Efficient Deep Learning for Massive MIMO Channel State Estimation

Mason del Rosario

Doctoral Qualifying Examination

May 2021

Outline

Background

Why MIMO?

Massive MIMO is a key enabling technology for future wireless communications networks.

- 5G, Ultra-Dense Networks, IoT

S. Marek, "Sprint Spent $1B on Massive MIMO for Its 5G Network in Q2," *SDxCentral*, `https://www.sdxcentral.com/articles/news/sprint-spent-1b-on-massive-mimo-for-its-5g-network-in-q2/2018/06/`. Accessed: Feb 22, 2020.

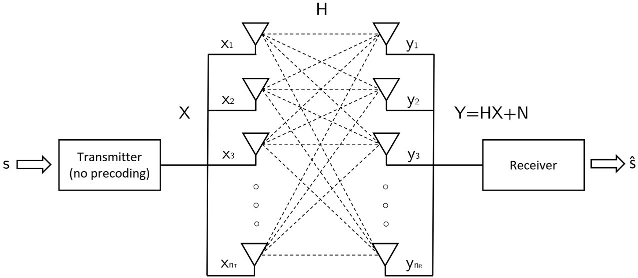
Massive MIMO is a key enabling technology for future wireless communications networks.

- 5G, Ultra-Dense Networks, IoT

The efficacy of MIMO depends on accurate *Channel State Information (CSI)*.

S. Marek, "Sprint Spent $1B on Massive MIMO for Its 5G Network in Q2," *SDxCentral*, `https://www.sdxcentral.com/articles/news/sprint-spent-1b-on-massive-mimo-for-its-5g-network-in-q2/2018/06/`. Accessed: Feb 22, 2020.

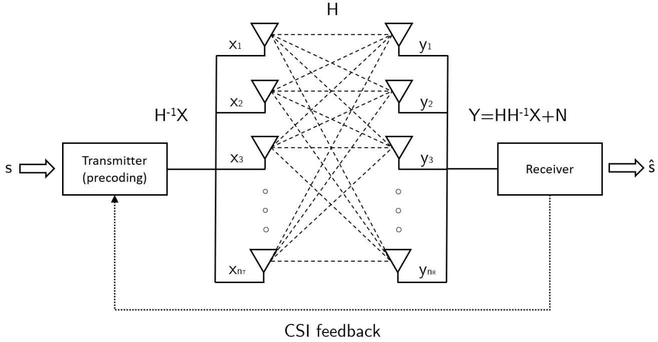
Massive MIMO uses numerous antennas to endow transceivers with spatial diversity.



The fading coefficients between each set of Tx/Rx antennas constitute **Channel State Information (CSI)**, $\mathbf{H}$. For n_T, n_R antennas,

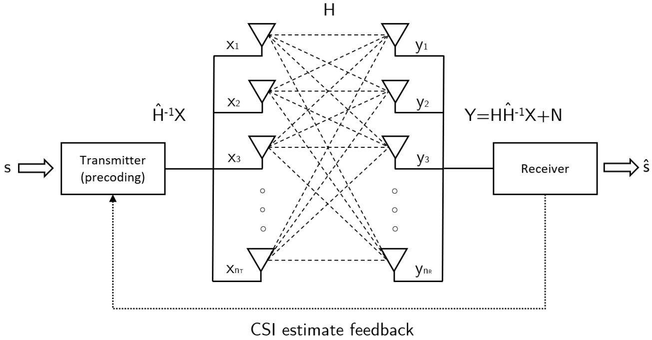
$$\mathbf{H} = \begin{bmatrix} h_{1,1} & h_{1,2} & \dots & h_{1,n_T} \\ h_{2,1} & h_{2,2} & \dots & h_{2,n_T} \\ \vdots & \vdots & \vdots & \vdots \\ h_{n_R,1} & h_{n_R,2} & \dots & h_{n_R,n_T} \end{bmatrix}$$

Perfect CSI (i.e., exact knowledge of the channel, **H**) allows us to maximize the power of the received symbol by precoding.



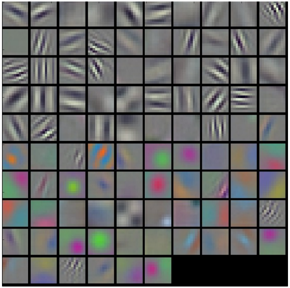
MIMO and CSI Estimation

However, transmitting $\mathbf{H}$ is costly. Instead, generate **CSI Estimates**, $\hat{\mathbf{H}}$, based on **compressed feedback**.



