THE UNIVERSITY OF TEXAS AT AUSTIN
WHAT STARTS HERE CHANGES THE WORLD

6G @ UT
WNCG

High Dimensional 6G Networks via Site-Specific Deep Learning

Jeffrey G. Andrews
Department of Electrical and Computer Engineering
The University of Texas at Austin

IEEE Wireless Communications Technical Committee Seminar
Dec. 13, 2021

This work has been supported by NSF and Samsung research

Outline

- Introduction
  - From 5G to 6G: Brief Reflections and 6G@UT Research Directions
  - Machine Learning's Role in 5.5G and 6G
- Technical Example 1: Site-Specific Learned Probing Beams for Fast Beam Alignment
- Technical Example 2: Ultra-high Dimensional Channel Estimation via Deep Generative Models (DGMs)
- Parting Remarks

From LTE to 5G

Key new features of 5G vs. LTE:

- Millimeter wave (mmWave), esp. beam management and alignment features
- Variable bandwidths (“bandwidth parts”) and scalable OFDM subcarrier widths
- Flexible self-contained slots and control channels, more TDD emphasis
- Ultra-reliable low latency communication (URLLC) support
- Designed for max “forward compatibility”

Inherited/modified features from LTE:

- OFDMA with most data and control channel structures preserved
- Carrier aggregation including unlicensed spectrum and now mmWave
- Most of the multi-antenna (MIMO) and CoMP (multi-transmission point) frameworks
- **Overall** – 5G can be viewed as largely an evolution from 4G (LTE), as opposed to a clean break as in previous G’s.

Predictions:

1. 5G will have a longer life cycle than previous Gs
2. 6G will not be a clean break from 5G (so defn. of 6G is “open”)

6G@UT: UT Austin's New 6G Research Center

Four Main Research Directions. More info: http://6g-ut.org/

1. **Deeply Embedded Machine Learning**
   a. At PHY, MAC, Network layers – focusing on disruptive approaches
   b. From modem up to a network-level scale, leveraging sensing
2. **Pervasive Sensing**
   a. High integrity localization and mapping via 6G network infrastructure
   b. Sensing as a service; sensing as an input to ML algorithms
3. **New Spectrum and Topologies**
   a. New spectrum (e.g. > 100 GHz) and new spectrum access modalities
   b. Non-terrestrial network integration (esp. LEO) for global coverage
4. **Network Architectures, Slicing and Sharing**
   a. True network slicing, separation of data and control planes
   b. Intensive softwarization; cellular in the cloud

ML's Role in Future Wireless

- ML is a broad set of ever-evolving techniques
  - Determining the most appropriate approach is often the key research problem
  - Incredibly, in some cases, you may even determine you don't need or want ML!

- It is useful to think in terms of time scales and the training procedure
  - Deep Neural Networks (DNNs)
    - Often require considerable training, but then can make fast inferences or classifications and fully leverage GPU architectures.
    - A DNN is in my mind basically a powerful form of adaptive signal processing
  - Reinforcement learning (RL)
    - Requires considerable "start up" (offline) training, and then can learn and adapt to slow changes in the environment (online phase), very powerful for complex time-varying problems
    - However – we've found RL for wireless systems to often have convergence problems, and to be quite data hungry and slow.
  - A DNN is typically more suitable for the physical layer (PHY).
  - RL for the upper layers and at a network level (with above caveats), although also for some "trial and error" type problems at the PHY/MAC

- In this talk we focus on Deep Learning at the PHY for two different (but related) problems; and we use two very different architectures

Deep Learning at the PHY

- The 5G and WiFi PHY is highly optimized in both theory and practice
  - Guided by information theory and decades of implementation, its very tough to beat state of the art – e.g. MIMO-OFDMA + LDPCs in a Qualcomm ASIC – in any meaningful way in a generic setting.
  - For example, [AuoHoy21, ZhaDos21] recently show one can learn from scratch a competitive DNN-based transceiver, that even has some possible advantages. Impressive, but is it compelling?
- Instead of supplanting known PHY principles, I see the role of DL as more specific.
  1. For **nonlinear physical realities** that defy good models

     Examples: Low resolution A/D [BalAnd19], highly nonlinear RF/power amps [AuoHoy21], MIMO channel estimation with insufficient pilots and/or feedback [today's example 2]
  2. Finding approximate solutions to **open problems in information theory**

     Examples: feedback channel codes [Kim20], many-user interference channels at moderate SINR [Mis21]
  3. **Site-specific learning & design** where a "one-size fits all" approach is highly suboptimal

     Examples: BS parameter optimization, beam alignment in a specific environment [today's example 1], learned MIMO transceivers [O'Shea17], multiuser MIMO user selection

[AouHoy21] F. Aoudia, J. Hoydis, "Waveform Learning for Next-Generation Wireless Communication Systems", Sep. 2021, https://arxiv.org/abs/2109.00998
[ZhaDos21] Y. Zhang, A. Doshi, et al, "DeepWiPHY: Deep Learning-Based IEEE 802.11ax Receiver", IEEE Trans. on Wireless Comm, March 2021.
[Kim20] H. Kim et al, "Deepcode: Feedback Codes via Deep Learning", IEEE J. on Sel. Areas in Info. Theory, May 2020.
[Mis21] R. Mishra et al, "Distributed Interference Alignment for K-user Interference Channels via Deep Learning", IEEE ISIT, July 2021.
[BalAnd19] E. Balevi and J. G. Andrews, "One-Bit OFDM Receivers via Deep Learning", IEEE Trans. on Communications, June 2019.
[O'Shea17] T. O'Shea, T. Erpek, and T. C. Clancy, "Deep Learning Based MIMO Communications", https://arxiv.org/abs/1707.07980

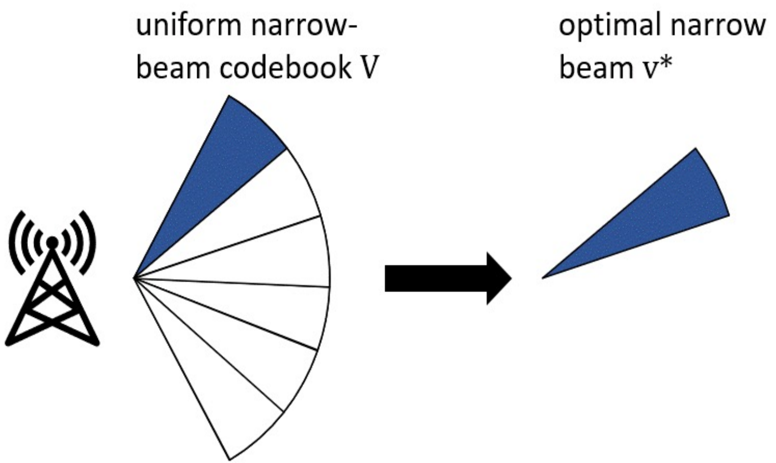
Example 1: Learning Site-Specific Probing Beams for Fast mmWave Beam Alignment

1. Y. Heng, J. Mo, J. G. Andrews, “Learning Site-Specific Probing Beams for Fast mmWave Beam Alignment”, under revision, *IEEE Trans. on Wireless Comm*. Available: https://arxiv.org/abs/2107.13121
2. Y. Heng, J. Mo, and J. G. Andrews, “"Learning Probing Beams for Fast mmWave Beam Alignment”, *IEEE Globecom,* Madrid, Spain, Dec. 2021.

This work has been done in collaboration with, and is supported by, Samsung research

Beam Alignment in 5G

- Wireless systems operating at a carrier frequency above roughly 15 GHz – ”mmWave” and Sub-TeraHz (THz) – need increasingly directional beamforming (BF) to achieve viable received signal strength
- 5G mmWave base stations (BS) and user equipment (UE) at have large – 64 or 128 at BS side – codebooks of indexed analog beams, from which a good beam pair needs to be selected.
- A typical approach is an exhaustive search over a “DFT” codebook of 64 evenly spaced beams over a 3D cone or pyramid shape: slow and does not scale well to higher frequencies or mobile scenarios

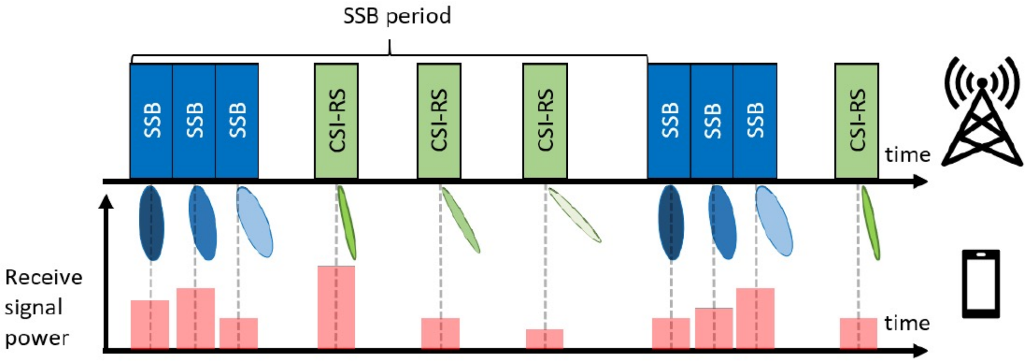


For an overview of beam alignment in 5G and a view to the future, we have a recent paper with Samsung research:

Y. Heng, J. G. Andrews, J. Mo, V. Va, A. Ali, B. Ng, and C. Zhang, “Six Key Challenges for Beam Management in 5.5G and 6G Systems”, *IEEE Communications Magazine*, July 2021.

