Precoding-oriented Massive MIMO CSI Feedback Design

Fabrizio Carpi[1], Sivarama Venkatesan[2], Jinfeng Du[2],
Harish Viswanathan[2], Siddharth Garg[1], Elza Erkip[1]

[1]New York University, [2]Nokia Bell Labs

IEEE International Conference on Communications (ICC)
28 May – 01 June 2023 — Rome, Italy

Outline

Motivation

- Downlink frequency division duplexing (FDD) MIMO: the channel state information (CSI) has to be fed back on the uplink

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.

Motivation

- Downlink frequency division duplexing (FDD) MIMO: the channel state information (CSI) has to be fed back on the uplink
- Classical CSI compression techniques are not practical for large systems or rely on assumptions on the channel model $\Longrightarrow$ deep learning techniques achieve good performance with fewer assumptions (data-driven)

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.

Motivation

- Downlink frequency division duplexing (FDD) MIMO: the channel state information (CSI) has to be fed back on the uplink
- Classical CSI compression techniques are not practical for large systems or rely on assumptions on the channel model $\Longrightarrow$ deep learning techniques achieve good performance with fewer assumptions (data-driven)

- Most previous works focus on the CSI compression for channel reconstruction
  But... the ultimate goal is to achieve high spectral efficiency!

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.

Motivation

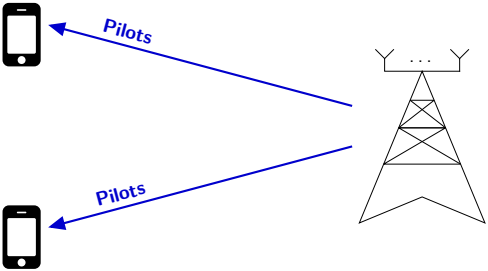
- Downlink frequency division duplexing (FDD) MIMO: the channel state information (CSI) has to be fed back on the uplink
- Classical CSI compression techniques are not practical for large systems or rely on assumptions on the channel model $\Longrightarrow$ deep learning techniques achieve good performance with fewer assumptions (data-driven)

- Most previous works focus on the CSI compression for channel reconstruction
  But... the ultimate goal is to achieve high spectral efficiency!

$\Longrightarrow$ **We focus on task-oriented CSI compression**

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.