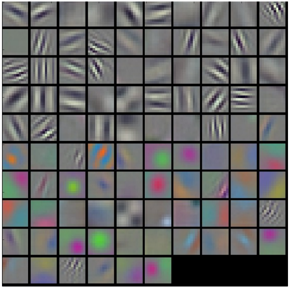
Goal: Find low-dimensional representation, feed back to transmitter for recovery of $\hat{\mathbf{H}}$ which is an accurate approximation of $\mathbf{H}$ in MSE sense.

- CNNs = state-of-the art performance in image processing applications
- Capable of extracting features from 2D, grid-like data



A. Karpathy, “Visualizing What ConvNets Learn," `http://cs231n.github.io/understanding-cnn/`. Accessed: Feb 24, 2020.

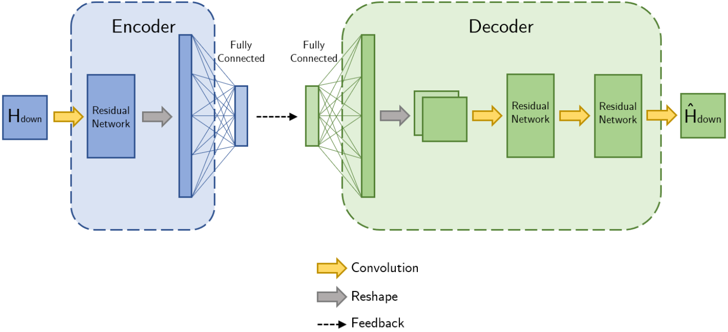
- CNNs = state-of-the art performance in image processing applications
- Capable of extracting features from 2D, grid-like data



- **Recently, CNNs applied to CSI estimation**

A. Karpathy, "Visualizing What ConvNets Learn," `http://cs231n.github.io/understanding-cnn/`. Accessed: Feb 24, 2020.

- CNN-based autoencoder for learned CSI compression and feedback [1]



C. Wen, W. Shih, and S. Jin, "Deep learning for massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 7, pp. 748–751, Oct 2018

Spherical Normalization

Power-based normalization for improved CSI reconstruction accuracy.

Most works perform **minmax scaling** – Take the extrema $(\mathbf{H}_{\text{min}}, \mathbf{H}_{\text{max}})$ of the real and imaginary channels,

$$\mathbf{H}_{n,\text{minmax}}(i,j) = \frac{\mathbf{H}_n(i,j) - \mathbf{H}_{\text{min}}}{\mathbf{H}_{\text{max}} - \mathbf{H}_{\text{min}}} \in [0,1],$$

for $n \in [1, \ldots, N]$ given N samples and i, j indexing rows/columns of CSI matrices.

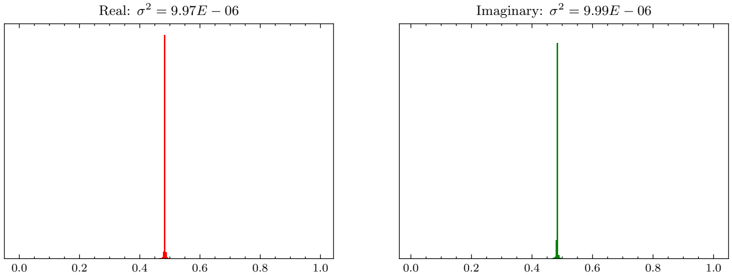


Figure: Distribution and variance of minmax-normalized COST2100 real/imaginary channels ($N = 99000$) images.

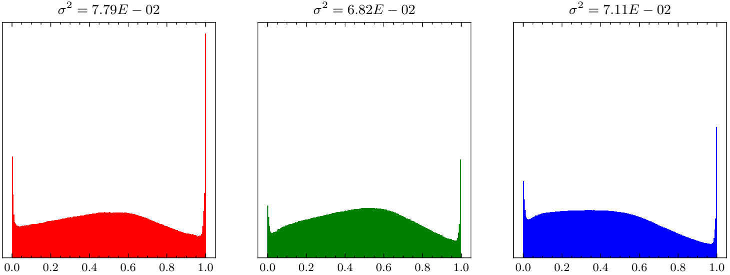


Figure: Distribution and variance of minmax-normalized ImageNet color channels ($N = 50000$) images.

Spherical normalization – scale each channel sample by its power,

$$\check{\mathbf{H}}_d^n = \frac{\mathbf{H}_d^n}{\|\mathbf{H}_d^n\|_2}. \tag{1}$$

After applying (1) to each sample, minmax scaling is applied to the entire dataset.

The resulting dataset under spherical normalization can exhibits a larger variance than the same dataset under minmax scaling.