Beam Alignment in 5G

- Downlink Beam alignment in 5G is based on this beam sweep (exhaustive search) approach.
  1. The BS sweeps through the transmit (Tx) codebook using Synchronization Signal Block (SSB) “wide” beams, UE transmits a random access preamble back to BS corresponding to the best SSB (beam)
  2. Channel State Information Reference Signal (CSI-RS) “narrow beams" are used for beam refinement, up to 4 signal strength measurements can be fed back by the UE
  3. The UE may also need to sweep its receive (Rx) codebook – causing a multiplicative increase in latency
  4. Eventually/hopefully, the best beam pair is selected.
- Although the limitations of this brute-force approach are obvious, it is hard to use many of the more “intelligent” methods, which may miss detecting new UEs or new UE positions

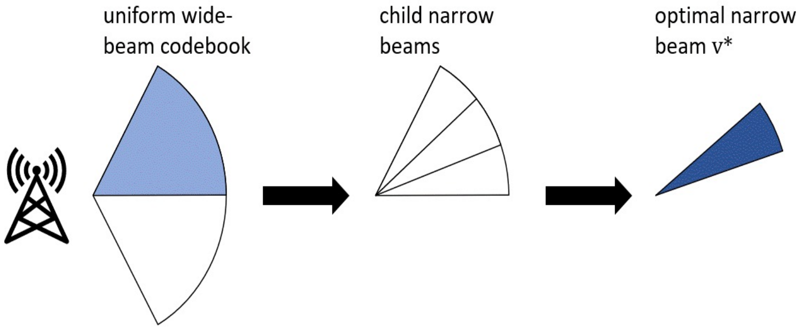


Beam alignment is the #1 bottleneck to mmWave and THz communication.

Enhancements to Beam Alignment

- **Hierarchical beam search** iteratively reduce the search space by sweeping wide beams first, then narrower child beams [1].
  - Reduces the beam sweeping overhead compared to the exhaustive search
  - Prone to search errors caused by noisy measurements
  - We will use this as a baseline
- **Context information** such as location [2], out-of-band information [3] and vision [4] can assist beam alignment.
  - Feedback of context information requires additional standards support
  - Possible privacy issues for localization and vision
- **Site-specific codebooks** can reduce beam sweeping overhead [5]. This is more related to our approach.
- Many other clever methods for beam alignment have been proposed, e.g. [6], but nearly all have important limitations in a real-world cellular system with many mobile users.
- **Our goal is to develop a practical technique that fits into the 5G framework, yet achieves big gains over the current exhaustive and hierarchical search methods**

Hierarchical search



[1] C. Qi, K. Chen, O. A. Dobre and G. Y. Li, "Hierarchical Codebook-Based Multiuser Beam Training for Millimeter Wave Massive MIMO," in IEEE Trans. Wireless Commun., 2020.

[2] Y. Heng and J. G. Andrews, "Machine Learning-Assisted Beam Alignment for mmWave Systems", to appear, IEEE Trans. on Cognitive Comm. and Networking. (early access on IEEExplore)

[3] A. Ali, N. Gonzalez-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," IEEE Trans. Wireless Commun., 2018.

[4] W. Xu, et al. "3D Scene-Based Beam Selection for mmWave Communications." IEEE Wireless Commun. Letters, 2020

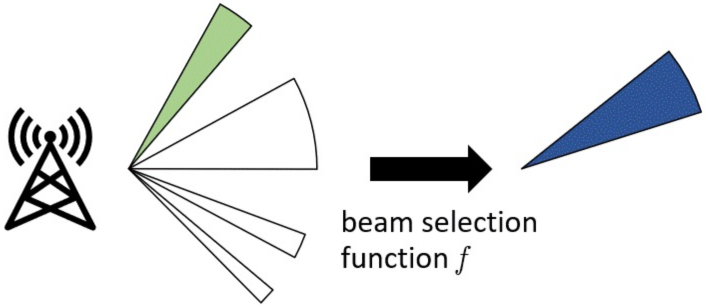
[5] M. Alrabeiah, Y. Zhang, and A. Alkhateeb. "Neural Networks Based Beam Codebooks: Learning mmWave Massive MIMO Beams that Adapt to Deployment and Hardware." arXiv preprint arXiv:2006.14501, 2020

[6] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi and R. W. Heath, "Spatially Sparse Precoding in Millimeter Wave MIMO Systems," in *IEEE Transactions on Wireless Communications*, March 2014.

Overview of our proposed beam alignment method

Our method utilizes two neural networks:

1. A **probing codebook W,** which is learned via site-specific training. **W** consists of a small number (~10) of beams that learn to efficiently cover the key parts of the whole angular space.
2. A **beam selection function** *f*() that uses UE measurements & feedback on **W**'s probing beams to and predict the optimal narrow beam in a standard codebook **V** (e.g. DFT, size 128)



The proposed method does not require additional context information, *and is compatible with the 5G beam alignment framework.*

How it works:

1. BS sweeps probing codebook **W**
2. UE measures and reports the received power of each beam in **W**
3. BS predicts the optimal narrow beam in **V** using *f*() based on the UE feedback
4. BS transmits data to a UE using its predicted **v***

Publications:

1. Y. Heng, J. Mo, J. G. Andrews, "Learning Site-Specific Probing Beams for Fast mmWave Beam Alignment", under revision, *IEEE Trans. on Wireless Comm*. Available: https://arxiv.org/abs/2107.13121
2. Y. Heng, J. Mo, and J. G. Andrews, ""Learning Probing Beams for Fast mmWave Beam Alignment", *IEEE Globecom,* Madrid, Spain, Dec. 2021.

System and Signal Model (Baseline)

- DL multiple-input single-output (MISO) channel model, ULA (planar also possible), analog BF only:

$$\mathbf{h} = \sum_{\ell=1}^{N_P} \alpha_\ell \mathbf{a}(\phi_\ell). \qquad \mathbf{a}_{\text{ULA}}(\phi_l) = \frac{1}{\sqrt{N_t}} \left[1 \quad e^{j\frac{2\pi}{\lambda} d \sin\phi_l} \quad \dots \quad e^{j(N_t-1)\frac{2\pi}{\lambda} d \sin\phi_l} \right]^T$$

- The BS will transmit a data symbol s using a BF vector **v**, and the received signal can be written as:

$$y = \sqrt{P_T}\mathbf{h}^H \mathbf{v} s + n$$

- The BS has a narrow beam codebook **V** with N_V Tx beams, our goal is to use beam **v** achieving the best SNR:

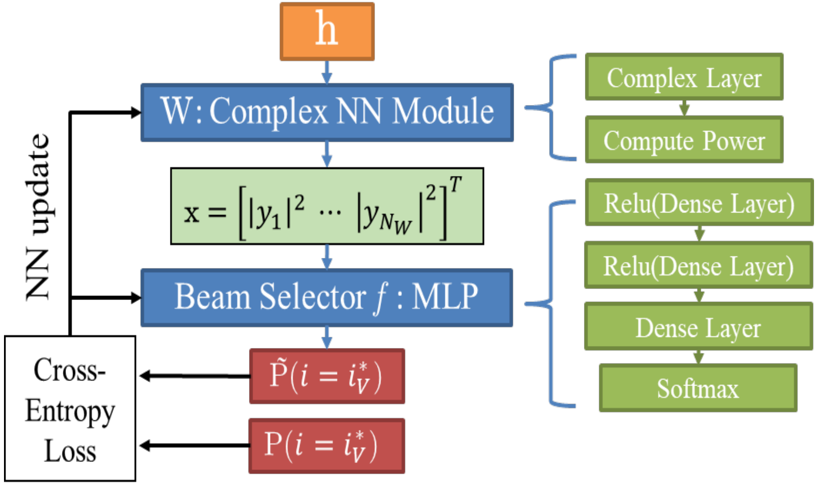
$$i_{\mathbf{v}}^* = \underset{i \in \{1,2,\cdots,N_{\mathbf{V}}\}}{\arg\max} \left(\frac{|\mathbf{h}^H \mathbf{v}_i|^2 P_T}{\sigma_n^2} \right) = \underset{i \in \{1,2,\cdots,N_{\mathbf{V}}\}}{\arg\max} \left(|\mathbf{h}^H \mathbf{v}_i|^2 \right)$$

- To do so, after sweeping the small (learned) probing codebook **W,** with N_W **<<** N_V beams, the received power of each beam in **W** is measured and reported to form the feature vector **x**:

$$\mathbf{x} = \left[|y_1|^2 \quad \dots \quad |y_{N_\mathbf{W}}|^2 \right]^T, \quad y_i = \sqrt{P_T}\mathbf{h}^H \mathbf{w}_i s + n_i$$

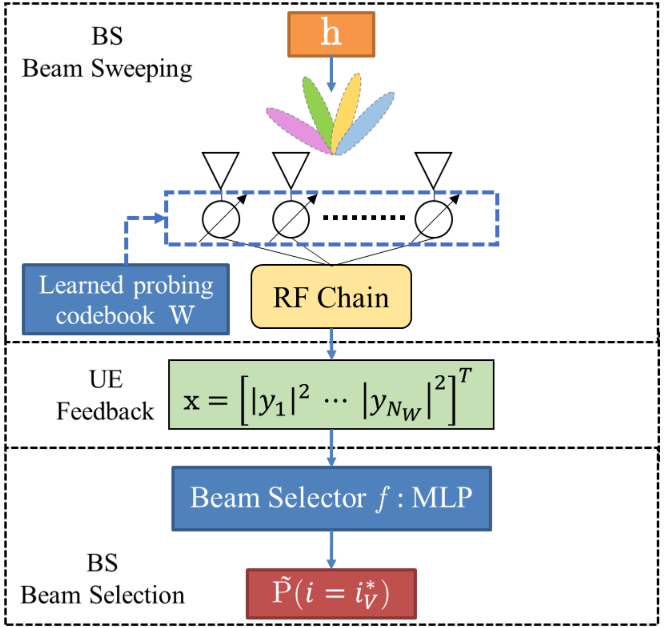
Neural Network Architecture: Learning (Training) Phase

- The entire probing-beam sweeping and narrow-beam selection procedure are stacked and trained as an end-to-end deep neural network (NN).
  - The probing codebook **W** is a complex-valued NN.
  - Beam selector *f()* is a multilayer perceptron (MLP) classifier.
- During training, the channel vector **h** is the input
- The output is the probability of each beam in **V** being the optimal narrow beam.
  - The loss function is the cross-entropy between this predicted distribution and the true optimal beam.
  - Both NN's are updated via the same loss function
- The proposed method requires an offline training phase.
  - The training data consists of measured/estimated channel vectors throughout the cell area.
  - These can be obtained through ray-tracing simulation (our approach) before deployment or through channel estimation in an actual deployment.