Conventional CSI feedback



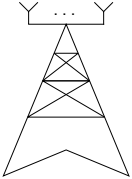
Conventional CSI feedback

1) Channel Estimation

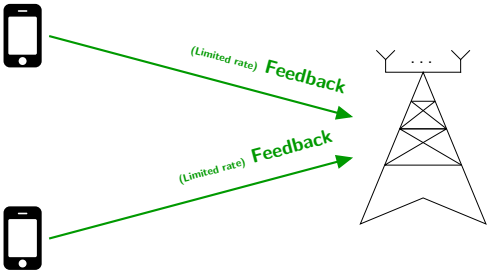
2) CSI compression

1) Channel Estimation

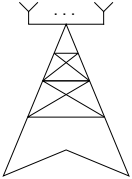
2) CSI compression



Conventional CSI feedback



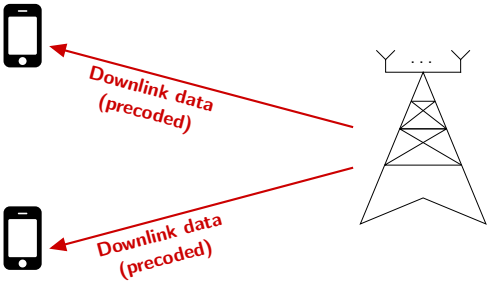
Conventional CSI feedback



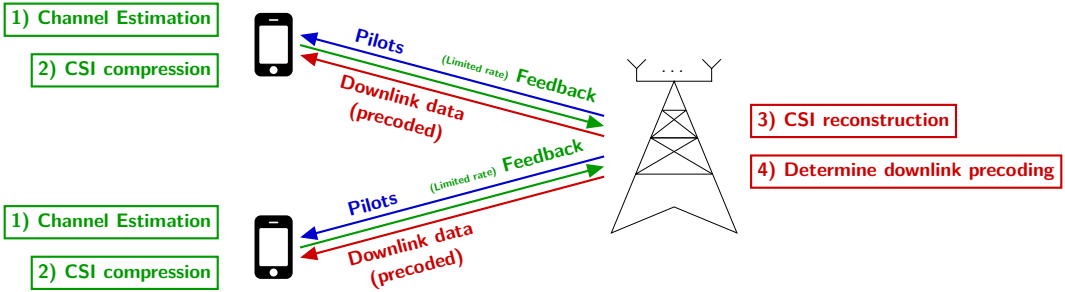
3) CSI reconstruction

4) Determine downlink precoding

Conventional CSI feedback



Conventional CSI feedback

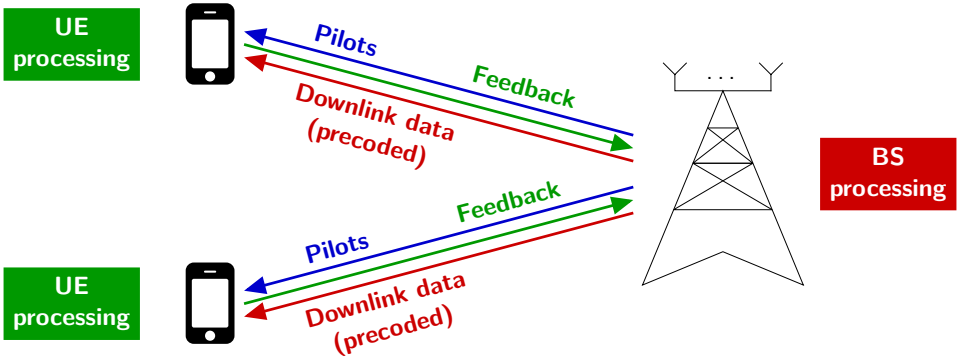


Precoding-oriented CSI feedback

Task: MIMO precoding
Goal: max achievable rate with limited feedback overhead

Precoding-oriented CSI feedback

Task: MIMO precoding
Goal: max achievable rate with limited feedback overhead



System Model: definitions

- One BS with N_t antennas, K single-antennas users
- Downlink transmitted signal with linear precoding:

$$\mathbf{x} = \sum_{k=1}^{K} \mathbf{v}_k s_k = \mathbf{V}\mathbf{s}$$

- Received signal at k-th user:

$$y_k = \mathbf{h}_k^H \mathbf{v}_k s_k + \sum_{j \neq k} \mathbf{h}_k^H \mathbf{v}_j s_j + z_k$$

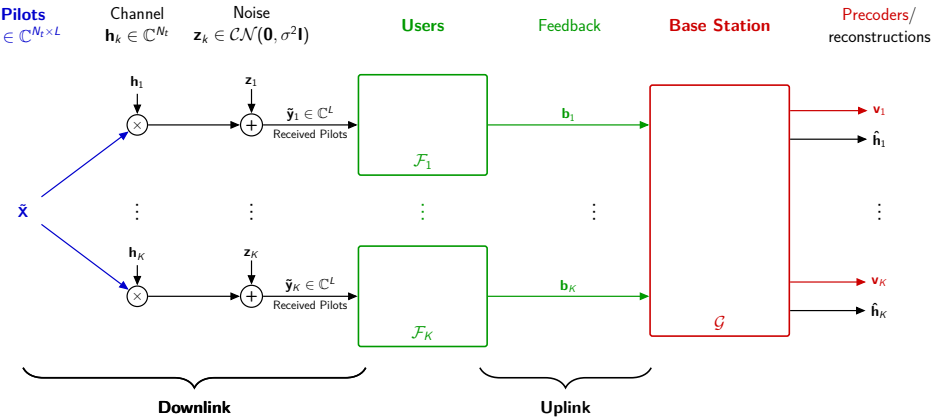
- **Metric**: sum of achievable rates:

$$R = \sum_{k=1}^{K} R_k = \sum_{k=1}^{K} \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{v}_k|^2}{\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{v}_j|^2 + \sigma^2} \right) \quad (1)$$

System Model: block diagram

We use neural networks to design pilots, feedback scheme $\mathcal{F}$, BS processing $\mathcal{G}$

