Channel Prediction in High-Mobility Massive MIMO: From Spatio-Temporal Autoregression to Deep Learning

Chi Wu, *Student Member, IEEE*, Xinping Yi, *Member, IEEE*, Yiming Zhu, *Student Member, IEEE*, Wenjin Wang, *Member, IEEE*, Li You, *Member, IEEE*, and Xiqi Gao, *Fellow, IEEE*

***Abstract*—While massive multiple-input multiple-output (MIMO) has achieved tremendous success in both theory and practice, it faces a crisis of sharp performance degradation in moderate or high-mobility scenarios (e.g., 30 km/h), due to the breach of uplink-downlink channel duality. Such a "curse of mobility" has spurred the research on channel prediction in high-mobility scenarios. Instead of predicting channel response matrix in the space-frequency domain, we investigate it in the angle-delay domain by utilizing the high angle-delay resolution of wideband massive MIMO systems. Specifically, we study the general angle-delay domain channel characterization and obtain that: 1) the correlations between the angle-delay domain channel response matrix (ADCRM) elements are decoupled significantly; 2) when the number of antennas and bandwidth are limited, the decoupling is insufficient and residual correlations between the neighboring ADCRM elements exist. Then focusing on the ADCRM, we propose two channel prediction methods: a spatio-temporal autoregressive (ST-AR) model-driven unsupervised-learning method and a deep learning (DL) based data-driven supervised-learning method. While the model-driven method provides a principled way for channel prediction, the data-driven method is generalizable to various channel scenarios. In particular, ST-AR exploits the residual spatio-temporal correlations of the channel element with its most neighboring elements, and DL realizes element-wise angle-delay domain channel prediction utilizing a complex-valued neural network (CVNN). Simulation results under the 3GPP non-line-of-sight (NLOS) scenarios indicate that, compared to the state-of-the-art Prony-based angular-delay domain (PAD) prediction method [1], both the proposed ST-AR and the CVNN-based channel prediction methods can enhance the channel prediction accuracy.**

***Index Terms*—Massive MIMO, channel prediction, machine learning, autoregressive model, angle-delay domain.**

Manuscript received December 31, 2020; revised March 17, 2021; accepted April 21, 2021. The work of C. Wu, Y. Zhu, W. Wang, L. You, and X. Q. Gao was supported in part by the National Natural Science Foundation of China under Grants 61631018, 61761136016, and 61801114, the Natural Science Foundation of Jiangsu Province under Grant BK20170688, and the National Key R&D Program of China under Grant 2019YFB1803102. The work of X. Yi was supported in part by the Royal Society International Exchanges Scheme under award IEC\NSFC\201080. *(Corresponding author: Wenjin Wang.)*

C. Wu, Y. Zhu, W. Wang, L. You, and X. Q. Gao are with the National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China, and also with the Purple Mountain Laboratories, Nanjing 211111, China (e-mail: chiwu@seu.edu.cn; zhuym@seu.edu.cn; wangwj@seu.edu.cn; liyou@seu.edu.cn; xqgao@seu.edu.cn).

X. Yi is with the Department of Electrical Engineering and Electronics, University of Liverpool, L69 3BX, United Kingdom (email: xinping.yi@liverpool.ac.uk).

I. Introduction

Massive multiple-input multiple-output (MIMO) is a key technique for the fifth generation (5G) wireless communication. By deploying a large number of antennas on the base station (BS), massive MIMO can provide high angular resolution and boost spectral and energy efficiency [2]–[4]. Then combined with the orthogonal frequency-division multiplexing (OFDM), it can reduce the intersymbol interference (ISI) in multipath environment [5]. For massive MIMO-OFDM wireless transmission, many techniques highly rely on accurate channel state information (CSI) acquisition, e.g., beamforming [6] and precoding [7]. However, in the moderate to high mobility scenarios, the acquired channel at the BS tends to be outdated, provided the short coherent time against the data processing and feedback delay. Even worse, the CSI staleness becomes more severe in the high-mobility scenarios, e.g., unmanned aerial vehicle (UAV) and vehicle-to-everything (V2X) communications, where the channel changes rapidly due to the Doppler effect and multipath propagation, making the estimated CSI unusable.

To resolve the transmission issues caused by CSI staleness, several techniques have been proposed, e.g., orthogonal time frequency space (OTFS) modulation [8], robust precoding [9], and channel prediction [1]. Among them, channel prediction is the most natural way which predicts the current and future channels by exploiting the possible temporal correlation from those in the past. This is underpinned by the fact that, when modeling the time-varying channel due to stationary scattering environment as a continuous-time process, there should exist certain temporal correlations between the sampled channels across time. A variety of channel prediction methods have been investigated in the literatures, e.g., the autoregressive (AR) model-based method [10]–[13], the sum-of-sinusoids model-based method [14], and the linear extrapolation-based method [15]. As an unsupervised statistical learning method, the AR model is commonly used for time series analysis that treats time-varying variables (e.g., channel states) as a wide-sense stationary (WSS) stochastic process [16]. When treating time-varying channel as a time series, machine learning is a natural choice for channel prediction in either an unsupervised or a supervised manner, by exploiting channel temporal correlation explicitly or implicitly from the previous channel data [17]. Compared with other methods that require accurate model parameter estimation, the AR can obtain model parameters by

a learning-to-predict strategy. In particular, as an unsupervised learning method, the AR establishes a model with parameters, which can be learned by fitting the model with the previous channel data and the underlying temporal correlation, thereby avoiding the estimation error in mathematical modeling. Nevertheless, due to the complex caused by the superposition of multipath, it is difficult to implement accurate channel prediction in conventional communication systems with a small number of antennas or narrow bandwidth. When it comes to modern massive MIMO systems, accurate channel prediction becomes more feasible thanks to the large number of antennas and wide bandwidth.

Configured with large antenna array, massive MIMO-OFDM systems bring a high resolution of multipath angle and delay, which offer new opportunities for accurate channel prediction. Very recently, various channel prediction methods have been proposed for massive MIMO/MIMO-OFDM systems [1], [10], [11], [18]. In [10], two AR-based channel prediction methods are proposed which utilize the spatio-temporal correlation across antennas. In [18], the fading coefficient of each path is first tracked, followed by a first Taylor (FIT) model to predict the variations of fading coefficients. The massive MIMO-OFDM channel in the space-frequency domain can be transformed into that in the angle-delay domain via inverse discrete Fourier transform (IDFT) [3], [19], thus the multipaths with different angles or delays can be decoupled significantly. Based on such transformation, [1] assumes that the channel is superimposed by multipath sine waves with determined path gains and proposes a Prony-based angle-delay domain (PAD) prediction method. It significantly improves the channel prediction accuracy and relieves the "curse of mobility" when the number of antennas and bandwidth are large. Meanwhile, by utilizing channel sparsity in the angle-delay domain, an AR-based channel prediction method with reduced computational complexity is proposed in [11]. However, the previous angle-delay domain prediction methods do not consider the IDFT leakage introduced by the limited number of antennas and bandwidth, which causes aliasing in the angle and delay domain and therefore degrades the prediction accuracy. While the above model-based methods provide analytical guarantees, it requires to frequently update model parameters to fit the change of environment, which incurs huge extra computational complexity. To make the channel prediction more adaptive to various environments, it calls for robust and generalizable approaches.

Deep learning (DL) is known for its excellent generalization performance in the sense that once a deep neural network is well-trained with sufficient data, the trained model is able to work well for unseen data [20]. This is most suitable for adaptive designs of physical layer wireless communications [21]. By treating the time-varying channel as a time series, the channel prediction can be cast into a sequence learning problem, thereby motivating many DL-based prediction methods [13], [22]–[28]. In [22], a multi-layer perceptron (MLP)-based channel predictor is proposed for massive MIMO systems. [25] proposes a recurrent neural network (RNN)-based channel predictor using the whole vectorized channel matrix as the input and [23] proposes a multi-step RNN that can flexibly tune the number of prediction steps for massive MIMO systems. In [13], a convolutional neural network (CNN)-AR/RNN channel predictor is proposed, where the CNN extracts the autocorrelation function (ACF) pattern to determine the coefficients of the AR or RNN channel predictor. However, the above learning-based methods are implemented in the space/frequency domain channel without exploiting the angle-delay resolution from large-scale antenna array and wide bandwidth, thereby still facing the accuracy problem brought by complicated multipath superposition.

In this paper, we address the above problems by exploiting the structural properties of channel characterization in the angle-delay domain under the massive MIMO-OFDM configurations. Specifically, with the transformed channel matrix from space-frequency to angle-delay domain via IDFT, we propose a spatio-temporal AR (ST-AR)-based channel prediction method and a complex-valued neural network (CVNN)-based channel prediction method. The main contributions of our paper are as follows:

- Considering the impact of IDFT leakage, we develop a general angle-delay domain channel characterization for a finite number of antennas and bandwidth. We prove that: 1) When the number of antennas and bandwidth are infinite, the different angle-delay domain channel response matrix (ADCRM) elements are uncorrelated with each other in the time domain. 2) When the number of antennas and bandwidth are limited, there exist residual temporal correlations between the adjacent elements of ADCRM brought by IDFT leakage.
- According to the sparse structure of ADCRM, we propose an adaptive selection method of the significant elements of ADCRM according to the signal-to-noise ratio (SNR) to eliminate the influence of noise contamination, so that the predictions can only be implemented on the significant elements to reduce the computational complexity.
- We propose a ST-AR-based channel prediction method by leveraging the residual temporal correlations between the neighboring elements of the ADCRM. Differently from the channel modeling with deterministic multipath gains as in [1], [14], we treat multipath gain as a stochastic process, which captures both randomness of path gain and temporal correlation over time. By such modeling, we are able to decompose the channel into independent multipaths distinguished by delay or angle in the ADCRM. Going beyond the conventional AR model, we propose a ST-AR channel prediction method which comprehensively considers the IDFT leakage for finite antennas and bandwidth. The ST-AR works in an unsupervised learning way—the predicted current or future channels are directly learned from the previous channel data without training.
- We propose a data-based channel prediction method by constructing a CVNN architecture, leveraging the advantageous generalization performance of DL. Motivated by the path decoupling of the angle-delay domain channel, we construct a unified CVNN model to fit the pattern of changes of all ADCRM significant elements. Once trained

efficiently, the CVNN with high generalization ability can realize prediction in different scenarios. Notably, with the proposed CVNN-based channel prediction, it is ready to realize flexible predictions, e.g., consecutive channel prediction, by simply changing the output size of CVNN.

The rest of this paper is organized as follows: In Section II, we introduce the massive MIMO-OFDM channel model. In Section III, we investigate the angle-delay domain channel prediction. Then we propose the ST-AR and CVNN-based channel prediction methods in Sections IV and V, respectively. Finally, simulation results are shown in Section VI.