Our Architecture: Deployment Phase

- For an actual BS deployment, after training, the learned probing codebook **W** is extracted and implemented in RF e.g. as phase shifters
  1. The BS periodically sweeps through its site-specific learned probing codebook **W**
  2. The UEs measure and feed back the received power of each probing beam, forming the input feature **x**.
  3. The BS predicts the optimal (top-1) narrow beam or the top-k candidate beams in **V** to try using the learned MLP beam selection function *f*().
- If the environment or the overall UE location distribution changes (occurs slowly, on the order of hours or more), we can re-enter the training phase to update **W** and *f()*.



Experimental Setup

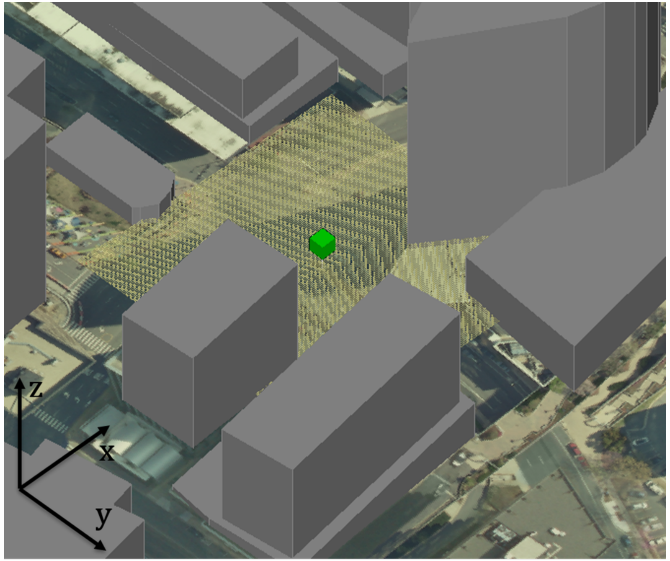
- Channel data is from our own Rosslyn dataset [2] and the public DeepMIMO dataset [7]:
  - Generated using a commercial ray-tracing by “Wireless InSite”
  - 4 different environments containing LOS and NLOS UEs.
  - 0.6/0.2/0.2 is training/validation/testing split
- NN parameters:
  - MLP has 2 hidden layers with ReLu activation
  - NN Trained for 200 epochs using the Adam optimizer
- We train models with different sizes of **W***:* N_W = [6, 8, 10, 12, 16,20]
- Baselines:
  1. **Genie** (true optimum, a hard upper bound): picks the beam in **V** with the highest BF gain
  2. **Exhaustive search**: picks the beam in **V** with the highest received power (measurement is degraded by noise)
  3. **2-stage hierarchical search**: first searches N_W wide beams covering the entire angular space, then searches all child beams of the best wide beam, finally selects the child beam with the highest received power.
  4. **Binary search:** repeatedly splits the search space into two equal partitions and search two wide beams covering each partition until reaching the final narrow beam

BS Antenna	64×1 ULA
UE Antenna	Single
Narrow beam codebook size N_V	128
Carrier Frequency	Rosslyn, DeepMIMO O1_28, O1_28B: 28 GHz DeepMIMO I3: 60 GHz
Bandwidth (B)	100 MHz
Transmit Power (P_T)	Rosslyn, DeepMIMO O1_28, I3: 10 dBm DeepMIMO O1_28B: 20 dBm
Noise power spectral density (PSD)	-161 dBm / Hz
Number of Rays	25

[2] Y. Heng and J. G. Andrews, “Machine Learning-Assisted Beam Alignment for mmWave Systems”, to appear, IEEE Trans. on Cognitive Comm. and Networking.
Dataset: https://github.com/YuqiangHeng/ML-mmWave-Beam-Alignment
[7] A. Alkhateeb,"DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive MIMO Applications ,“ in Proc. ITA, Feb. 2019.

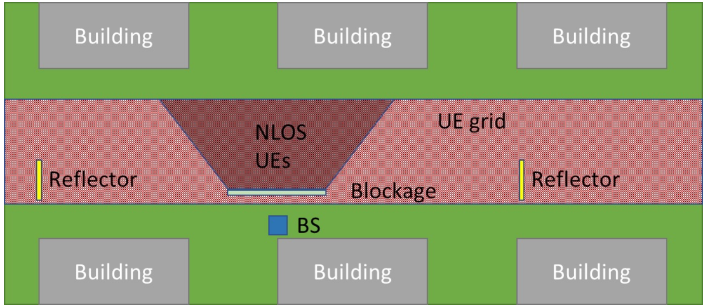
Rosslyn Ray-tracing Dataset

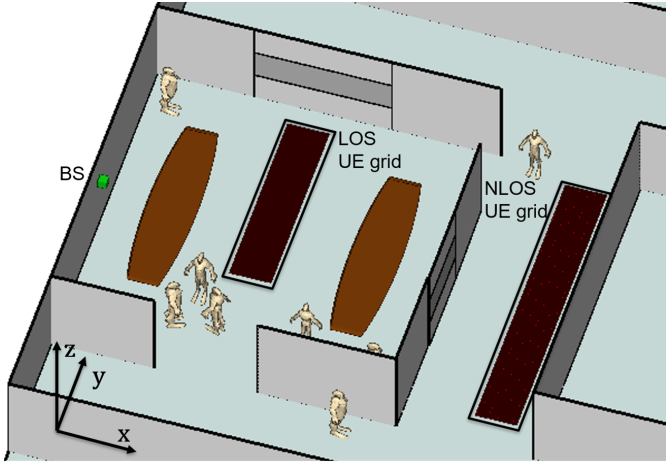
- The ray-tracing environment is a 3-D reconstruction of an urban outdoor area of “Rosslyn” in Arlington, VA
  - The buildings and terrain are modeled with concrete material with the appropriate dielectric properties
- BS placed at the center of an intersection with 10-meter elevation, with 64 antennas
- 73,884 UEs are placed uniformly around the AP on the terrain surface in a roughly (90 meters)^2 grid with 0.35 meter spacing and 2-meter elevation.
- 28 GHz carrier, 100 MHz bandwidth



[2] Y. Heng and J. G. Andrews, “Machine Learning-Assisted Beam Alignment for mmWave Systems”, to appear, IEEE Trans. on Cognitive Comm. and Networking.
Dataset: https://github.com/YuqiangHeng/ML-mmWave-Beam-Alignment

DeepMIMO Datasets





DeepMIMO O1_28 & O1_28B

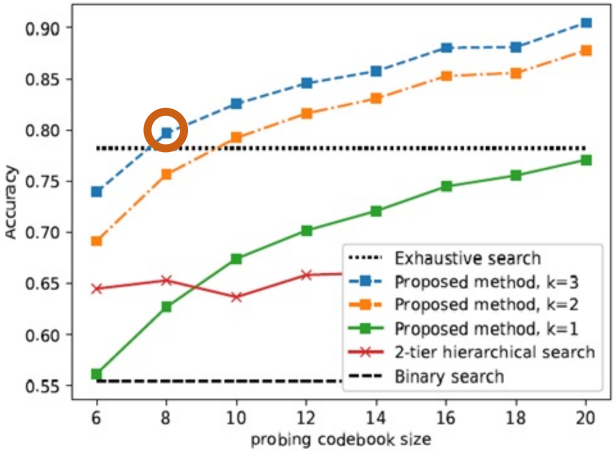
- **Outdoor street environment (O)** with buildings on both sides.
- O1_28 contains 72,581 LOS UEs.
- O1_28B includes an additional metal screen in front of the BS and reflectors on both sides.
  - Contains 497,931 LOS + NLOS UEs
- Both are at 28 GHz carrier

DeepMIMO I3

- **Indoor office environment (I)** with a grid of LOS UE in the room and NLOS UEs in the corridor.
  - Contains 118,959 LOS + NLOS UEs
- 60 GHz carrier

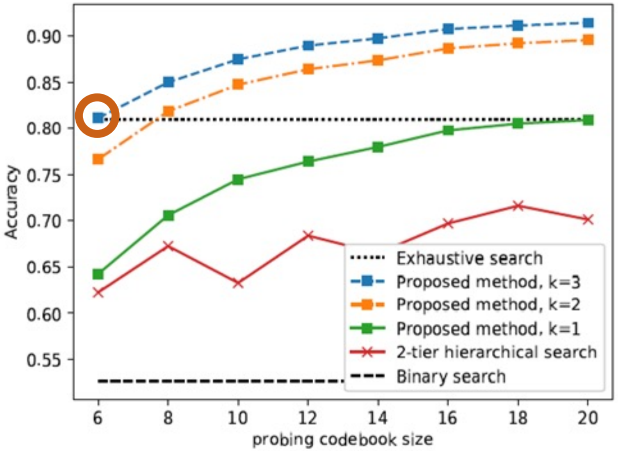
[7] A. Alkhateeb,"DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive MIMO Applications ,“ in Proc. ITA, Feb. 2019.