Goal: max achievable rate with limited feedback overhead



Previous work

Ref	Task-oriented?	Feedback overhead optimization?
[1]	Yes, precoding	No, overhead is fixed with the architecture
[2]	No, channel reconstruction	Yes, Lagrangian loss and entropy coding [3, 4]
Ours	Yes, precoding	Yes, Lagrangian loss and entropy coding [3, 4]

- [3, 4]: image compression with neural networks (autoencoder), the loss function includes a tradeoff between feedback overhead (rate) and image reconstruction performance

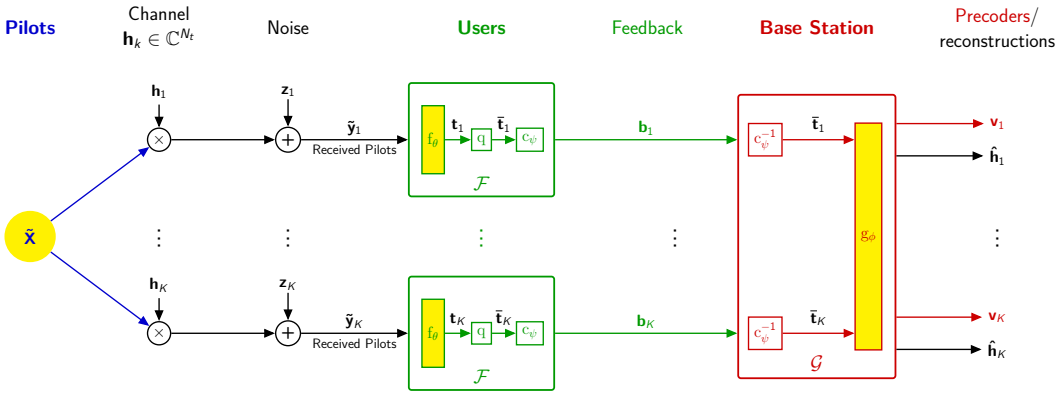
[1] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4044–4057, 2021.

[2] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2621–2633, 2021.

[3] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017.

[4] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

Neural Networks



Neural networks

Pilots $\tilde{\mathbf{X}} \in \mathbb{C}^{N_t \times L}$

- Fully-connected (FC) layer with linear activation and zero bias
- Power constraint: $\|\tilde{\mathbf{x}}_\ell\|_2^2 = P$

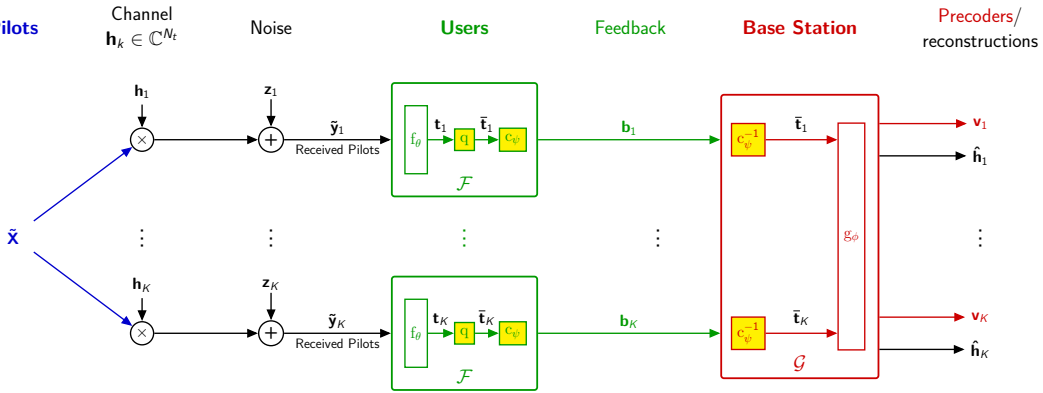
User network f_θ

- DNN with 4 FC layers
- Extracts *semantic features* $\mathbf{t}_k$ from the noisy pilots (per user)

BS network g_ϕ

- DNN with 5 FC layers
- Determines the K precoders from the K semantic features
- Power constraint: $\mathrm{Tr}(\mathbf{V}\mathbf{V}^H) = P$