Notations: We use $\bar{\jmath} = \sqrt{-1}$ to represent the imaginary unit. Vectors and matrices are in lowercase boldface and uppercase boldface, respectively (e.g., vector $\mathbf{a}$ and matrix $\mathbf{A}$), with the element's index starts from 0. $[\cdot]_i$ denotes the ith element of a vector or the ith column of a matrix, and $[\cdot]_{i,j}$ denotes the (i,j)th element of a matrix. $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^H$ represent the transpose, conjugate, and conjugate-transpose of a vector or matrix, respectively. The complex number field is represented by $\mathbb{C}$. The symbol $\otimes$ denotes the Kronecker product. $E\{\cdot\}$ denotes the expectation operation. $\lfloor x \rfloor$ represents the maximum integer not greater than x, and $\langle \cdot \rangle_N$ represents the modulo-N operation. $\delta(\cdot)$ is the delta function. Given an integer M, we define $[M] \triangleq \{0, 1, \cdots, M-1\}$.

II. Massive MIMO-OFDM Channel Model in Mobile Environment

A. Channel Model

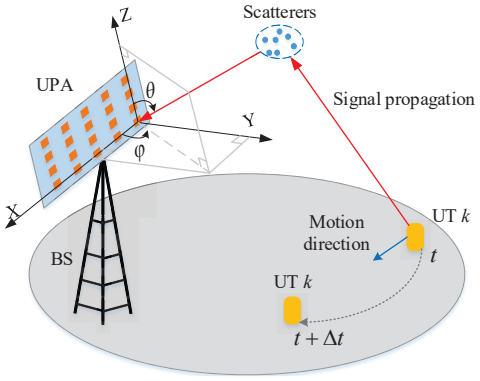


Fig. 1: Massive MIMO-OFDM system with moving UT.

We consider the uplink transmission shown in Fig. 1, where the BS is equipped with a uniform planar array (UPA) with one-half wavelength antenna spacing, and the user terminal (UT) is equipped with an omni-directional antenna. The UPA comprises M antennas in each column and N antennas in each row with the number of antennas $N_a = MN$. Given a certain angle of arrival (AOA) which can be decomposed into the elevation AOA $\theta \in [0, \pi]$ and the azimuth AOA $\varphi \in [0, \pi]$, the corresponding three dimensional (3D) array response vector $\mathbf{a}(\theta, \varphi)$ can be written as [2], [29], [30]

$$\mathbf{a}(\theta, \varphi) = \mathbf{a}^{(v)}(\theta) \otimes \mathbf{a}^{(h)}(\varphi; \theta), \tag{1}$$

where

$$\mathbf{a}^{(v)}(\theta) = [1, e^{-\bar{\jmath}\pi\cos\theta}, \cdots, e^{-\bar{\jmath}\pi(M-1)\cos\theta}]^T, \tag{2}$$

$$\mathbf{a}^{(h)}(\varphi; \theta) = [1, e^{-\bar{\jmath}\pi\sin\theta\cos\varphi}, \cdots, e^{-\bar{\jmath}\pi(N-1)\sin\theta\cos\varphi}]^T. \tag{3}$$

Before transmission, the signal is modulated by OFDM comprising N_c sub-carriers and N_g-length ($N_g < N_c$) cyclic prefix (CP). Denote T_s as the system sampling interval and $B = \frac{1}{T_s}$ as the signal bandwidth. We use $T_c = N_c T_s$ and $T_g = N_g T_s$ to represent the OFDM symbol duration and the CP duration, respectively. Therefore the frequency spacing is $\Delta f = \frac{1}{T_c}$. Generally, we assume the CP duration T_g is larger than the maximum UT's channel delay, i.e., $T_g > \tau_{\max}$ [3], [31]. Given a certain delay $\tau \in (0, \tau_{\max}]$, the delay response vector can be defined as [1]

$$\mathbf{b}(\tau) = \left[1, e^{-\bar{\jmath}2\pi\frac{1}{T_c}\tau}, \cdots, e^{-\bar{\jmath}2\pi\frac{N_c-1}{T_c}\tau}\right]^T. \tag{4}$$

Consider the time-varying MIMO-OFDM channel where the variation of the channel is caused by the movement of the UT, and the signal from the UT to the BS propagates through the multipath environment. We first define the AOA-delay tuple as (θ, ϕ, τ), by which we specify a certain path. Different from the channel modeling in [1] that considers each physical path separately, we merge multiple physical paths with similar AOA-delay tuples into one single virtual path. The virtual path can be seen as a stochastic process with realizations being those merged physical paths. By doing so, the AOA-delay tuples are distinguishable for different virtual paths, so that the physical paths are grouped and then decoupled to a certain extent. The complex gain, Doppler shift, and initial phase of a certain path of the kth UT can be defined as the functions of the tuple (θ, ϕ, τ) represented by $c_k(\theta, \varphi, \tau)$, $\nu_k(\theta, \varphi, \tau)$, and $\phi_k(\theta, \varphi, \tau)$, respectively. We assume the uncorrelated Rayleigh scattering environment in which the different paths are mutually independent, so that the complex path gain satisfies [3], [32]

$$\begin{aligned} &E\left\{c_k(\theta, \varphi, \tau) {c_k}^*(\theta', \varphi', \tau')\right\} \\ &= S_k(\theta, \varphi, \tau)\, \delta(\theta - \theta')\, \delta(\varphi - \varphi')\, \delta(\tau - \tau'), \end{aligned} \tag{5}$$

where $S_k(\theta, \varphi, \tau)$ is the power angle-delay spectrum. Since the scattering environment is stationary in a period of time, the angle between the path with respect to the AOA-delay tuple (θ, ϕ, τ) and the motion direction of the kth UT $\psi_k(\theta, \varphi, \tau)$ remains unchanged. Given the velocity of the kth UT as v_k, the Doppler shift $\nu_k(\theta, \varphi, \tau)$ can be represented as [33]

$$\nu_k(\theta, \varphi, \tau) = \frac{f_c v_k \cos\left(\psi_k(\theta, \varphi, \tau)\right)}{c}, \tag{6}$$

where f_c is the carrier frequency, and c is the light speed. We assume that the initial phase $\phi_k(\theta, \varphi, \tau)$ is uniformly distributed in $[-\pi, \pi)$, i.e., $\phi_k(\theta, \varphi, \tau) \sim U[-\pi, \pi)$, and independent of $\phi_k(\theta', \varphi', \tau')$ for $(\theta, \varphi, \tau) \neq (\theta', \varphi', \tau')$.

Generally, the variations of the multipath angles and delays are much slower than that of the fading coefficients [1], [18]. We assume the multipath angles and delays can be regarded to be time-invariant compared with the fading coefficients. Then based on the above modeling of the complex gain, Doppler shift, and initial phase, we describe the time-varying space-frequency domain channel response matrix (SFCRM) of the

kth UT $\mathbf{G}_k(t) \in \mathbb{C}^{N_a \times N_c}$ as[1]

$$\begin{aligned}\mathbf{G}_k(t) = &\int_0^{\pi}\int_0^{\pi}\int_0^{\tau_{\max}} c_k(\theta,\varphi,\tau)\mathbf{a}(\theta,\varphi) \\ &\cdot \mathbf{b}^T(\tau) e^{\bar{\jmath}2\pi\nu_k(\theta,\varphi,\tau)t+\bar{\jmath}\phi_k(\theta,\varphi,\tau)} d\tau d\varphi d\theta.\end{aligned} \tag{7}$$

B. General Angle-Delay Domain Channel Characterization

Based on the time-varying SFCRM in (7), we next investigate the general angle-delay domain channel characterization for massive MIMO systems with general configurations, i.e., a large yet finite number of antennas and bandwidth. Let us define the IDFT of the array response vectors as

$$\begin{aligned}\tilde{\mathbf{a}}(\theta,\varphi) &\triangleq \frac{1}{\sqrt{MN}}(\mathbf{V}_M \otimes \mathbf{V}_N)^H \mathbf{a}(\theta,\varphi) \\ &\overset{(a)}{=} \underbrace{\left(\frac{1}{\sqrt{M}}\mathbf{V}_M^H \mathbf{a}^{(v)}(\theta)\right)}_{\triangleq \tilde{\mathbf{a}}^{(v)}(\theta)} \otimes \underbrace{\left(\frac{1}{\sqrt{N}}\mathbf{V}_N^H \mathbf{a}^{(h)}(\varphi;\theta)\right)}_{\triangleq \tilde{\mathbf{a}}^{(h)}(\varphi;\theta)},\end{aligned} \tag{8}$$

where (a) follows from (1) and uses the Kronecker product property $(\mathbf{A}\otimes\mathbf{B})(\mathbf{C}\otimes\mathbf{D}) = (\mathbf{AC})\otimes(\mathbf{BD})$, $\mathbf{V}_M \in \mathbb{C}^{M\times M}$ and $\mathbf{V}_N \in \mathbb{C}^{N\times N}$ are phase-shift DFT matrices with the (i,j)th elements given by $[\mathbf{V}_M]_{i,j} \triangleq \frac{1}{\sqrt{M}} e^{-\bar{\jmath}2\pi\frac{i(j-M/2)}{M}}$ and $[\mathbf{V}_N]_{i,j} \triangleq \frac{1}{\sqrt{N}} e^{-\bar{\jmath}2\pi\frac{i(j-N/2)}{N}}$, respectively. Then the ith elements of $\tilde{\mathbf{a}}^{(v)}(\theta)$ and $\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)$ are calculated as

$$\begin{aligned}\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i &= \frac{1}{\sqrt{M}}\sum_{m=0}^{M-1}\frac{1}{\sqrt{M}} e^{\bar{\jmath}2\pi\frac{m(i-M/2)}{M}} e^{-\bar{\jmath}\pi m\cos\theta} \\ &= f_M\left(\frac{1}{2}(\cos\theta+1) - \frac{i}{M}\right),\end{aligned} \tag{9}$$

and

$$\begin{aligned}\left[\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)\right]_i &= \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}\frac{1}{\sqrt{N}} e^{\bar{\jmath}2\pi\frac{n(i-N/2)}{N}} e^{-\bar{\jmath}\pi n\sin\theta\cos\varphi} \\ &= f_N\left(\frac{1}{2}(\sin\theta\cos\varphi+1) - \frac{i}{N}\right),\end{aligned} \tag{10}$$

respectively, with the following definition

$$f_M(x) \triangleq e^{-\bar{\jmath}\pi(M-1)x} \cdot \frac{\sin(\pi M x)}{M\sin(\pi x)}. \tag{11}$$

Similarly, we also define the IDFT of the delay response vector

$$\tilde{\mathbf{b}}(\tau) \triangleq \frac{1}{\sqrt{N_c}}\mathbf{F}_{N_c\times N_g}^H \mathbf{b}(\tau), \tag{12}$$

where $\mathbf{F}_{N_c\times N_g} \in \mathbb{C}^{N_c\times N_g}$ is the matrix composed of the first N_g columns of DFT matrix $\mathbf{F}_{N_c}$ with the (i,j)th element $[\mathbf{F}_{N_c\times N_g}]_{i,j} \triangleq \frac{1}{\sqrt{N_c}} e^{-\bar{\jmath}2\pi\frac{ij}{N_c}}$. Then the jth element of $\tilde{\mathbf{b}}(\tau)$ is

$$\begin{aligned}\left[\tilde{\mathbf{b}}(\tau)\right]_j &= \frac{1}{\sqrt{N_c}}\sum_{l=0}^{N_c-1}\frac{1}{\sqrt{N_c}} e^{\bar{\jmath}2\pi\frac{jl}{N_c}} e^{-\bar{\jmath}2\pi\frac{l\tau}{N_c T_s}} \\ &= f_{N_c}\left(\frac{\tau}{T_s N_c} - \frac{j}{N_c}\right).\end{aligned} \tag{13}$$

[1]Such a channel modeling approach is also generally used in previous literature, refer to [3], [34]–[36].