Evaluation: beam alignment accuracy in NLOS scenarios



DeepMIMO I3

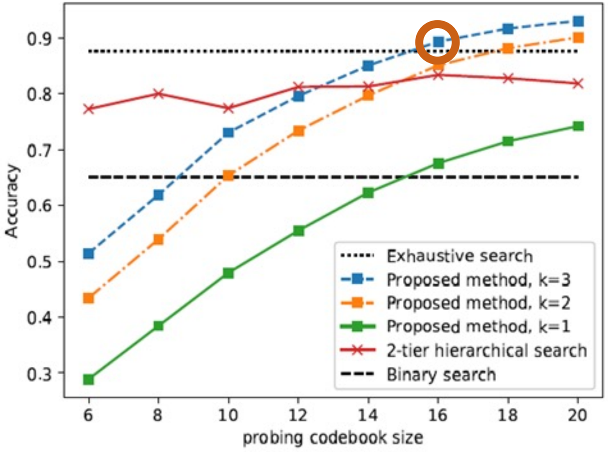
Beats exhaustively sweeping 128 beams with just 9 or 11 beams (<u>14x or 11x gain</u>)



DeepMIMO O1_28B

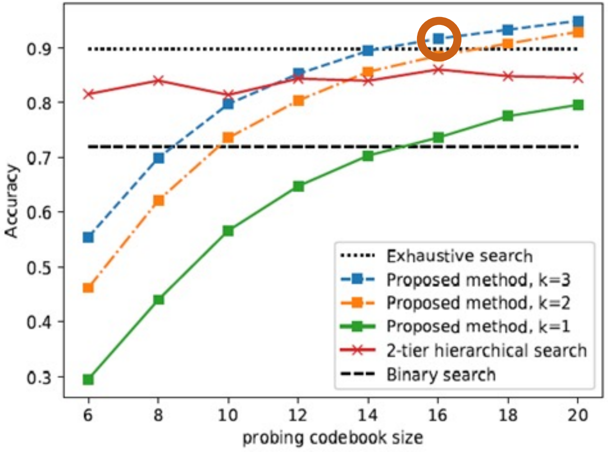
- The beam alignment accuracy is the probability (relative frequency) that the BS selects the optimal narrow beam from **V**.
- The genie (UB) has probability 1 of selecting the best beam.
- The proposed method outperforms the hierarchical searches with just 6 probing beams and no additional beam sweeping (k = 1).
- By trying the top-3 predicted candidate beams, the proposed method quickly outperforms even the exhaustive search!

Evaluation: beam alignment accuracy in LOS scenarios



Rosslyn

Beats exhaustively sweeping 128 beams by sweeping 19 beams (6x gain)



DeepMIMO O1_28

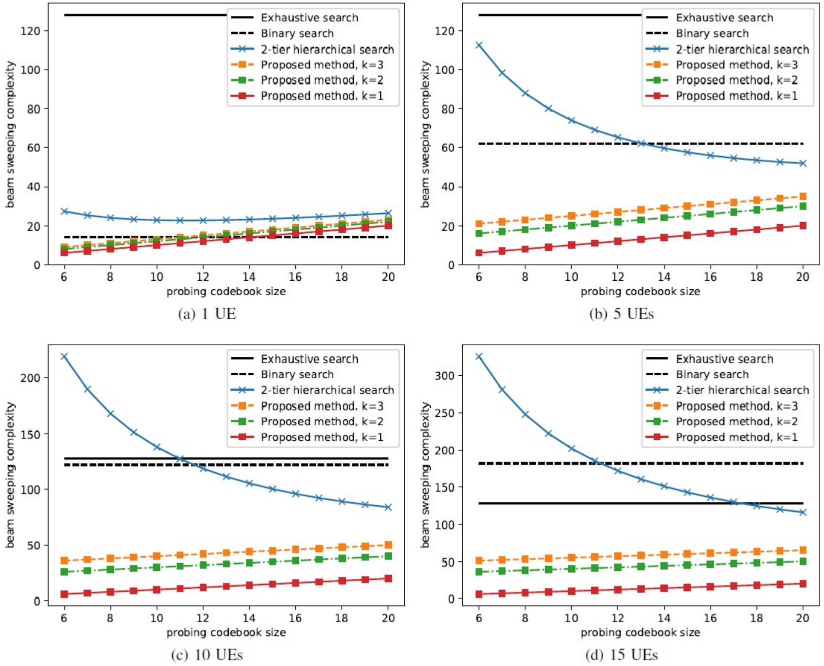
- There is less gain compared to in the NLOS scenarios, since there is considerably less structure to learn and the angle of arrival (AoA) distribution is more uniform
- The proposed method can still beat hierarchical searches with 14 probing beams and the exhaustive search with 16.

Beam sweeping complexity/latency for many UEs

- When considering simultaneous beam alignment for multiple UEs (5, 10 and 15 UEs) the proposed method still achieves much lower beam sweeping complexity (sweeps far fewer total beams).
- For 10 UEs, the proposed method reduces the beam sweeping overhead of the hierarchical search methods by around **3x** with 12 probing beams and k=3 and by around **10x** with 12 probing beams and k=1.
- Note: the multiple UE scenario is when the apparent gains of many other approaches evaporate

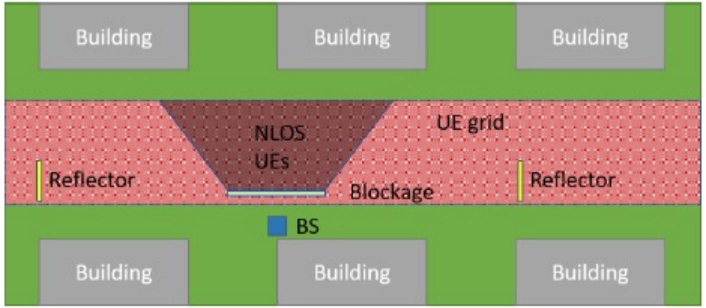
Beam alignment method	**Beam sweeping complexity**	**Feedback complexity**
Proposed method	$N_W + K \cdot k\mathbb{1}_{\{k>1\}}$	KN_W received signal power + $K \cdot \mathbb{1}_{\{k>1\}}$ beam indices
2-tier hierarchical search	$N_W + K\frac{N_V}{N_W}$	$2K$ beam indices
Binary hierarchical search	$2 + 2K\log_2\frac{N_V}{2}$	$K\log_2 N_V$ beam indices
Exhaustive search	N_V	K beam indices

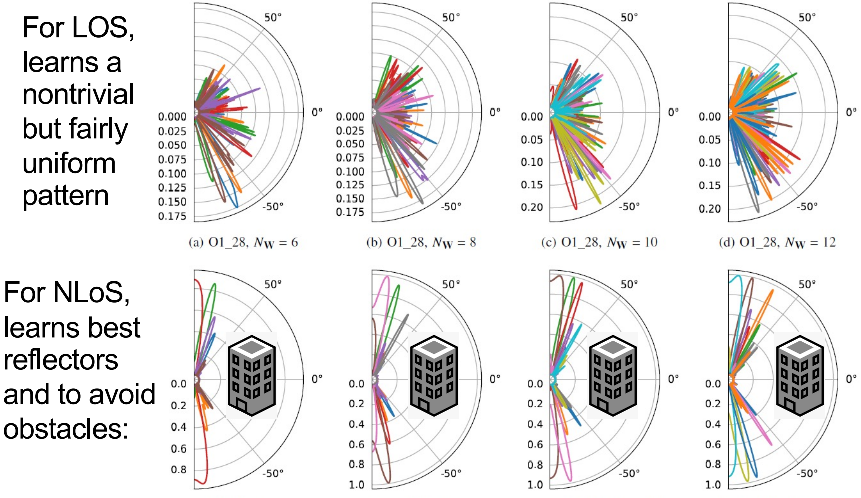
Beam Sweeping Complexity for K UEs



Probing codebook beams: what do they look like?

- The architecture consistently learns probing beam patterns that “make sense” in the context of the propagation environment.
- NLOS environment:
  - Strong beams tend to point towards the reflectors and LOS UEs on either side
  - Little energy directed towards blockage areas or dead zones
- Important to note that we do <u>not</u> optimize the probing codebook **W** for average BF gain or SNR.
- Rather, the probing codebook is optimized to learn the propagation environment to benefit the downstream MLP beam selector *f()*





(a) O1_28, $N_W = 6$ (b) O1_28, $N_W = 8$ (c) O1_28, $N_W = 10$ (d) O1_28, $N_W = 12$

(e) O1_28B, $N_W = 6$ (f) O1_28B, $N_W = 8$ (g) O1_28B, $N_W = 10$ (h) O1_28B, $N_W = 12$

Wrap-up of Example 1

- We proposed a promising, 5G-plausible, deep-learning based beam alignment method that predicts the optimal narrow beam(s) after sweeping a learned site-specific probing codebook
- There is significant gain in the challenging NLOS scenarios
  - The proposed method can reach or even exceed the exhaustive search accuracy, while reducing the search latency **by over 10x.**
  - We conjecture the gains could be even larger for narrower beams at higher carrier frequencies (THz)
- My take on why it works so well (in light of the "no free lunch" principle):
  - Instead of a one-size-fits-all solution, our probing codebook exploits the unique propagation and UE clustering in each cell site, which avoids most wasteful searches
  - The end-to-end training of both the probing codebook and the beam selector f() allows synergies between them to develop
  - Our scheme does increase the UL feedback per UE – we send feedback on <u>all</u> N_W probing beams, instead of just the best beam(s). However, this is probably a great tradeoff in most cases, since we achieve much faster downlink beam alignment.
- Considerable scope for future work and generalizations of this framework

Example 2:
Ultra High Dimensional Channel Estimation leveraging Deep Generative Networks

1. E. Balevi, A. Doshi, A. Jalal, A. Dimakis, and J. G. Andrews, “High Dimensional Channel Estimation Using Deep Generative Networks”, *IEEE Journal on Sel. Areas in Communications*, Vol. 39, No. 1, pp. 18-30, Jan. 2021
2. E. Balevi and J. G. Andrews, “Wideband Channel Estimation with A Generative Adversarial Network”, *IEEE Trans. on Wireless Comm,* Vol. 20, No. 5, pp. 3049-60, May 2021.

This work has been supported by NSF (and preliminary work by Intel)

Motivation

(I thought Channel Estimation was a solved problem?)

- At large bandwidths (e.g. > 1 GHz) and high frequencies (eventually > 100 GHz):
  - Could have antenna spacings on the order of 2-3 mm (100-150 GHz carrier).
  - 6G base stations could have ~10,000 antenna elements in a compact planar array
- Channel estimation and high gain beam alignment will be the most challenging problems with dimensionality on the order of:
  - 1,000-10,000 x 100-1,000 spatial channel dimensions (correlated)
  - 1,000 subcarriers over 10+ coherence bandwidths
  - 10^6-10^9 total (correlated) dimensions – *ultra high dimensional* (UHD)
  - Current approaches won’t scale: too many pilots, too much computation
- Meanwhile, deep learning approaches for efficiently *approximating* large inverse problems are experiencing rapid advancement, e.g. deep generative models (DGMs)
- From Ex. 1, also recall we need channel estimates to learn a probing codebook.