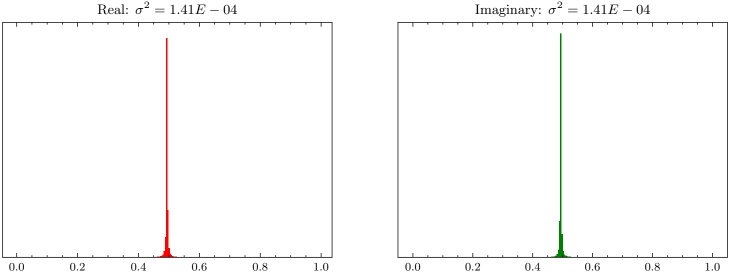


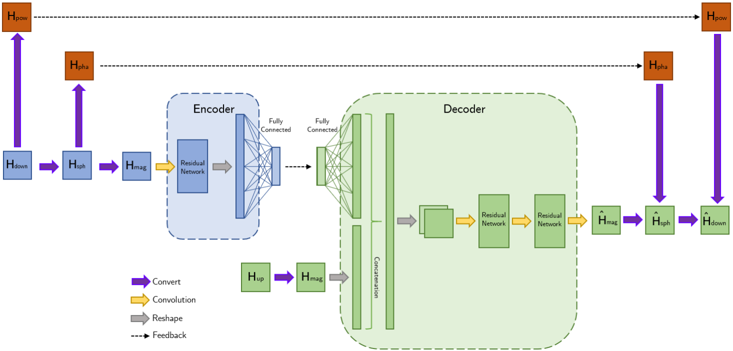
Figure: Distribution and variance of COST2100 real/imaginary channels under spherical normalization ($N = 99000$) images.

Under spherical normalization, MSE loss becomes equivalent to NMSE. Recall the definitions,

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{N} \|\mathbf{H}_k - \hat{\mathbf{H}}_k\|^2, \quad \mathrm{NMSE} = \frac{1}{N}\sum_{k=1}^{N} \frac{\|\mathbf{H}_k - \hat{\mathbf{H}}_k\|^2}{\|\mathbf{H}_k\|^2}$$

The MSE of the spherically normalized estimator is equivalent to the NMSE of the regular estimator, i.e.

$$\begin{aligned} \mathrm{MSE}_{\mathrm{Sph}} &= \frac{1}{N}\sum_{k=1}^{N} \|\check{\mathbf{H}}_k - \hat{\check{\mathbf{H}}}_k\|^2 \\ &= \frac{1}{N}\sum_{k=1}^{N} \left\| \frac{\mathbf{H}_k}{\|\mathbf{H}_k\|^2} - \frac{\hat{\mathbf{H}}_k}{\|\mathbf{H}_k\|^2} \right\|^2 \\ &= \frac{1}{N}\sum_{k=1}^{N} \frac{\|\mathbf{H}_k - \hat{\mathbf{H}}_k\|^2}{\|\mathbf{H}_k\|^2} \ \square. \end{aligned}$$



Hpow
Hpha
Encoder
Decoder
Fully Connected
Fully Connected
Hdown
Hsph
Hmag
Residual Network
Concatenation
Residual Network
Residual Network
Ĥmag
Ĥsph
Ĥdown
Hpha
Hpow
Hup
Hmag
Convert
Convolution
Reshape
Feedback

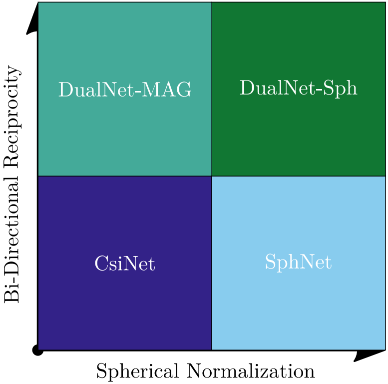


Figure: Illustration of techniques used in different models.

Z. Liu, **M. del Rosario**, X. Liang, L. Zhang, and Z. Ding, "Spherical normalization for learned compressive feedback in massive mimo csi acquisition," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2020

Two MIMO scenarios using COST 2100 model with 32 antennas at gNB and single UE (single antenna), 1024 subcarriers.

1. **Indoor** environment using 5.3GHz, 0.1 m/s UE mobility, square area of length 20m
2. **Outdoor** environment using 300MHz, 1 m/s UE mobility, square area of length 400m

Dataset: 10^5 channel samples – 70%/30% training/test split.