Feedback Overhead Optimization



Feedback Overhead Optimization

Feedback quantization: DNN output to bitstream — from [Ballé et al., 2017, 2018]

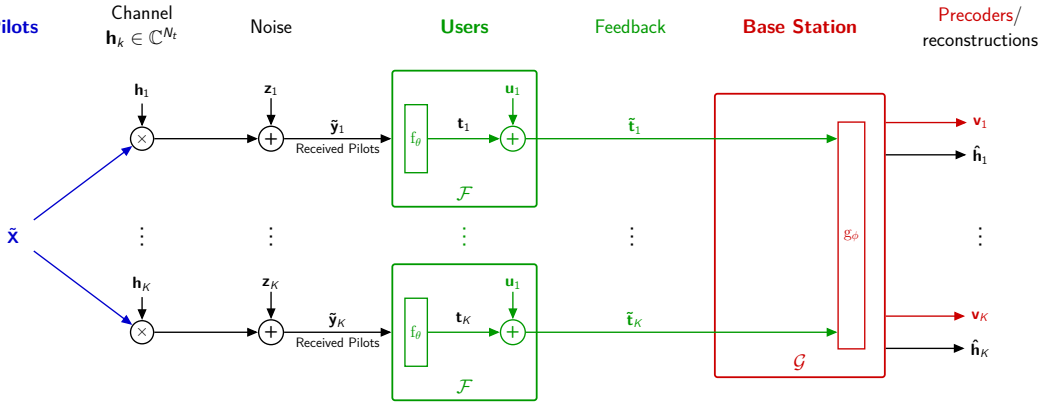
During test time:

- Quantizer q: scalar quantization to closest integer $\Longrightarrow$ quantized features $\bar{\mathbf{t}}_k$
- Entropy coder c_ψ: converts $\bar{\mathbf{t}}_k$ into bit streams $\mathbf{b}_k$
- Entropy decoder c_ψ^{-1}: (lossless) reconstruction $\mathrm{c}_\psi^{-1}(\mathbf{b}_k) = \bar{\mathbf{t}}_k$
- ψ: parameters learned during training $\Longrightarrow \sim$ probability distribution of $\bar{\mathbf{t}}_k$

During training:

- Simulate quantization noise: add $u_k^1, \ldots, u_k^{N_b} \sim \mathcal{U}[-0.5, +0.5]$
- Pseudo-quantized features: $\tilde{\mathbf{t}}_k = \mathbf{t}_k + \mathbf{u}_k$
- Note: the probability distribution of $\tilde{\mathbf{t}}_k$ is a continuous relaxation of the one of $\bar{\mathbf{t}}_k$

Loss Function



$$\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$$

$$\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$$
$$\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_K]$$

Loss Function

Loss function including three possible metrics:

$$\mathcal{L}(\theta, \phi, \psi) = \mathcal{O} - \lambda\mathcal{R} + \gamma\mathcal{D}, \tag{2}$$

- Feedback overhead $\mathcal{O}$: entropy (rate) of the pseudo-quantized features $\tilde{\mathbf{t}}_k$
- Performance $\mathcal{R}$: achievable rates with precoding $\mathbf{V}$
- Distortion $\mathcal{D}$: reconstruction loss when estimating channels $\hat{\mathbf{H}}$
- λ and γ determine the tradeoff between the three components

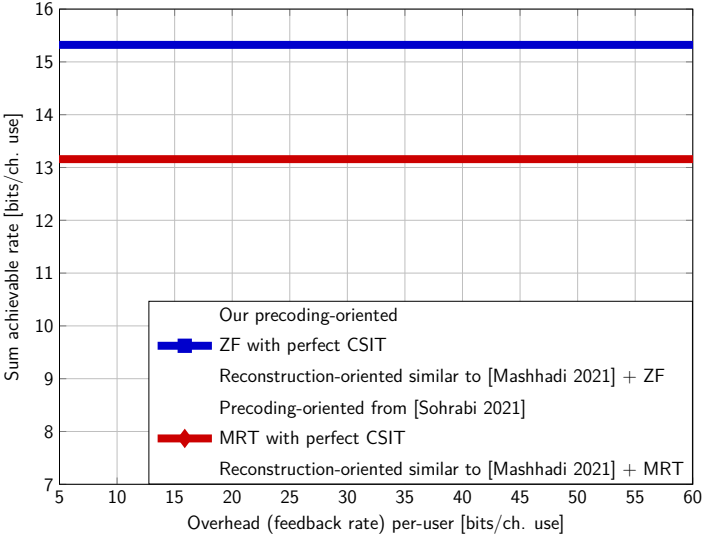
Loss Function

Loss function including three possible metrics:

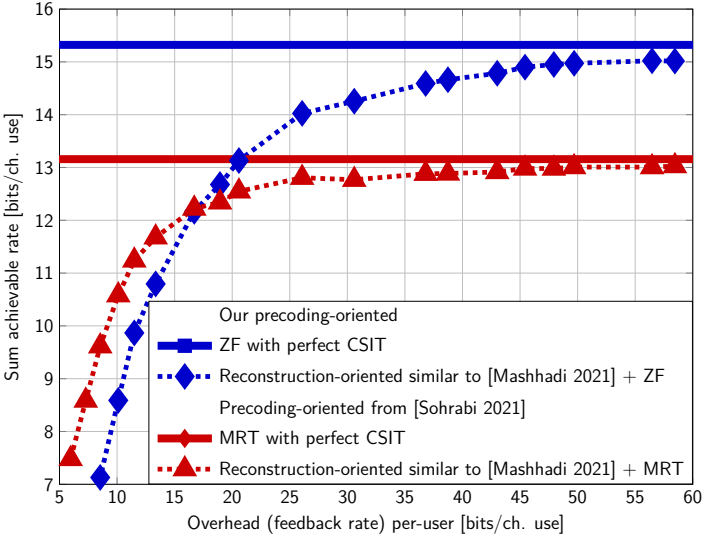
$$\mathcal{L}(\theta, \phi, \psi) = \mathcal{O} - \lambda \mathcal{R} + \gamma \mathcal{D}, \tag{2}$$

- Feedback overhead $\mathcal{O}$: entropy (rate) of the pseudo-quantized features $\tilde{\mathbf{t}}_k$
  $\mathcal{O}_k(\theta, \psi) = \mathbb{E}_{\mathbf{h}_k, \mathbf{u}_k, \mathbf{z}_k} \left[-\log p_{\tilde{\mathbf{t}}}(\tilde{\mathbf{t}}_k; \psi) \right]$
- Performance $\mathcal{R}$: achievable rates with precoding $\mathbf{V}$
  $\mathcal{R}_k(\theta, \psi, \phi) = \mathbb{E}_{\mathbf{h}_k, \mathbf{U}, \mathbf{Z}} \log_2 \left(1 + \frac{|\mathbf{h}_k^H \mathbf{v}_k|^2}{\sum_{j \neq k} |\mathbf{h}_k^H \mathbf{v}_j|^2 + \sigma^2} \right)$
- Distortion $\mathcal{D}$: reconstruction loss when estimating channels $\hat{\mathbf{H}}$
  $\mathcal{D}(\theta, \psi, \phi) = \mathbb{E}_{\mathbf{H}, \mathbf{U}, \mathbf{Z}} \left\| \mathbf{H} - \hat{\mathbf{H}} \right\|_2^2$
- λ and γ determine the tradeoff between the three components

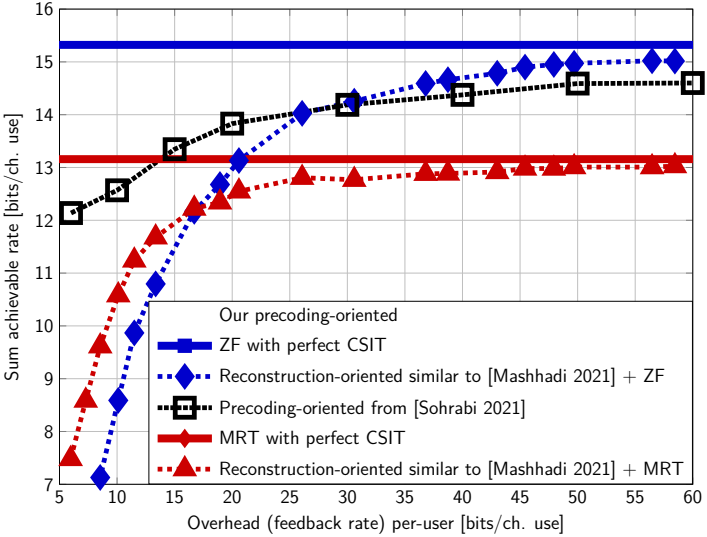
Results: 2 users, 64 TX antennas, 2-path channel, 8 pilots, SNR = 10dB