High Dimensional Channel Estimation

- Traditional channel estimators such as LMMSE are near-optimal for rich multipath channels
  - However, ultra high-dimensional channels tend to exhibit extremely sparse structures [Bajwa10], which these estimators cannot directly exploit
  - Moreover, LS-type estimators require many pilots: at least equal to the number of transmit antennas, as well as priors like the correlation matrix
- As a remedy, sparsity has been exploited via compressed sensing (CS), e.g. in underwater acoustic channels [Berger10] and mmWave channels (e.g. [Alk14, Ven17]).
- High dimensional channels are often very sparse/low rank [Rappaport19], [Eliasi17], but not necessarily in a known basis: basis can vary based on the environment
- Our approach: a site-specific DGM which learns the propagation environment via a GAN

[Bajwa10] W. Bajwa, J. Haupt, A. M. Sayeed, R. Nowak, “Compressed Channel Sensing: A New Approach to Estimating Sparse Multipath Channels’’, Proc. IEEE 2010.
[Berger10] C. R. Berger, Z. Wang, J. Huang, S. Zhou, “Application of Sensing to Sparse Channel Estimation’’, IEEE Comm. Magazine, Nov. 2010.
[Alk14] A. Alkhateeb, et al, “Compressive Channel estimation and hybrid precoding for millimeter wave cellular systems,” *IEEE JSTSP*, Oct. 2014.
[Rappaport19] T. Rappaport et al. "Wireless communications and applications above 100 GHz: Opportunities and challenges for 6G and beyond." *IEEE Access*, Jun. 2019.
[Eliasi17] P. Eliasi et al, “Low-rank spatial channel estimation for millimeter wave cellular systems,” *IEEE Trans. on Wireless Comm.*, 2017
[Ven17] K. Venugopal et al, “Channel estimation for hybrid architecture-based wideband millimeter wave systems,” IEEE JSAC, Sep. 2017.

State of the art in high dimensional channel estimation

(not an exhaustive list)

- Compressed Sensing (CS) using Matching Pursuit (MP) algorithms [Alk14] [Lee16]
  - Need to find appropriate sparsifying basis
  - Solve complex optimization problem at each coherence interval
- Message Passing (EM-GM-AMP, VAMP) [VilaSchniter13] [Rangan19]
  - Works well for a large class of random sensing matrices, but require a sparsifying basis
- Recent Deep Learning techniques [Wen18] [Yang19] [Gao19] [Dong19]
  - Supervised, excessive time required to generate the necessary labeled data and then train
  - Unsupervised techniques would in general be far preferable

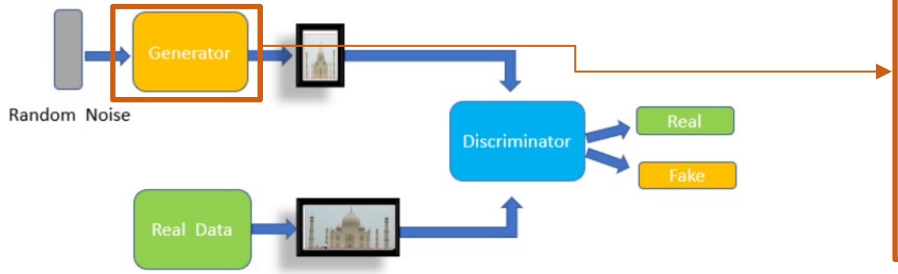
Our idea: use the structure captured by a deep generative model as a prior, eliminating the need for a sparsifying basis or supervised learning

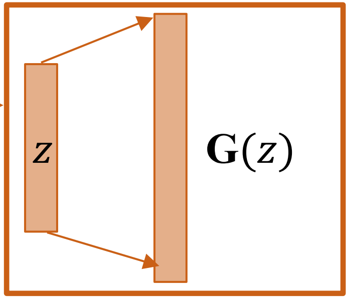
[Lee16] J. Lee et al. "Channel estimation via orthogonal matching pursuit for hybrid MIMO systems in millimeter wave communications." *IEEE T. Comm.*, Apr. 2016.
[VilaSchitner13] J. Vila and P. Schniter. "Expectation-maximization Gaussian-mixture approximate message passing." *IEEE Trans. on Signal Process.,* Jul. 2013.
[Rangan19] S. Rangan et al. "Vector approximate message passing." *IEEE Trans. on Info. Theory*, May 2019.
[Wen 18] C. K. Wen et al., "Deep learning for massive MIMO CSI feedback." *IEEE Wireless Communications Letters*, Oct. 2018.
[Yang19] Y. Yang et al., "Deep learning-based channel estimation for doubly selective fading channels." *IEEE Access*, *Mar. 2019*
[Gao19] S. Gao et al, "Deep learning based channel estimation for massive MIMO with mixed-resolution ADCs," *IEEE Communications Letters*, Aug. 2019
[Dong19] P. Dong et al, "Deep CNN-Based Channel Estimation for mmWave Massive MIMO Systems," *IEEE Jour of Sel Top in Signal Processing , Sep 2019.*

Deep Generative Models

- A deep generative model is a feed-forward NN
  - Input vector $z \in \mathbb{R}^d$, output vector $G(z) \in \mathbb{R}^n$ where $d \ll n$.
  - For small images, perhaps $d = 100$ and $n = 64\times64\times3$ ($= 12{,}288$).
- This NN can be trained to take a iid Gaussian input z and produce samples of complicated distributions, e.g. human faces [Radford16]
- One powerful method for training generative models is **Generative Adversarial Nets (GANs).**

GAN faces [Karras18]





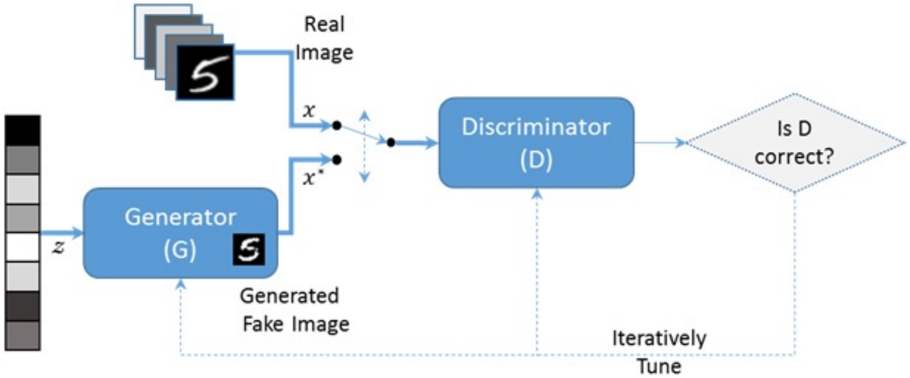
[Radford16] Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," ICLR, May 2016.
[Karras18] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," ICLR 2018

Generative Adversarial Net (GAN)

- A GAN [Goodfellow2014] consists of two feed-forward NNs, a generator $\mathbb{G}$ & discriminator $\mathbb{D}$ engaging in an iterative minimax game:

$$\min_G \max_D V(D,G) = E_{x \sim \mathbb{P}_r(x)}[\log D(x)] + E_{z \sim \mathbb{P}_z(z)}[\log(1 - D(G(z)))]$$

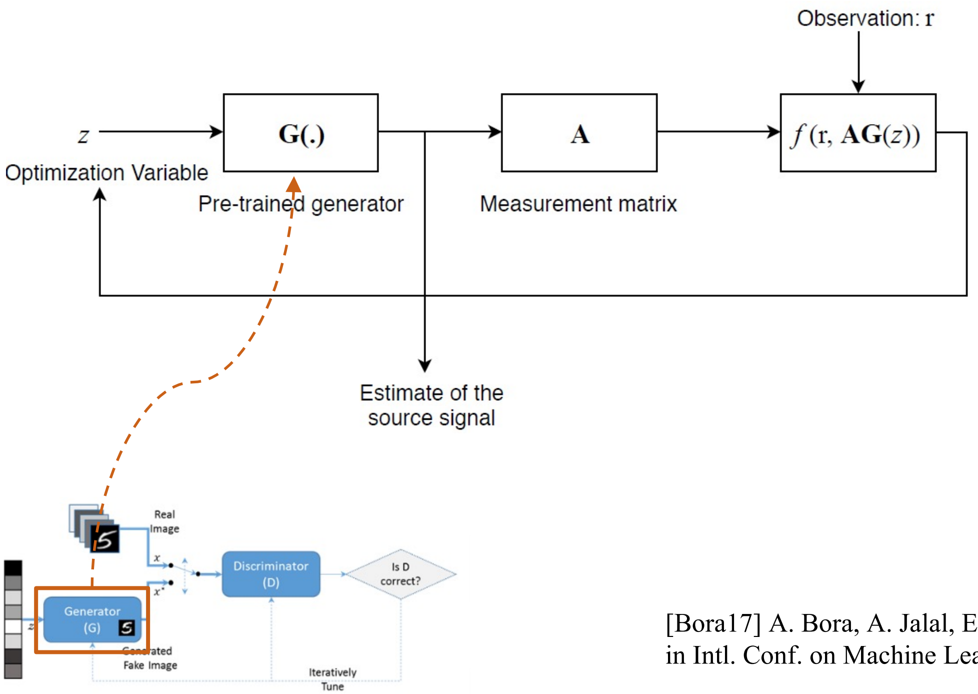
- $\mathbb{G}$ attempts to learn the data distribution $\mathbb{P}_r$, while $\mathbb{D}$ learns to discriminate between real data samples $\sim \mathbb{P}_r$ and fake ones from $\mathbb{G} \sim \mathbb{P}_g$.



https://cntk.ai/pythondocs/CNTK_206A_Basic_GAN.html

[Goodfellow14] Goodfellow, Ian, et al. "Generative adversarial nets." *Advances in neural information processing systems*. 2014.