Hyperparameters: Adam optimizer with learning rate 10^{-3}, batch size 200, 1000 epochs, MSE loss

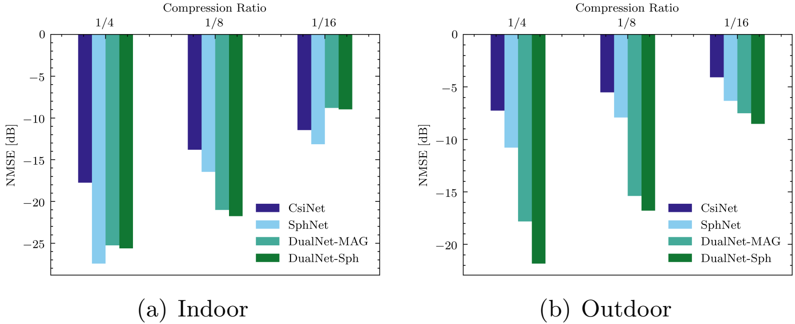


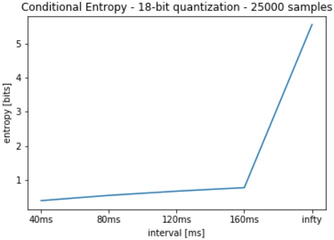
Figure: NMSE (lower is better) comparison of bidirectional reciprocity and spherical normalization against CsiNet for increasing compression ratio [2]

Z. Liu, **M. del Rosario**, X. Liang, L. Zhang, and Z. Ding, "Spherical normalization for learned compressive feedback in massive mimo csi acquisition," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2020

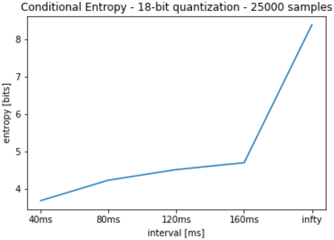
MarkovNet

A deep differential autoencoder for efficient temporal learning.

CSI estimation techniques benefit from temporal information.



(a) Indoor



(b) Outdoor

Figure: Conditional entropy between CSI matrices for different feedback intervals. COST2100 model used for (a) Indoor and (b) Outdoor network

Recurrent neural networks (RNNs) contain trainable long short-term memory (LSTM) cells which learn temporal relationships.

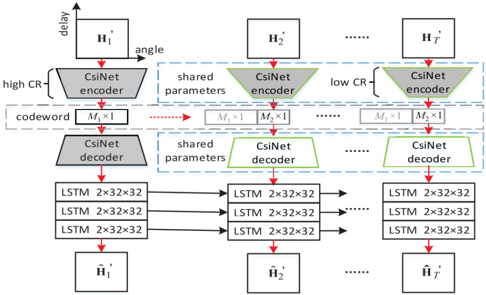


Figure: CsiNet-LSTM network architecture [3].

T. Wang, C. Wen, S. Jin, and G. Y. Li, "Deep Learning-Based CSI Feedback Approach for Time-Varying Massive MIMO Channels," *IEEE Wireless Comm. Letters*, vol. 8, pp. 416–419, April 2019

The number of parameters/FLOPs for RNNs is large.

Table: Model size and computational complexity of CsiNet-LSTM and CsiNet. M: million.

	Parameters		**FLOPs**	
CR	**CsiNet-LSTM**	**CsiNet**	**CsiNet-LSTM**	**CsiNet**
1/4	132.7 M	2.1 M	412.9 M	7.8 M
1/8	123.2 M	1.1 M	410.8 M	5.7 M
1/16	118.5 M	0.5 M	409.8 M	4.7 M
1/32	116.1 M	0.3 M	409.2 M	4.1 M
1/64	115.0 M	0.1 M	409.0 M	3.9 M

T. Wang, C. Wen, S. Jin, and G. Y. Li, "Deep Learning-Based CSI Feedback Approach for Time-Varying Massive MIMO Channels," *IEEE Wireless Comm. Letters*, vol. 8, pp. 416–419, April 2019

Instead of learning a temporal dependency across multiple timeslots, we proposed a one-step differential encoder.

For short enough time intervals between t and $t-1$, we view CSI data as a Markov chain, i.e.

$$\mathbf{H}_t = \gamma \mathbf{H}_{t-1} + \mathbf{V}_t,$$

with $\gamma \in \mathbb{R}^+$ and i.i.d $\mathbf{V}_t$ such that $\mathbf{V}_t \sim \mathcal{CN}(\mathbf{0}, \Sigma_V)$.

Z. Liu †, **M. del Rosario †**, and Z. Ding, "A Markovian Model-Driven Deep Learning Framework for Massive MIMO CSI Feedback," *arXiv e-prints*, Sept. 2020. Submitted to IEEE Transactions on Wireless Communications

The ordinary least-squares solution, γ, is given as

$$\gamma = \frac{\mathrm{Trace}(\mathbb{E}\left[\mathbf{H}_{t-1}^H \mathbf{H}_t\right])}{\mathbb{E}\|\mathbf{H}_t^H \mathbf{H}_t\|^2}.$$

We utilize the estimator, $\hat{\gamma}$, based on the second-order statistics of the CSI matrices,

$$\hat{\gamma} = \frac{\sum_{i=1}^N \mathrm{Trace}(\left[\mathbf{H}_{t-1}^H(i)\mathbf{H}_t(i)\right])}{\sum_{i=1}^N \|\mathbf{H}_t^H(i)\mathbf{H}_t(i)\|^2},$$

for training set of size N.