To summarize, the IDFT counterparts of the three response vectors, i.e., $\tilde{\mathbf{a}}^{(v)}(\theta)$, $\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)$, and $\tilde{\mathbf{b}}(\tau)$ share the same representation of (11), and hence possess some common properties which will be described in Lemma 1.

Lemma 1: For $i \in [M]$, $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i$ is a function of the elevation angle θ with the following properties:

P1: The peak point of the function $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i$ is $(\theta_i, 1)$, i.e., $\left[\tilde{\mathbf{a}}^{(v)}(\theta_i)\right]_i = 1$, where θ_i is defined as

$$\theta_i \triangleq \arccos\left(\frac{2i}{M} - 1\right). \tag{14}$$

P2: The two zero points closest to the peak point of the function $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i$ are $(\theta_{i+1}, 0)$ and $(\theta_{i-1}, 0)$, respectively, i.e., $\left[\tilde{\mathbf{a}}^{(v)}(\theta_{i+1})\right]_i = \left[\tilde{\mathbf{a}}^{(v)}(\theta_{i-1})\right]_i = 0$.

P3: The intersection points of the amplitude of the function $\left|\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i\right|$ and those of the nearest functions, i.e., $\left|\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_{i+1}\right|$ and $\left|\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_{i-1}\right|$, are $\left(\theta_{i-1/2}, \frac{2}{\pi}\right)$ and $\left(\theta_{i-1/2}, \frac{2}{\pi}\right)$, respectively, when M is sufficiently large[2], i.e., $\left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i+1/2}\right)\right]_i\right| = \left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i+1/2}\right)\right]_{i+1}\right| = \frac{2}{\pi}$ and $\left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i-1/2}\right)\right]_i\right| = \left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i-1/2}\right)\right]_{i-1}\right| = \frac{2}{\pi}$.

P4: $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i$ is asymptotically equal to $\delta(\theta-\theta_i)$ when M tends to infinite, i.e.,

$$\lim_{M\to\infty}\left(\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i - \delta(\theta-\theta_i)\right) = 0. \tag{15}$$

For $q \in [N]$ and $j \in [N_c]$, the above four properties also apply to $\left[\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)\right]_q$ as a function of the azimuth angle ϕ, with θ_i replaced by $\varphi_q \triangleq \arccos\left(\left(\frac{2q}{N}-1\right)\frac{1}{\sin\theta}\right)$, and $\left[\tilde{\mathbf{b}}(\tau)\right]_j$ as a function of the delay τ, with θ_i replaced by $\tau_j \triangleq jT_s$.

Proof: See Appendix A. □

Combining the IDFTs of the array response vector and delay response vector, we transform the SFCRM $\mathbf{G}_k(t)$ into the ADCRM $\mathbf{H}_k(t) \in \mathbb{C}^{N_a\times N_g}$, given as

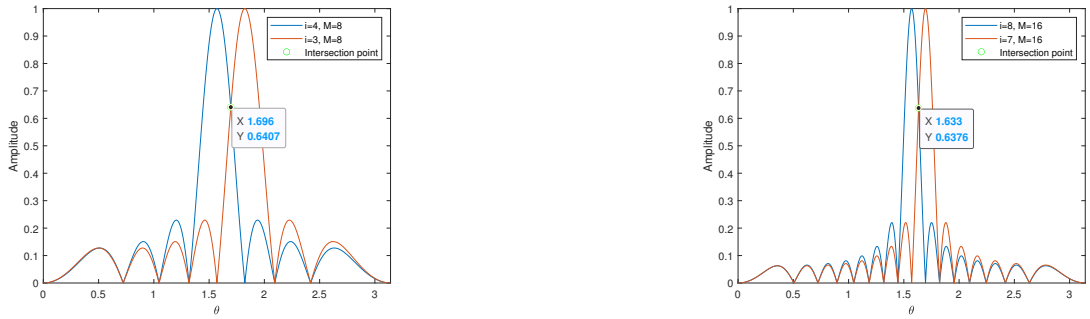
$$\mathbf{H}_k(t) \triangleq \frac{1}{\sqrt{MNN_c}}(\mathbf{V}_M\otimes\mathbf{V}_N)^H \mathbf{G}_k(t)\mathbf{F}_{N_c\times N_g}^*. \tag{16}$$

By substituting (7), (8), and (12) into (16), we calculate the (i,j)th element of the ADCRM as

$$\begin{aligned}&[\mathbf{H}_k(t)]_{i,j} \\ &= \int_0^{\pi}\int_0^{\pi}\int_0^{\tau_{\max}} c_k(\theta,\varphi,\tau)[\tilde{\mathbf{a}}(\theta,\varphi)]_i \\ &\quad\cdot\left[\tilde{\mathbf{b}}^T(\tau)\right]_j e^{j2\pi\nu_k(\theta,\varphi,\tau)t+j\phi_k(\theta,\varphi,\tau)} d\tau d\varphi d\theta \\ &\overset{(a)}{=} \int_0^{\pi}\int_0^{\pi}\int_0^{\tau_{\max}} c_k(\theta,\varphi,\tau)\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_{i_M}\left[\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)\right]_{i_N} \\ &\quad\cdot\left[\tilde{\mathbf{b}}(\tau)\right]_j e^{j2\pi\nu_k(\theta,\varphi,\tau)t+j\phi_k(\theta,\varphi,\tau)} d\tau d\varphi d\theta,\end{aligned} \tag{17}$$

where (a) follows from the definition of Kronecker product with the auxiliary symbols $i_M \triangleq \lfloor i/N \rfloor$ and $i_N \triangleq \langle i \rangle_N$. Then the representation of (17) asymptotically converges to a closed-form, as demonstrated in Lemma 2 below.

[2]In practice, $M \geq 8$ is a sufficiently large number according to simulations.



(a) (b)

Fig. 2: The curve of the amplitude $\left|\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i\right|$ varying with θ. (a) M=8 with $i=3$ and 4. (b) $M=16$ with $i=7$ and 8.

Lemma 2: When M, N, and N_c tend to infinity, the (i,j)th element of the ADCRM $\mathbf{H}_k(t)$ is asymptotically equal to $\left[\mathbf{H}_k^{\text{inf}}(t)\right]_{i,j}$, i.e.,

$$\lim_{M,N,N_c\to\infty}\left(\left[\mathbf{H}_k(t)\right]_{i,j}-\left[\mathbf{H}_k^{\text{inf}}(t)\right]_{i,j}\right)=0, \tag{18}$$

where $\left[\mathbf{H}_k^{\text{inf}}(t)\right]_{i,j}$ is defined as

$$\begin{aligned}\left[\mathbf{H}_k^{\text{inf}}(t)\right]_{i,j}=&S_k(\theta_{i_M},\varphi_{i_N},\tau_j)\\&\cdot e^{j2\pi\nu_k\left(\theta_{i_M},\varphi_{i_N},\tau_j\right)t+j\phi_k\left(\theta_{i_M},\varphi_{i_N},\tau_j\right)}.\end{aligned} \tag{19}$$

Proof: By substituting P4 in Lemma 1 into (17), (19) can be immediately obtained and hence omitted here. □

Provided with the insights of the asymptotic results in Lemma 2, we take a closer look at (19) in the angle-delay domain for finite M, N, and N_c. Fig. 2(a) and Fig. 2(b) show the curves of the amplitude $\left|\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i\right|$ with different M and i. The function $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i$ is constituted with one main lobe and many side lobes, and as the number of the column antennas M increases, $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i$ is concentrated on the main lobe more distinctly. Inspired by this, we approximate $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_{i_M}$ in (17) by confining its value within the main lobe. For an integer $i_M\in[M]$, we define the following partition of the elevation angle θ as

$$O_{\theta,i_M}\triangleq\left[\theta_{i_M+\Delta_1},\theta_{i_M-\Delta_1}\right). \tag{20}$$

where Δ_1 is defined to partition the elevation angle domain. According to P2 in Lemma 1, the main lobe of $\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_{i_M}$ exists with $\theta\in[\theta_{i_M+1},\theta_{i_M-1})$. Hence, Δ_1 is confined as $0<\Delta_1\le 1$. Note that $\left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i_M+\Delta_1}\right)\right]_{i_M}\right|=\left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i_M-\Delta_1}\right)\right]_{i_M}\right|$ holds due to the definitions in (9), (11), and (14). Given a threshold $0\le\zeta<1$, Δ_1 is determined by the following equation

$$\left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i_M+\Delta_1}\right)\right]_{i_M}\right|=\left|\left[\tilde{\mathbf{a}}^{(v)}\left(\theta_{i_M-\Delta_1}\right)\right]_{i_M}\right|=\zeta. \tag{21}$$

Analogously, we also make the following partitions of the azimuth angle φ and delay τ. For an integer $i_N\in[N]$,

$$O_{\varphi,i_N}\triangleq\left[\varphi_{i_N+\Delta_2},\varphi_{i_N-\Delta_2}\right), \tag{22}$$

and for an integer $j\in[N_c]$,

$$O_{\tau,j}\triangleq\left[\tau_{j-\Delta_3},\tau_{j+\Delta_3}\right), \tag{23}$$

where $0<\Delta_2\le 1$ and $0<\Delta_3\le 1$ are defined to partition the azimuth angle and delay domains, respectively, and determined by the following equations

$$\left|\left[\tilde{\mathbf{a}}^{(h)}\left(\varphi_{i_N+\Delta_2};\theta\right)\right]_{i_N}\right|=\left|\left[\tilde{\mathbf{a}}^{(h)}\left(\varphi_{i_N-\Delta_2};\theta\right)\right]_{i_N}\right|=\zeta, \tag{24}$$

and

$$\left|\left[\tilde{\mathbf{b}}\left(\tau_{j-\Delta_3}\right)\right]_j\right|=\left|\left[\tilde{\mathbf{b}}\left(\tau_{j+\Delta_3}\right)\right]_j\right|=\zeta, \tag{25}$$

respectively. Specifically, Δ_1, Δ_2, and Δ_3 have the following asymptotic representations.