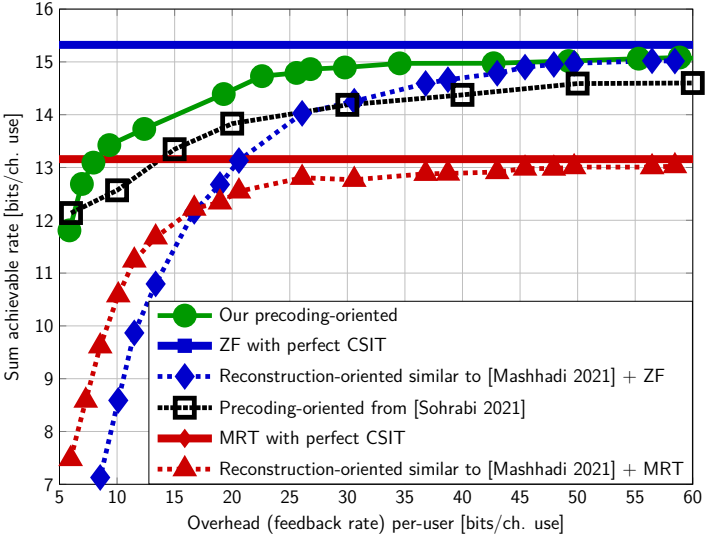
Results: 2 users, 64 TX antennas, 2-path channel, 8 pilots, SNR = 10dB



Results: 2 users, 64 TX antennas, 2-path channel, 8 pilots, SNR = 10dB



Results: 2 users, 64 TX antennas, 2-path channel, 8 pilots, SNR = 10dB



Conclusion

- Analysis of the tradeoff between feedback overhead, system performance, and channel reconstruction
- Focus on performance as the final metric for our task (MIMO precoding)
- Precoding-oriented works best for the small overhead regime
  - Optimized for overhead-performance
  - Good news for beyond 5G! Precoding-oriented may unlock the potential of *extreme* massive MIMO arrays
  - Moreover: integration with ML/AI is on the table for the next 3GPP releases
- If the feedback overhead budget is large, then all methods are equivalent

Thank you! Q&A?

Precoding-oriented Massive MIMO CSI Feedback Design

Fabrizio Carpi[1], Sivarama Venkatesan[2], Jinfeng Du[2],
Harish Viswanathan[2], Siddharth Garg[1], Elza Erkip[1]

fabrizio.carpi@nyu.edu

NYU TANDON SCHOOL OF ENGINEERING

NYU WIRELESS

NOKIA BELL LABS

https://fabriziocarpi.github.io/

References

[1] F. Sohrabi, K. M. Attiah, and W. Yu, "Deep learning for distributed channel feedback and multiuser precoding in FDD massive MIMO," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4044–4057, 2021.

[2] M. B. Mashhadi, Q. Yang, and D. Gündüz, "Distributed deep convolutional compression for massive MIMO CSI feedback," *IEEE Transactions on Wireless Communications*, vol. 20, no. 4, pp. 2621–2633, 2021.

[3] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017.

[4] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.

Simulation setup

Channel model: multipath, BS with uniform linear array.

Compare with:

- MRT/ZF with CSIT: full CSI $\mathbf{H}$ is available at the transmitter (BS), then compute maximal-ratio transmission (maximize per-user received power) or zero-forcing (minimize inter-user interference)
- Precoder-oriented from [1]: similar setup, but with a different feedback overhead optimization. The loss function is the performance and the feedback overhead is determined by the number of feedback taps (fixed).
- Reconstruction-oriented similar to [2] + MRT/ZF: system trained for overhead-distortion loss, followed by MRT/ZF (non precoding-oriented)