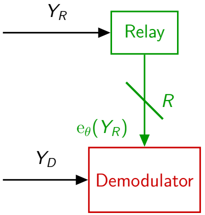
# Quantization and decisions for 4-PAM, $\gamma_D = \gamma_R = 13$ dB, $R \approx 1$



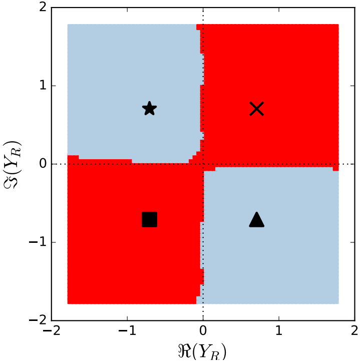



# Quantization and decisions for 4-QAM, $\gamma_D = \gamma_R = 7$ dB, $R \approx 1$

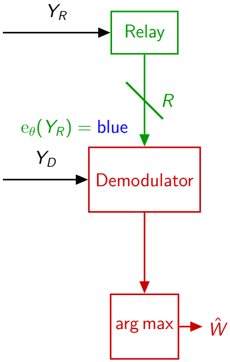
Destination's Demodulator

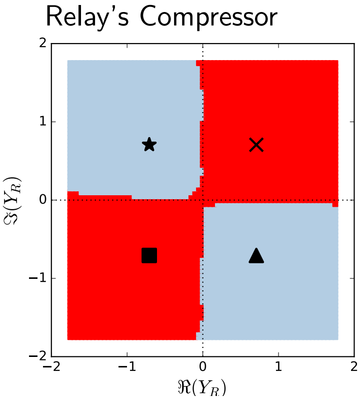


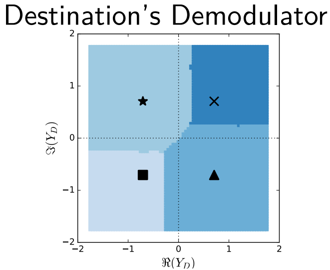

Relay's Compressor



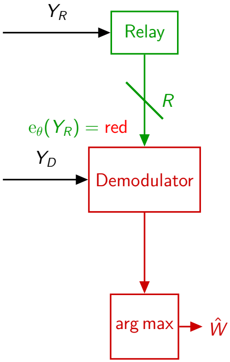
# Quantization and decisions for 4-QAM, $\gamma_D = \gamma_R = 7$ dB, $R \approx 1$



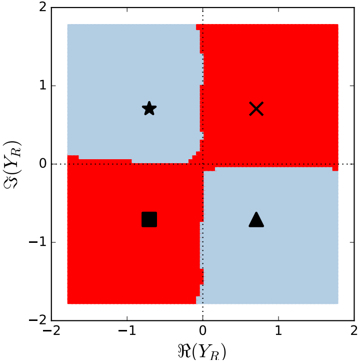



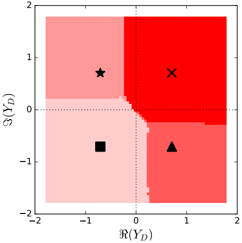
# Quantization and decisions for 4-QAM, $\gamma_D = \gamma_R = 7$ dB, $R \approx 1$



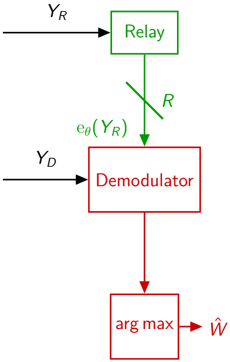

Relay's Compressor



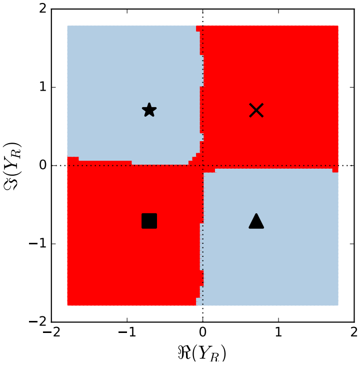

Destination's Demodulator



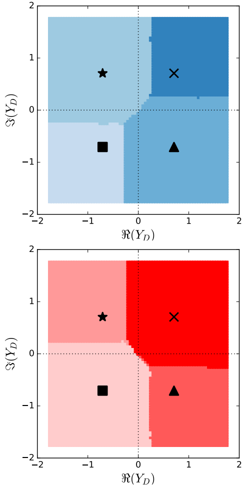
# Quantization and decisions for 4-QAM, $\gamma_D = \gamma_R = 7$ dB, $R \approx 1$




Relay's Compressor




Destination's Demodulator



# Summary on the Detection-Oriented Neural Relays

- End-to-end optimization of relay rate and communication rate
- Proof-of-concept towards **practical** CF schemes
- Final output is a look-up table, learned through neural relays
- Training over a range provides robustness in the SNR

# Summary on the Detection-Oriented Neural Relays

- End-to-end optimization of relay rate and communication rate
- Proof-of-concept towards **practical** CF schemes
- Final output is a look-up table, learned through neural relays
- Training over a range provides robustness in the SNR

**Future directions**:

- Extend to general relay channel (destination decodes the compressed index first)
- Extend to half- and full-duplex channels, including different channel models
- Consider multi-hop networks and MIMO relay channels

# Outline

# Conclusion

We proposed two examples of **task-aware** design of communication systems for **general-purpose networks**:

1. Precoding-oriented CSI feedback
2. Detection-oriented CF relays

Task-aware design = learned neural compression + domain-inspired loss function

# Conclusion

We proposed two examples of **task-aware** design of communication systems for **general-purpose networks**:

1. Precoding-oriented CSI feedback
2. Detection-oriented CF relays

Task-aware design = learned neural compression + domain-inspired loss function

Common directions for further investigation:

- Robustness w.r.t. considered scenario
- Scalability when increasing system dimensions (users, cells, relays, ...)
- Neural network architecture choice and training methodologies

# *Thank you! Q&A?*

**Learned Task-Aware Compression Methods in Communication Systems**

Fabrizio Carpi

Co-Advisors: Prof. Elza Erkip and Prof. Siddharth Garg

NYU TANDON SCHOOL OF ENGINEERING

NYU WIRELESS

**fabrizio.carpi@nyu.edu** **https://fabriziocarpi.github.io/**

# List of Papers

- Channel state information (CSI) feedback [1] [2]
- Compress-and-Forward (CF) relaying [3] [4]

Other topics I have worked on

- PAPR reduction for DFT-s-OFDM systems [5]
- Compression for Hypothesis Testing[6]

[1] F. Carpi, S. Venkatesan, J. Du, H. Viswanathan, S. Garg, E. Erkip, "Precoding-oriented massive MIMO CSI feedback design," ICC 2023
[2] F. Carpi, S. Garg, E. Erkip, "Learned Precoding-Oriented CSI Feedback in Multi-Cell Multi-User MIMO Systems," in preparation
[3] E. Ozyilkan*, F. Carpi*, S. Garg, E. Erkip, "Neural Compress-and-Forward for the Relay Channel," SPAWC 2024
[4] E. Ozyilkan*, F. Carpi*, S. Garg, E. Erkip, "Learning-Based Compress-and-Forward Schemes for the Relay Channel," arxiv 2024
[5] F. Carpi, S.Rostami, J.Cho, S.Garg, E.Erkip, C.Zhang,"Learned Pulse Shaping Design for PAPR Reduction in DFT-s-OFDM,"SPAWC 2024
[6] F. Carpi, S. Garg, E. Erkip, "Single-Shot Compression for Hypothesis Testing," SPAWC 2021