Lemma 3: Given a certain threshold ζ, Δ_1, Δ_2, and Δ_3 are the same for all $i_M\in[M]$, $i_N\in[N]$, and $j\in[N_c]$, respectively, and equal to $\text{sinc}^{-1}(\zeta)\in(0,1]$ when M, N, and N_c tend to infinity, i.e.,

$$\Delta_1=\Delta_2=\Delta_3=\text{sinc}^{-1}(\zeta)\in(0,1]\ \text{as}\ M,N,N_c\to\infty, \tag{26}$$

where $\text{sinc}^{-1}(\cdot)\in(0,1]$ is the inverse of the function $\text{sinc}(x)\triangleq\frac{\sin(\pi x)}{\pi x}$ with $x\in(0,1]$.

Proof: See Appendix B. □

For simplicity, we define the set

$$\bar{O}_{i,j}\triangleq\left\{(\theta,\varphi,\tau):\theta\in O_{\theta,i_M},\varphi\in O_{\varphi,i_N},\text{and}\ \tau\in O_{\tau,j}\right\}, \tag{27}$$

where O_{θ,i_M}, O_{φ,i_N}, and $O_{\tau,j}$ are defined in (20), (22), and (23), respectively. Then by setting the appropriate threshold ζ, we can approximate the (i,j)th element of the ADCRM as

$$\begin{aligned}\left[\mathbf{H}_k(t)\right]_{i,j}\approx&\iiint\limits_{\bar{O}_{i,j}}c_k(\theta,\varphi,\tau)\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_{i_M}\left[\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)\right]_{i_N}\\&\cdot\left[\tilde{\mathbf{b}}(\tau)\right]_j e^{\bar{\jmath}2\pi\nu_k(\theta,\varphi,\tau)t+\bar{\jmath}\phi_k(\theta,\varphi,\tau)}d\tau d\varphi d\theta.\end{aligned} \tag{28}$$

It is worth noting that the above approximation is also considered for sufficiently large M, N, and N_c in [4], [32], in which Δ_1, Δ_2, and Δ_3 are all set to $\frac{1}{2}$ directly with the corresponding $\zeta=\frac{2}{\pi}$ according to Lemma 3, as such the paths contained in different ADCRM elements are not aliased according to P3 in

Lemma 1. In addition, (19) in Lemma 2 is also a special case of (28) by setting $\zeta \to 1$ when M, N, and N_c tend to infinity. The approximation in (28) is established for general cases of finite M, N, and N_c, when the IDFT leakage leads to aliasing between $[\mathbf{H}_k(t)]_{i,j}$ and $[\mathbf{H}_k(t)]_{i',j'}$ in the angle-delay domain with $(i,j) \neq (i',j')$.

III. Channel Prediction in the Angle-Delay Domain

A. Channel Temporal Correlation in Angle-Delay Domain

As channel temporal correlation plays a crucial role in channel predictin, we start with the investigation of temporal correlation of ADCRM. Given the definition of temporal correlation between the (i,j)th and (i',j')th elements of ADCRM as

$$R^{\mathrm{AD}}_{i,j,i',j'}(\Delta t) \triangleq E\left\{[\mathbf{H}_k(t+\Delta t)]_{i,j}[\mathbf{H}_k(t)]^*_{i',j'}\right\}, \quad (29)$$

we show that $R^{\mathrm{AD}}_{i,j,i',j'}$ has a compact support, as in Theorem 1.

Theorem 1: Given any two elements of time-varying ADCRM, e.g., $[\mathbf{H}_k(t+\Delta t)]_{i,j}$ and $[\mathbf{H}_k(t)]_{i',j'}$, the temporal correlation between them yields

$$\begin{aligned} &R^{\mathrm{AD}}_{i,j,i',j'}(\Delta t) = 0, \text{ if } |i_M - {i'}_M| \geq 2\Delta_1 \\ &\text{or } |i_N - {i'}_N| \geq 2\Delta_2 \text{ or } |j-j'| \geq 2\Delta_3, \end{aligned} \quad (30)$$

where Δ_1, Δ_2, and Δ_3 are defined as the partitions in (20), (22), and (23), respectively. When $M \to \infty$, $N \to \infty$, and $N_c \to \infty$, channel temporal correlation asymptotically concentrates on

$$\begin{aligned} \lim_{M.N,N_c\to\infty} \Big(&R^{\mathrm{AD}}_{i,j,i',j'}(\Delta t) - S_k(\theta_{i_M},\varphi_{i_N},\tau_j) \\ &\cdot e^{\bar{\jmath}2\pi\nu_k(\theta_{i_M},\varphi_{i_N},\tau_j)\Delta t}\delta(i-i')\delta(j-j')\Big) = 0, \end{aligned} \quad (31)$$

which indicates that the temporal correlation of each element of ADCRM is disentangled from other elements.

Proof: See Appendix C. □

Theorem 1 reveals that the temporal correlation $R^{\mathrm{AD}}_{i,j,i',j'}$ is bounded within a compact support. As such, the temporal correlation between two ADCRM elements that are far enough from each other is zero. Theorem 1 also gives the asymptotic representation of $R^{\mathrm{AD}}_{i,j,i',j'}$ for the infinite number of antennas and bandwidth. For the finite case, we further give explicit compact support of $R^{\mathrm{AD}}_{i,j,i',j'}$ in Corollary 1.

Remark 1: From Theorem 1, when $M \to \infty$, $N \to \infty$, and $N_c \to \infty$, the temporal correlations between different ADCRM elements are decoupled. Thus, the prediction of one ADCRM element does not rely on other ADCRM elements, which inspires the element-wise prediction method for the massive MIMO-OFDM channel.

Remark 2: Compared with the SFCRM, the ADCRM has a much smaller dimension. In addition, due to the finite number of multipath components, most of the elements in the ADCRM are close to zeros which yields the property of channel sparsity. Therefore the number of elements that need to be predicted is even smaller.

Corollary 1: When M, N, and N_c are limited/finite, the prediction of the (i,j)th element of the ADCRM is uncorrelated with the historical values of the (i',j')th element of the ADCRM for all $(i',j') \notin \Lambda_{i,j}$, i.e.,

$$R^{\mathrm{AD}}_{i,j,i',j'}(\Delta t) = 0, \quad \forall (i',j') \notin \Lambda_{i,j}, \quad (32)$$

where $\Lambda_{i,j}$ is defined as

$$\begin{aligned} \Lambda_{i,j} \triangleq \big\{(i',j') : &|i'_M - i_M| \leq 1, \\ &|i'_N - i_N| \leq 1, \text{and } |j'-j| \leq 1\big\}. \end{aligned} \quad (33)$$

Proof: For finite M, N, and N_c, Δ_1, Δ_2, and Δ_3 are bounded no larger than 1, hence (32) can be obtained by substituting $\Delta_1, \Delta_2, \Delta_3 \leq 1$ into (30). □

Remark 3: Corollary 1 reveals that when M, N, and N_c are finite, the temporal correlation only exists between one ADCRM element and a few of its neighboring ones. Hence, we can exploit such a property of locality in temporal correlation to promote the accuracy of the prediction of the current ADCRM element by only considering the neighboring ones.

B. Channel Prediction Using Channel Sparsity

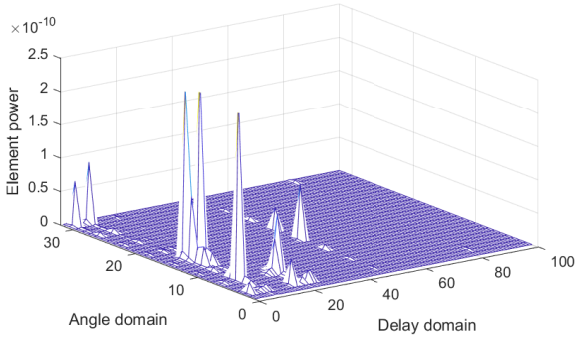


Fig. 3: The power density of ADCRM in the case of $M = 8$, $N = 16$, and $B = 30$ MHz.

From Fig. 3, we observe that the power of the most elements of ADCRM equals or approaches to zero in the practical scenarios. It is evidenced that the channel sparsity still exists when the number of antennas and bandwidth are limited. It suggests that the prediction could be dedicated to the significant elements, while the non-significant ones (those close or equal to zero) can be ignored without much loss of prediction accuracy. Besides, the changes of the delay and angle over time are much slower than the change of the fading coefficient over time. Therefore, in the process of channel prediction, the indices of the significant element determined by the channel delays and angles are fixed. To this end, we introduce a criterion to distinguish the significant elements from the non-significant ones. Define $\Omega_s = \{(i_s,j_s)\}_{s=0}^{N_s-1}$ as the set of the indices of the significant elements of the ADCRM, where $N_s = |\Omega_s|$ denotes the number of the significant elements. By selecting the ADCRM elements with

the largest power, the set Ω_s is determined by

$$\begin{aligned}\Omega_s =& \arg\min_{\Omega_s} |\Omega_s| \\ \text{s.t.} & \sum_{(i_s,j_s)\in\Omega_s} E\left\{\left|\left[\mathbf{H}_k\left(t\right)\right]_{i_s,j_s}\right|^2\right\} \\ & \geqslant \lambda \sum_{i=0}^{N_a-1}\sum_{j=0}^{N_g-1} E\left\{\left|\left[\mathbf{H}_k\left(t\right)\right]_{i,j}\right|^2\right\},\end{aligned} \tag{34}$$

where λ is the percentage of the sum power of the significant elements accounting for that of the overall elements in ADCRM. Setting λ to be close to 1 will still yield a small N_s in practice, because the power of ADCRM is dominated by a few most significant ones. Generally, N_s is much smaller than original size of the ADCRM $\mathbf{H}_k\left(t\right)$, i.e., $N_s \ll N_a N_g$. Thus the computational complexity can be greatly reduced by implementing prediction only on these significant elements.

In addition, when the acquired channel is contaminated with noise, the ratio λ can be set adaptively according to the SNR. We represent the estimated channel with noise as

$$\tilde{\mathbf{G}}_k\left(t\right) = \mathbf{G}_k\left(t\right) + \mathbf{Z}_k\left(t\right), \tag{35}$$

where $\mathbf{Z}_k\left(t\right) \in \mathbb{C}^{N_a\times N_c}$ denotes the noise contamination, whose elements are assumed to be identically and independently distributed (i.i.d.) as $\mathcal{CN}\left(0,\sigma_z^2\right)$ with σ_z^2 being the noise power. The SNR in dB of the estimated channel is given by

$$\mathrm{SNR} = 10\log\left(\frac{\sum_{i=0}^{N_a-1}\sum_{j=0}^{N_c-1} E\left\{\left|\left[\mathbf{G}_k\left(t\right)\right]_{i,j}\right|^2\right\}}{N_a N_c \sigma_z^2}\right). \tag{36}$$

Proposition 1: For the noisy ADCRM with SNR in dB, a proper percentage value of λ can be approximately set to

$$\lambda \approx \frac{N_c \cdot 10^{\frac{\mathrm{SNR}}{10}}}{N_g + N_c \cdot 10^{\frac{\mathrm{SNR}}{10}}}. \tag{37}$$

Proof: See Appendix D. □

From Proposition 1, the ratio λ can be chosen adaptively using (37). By doing so, it not only reduces computational complexity of the prediction phase, but also mitigates noise contamination in ADCRM, resulting in robust channel prediction performance against noise.