With the one-step estimator $\hat{\gamma}$, we propose train an encoder for the estimation error as

$$\mathbf{s}_t = f_{e,t}(\mathbf{H}_t - \hat{\gamma}\hat{\mathbf{H}}_{t-1}),$$

and we jointly train a decoder,

$$\hat{\mathbf{H}}_t = f_{d,t}(\mathbf{s}_t) + \hat{\gamma}\hat{\mathbf{H}}_{t-1}$$

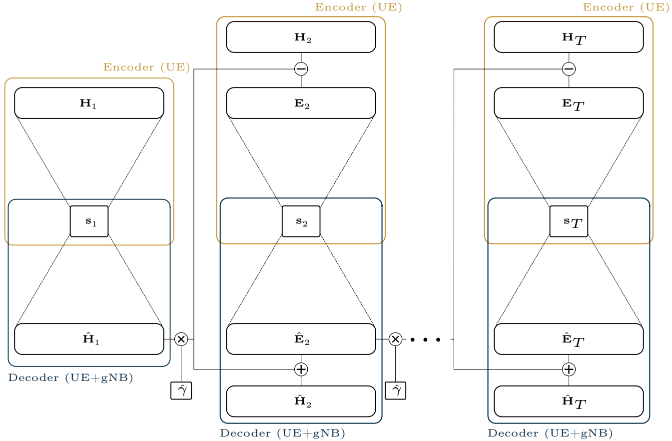
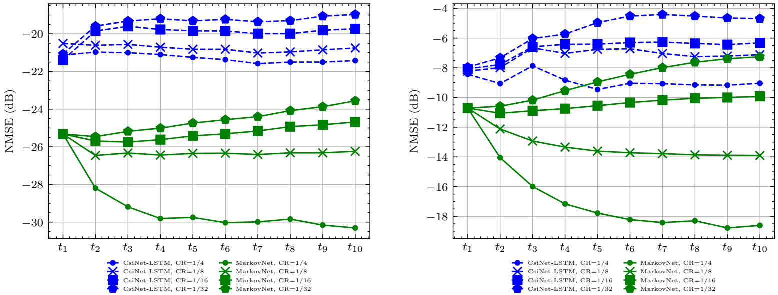


Figure: Abstract architecture for MarkovNet. Networks at $t \geq 2$ are trained to predict the estimation error, $\mathbf{E}_t$.

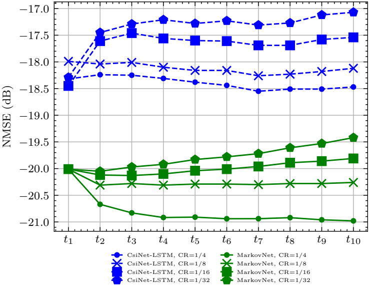
MarkovNet: Results ($\text{NMSE}_{\text{truncated}}$)



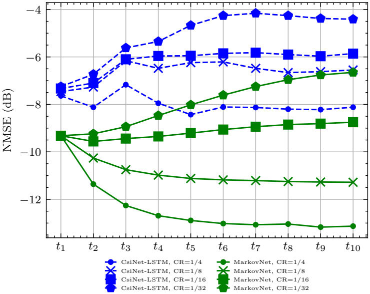
(a) Indoor (b) Outdoor

Figure: $\text{NMSE}_{\text{truncated}}$ comparison of MarkovNet and CsiNet-LSTM at various compression ratios (CR).

MarkovNet: Results ($\mathrm{NMSE_{all}}$)



(a) Indoor



(b) Outdoor

Figure: $\mathrm{NMSE_{all}}$ comparison of MarkovNet and CsiNet-LSTM at various compression ratios (CR).