Compressed Sensing (for image reconstruction) using Generative NNs



<u>Prior Algorithm [Bora17]:</u>

1. Train a GAN using an image dataset.
2. Extract the trained generator **G**.
3. Given a noisy compressed observation **r**, reconstruct the true image **r** encodes by solving the following optimization problem using gradient descent:
$$z^* = \arg\min_z f(\mathbf{r}, \mathbf{AG}(z)),$$
where f is a loss function. For example,
$$f(\mathbf{r}, \mathbf{AG}(z)) = ||\mathbf{r} - \mathbf{AG}(z)||_2^2$$
4. The reconstructed image is then $\mathbf{G}(z^*)$.

[Bora17] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, “Compressed sensing using generative models,” in Intl. Conf. on Machine Learning (ICML), Aug. 2017, pp. 537–546.

Part 1: Narrowband Channel Estimation

- Training-based channel estimation approach for narrowband point-to-point DL MIMO setup with N_p pilots & N_t transmit antennas & N_r receive antennas.
- In each time slot, BS employs a training beamformer $\mathbf{p} \in \mathbb{C}^{N_t \times 1}$ to transmit a pilot symbol $x = 1$, while the UE makes N_r measurements.
- N_p distinct beamforming vectors are employed during training. Denote $\mathbf{P} = [\mathbf{p}_1, ..., \mathbf{p}_{N_p}] \in \mathbb{C}^{N_t \times N_p}$
- Assuming the spatial channel matrix $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$ remains constant over the N_p time slots, received training signal $\mathbf{Y} \in \mathbb{C}^{N_r \times N_p}$ at UE:

$$\mathbf{Y} = \mathbf{H}\mathbf{P} + \mathbf{N}$$

- Vectorizing, and utilizing Kronecker products:

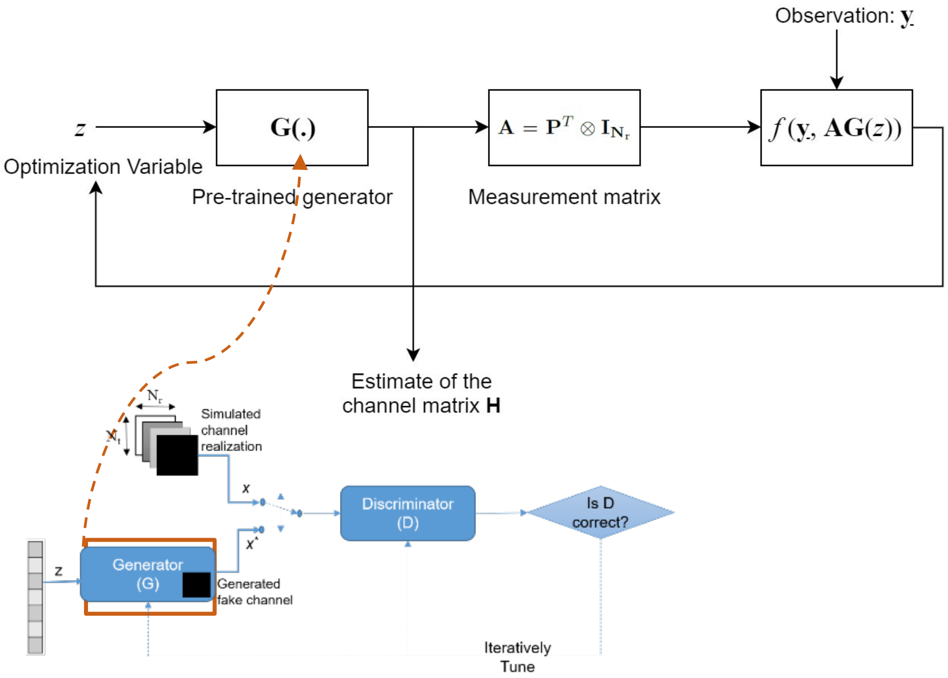
$$\underline{\mathbf{y}} = (\mathbf{P}^T \otimes \mathbf{I}_{\mathbf{N}_r})\underline{\mathbf{H}} + \underline{\mathbf{n}}, \quad \text{where}$$

$$\underline{\mathbf{y}}, \underline{\mathbf{n}} \in \mathbb{C}^{N_r N_p \times 1} \quad \underline{\mathbf{H}} \in \mathbb{C}^{N_r N_t \times 1}$$

<u>Challenge:</u> Assuming $N_p < N_t$, finding $\underline{\mathrm{H}}$ from $\underline{\mathrm{y}}$ is an ill-posed inverse problem.

<u>Our idea:</u> Use the structure captured by a pretrained deep generative model as a prior.

Narrowband Generative Channel Estimator (GCE)



<u>Adapting [Bora17] to channel estimation:</u>

1. Train a GAN using a set of "real" channel realizations (details shortly).
2. Extract the trained generator **G**.
3. Given noisy pilot measurements $\underline{y}$, reconstruct the channel by solving the following optimization problem using gradient descent:

$$z^* = \arg\min_{z \in \mathbb{R}^d} \; ||\underline{\mathbf{y}} - (\mathbf{P}^T \otimes \mathbf{I}_{N_r})\underline{\mathbf{G}}(z)||_2^2 + \lambda_{\text{reg}}||z||_2^2,$$

where d is the GAN's input vector dimension, and λ_{reg} is a regularization parameter.

4. The reconstructed channel estimate is then $\mathbf{G}(z^*)$, which is $N_r \times N_t$

We refer to this framework as the **Generative Channel Estimator (GCE).**

Wireless System Parameters

Delay Profile	CDL-D*
N_t	64
N_r	16
Antenna Array	ULA
Carrier Frequency	40 GHz
Antenna Spacing	$\lambda/10$

Before using the simulated channel matrices for training the GAN, we normalize them element-wise:

$$\mu_i = \mathbf{E}[\mathbf{H_G}_i] \quad \sigma_i^2 = \mathbf{E}[(\mathbf{H_G}_i - \mu_i)^2]$$

$$\mathbf{H_G}_{i,norm} = \frac{\mathbf{H_G}_i - \mu_i}{\sigma_i}$$

For generating the baselines, considering the clustered channel model (CDL), we use 2D DFT array response matrices $\mathbf{A}_\mathrm{T}$ and $\mathbf{A}_\mathrm{R}$ to obtain the sparse channel representation $\mathbf{H}_\mathrm{v} \in \mathbb{C}^{N_\mathrm{r} \times N_\mathrm{t}}$

$$\underline{\mathbf{H}} = ((\mathbf{A}_\mathrm{T}^H)^T \otimes \mathbf{A}_\mathrm{R})\underline{\mathbf{H}_\mathrm{v}}$$

The received signal at the UE is: $\underline{\mathbf{y}} = ((\mathbf{A}_\mathrm{T}^H P)^T \otimes \mathbf{A}_\mathrm{R})\underline{\mathbf{H}_\mathrm{v}} + \underline{\mathbf{n}}$

Denote by $\mathbf{A}_\mathrm{sp} = ((\mathbf{A}_\mathrm{T}^H P)^T \otimes \mathbf{A}_\mathrm{R})$.

* From 3GPP specs TR 38.901, a spatial LOS channel model used for UMi scenarios. CDL – Clustered Delay Line

Performance Benchmarks

OMP Channel Estimation [Méndez-Rial16]: Solves this non-convex combinatorial problem:

$$\underset{\underline{\mathbf{H}}_{\mathrm{v}} \in \mathbb{C}^{N_{\mathrm{r}} N_{\mathrm{t}}}}{\operatorname{minimize}} \; ||\underline{\mathbf{H}}_{\mathrm{v}}||_0 \text{ subject to } ||\underline{y} - \mathbf{A}_{\mathrm{sp}} \underline{\mathbf{H}}_{\mathrm{v}}||_2 \leq \sigma$$

The OMP stopping criterion is based on residual error power, chosen to be the noise variance.

Lasso Channel Estimation: Considering the $\mathcal{L}1$ convex relaxation of the OMP problem (Basis Pursuit Denoising), we solve the following Lagrangian form using a convex solver:

$$\underset{\underline{\mathbf{H}}_{\mathrm{v}} \in \mathbb{C}^{N_{\mathrm{r}} N_{\mathrm{t}}}}{\operatorname{minimize}} \; ||\underline{\mathbf{H}}_{\mathrm{v}}||_1 + \lambda_{sp} ||\underline{y} - \mathbf{A}_{\mathrm{sp}} \underline{\mathbf{H}}_{\mathrm{v}}||_2$$

EM-GM-AMP Channel Estimation [VilaSchniter13]: Given y & $\mathbf{A}_{\mathrm{sp}}$, EM-GM-AMP recovers $\mathbf{H}_{\mathrm{v}}$ from which we can recover $\mathbf{H}$.

[Méndez-Rial16] R. Méndez-Rial et al. “Hybrid MIMO architectures for millimeter wave communications: Phase shifters or switches?”. *IEEE Access*, Jan. 2016
[VilaSchitner13] J. Vila and P. Schniter. "Expectation-maximization Gaussian-mixture approximate message passing." *IEEE Trans. on Signal Process.,* Jul. 2013

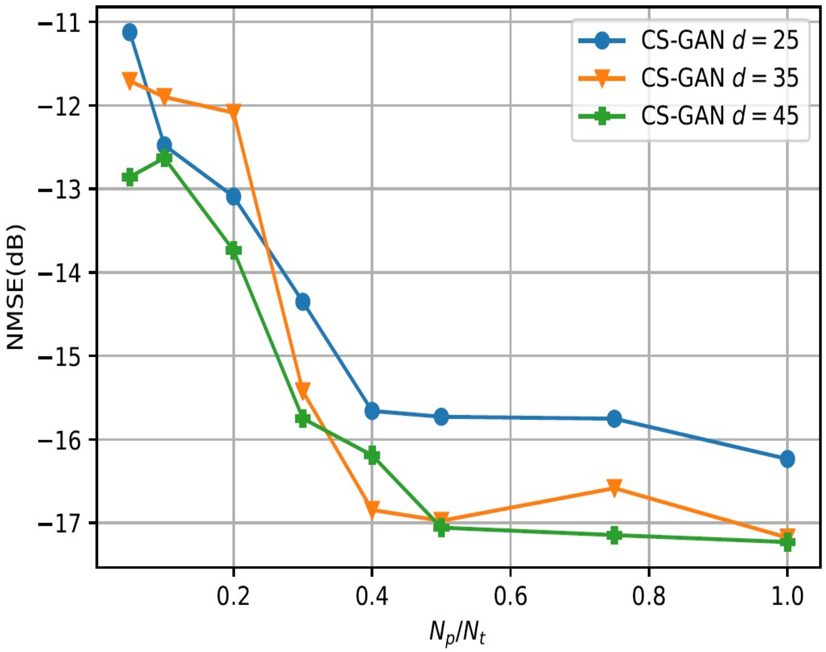
GAN Model and Training Details

Training data size	3654
Testing data size	12
Optimizer	RMSProp
Learning Rate	0.00005
Batch Size	200
Epochs	3000
λ_{reg}	0.001

- Our Wasserstein GAN [Arjovsky17] was trained with simulated channel realizations.
- The generator G(z) is a Deep Convolutional neural network, which is then extracted to use in the Algorithm given earlier.
- G takes an input $z \epsilon \mathbb{R}^d$, passes it through a dense layer with output size $128 N_t N_r/16$, and reshapes it to $((N_t/4), (N_r/4), 128)$.
- This latent representation is passed through $k = 2$ layers, each consisting of the following units: up-sampling, 2D Convolution with a kernel size of 4 and Batch Normalization.
- Finally passed through a 2D Convolutional layer with linear activation to obtain $\mathbf{G}(z)$, the $N_r \times N_t$ channel estimate

[Arjovsky17] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in Intl. Conf. on Machine Learning (ICML), 2017, pp.214–223.