IV. Spatio-temporal Autoregressive Channel Prediction Method

By modeling the time-varying ADCRM element as a stochastic process in Section II, we adopt AR model to capture the channel fading dynamics [10], [37]. Theorem 1 and Corollary 1 reveal that the temporal correlations between different ADCRM elements are decoupled significantly, and only exist between the neighboring ADCRM elements when the number of antennas and bandwidth are limited. Using these temporal correlations between neighboring ADCRM elements can help improve the accuracy of channel prediction. Therefore, we propose an element-wise ST-AR channel prediction method in the angle-delay domain which predicts one ADCRM element by comprehensively using its neighboring ones.

In practice, the channel samples are usually obtained using the sounding reference signal (SRS). Given the period of SRS ΔT, the ADCRM of the kth UT at time $t = n\Delta T$ is

$$\mathbf{H}_k\left[n\right] \triangleq \mathbf{H}_k\left(n\Delta T\right). \tag{38}$$

When the BS is equipped with a UPA, the ADCRM $\mathbf{H}_k$ actually contains three physical dimensions: delay, elevation angle, and azimuth angle. To determine the closest neighboring elements of the ADCRM in these three dimensions, we reshape the two-dimensional (2D) ADCRM into a 3D angle-delay domain channel response tensor (ADCRT) $\mathbf{T}_k \in \mathbb{C}^{M\times N\times N_g}$, which is expressed by the following formula

$$\left[\mathbf{T}_k\left[n\right]\right]_{q,i,j} \triangleq \left[\mathbf{H}_k\left[n\right]\right]_{qN+i,j}, \tag{39}$$

where the index (q,i,j) represents the qth elevation angle, ith azimuth angle, and jth tap of delay, respectively. Correspondingly, the set of the indices of significant elements of ADCRT is denoted as $\bar{\Omega}_s = \{(\bar{q}_s,\bar{i}_s,\bar{j}_s)\}_{s=0}^{N_s-1}$ with $\bar{q}_s = \lfloor i_s/N \rfloor$, $\bar{i}_s = \langle i_s\rangle_N$, and $\bar{j}_s = j_s$ for $(i_s,j_s)\in\Omega_s$.

To describe the proposed ST-AR channel predictor, we take the prediction of the (q,i,j)th element of the ADCRT for the kth UT at the nth SRS period $\left[\mathbf{T}_k\left[n\right]\right]_{q,i,j}$ as an example. We use V_1, V_2, H_1, H_2, D_1, and D_2 to define the ranges of the indices of the neighboring elements. According to Corollary 1 and verified by simulation, it is sufficient for the ST-AR to select V_1, H_1, and D_1 from $\{-1,0\}$ and V_2, H_2, and D_2 from $\{0,1\}$. Combining the P-order AR model [37], the proposed ST-AR channel predictor is represented as

$$\begin{aligned}\left[\hat{\mathbf{T}}_k[n]\right]_{q,i,j} = & \sum_{v=V_1}^{V_2}\sum_{h=H_1}^{H_2}\sum_{d=D_1}^{D_2}\sum_{p=1}^{P} a_{q,i,j,v,h,d,p} \\ & \cdot \left[\mathbf{T}_k[n-p]\right]_{\langle q+v\rangle_M,\langle i+h\rangle_N,\langle j+d\rangle_{N_c}},\end{aligned} \tag{40}$$

where $\{a_{q,i,j,v,h,d,p}\}$ is the parameters of the ST-AR channel predictor, and the number of the parameters is denoted by $P' = (V_2 - V_1 + 1)(H_2 - H_1 + 1)(D_2 - D_1 + 1)P$.

For simplicity, we rewrite (40) into a matrix form, given as

$$\left[\mathbf{T}_k[n]\right]_{q,i,j} = \mathbf{a}_{q,i,j}^T \mathbf{h}_{q,i,j}\left[n\right], \tag{41}$$

where $\mathbf{a}_{q,i,j} \in \mathbb{C}^{P'\times 1}$ and $\mathbf{h}_{q,i,j}\left[n\right] \in \mathbb{C}^{P'\times 1}$ are defined as the ST-AR parameters vector and the historical channel elements vector with respect to the (q,i,j)th element. For $v\in[V_1,V_2]$, $h\in[H_1,H_2]$, $d\in[D_1,D_2]$, and $p = 1,2,\cdots,P$, the elements of $\mathbf{a}_{q,i,j}$ and $\mathbf{h}_{q,i,j}\left[n\right]$ are given by

$$\left[\mathbf{a}_{q,i,j}\right]_{p'} \triangleq a_{q,i,j,v,h,d,p}, \tag{42}$$

$$\left[\mathbf{h}_{q,i,j}\left[n\right]\right]_{p'} \triangleq \left[\mathbf{T}_k[n-p]\right]_{\langle q+v\rangle_M,\langle i+h\rangle_N,\langle j+d\rangle_{N_c}}, \tag{43}$$

where p' is the auxiliary index, given by

$$\begin{aligned}p' = & (v-V_1)(H_2-H_1+1)(D_2-D_1+1)P + (h \\ & - H_1)(D_2-D_1+1)P + (d-D_1)P + p - 1.\end{aligned} \tag{44}$$

In what follows, we introduce the parameters learning process for the ST-AR model using the latest S channel samples. We use the minimum mean square error (MMSE) criterion to calculate the ST-AR parameters vector $\mathbf{a}_{q,i,j}$, given

by

$$\hat{\mathbf{a}}_{q,i,j}^{T}=\underset{\mathbf{a}_{q,i,j}^{T}}{\arg\min}\, E\left\{\left\|\left[\mathbf{T}_k[n]\right]_{q,i,j}-\mathbf{a}_{q,i,j}^{T}\mathbf{h}_{q,i,j}\left[n\right]\right\|^2\right\}. \quad (45)$$

According to the orthogonal principle [38], the optimal ST-AR parameters vector $\hat{\mathbf{a}}_{q,i,j}$ satisfies

$$E\left\{\left(\left[\mathbf{T}_k[n]\right]_{q,i,j}-\hat{\mathbf{a}}_{q,i,j}^{T}\mathbf{h}_{q,i,j}\left[n\right]\right)\mathbf{h}_{q,i,j}^{H}\left[n\right]\right\}=\mathbf{0}. \quad (46)$$

After some straightforward operations, we solve that

$$\hat{\mathbf{a}}_{q,i,j}^{T}=\mathbf{v}_{q,i,j}^{T}\mathbf{R}_{q,i,j}^{-1}, \quad (47)$$

where

$$\mathbf{R}_{q,i,j}\triangleq E\left\{\mathbf{h}_{q,i,j}\left[n\right]\mathbf{h}_{q,i,j}^{H}\left[n\right]\right\}\in\mathbb{C}^{P'\times P'}, \quad (48)$$

and

$$\mathbf{v}_{q,i,j}\triangleq E\left\{\left[\mathbf{T}_k[n]\right]_{q,i,j}\mathbf{h}_{q,i,j}^{*}\left[n\right]\right\}\in\mathbb{C}^{P'\times 1}. \quad (49)$$

As shown in (48) and (49), $\mathbf{R}_{q,i,j}$ and $\mathbf{v}_{q,i,j}$ are composed of the ACFs of the element $\left[\mathbf{T}_k[n]\right]_{q,i,j}$ and the cross-correlation functions (CCFs) between $\left[\mathbf{T}_k[n]\right]_{q,i,j}$ and its neighboring elements. To approximate them, we use the historical time series samples $\{[\mathbf{T}_k[m]]\}_{m=n-S}^{n-1}$ of length S, where $S>P$. Then by averaging the $(S-P)$ time samples, the estimation of ACF/CCF is given in (50), shown at the bottom of the page, where $(q,i,j)=(q',i',j')$ is for the ACFs and $(q,i,j)\neq(q',i',j')$ is for the CCFs. As such, each element of $\mathbf{R}_{q,i,j}$ and $\mathbf{v}_{q,i,j}$ can be obtained by estimating the corresponding ACF/CCF.

After calculating the parameters of the ST-AR model, we use (40) to predict the channel at the next moment. To reduce the computational complexity, the predictions are only implemented on the significant elements while the other elements are directly set to zero. Further, through iterative prediction, the ST-AR model can also realize the consecutive channel prediction. Suppose the number of consecutive predicted channels is L. Firstly, the channels $\{\mathbf{T}_k[n-P],\mathbf{T}_k[n-P+1],\cdots,\mathbf{T}_k[n-1]\}$ at the first P SRS periods are inputted to predict the channel $\hat{\mathbf{T}}_k[n]$ at the nth SRS period. Next, the channels at the previous $(P-1)$ SRS periods and the predicted channel at the nth SRS period $\left\{\mathbf{T}_k[n-P+1],\cdots,\mathbf{T}_k[n-1],\hat{\mathbf{T}}_k[n]\right\}$ are inputted to predict the channel $\hat{\mathbf{T}}_k[n+1]$ at the $(n+1)$th SRS period. Each iteration uses the predicted channels of previous iterations until $\hat{\mathbf{T}}_k[n+L-1]$ is predicted. To summarize, we present the ST-AR channel prediction method in Algorithm 1.

Compared with the linear model PAD, the ST-AR comprehensively considers the temporal correlation between adjacent elements, at the cost of more channel samples to learn model parameters. It is worth noting that although PAD only uses twice the order of channel samples to learn the parameters [1], the model order is dependent of the number of antennas and bandwidth. This directly affects the performance of prediction which will be shown in Section VI.