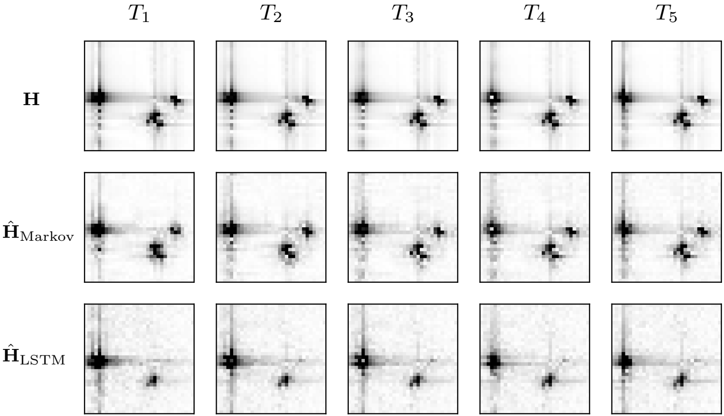


Figure: Ground truth CSI ($\mathbf{H}$), MarkovNet estimates ($\hat{\mathbf{H}}_{\text{Markov}}$), and CsiNet-LSTM estimates ($\hat{\mathbf{H}}_{\text{LSTM}}$) for five timeslots ($T_1$ through T_5) on one outdoor sample from the test set (both networks at CR $= \frac{1}{4}$).

SphNet-Quant

An end-to-end trained autoencoder with learned feedback quantization.

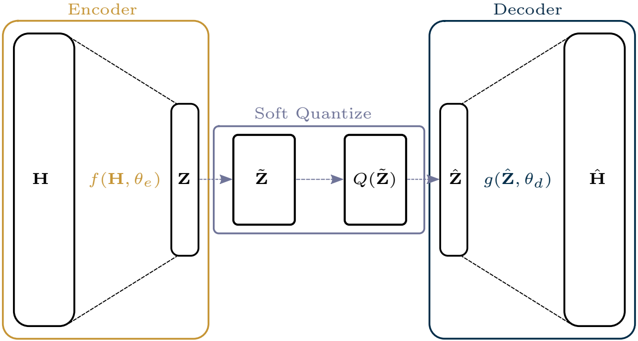


Figure: Abstract architecture for CsiNet-Quant. SoftQuantize layer ($Q(\tilde{\mathbf{Z}})$) is a continuous, softmax-based relaxation of a d-dimensional quantization of the latent layer $\mathbf{Z}$.

Define the m-dimensional codebook of size L as $\mathbf{C} \in \mathbb{R}^{m\times L}$. The soft assignments of the j-th latent vector $\tilde{\mathbf{z}}_j$ can be written as,

$$\phi(\tilde{\mathbf{z}}_j) = \left[\frac{\exp(-\sigma\|\tilde{\mathbf{z}}_j - \mathbf{c}_\ell\|^2)}{\sum_{i=1}^{L}\exp(-\sigma\|\tilde{\mathbf{z}}_j - \mathbf{c}_i\|^2)}\right]_{\ell\in[L]} \in \mathbb{R}^L, \qquad (2)$$

which is referred to as the ‘softmax’ function. σ is a *temperature* or *annealing* parameter which controls the degree of quantization,

$$\lim_{\sigma\to\infty}\phi(\tilde{\mathbf{z}}_j) = \mathrm{onehot}(\tilde{\mathbf{z}}_j) = \begin{cases} 1 & \ell = \operatorname*{argmax}_{\ell}\, \phi(\tilde{\mathbf{z}}_j)[\ell] \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

The soft assignments ϕ admit probability masses over the codewords,

$$q_j = \phi(\tilde{\mathbf{z}}).$$

Based on finite samples, we define the histogram probability estimates p_j

$$p_j = \frac{|\{e_l(\mathbf{z}_i) | l \in [m], i \in [N], e_l(\mathbf{z}_i) = j\}|}{mN}.$$

Our target for the rate loss is the crossentropy between p_j and q_j term,

$$H(\phi) := H(p, q) = -\sum_{j=1}^{L} p_j \log q_j = H(p) + D_{\mathrm{KL}}(p \| q).$$

Rate-Distortion Loss

Loss function for soft quantization = regularized rate-distortion function.

$$\underset{\theta_e,\theta_d,\mathbf{C}}{\operatorname{argmin}}\ L_d(\mathbf{H},\hat{\mathbf{H}}) + \lambda L_{\ell^2}(\theta_e,\theta_d,\mathbf{C}) + \beta L_r(\theta_e,\mathbf{C}) \tag{4}$$

Where the different loss terms are

Term	**Definition**	**Description**
$L_d(\mathbf{H},\hat{\mathbf{H}})$	$\frac{1}{N}\sum_{i=1}^{N} \lVert \mathbf{H}_i - g(Q(f(\mathbf{H}_i,\theta_e),\mathbf{C}),\theta_d) \rVert^2$	distortion loss
$L_{\ell^2}(\theta_e,\theta_d,\mathbf{C})$	$\lVert\theta_e\rVert^2 + \lVert\theta_d\rVert^2 + \lVert\mathbf{C}\rVert^2$	ℓ^2 penalty
$L_r(\theta_e,\mathbf{C})$	$m\beta H(\phi)$	rate loss