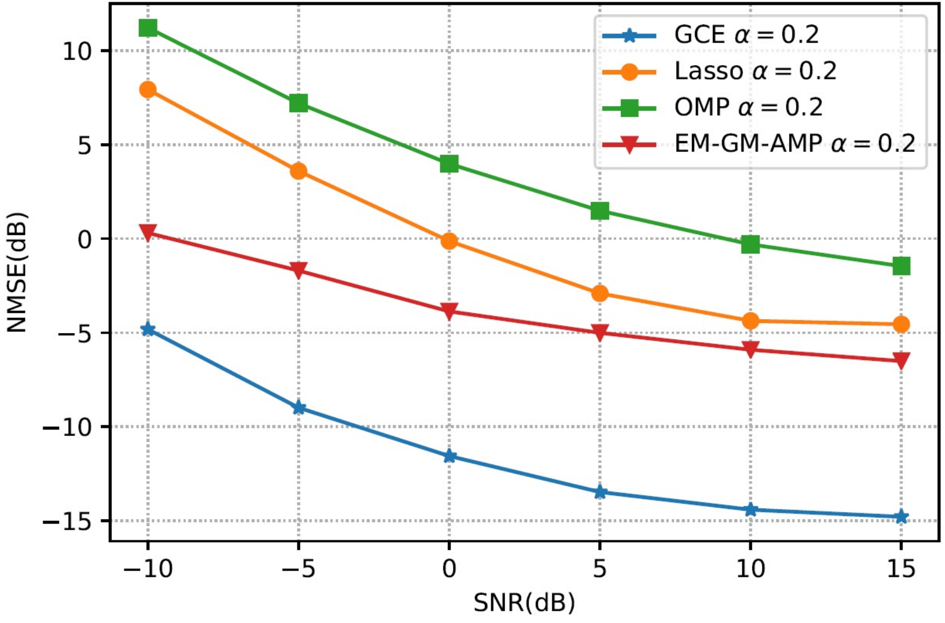
GCE requires under 40 parameters to represent a 16x64 complex CDL channel



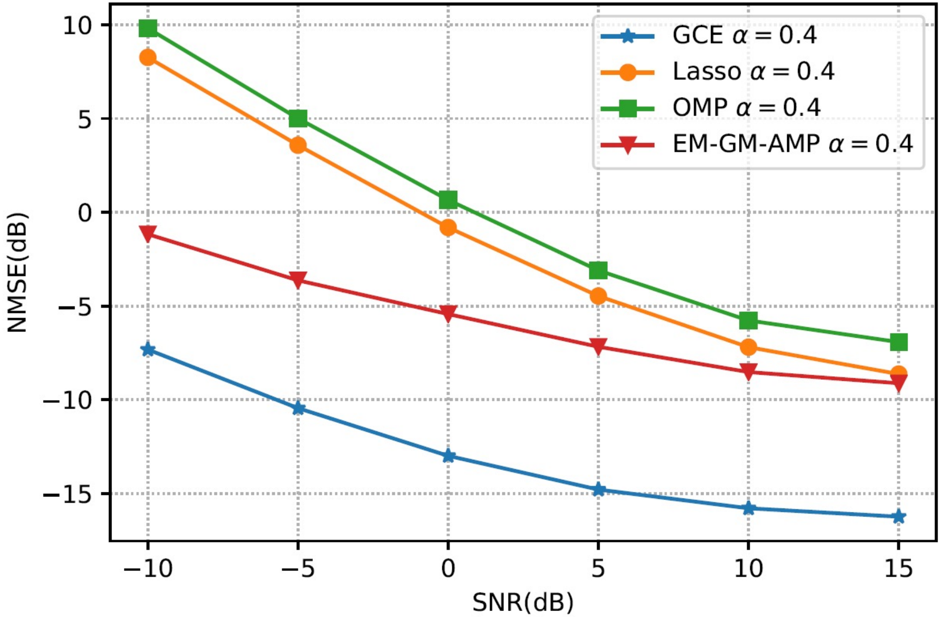
- We determine the optimal dimension d of the input z to the generator in absence of noise, plotting NMSE as a function of N_p/N_t.
- $d = 35$ appears sufficient, and increasing N_p/N_t beyond 0.4 does not impact the NMSE:
  - **Thus, we get over 50x compression** (very useful for channel feedback, if needed)
- Using the input vector z^* (of size d), we can recover the channel estimate without knowing that the channel is sparse in any particular (e.g. DFT) basis.
- GCE provides a model-free approach for efficiently representing inherently sparse or structured channels.

GCE outperforms OMP, Lasso & EM-GM-AMP

$\alpha = N_p/N_t = 0.2$



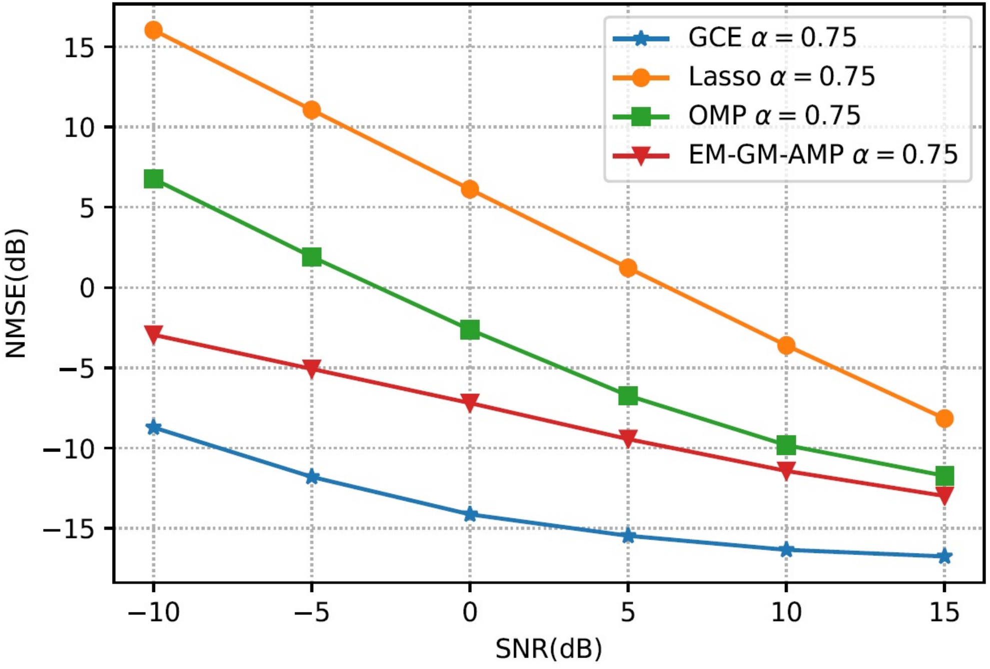
$\alpha = N_p/N_t = 0.4$



GCE achieves about 8 dB NMSE gain over EM-GM-AMP at SNR = 15 dB

GCE outperforms OMP, Lasso & EM-GM-AMP

$\alpha = N_p/N_t = 0.75$



Explanations/Comments:

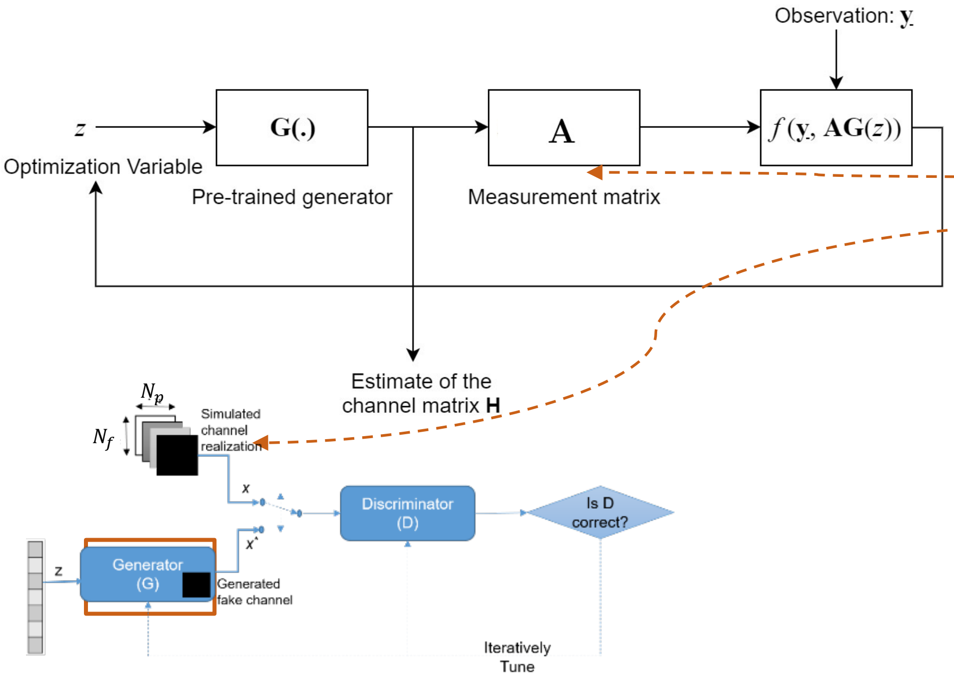
1. GCE has prior information about the channel distribution, which the other techniques do not
2. However, the other techniques do know and exploit the sparsifying (e.g. DFT) basis: GCE does not
3. We exploit reduced antenna spacing = high spatial correlation to learn channel distribution