Algorithm 1 The ST-AR channel prediction method

1: Obtain the time series of ADCRM $\{\mathbf{H}_k[n-P],\mathbf{H}_k[n-P+1],\cdots,\mathbf{H}_k[n-1]\}$ at the first P SRS periods according to (16);
2: Transform the ADCRM $\mathbf{H}_k$ into its 3D counterpart $\mathbf{T}_k$ using (39);
3: Determine the set $\bar{\Omega}_s=\{(\bar{q}_s,\bar{i}_s,\bar{j}_s)\}_{s=0}^{N_s-1}$ of the indices of the significant elements of the ARCRT using (37) and (34);
4: **for all** $(q,i,j)\in\bar{\Omega}_s$ **do**
5: Determine the ACFs and CCFs contained in $\mathbf{R}_{q,i,j}$ and $\mathbf{v}_{q,i,j}$ according to (48) and (49);
6: Estimate the ACFs and CCFs using (50), then obtain $\mathbf{R}_{q,i,j}$ and $\mathbf{v}_{q,i,j}$;
7: Calculate the MMSE estimate of the ST-AR parameters vector $\hat{\mathbf{a}}_{q,i,j}$ using (47);
8: Predict the channel $\left[\hat{\mathbf{T}}_k[n]\right]_{q,i,j}$ at the nth SRS period using (40);
9: **end for**
10: Set the non-significant elements to zero and obtain the prediction $\hat{\mathbf{T}}_k[n]$;
11: **if** The L-length consecutive channel prediction is needed **then**
12: **for** $l=1,\cdots,L-1$ **do**
13: Update time series of ADCRT by removing the oldest one and appending the newest predicted one;
14: **for all** $(q,i,j)\in\bar{\Omega}_s$ **do**
15: Predict the ADCRT element $\left[\hat{\mathbf{T}}_k[n+l]\right]_{q,i,j}$ at the $(n+l)$th SRS period using (40);
16: **end for**
17: Set the non-significant elements to zero and obtain the prediction $\hat{\mathbf{T}}_k[n+l]$;
18: **end for**
19: **end if**

V. Complex-valued Neural Network Channel Prediction Method

In this section, we propose a DL-based angle-delay channel predictor for massive MIMO-OFDM systems. Compared to

$$\hat{R}_{q,i,j,q',i',j'}^{\mathrm{AD}}\left[\Delta n\right]=\begin{cases}\dfrac{1}{S-P}\displaystyle\sum_{m=n-S}^{n-P-1}\left[\mathbf{T}_k[m+\Delta n]\right]_{q,i,j}\left[\mathbf{T}_k[m]\right]_{q',i',j'}^{*}, & \Delta n=1,2,\cdots,P,\\ \dfrac{1}{S-P}\displaystyle\sum_{m=n-S}^{n-P-1}\left[\mathbf{T}_k[m]\right]_{q,i,j}\left[\mathbf{T}_k[m-\Delta n]\right]_{q',i',j'}^{*}, & \Delta n=0,-1,\cdots,-P+1,\end{cases} \quad (50)$$

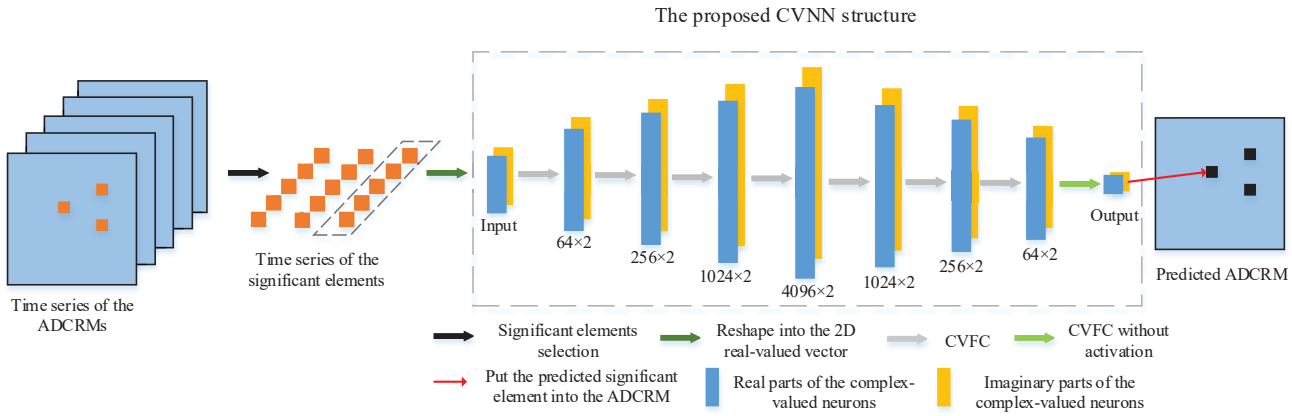


Fig. 4: The overall structure of the CVNN predictor.

the conventional model-driven channel predictors, the data-driven channel predictors have two benefits: 1) By leveraging the advantageous generalization performance of DL, the well-trained network can realize prediction in different scenarios without retraining; 2) The length of the network output can be tuned to realize more flexible prediction, e.g., consecutive channel prediction, as such the error brought by iterative prediction is avoided.

According to Theorem 1, the temporal correlations between different ADCRM elements are decoupled. The inherent reason is that the superimposed multipath is decoupled into different ADCRM elements. As such, we propose a unified neural network to fit the pattern of all ADCRM significant elements and implement the element-wise channel prediction. Sincethe CVNN has advantages when dealing with the complex-valued input and output (e.g., the complex-valued channel data), as suggested and verified in [39], we adopt the CVNN to realize the data-driven channel prediction. Specifically, given the (i,j)th element of the ADCRM for the kth UT at the nth SRS period $[\mathbf{H}_k[n]]_{i,j}$, we first disassemble and cascade the imaginary and the real part of $[\mathbf{H}_k[n]]_{i,j}$ into a 2D real-valued vector $\mathbf{b}_{k,i,j}[n]$, given as

$$\mathbf{b}_{k,i,j}[n] = \left[\Re\left([\mathbf{H}_k[n]]_{i,j}\right), \Im\left([\mathbf{H}_k[n]]_{i,j}\right)\right]^T. \tag{51}$$

Then we take time series $\{\mathbf{b}_{k,i,j}[n-P], \cdots, \mathbf{b}_{k,i,j}[n-1]\}$ at the first P SRS periods as the input of the CVNN, and the output is the predicted value of the time series at the nth SRS period $\hat{\mathbf{b}}_{k,i,j}[n]$. The CVNN itself can be regarded as a nonlinear function represented by $f_{\text{CVNN}}(\cdot)$, such that the process of the channel prediction can be written as

$$\hat{\mathbf{b}}_{k,i,j}[n] = f_{\text{CVNN}}(\mathbf{b}_{k,i,j}[n-P], \cdots, \mathbf{b}_{k,i,j}[n-1]). \tag{52}$$

The overall structure of the proposed CVNN $f_{\text{CVNN}}(\cdot)$ is illustrated in Fig. 4. The network consists of two parts: a generator, to gradually represent the input data in a higher dimensional space so that the spatial correlation can be gradually decoupled, and a compressor, to discard irrelevant information and crystallize the most representative features. The generator is made up of four complex-valued fully connected layers (CVFLs) with increasing number of neurons from 64, to 256, 1024, and finally 4096. The complex-valued fully connection (CVFC) contains a complex-valued multiplication and complex-valued activation function, which will be detailed later. The compressor is comprised of three CVFLs with decreasing number of neurons from 4096, to 1024, 256, and finally 64. The gradual representation in the generator/compressor is to capture/retain as much information of the input as possible, where the nonlinear activation functions at each layer help increase the capacity of the generator/comperssor.

To describe the realization of complex-valued multiplication in CVNN, we first give a complex-valued weight matrix of $\mathbf{W} = \mathbf{A} + \bar{\jmath}\mathbf{B}$ and a complex-valued input feature vector of $\mathbf{h} = \mathbf{x} + \bar{\jmath}\mathbf{y}$. Then the complex-valued multiplication can be given as [39]

$$\mathbf{W}\cdot\mathbf{h} = \mathbf{A}\cdot\mathbf{x} - \mathbf{B}\cdot\mathbf{y} + \bar{\jmath}(\mathbf{A}\cdot\mathbf{y} + \mathbf{B}\cdot\mathbf{x}). \tag{53}$$

Inspired by (53), we constitute the real-valued weight matrix $\mathbf{W}'$ and the real-valued input vector $\mathbf{h}'$ to replace their complex-valued counterparts $\mathbf{W}$ and $\mathbf{h}$, and use the real-valued matrix multiplication to replace the complex-valued multiplication, given as [39]

$$\begin{bmatrix} \Re(\mathbf{W}\cdot\mathbf{h}) \\ \Im(\mathbf{W}\cdot\mathbf{h}) \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{A} & -\mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{bmatrix}}_{\text{weight matrix } \mathbf{W}'} \cdot \underbrace{\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}}_{\text{feature vector } \mathbf{h}'}. \tag{54}$$

After the complex-valued multiplication, a complex-valued activation is followed to introduce nonlinearity to the CVNN. We apply two rectified linear units (ReLU) on the imaginary and real part of the complex value respectively to implement the complex-valued activation, i.e., complex-valued ReLU. Specifically, given a complex value of z, the complex-valued ReLU can be written as [39]

$$\text{CReLU}(z) = \text{ReLU}(\Re(z)) + \bar{\jmath}\,\text{ReLU}(\Im(z)). \tag{55}$$

Since the non-significant elements of the ADCRM reveal little information of temporal correlation, we construct the training dataset using the significant channel elements to make the training of the CVNN efficient. The number of ADCRMs used in the training phase is K_{train} and the number of the significant elements is N_s. By letting $\boldsymbol{\theta}$ be the set of trainable

TABLE I: Complexity of the proposed channel predictors

Method	Parameter learning	On-line predicting
	On-line learning	
AR [11]	$\mathcal{O}(N_s(P+1)(S-P))+\mathcal{O}\left(N_sP^3\right)$	$\mathcal{O}\left(N_sPL\right)$
PAD [1]	$\mathcal{O}(N_sP^{2.37})$	$\mathcal{O}(N_sPL)$
ST-AR	$\mathcal{O}\left(N_s\left(J^2P+\frac{J^2+J}{2}\right)(S-P)\right)+\mathcal{O}\left(N_sP'^3\right)$	$\mathcal{O}(N_sP'L)$
	Off-line learning	
CVNN	$\mathcal{O}\left(N_{\text{epoch}}K_{\text{train}}N_s\left(Ph_1+\sum_{i=1}^{I-1}h_ih_{i+1}+h_IL\right)\right)$	$\mathcal{O}\left(N_s\left(Ph_1+\sum_{i=1}^{I-1}h_ih_{i+1}+h_IL\right)\right)$

parameters of the CVNN, we employ the mean absolute error (MAE) of the training dataset as the cost function, given as

$$J\left(\boldsymbol{\theta}\right)=\frac{1}{K_{\text{train}}N_s}\sum_{k=0}^{K_{\text{train}}-1}\sum_{s=0}^{N_s-1}\left\|\hat{\mathbf{b}}_{k,i_s,j_s}\left[n\right]-\mathbf{b}_{k,i_s,j_s}\left[n\right]\right\|_1, \tag{56}$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm, and $\{\mathbf{b}_{k,i_s,j_s}[n]\}_{s=0}^{N_s-1}$ is the set of the significant elements of kth ADCRM selected according to (34).

Since the length of the CVNN's output is flexible, we can exploit the CVNN-based predictor to realize the prediction of consecutive L channels. To be specific, by extending (52) we modify the length of the last layer as L, then the output of CVNN predictor is given as

$$\begin{aligned}&\left[\hat{\mathbf{b}}_{k,i,j}\left[n\right],\cdots,\hat{\mathbf{b}}_{k,i,j}\left[n+L-1\right]\right]\\&\quad=f_{\text{CVNN}-\text{L}}(\mathbf{b}_{k,i,j}\left[n-p\right],\cdots,\mathbf{b}_{k,i,j}\left[n-1\right]),\end{aligned} \tag{57}$$

where $f_{\text{CVNN}-\text{L}}(\cdot)$ is the function of CVNN for consecutive L channels prediction. The structure of $f_{\text{CVNN}-\text{L}}(\cdot)$ is the same with that of $f_{\text{CVNN}}(\cdot)$ except different output lengths when $L>1$.