Results: Rate-Distortion (Outdoor)

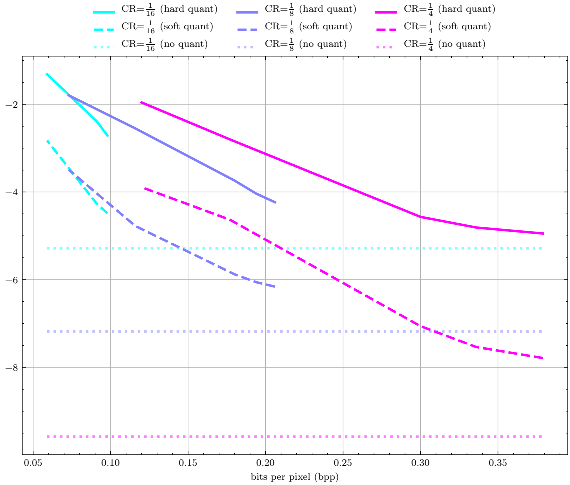


Figure: Rate distortion of CsiNet-Quant under minmax normalization using: $L = 1024$ centers, $d = 4$.

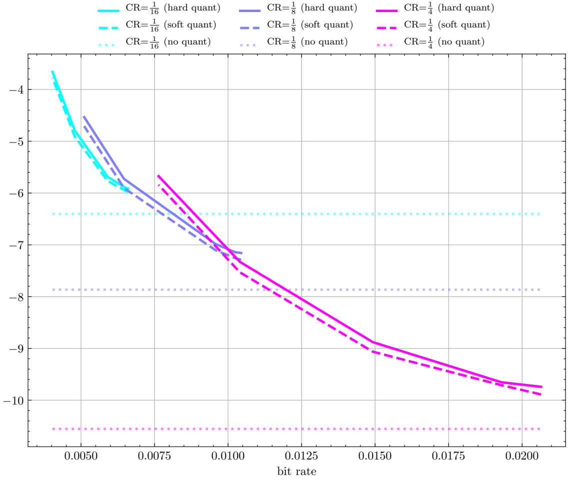


Figure: Rate distortion of SphNet-Quant using: $L = 1024$ centers, CR= $\frac{1}{4}$, $d = 4$. Bit rates are realized under arithmetic coding of quantized features.

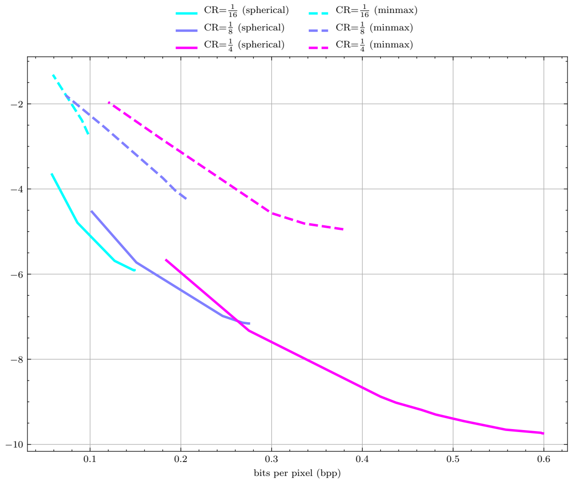


Figure: Rate distortion of CsiNet-Quant under both minmax (dotted line) and spherical (solid line) normalization using: $L = 1024$ centers, $d = 4$. Hard quantization performance shown for each CR.

Questions?

mdelrosa@ucdavis.edu

References

[1] C. Wen, W. Shih, and S. Jin, "Deep learning for massive mimo csi feedback," *IEEE Wireless Communications Letters*, vol. 7, pp. 748–751, Oct 2018.

[2] Z. Liu, **M. del Rosario**, X. Liang, L. Zhang, and Z. Ding, "Spherical normalization for learned compressive feedback in massive mimo csi acquisition," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6, 2020.

[3] T. Wang, C. Wen, S. Jin, and G. Y. Li, "Deep Learning-Based CSI Feedback Approach for Time-Varying Massive MIMO Channels," *IEEE Wireless Comm. Letters*, vol. 8, pp. 416–419, April 2019.

[4] Z. Liu †, **M. del Rosario †**, and Z. Ding, "A Markovian Model-Driven Deep Learning Framework for Massive MIMO CSI Feedback," *arXiv e-prints*, Sept. 2020.

Submitted to IEEE Transactions on Wireless Communications.

[5] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[6] G. Ver Steeg, G. Ballabio, E. Sennesh, M. Rebo, and D. Ulianych, "Non-parametric entropy estimation (npeet)," October 2019.