Part 2: Wideband Channel Estimation

- Thus far we have assumed narrowband channel estimation, to simplify the problem and focus on the spatial domain. However:
  - Large bandwidth channels are frequency selective.
  - Only the time and spatial domain correlation can be exploited for narrowband, so we used an artificially small antenna spacing ($\lambda/10$) to generate sufficient correlation (which presumably a real-world channel would also provide).
- GANs are utilized for wideband channel modeling [Dorner20], but this estimates only the conditional distribution, and is not a channel estimator.
- Wideband model:
  - N_p pilots, N_t transmit antennas, N_r receive antennas & now N_f subcarriers
  - Transmit pilot symbols from multiple RF chains as opposed to a single RF chain
- Same basic steps, i.e., vectorizing, and utilizing Kronecker product, then building upon and modifying the previous generative channel estimation (GCE) architecture.

[Dorner20] S. Dorner, M. Henninger, S. Cammerer, and S. ten Brink, "WGAN-based Autoencoder Training Over-the-air," *Arxiv:2003.02744*, March 2020.

Wideband Generative Channel Estimator



<u>The key differences vs. narrowband:</u>

a. The measurement matrix has a different (wideband) structure.

b. The channels are structured as N_tN_r distinct complex planes of size N_f x N_p Thus:
- We can exploit both frequency and time correlations.
- Antenna spacing is the standard $\lambda/2$.

Theoretical Result

- To guarantee an upper bound to the channel estimation error, the measurement matrix must have sub-Gaussian entries [Bora17].
- For channel estimation, there are 2 additional constraints for the measurement matrix due to:
  1. Total transmission power constraint
  2. Constant modulus constraint, due to the phase shifters in the analog precoder/combiner

- Recall that

$$\underbrace{(\mathbf{I}_{\mathrm{N_p}} \otimes \mathbf{A}[n])}_{\mathbf{A}} \quad \text{where} \quad \underbrace{(\mathbf{s}[n]^T(\mathbf{I}_{\mathrm{N_f}} \otimes \mathbf{F}_{\mathrm{RF}}^T) \otimes (\mathbf{I}_{\mathrm{N_f}} \otimes \mathbf{W}_{\mathrm{RF}}^H))}_{\mathbf{A}[n]}$$

where $\mathbf{s}[n]$: pilots $\mathbf{F}_{\mathrm{RF}}$: analog precoder $\mathbf{W}_{\mathrm{RF}}^H$: analog combiner

- **Theorem**. If the pilot symbols are zero mean bounded i.i.d. random variables, then the measurement matrix A has sub-Gaussian entries for a given total transmission power [Bal20].

[Bal20] E. Balevi and J. G. Andrews, “Wideband Channel Estimation with a Generative Adversarial Network,” IEEE Trans. Wireless, May 2021.

Wideband System Details

Channel parameters

Delay Spread	TDL-E*
N_t	64
N_r	16
N_f	64
Antenna Array	URA
Antenna Spacing	$\lambda/2$

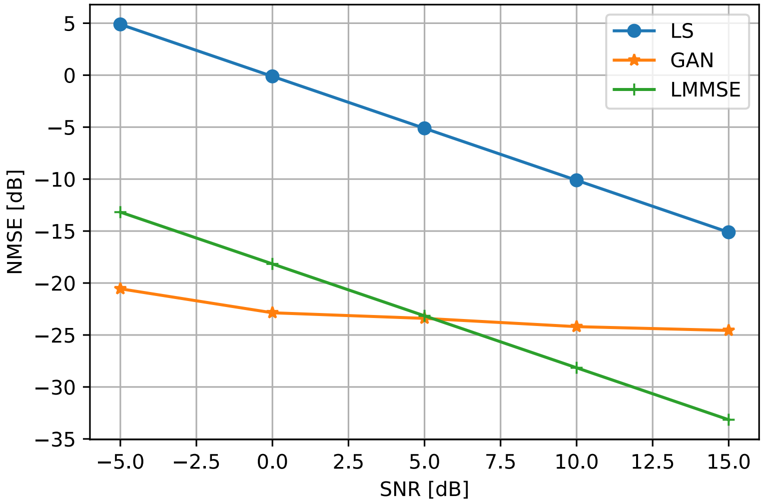
We use a Wasserstein GAN, whose architecture is the same as the narrowband estimator.

Training data size	5000
Testing data size	10
Optimizer	RMSProp
Learning Rate	0.00005
Batch Size	200
Epochs	3000
λ_{reg}	0

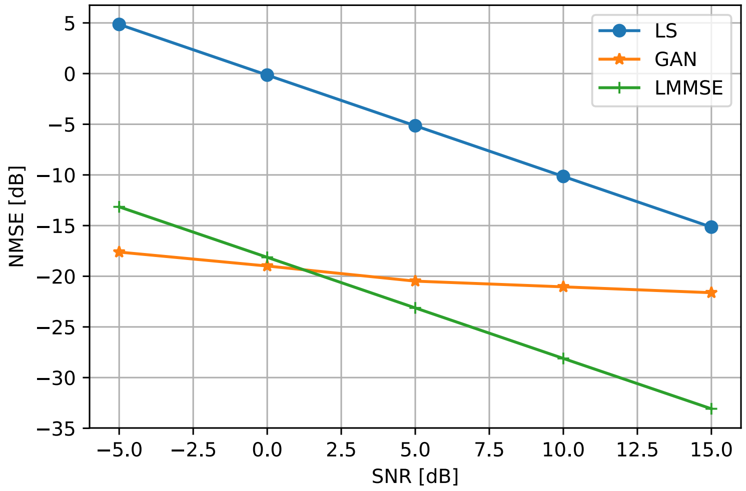
* From 3GPP specs TR 38.901

GCE outperforms LS and approaches LMMSE

- First, we benchmark with optimum performance, thus assume $N_p/N_t = 1$ for each coherence bandwidth, and compare it with:
  1. Practical low-complexity LS estimator
  2. Complex, but conventionally optimum* LMMSE estimator



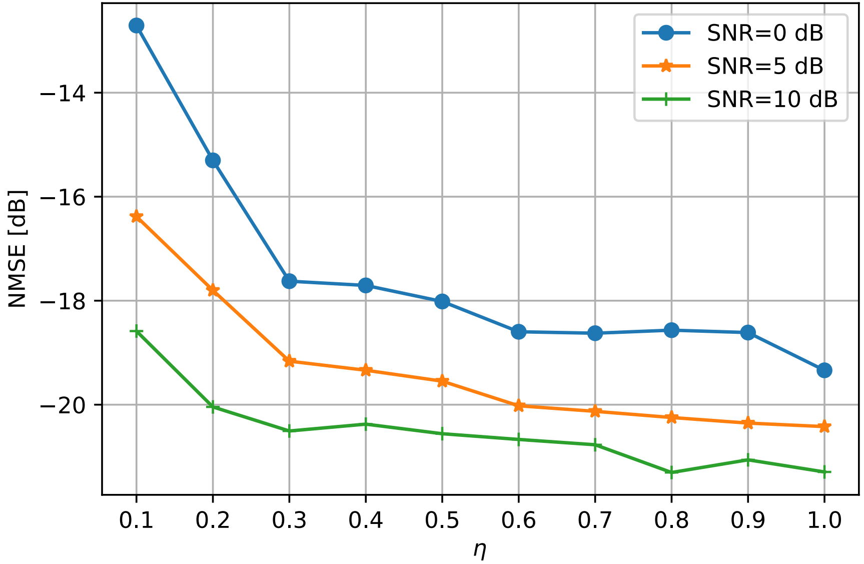
delay spread = 10ns



delay spread = 100ns

For low SNR, GCE achieves superior (even optimum*) performance.

GCE scales well with decreasing pilot overhead



$$\eta = N_p / N_f$$

The pilot overhead can be reduced by 70% with just ~1 dB loss in NMSE.

Wrap-up of GAN-based Channel Estimation

- Channel estimation is an important bottleneck for high dimensional wireless systems, such as mmWave and especially upper mmWave/THz
- Our novel generative channel estimator (GCE), leveraging deep generative networks, achieves impressive estimation accuracy and robustness
- GCE does not require knowledge of the sparsifying basis of the channel, immensely reduces the number of pilots required, and works especially well at low SNR
- The computational complexity of the proposed estimator is reasonable:
    - <u>Narrowband channel estimation:</u> O($N_t {N_r}^2$), where in downlink $N_r \ll N_t$. Better than OMP which has O(${N_t}^3 {N_r}^2$) and similar to EM-GM-AMP at O($N_t N_r \log(N_t N_r)$).
    - <u>Wideband channel estimation:</u> O($N_t N_r N_f {N_p}^2$), where $N_r, N_p \ll N_t, N_f$, i.e., increases linearly with the number of transmit antennas and subcarriers.

Parting Comments

- Deep Learning is a powerful tool for wireless systems – but not a panacea
  - Learning when and how to use it (and also when not to!) is a key research challenge for the next decade
  - Applying ML successfully for 5G/6G requires a strong communication theory and communication systems engineering background
- High dimensional channels are inherently "unknowable" (esp. with any mobility), complex/correlated, and require suboptimum RF electronics; so are a promising application space
- Industry is very excited about the potential of ML for 6G
  - They follow and support our research, and are actively doing their own studies
  - Desire for good datasets and simulators is substantial, and a big challenge right now – academia can help