After training, the CVNN-based channel predictions are implemented on the significant elements of the ADCRM while the non-significant elements are set to zero directly. Different from the SR-AR channel predictor, the CVNN channel predictor can realize the consecutive channel prediction directly without any iterative computations. To summarize, the CVNN-based channel prediction method is presented in Algorithm 2.

Algorithm 2 The CVNN-based channel prediction method

1: Obtain the time series of ADCRM $\{\mathbf{H}_k[n-P],\mathbf{H}_k[n-P+1],\cdots,\mathbf{H}_k[n-1]\}$ at the first P SRS periods according to (16);
2: Determine the set $\Omega_s=\{(\bar{i}_s,\bar{j}_s)\}_{s=0}^{N_s-1}$ of the indices of the significant elements of the ADCRM using (37) and (34);
3: **for all** $(q,i,j)\in\bar{\Omega}_s$ **do**
4: Obtain the time series $\{\mathbf{b}_{k,i,j}[n-P],\mathbf{b}_{k,i,j}[n-P+1],\cdots,\mathbf{b}_{k,i,j}[n-1]\}$ at the first P SRS periods according to (51);
5: Compute the predictions of current channel or consecutive L channels using the trained CVNN of (52) or (57);
6: **end for**
7: Set the non-significant elements to zero and obtain the prediction $\hat{\mathbf{H}}_k[n+l]$, $l=0,\cdots,L-1$.

VI. Complexity Analysis and Simulation Results

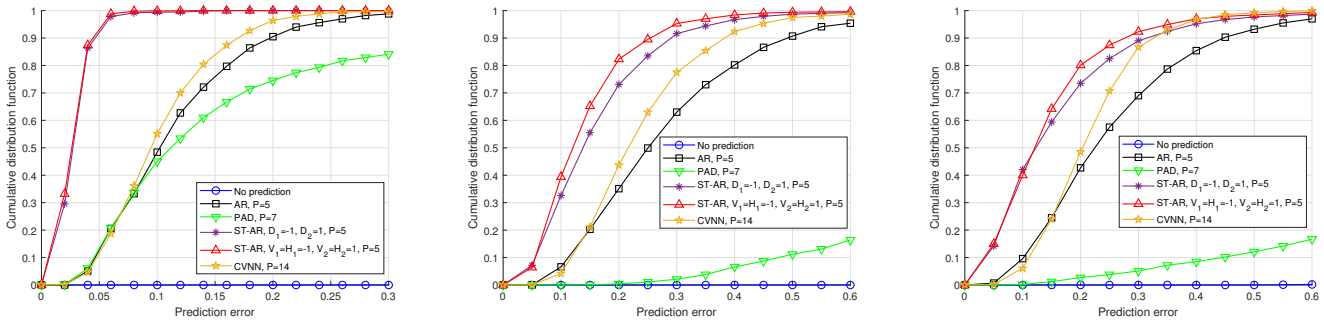
In this section, we first analyze the computational complexity of the proposed channel prediction methods. Then we give the simulation setup and evaluate the performance of the proposed ST-AR and CVNN-based channel prediction methods through intensive simulations.

A. Complexity Analysis

We first analyze the computational complexity of the ST-AR channel predictor. According to (48) and (49), it can be verified that the least number of CCFs/ACFs needed to be estimated in $\mathbf{R}_{q,i,j}$ and $\mathbf{v}_{q,i,j}$ is $\left(J^2P+\frac{J^2+J}{2}\right)$ where $J\triangleq(V_2-V_1+1)(H_2-H_1+1)(D_2-D_1+1)$ is the number of the neighboring ADCRM elements used in ST-AR channel predictor. Then the estimation of $\mathbf{R}_{q,i,j}$ and $\mathbf{v}_{q,i,j}$ using (50) has complexity of $\mathcal{O}\left(N_s\left(J^2P+\frac{J^2+J}{2}\right)(S-P)\right)$. The calculation of the ST-AR model parameters has complexity of $\mathcal{O}\left(N_sP'^3\right)$ due to the matrix inversion. Hence, the complexity of the ST-AR in the parameter learning phase is $\mathcal{O}\left(N_s\left(J^2P+\frac{J^2+J}{2}\right)(S-P)\right)+\mathcal{O}\left(N_sP'^3\right)$ where $P'=JP$. The prediction of L-length channels using the ST-AR has complexity of $\mathcal{O}(N_sP'L)$.

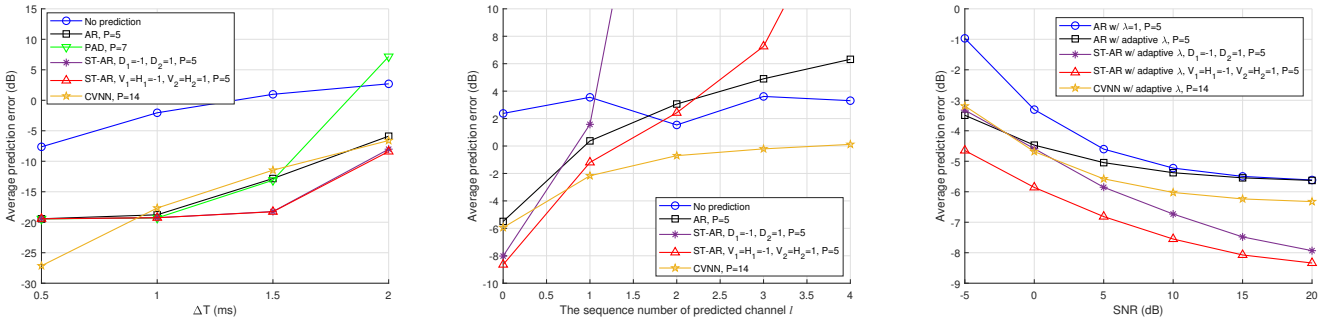
When it comes to the CVNN channel predictor, the complexity of CVNN can refer to that of the MLP in [22], because the CVFL has the same complexity order as the fully connected layer. With the number of training epochs N_{epoch}, the complexity of the CVNN in the parameters learning phase is given by $\mathcal{O}\left(N_{\text{epoch}}K_{\text{train}}N_s\left(Ph_1+\sum_{i=1}^{I-1}h_ih_{i+1}+h_IL\right)\right)$, where I is the number of hidden layers of CVNN and h_i is the number of complex-valued neurons in the ith hidden layer. The complexity of the prediction of L-length channels using CVNN is $\mathcal{O}\left(N_s\left(Ph_1+\sum_{i=1}^{I-1}h_ih_{i+1}+h_IL\right)\right)$.

We compare the complexity of the proposed channel predictors with some benchmark methods in Table I. It is worth noting that when applied to channel samples in new scenarios, the well-trained CVNN does not need re-training while the other methods need to frequently update the model parameters to adapt to the change of environment through on-line parameter learning. For simplification, we approximate the



(a) (b) (c)

Fig. 5: The CDF of the channel prediction error with different methods. (a) $v = 5$ km/h and $\Delta T = 20$ ms; (b) $v = 30$ km/h and $\Delta T = 4$ ms; (c) $v = 60$ km/h and $\Delta T = 2$ ms.



(a) (b) (c)

Fig. 6: The average channel prediction error versus (a) the SRS period ΔT, (b) the sequence number of predicted channel of consecutive $L = 5$ channel prediction, and (c) the SNR.

on-line computational complexity of AR, PAD, and ST-AR as $\mathcal{O}\left(N_s P^3\right)$, $\mathcal{O}\left(N_s P^{2.37}\right)$, and $\mathcal{O}\left(N_s P'^3\right)$, respectively. Through further network optimization, the number of neurons can be assigned by $h_i = \alpha_i P$ for $i = 1, 2, \cdots, I$, where α_i is a scaling constant. In this way, the on-line complexity of CVNN is approximated as $\mathcal{O}\left(N_s P^2\right)$. Therefore, by relegating computation to the off-line training, the CVNN has the lowest on-line complexity compared to other channel predictors.

B. Simulation Results

QuaDRiGa is a general channel simulator for simulating the time-varying channel and has been verified by field tests [40]. Besides, it can generate channels that meet 3GPP standards, e.g., 3GPP new radio (NR) [41]. Thus we employ QuaDRiGa for simulation. We consider the 3GPP urban macro (UMa) NLOS scenario. The carrier frequency f_c is set to 3.5 GHz and the signal bandwidth B is 30.72 MHz. The number of sub-carrier N_c is set to 1024, and the sub-carrier frequency spacing Δf is 30 kHz. The BS is equipped with a UPA with column antennas number $M = 4$ and row antennas number $N = 8$. The UT is equipped with an omni-directional antenna. Unless otherwise specified, the UT's velocity v and the SRS period ΔT are set to 60 km/h and 2 ms, respectively. To build the training dataset of CVNN, we regenerate the scattering environment 10 times, each scattering environment includes 30 UTs with each UT containing one training samples, so the size of training dataset is $K_{\text{train}} = 300$. Note that since one ADCRM training sample contains hundreds or thousands of significant elements, $K_{\text{train}} = 300$ is sufficient for the training of CVNN. The test dataset is generated in a new scattering environment with the same velocity and SRS period. It includes 10 UTs with each UT containing 100 test samples, so the size of test dataset is $K_{\text{test}} = 1000$.

To evaluate the performance of the proposed channel prediction methods, we first define the channel prediction error as

$$e = \frac{\left\| \mathbf{G} - \hat{\mathbf{G}} \right\|_F^2}{\|\mathbf{G}\|_F^2}, \tag{58}$$

where $\|\cdot\|_F$ represents the Frobenius norm, $\mathbf{G} \in \mathbb{C}^{N_a \times N_c}$ and $\hat{\mathbf{G}} \in \mathbb{C}^{N_a \times N_c}$ are the SFCRM and its prediction respectively. And the average channel prediction error in dB is calculated as $\bar{e} = 10 \log\{E\{e\}\}$. To demonstrate the superiority of the proposed channel prediction methods, we choose the PAD channel prediction method [1] and the AR channel prediction method [11] as benchmarks. For the sake of fairness, we implement all these methods in the angle-delay domain. As suggested in Section III-B, the prediction of the entire channel

Test dataset (v, ΔT)	**Error**
(60 km/h, 2 ms)	-6.60 dB
(45 km/h, 2 ms)	-8.15 dB
(30 km/h, 2 ms)	-8.93 dB
(15 km/h, 2 ms)	-11.00 dB
(60 km/h, 1.5 ms)	-8.67 dB
(60 km/h, 1 ms)	-9.21 dB
(60 km/h, 0.5 ms)	-10.86 dB

TABLE II: The average channel prediction error of the trained CVNN predictor on different test datasets.