†→ equal contribution

Appendix

Rather than scalar $\hat{\gamma} \in \mathbb{R}^+$, we can derive a multivariate p-step predictor, $\mathbf{W}_1, \ldots, \mathbf{W}_p$. Given p prior CSI samples, the mean-square optimal predictor $\hat{H}_t$ is a linear combination of these the prior CSI samples,

$$\hat{\mathbf{H}}_t = \mathbf{H}_{t-1}\mathbf{W}_1 + \cdots + \mathbf{H}_{t-p}\mathbf{W}_p + \mathbf{E}_t. \tag{5}$$

Error terms are uncorrelated with the CSI samples (i.e. $\mathbf{H}_{t-i}^{H}\mathbf{E}_t = 0$ for all $i \in [0, \ldots, p]$), and we pre-multiply by $\mathbf{H}_{t-i}^{H}$,

$$\begin{aligned} \mathbf{H}_{t-i}^{H}\hat{\mathbf{H}}_t &= \mathbf{H}_{t-i}^{H}\mathbf{H}_{t-1}\mathbf{W}_1 + \cdots + \mathbf{H}_{t-i}^{H}\mathbf{H}_{t-p}\mathbf{W}_p + \mathbf{H}_{t-i}^{H}\mathbf{E}_t \\ &= \mathbf{H}_{t-i}^{H}\mathbf{H}_{t-1}\mathbf{W}_1 + \cdots + \mathbf{H}_{t-i}^{H}\mathbf{H}_{t-p}\mathbf{W}_p. \end{aligned} \tag{6}$$

Denote the correlation matrix $\mathbf{R}_i = \mathbb{E}[\mathbf{H}_{t-i}^H \mathbf{H}_t]$. We presume the CSI matrices are generated by a stationary process, and consequently, they have the following properties:

1. $\mathbf{R}_i = \mathbb{E}[\mathbf{H}_{t-i}^H \mathbf{H}_t] = \mathbb{E}[\mathbf{H}_t^H \mathbf{H}_{t+i}]$
2. $\mathbf{R}_i = \mathbf{R}_{-i}^H$

Taking the expectation, we can write (6) as a linear combination of correlation matrices,

$$\mathbf{R}_{i+1} = \mathbf{R}_i \mathbf{W}_1 + \cdots + \mathbf{R}_{i-p+1} \mathbf{W}_p.$$

For p CSI samples, we can write a system of p equations, which admits the following,

$$\begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \dots \\ \mathbf{R}_p \end{bmatrix} = \begin{bmatrix} \mathbf{R}_0 & \mathbf{R}_1^H & \dots & \mathbf{R}_{p-1}^H \\ \mathbf{R}_1 & \mathbf{R}_0 & \dots & \mathbf{R}_{p-2}^H \\ \vdots & & \ddots & \vdots \\ \mathbf{R}_{p-1} & \mathbf{R}_{p-2} & \dots & \mathbf{R}_0 \end{bmatrix} \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \dots \\ \mathbf{W}_p \end{bmatrix}.$$

Solving for the coefficient matrices admits the solution

$$\begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \dots \\ \mathbf{W}_p \end{bmatrix} = \begin{bmatrix} \mathbf{R}_0 & \mathbf{R}_1^H & \dots & \mathbf{R}_{p-1}^H \\ \mathbf{R}_1 & \mathbf{R}_0 & \dots & \mathbf{R}_{p-2}^H \\ \vdots & & \ddots & \vdots \\ \mathbf{R}_{p-1} & \mathbf{R}_{p-2} & \dots & \mathbf{R}_0 \end{bmatrix}^+ \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \\ \dots \\ \mathbf{R}_p \end{bmatrix}, \tag{7}$$

where $[\cdot]^+$ denotes the Moore-Penrose pseudoinverse.

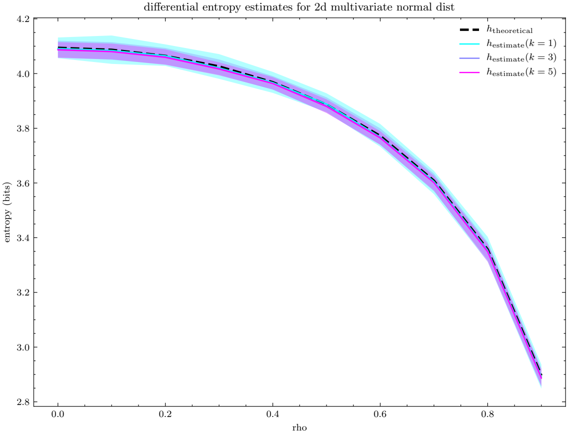


Figure: Differential entropy and estimates for 2d multivariate normal distribution. Estimates are based on the KL estimator [5] using the NPEET library [6].