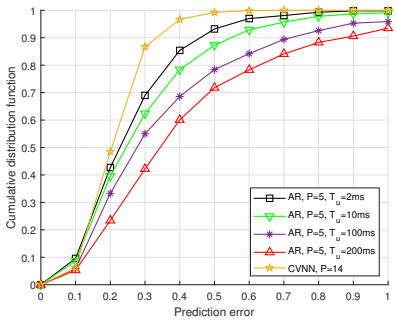


Fig. 7: The prediction error of the AR channel predictor with different T_u and CVNN channel predictor.

TABLE III: Comparisons of the channel prediction methods

Method	**Accuracy**	**Feasibility**	**On-line complexity**	**Generalization**
AR [11]	Low	Middle	High	Weak
PAD [1]	High	Low	Middle	Weak
ST-AR	High	Middle	High	Weak
CVNN	Middle	High	Low	Strong

is obtained by predicting the significant elements in the angle-delay domain. When channel prediction is not applied, the previous channel is usually treated as the current one, i.e., $\hat{\mathbf{H}}[n] = \mathbf{H}[n-1]$. So, we also simulate this treatment for the purpose of illustration. We consider the two ST-AR channel predictors with different scopes of the adjacent elements as follows:

1) The ST-AR with $D_1 = -1$ and $D_2 = 1$. The other parameters, i.e., V_1, V_2, H_1, and H_2 are all set to zero, such that only the nearest elements in the delay dimension are considered.
2) The ST-AR with $V_1 = H_1 = -1$ and $V_2 = H_2 = 1$. The other parameters, i.e., D_1 and D_2 are all set to zero, such that the nearest elements in both the vertical and horizontal angle dimensions are considered.

We consider that the length of available channel time series S for the above methods is limited, e.g., $S = 14$. For the PAD channel predictor, we set the order as $P = 7$ and the length of the time series it needed for parameter calculation is exactly $S' = 2P - 1 = 13$ according to [1]. For the AR and ST-AR channel predictors, we set a proper order $P = 5$ by considering the trade-off that at least $S - P = 9$ time samples are needed to calculate the model parameters accurately in our simulations. For the CVNN channel predictor, the channel time series are all put into the predictor, hence the order P equals to the length S, i.e., $P = S = 14$.

Fig. 5(a), Fig. 5(b), and Fig. 5(c) present the cumulative distribution function (CDF) of the prediction error with different methods. The UT's velocity and SRS period, i.e., (v, ΔT), are set to (5 km/h, 20 ms), (30 km/h, 4 ms), and (60 km/h, 2 ms) in Fig. 5(a), Fig. 5(b), and Fig. 5(c), respectively. It can be observed that the proposed two ST-AR channel predictors reach the highest prediction accuracy for all scenarios, and the proposed CVNN channel predictor also consistently obtains the higher accuracy compared to the AR and PAD channel predictors.

Fig. 6(a) shows the average prediction error versus the SRS period with different methods. The increase of SRS period leads to the increase of average prediction error of each channel predictor. Especially for the PAD channel predictor, when the SRS period reaches 2 ms, the PAD channel predictor with $P = 7$ can hardly achieve accurate prediction. Fig. 6(b) shows the performance of the consecutive $L = 5$ channel prediction with different methods. When the sequence number of predicted channel $l \geq 1$, the proposed CVNN channel predictor can achieve the best prediction accuracy compared with other predictors because it avoids the prediction error brought by the iterative calculation. The proposed ST-AR channel predictors obtain the worst prediction performance because they use the most channel elements in the prediction phase, which will aggravate the error introduced by iterative prediction. Fig. 6(c) shows the average channel prediction error versus SNR with different methods. Compared to setting fixed $\lambda = 1$, the error of AR channel predictor is significantly reduced at low SNRs by setting the ratio λ adaptively using (37) in Proposition 1. Combined with the adaptive λ, the proposed ST-AR and CVNN channel predictors can achieve higher prediction accuracy than the AR channel predictor with $\lambda = 1$ in all SNR regimes, and are more robust to noise contamination.

To demonstrate the generalization ability of the CVNN, we train the CVNN using the training dataset at $v = 60$ km/h and $\Delta T = 2$ ms, and evaluate the average channel prediction error of the trained CVNN on the test dataset at different v and ΔT, as shown in Table II. The trained CVNN can still predict the channel at different v and ΔT accurately, while the AR can only predict the channel at $v = 60$ km/h and

$\Delta T = 2$ ms with the average error of -5.90 dB as in Fig. 6(a), because its parameters are quite different when v and ΔT change. The generalization ability of CVNN is also reflected in the time stability of its parameters. In Fig. 7, we show the prediction error of the CVNN channel predictor without parameter update and AR channel predictor with different parameter update period T_u. It can be observed that when T_u increases, the accuracy of the AR predictor decreases gradually, while the CVNN does not require parameter update and always maintains higher prediction accuracy.

Finally, we summarize the performance of different channel prediction methods in Table III. In particular, "feasibility" is referred to the effectiveness of the channel predictor in different scenarios with, e.g., different SRS periods. In conclusion, the ST-AR reaches the highest accuracy at the cost of complexity, and is more feasible than the PAD method, whilst the CVNN is superior with respect to feasibility, online complexity, and generalization.

VII. Conclusion

In this paper, we studied the channel prediction problem for massive MIMO-OFDM systems with high-mobility user terminals, by leveraging unsupervised and supervised machine learning techniques. We first investigated a general angle-delay domain channel characterization and the temporal correlation of the angle-delay domain channel for massive MIMO-OFDM systems. Then we proposed two angle-delay domain channel prediction methods: a model-driven one, i.e., the ST-AR channel prediction method, and a data-driven one, i.e., the DL-based channel prediction method. Whilst the former provides a principled way for channel prediction, the latter is generalizable to a wide range of channel scenarios. Simulation results demonstrated that both the proposed ST-AR and the CVNN-based channel prediction methods can enhance the channel prediction performance compared with state-of-the-art channel prediction methods.

Appendix A
Proof of Lemma 1

P1, P2, and P3 can be derived immediately by substituting the definition in (14) into that in (9), so we omit the detailed derivations. Besides, it is worth noting that when $M \to \infty$, $\lim_{M\to\infty} f_M(x) \neq 0$ holds if and only if $\sin(\pi x) = 0$, i.e., $x = 0, \pm 1, \pm 2, \cdots$. Hence, put the constraint of $0 \leq i \leq M-1$ into consideration, $\lim_{M\to\infty} [\tilde{\mathbf{a}}(\theta)]_i \neq 0$ holds if and only if $\frac{1}{2}(\cos\theta + 1) - \frac{i}{M} = 0$. According to the definition in (14), we further have that $\lim_{M\to\infty} [\tilde{\mathbf{a}}(\theta)]_i \neq 0$ holds if and only if $\theta - \theta_i = 0$. Thus we derive

$$\begin{aligned}
&\lim_{M\to\infty} \left(\left[\tilde{\mathbf{a}}^{(v)}(\theta)\right]_i - \delta(\theta - \theta_i) \right) \\
&= \lim_{M\to\infty} f_M(0)\,\delta(\theta - \theta_i) - \lim_{M\to\infty} \delta(\theta - \theta_i) \\
&= \lim_{M\to\infty} \delta(\theta - \theta_i) - \lim_{M\to\infty} \delta(\theta - \theta_i) = 0.
\end{aligned} \tag{59}$$

Therefore (15) in P4 is obtained. The four properties of $\left[\tilde{\mathbf{a}}^{(h)}(\varphi;\theta)\right]_q$ and $\left[\tilde{\mathbf{b}}(\tau)\right]_j$ can be derived in a similar way. This completes the proof.

Appendix B
Proof of Lemma 3

We first calculate $\left|\left[\mathbf{a}^{(v)}(\theta_{i_M+\Delta_1})\right]_{i_M}\right|$ as

$$\begin{aligned}
\left|\left[\mathbf{a}^{(v)}(\theta_{i_M+\Delta_1})\right]_{i_M}\right| &\overset{(a)}{=} \frac{\sin\left(\pi M\left(\frac{1}{2}(\cos\theta_{i_M+\Delta_1}+1) - \frac{i_M}{M}\right)\right)}{M\sin\left(\pi\left(\frac{1}{2}(\cos\theta_{i_M+\Delta_1}+1) - \frac{i_M}{M}\right)\right)} \\
&\overset{(b)}{=} \frac{\sin(\pi\Delta_1)}{M\sin\left(\frac{\pi\Delta_1}{M}\right)},
\end{aligned} \tag{61}$$

where (a) follows from (9) and (11), and (b) follows from (14). When $M \to \infty$, we derive

$$\lim_{M\to\infty} M\sin\left(\frac{\pi\Delta_1}{M}\right) = \pi\Delta_1. \tag{62}$$

Substituting (62) into (61), we derive

$$\lim_{M\to\infty}\left|\left[\mathbf{a}^{(v)}(\theta_{i_M+\Delta_1})\right]_{i_M}\right| = \frac{\sin(\pi\Delta_1)}{\pi\Delta_1} = \mathrm{sinc}(\Delta_1). \tag{63}$$

Then substituting (63) into (21), when $M \to \infty$ yields

$$\Delta_1 = \mathrm{sinc}^{-1}(\zeta) \in [0,1]. \tag{64}$$

In addition, the representation in (61) does not contain i_M, which means Δ_1 is the same for all $i_M \in [M]$. The property of Δ_2 and Δ_3 can be proved similarly. This completes the proof.

Appendix C
Proof of Theorem 1

Substituting (5) and (28) into (29), since the initial phase $\phi_k(\theta,\varphi,\tau)$ is uniformly distributed in $[-\pi,\pi)$ and independent of $\phi_k(\theta',\varphi',\tau')$ for $(\theta,\varphi,\tau) \neq (\theta',\varphi',\tau')$, we calculate (60) shown at the bottom of the next page, where (a) is from (28), (b) is from (8), and (c) is from (5) and

$$\begin{aligned}
&E\left\{e^{\bar{\jmath}\phi_k(\theta,\varphi,\tau) - \bar{\jmath}\phi_k\left(\theta',\varphi',\tau'\right)}\right\} \\
&= \int_{-\pi}^{\pi} \frac{1}{2\pi} e^{\bar{\jmath}\phi_k(\theta,\varphi,\tau) - \bar{\jmath}\phi_k(\theta,\varphi,\tau)} d\phi_k\,\delta(\theta - \theta') \\
&\quad \cdot \delta(\varphi - \varphi')\,\delta(\tau - \tau') \\
&= \delta(\theta - \theta')\,\delta(\varphi - \varphi')\,\delta(\tau - \tau').
\end{aligned} \tag{65}$$

From the definition in (27), $\bar{O}_{i,j} \cap \bar{O}_{i',j'}$ can be represented as

$$\begin{aligned}
\bar{O}_{i,j} \cap \bar{O}_{i',j'} = \big\{ &(\theta,\varphi,\tau) : \theta \in O_{\theta,i_M} \cap O_{\theta,i'_M}, \\
&\varphi \in O_{\varphi,i_N} \cap O_{\varphi,i'_N}, \text{and } \tau \in O_{\tau,j} \cap O_{\tau,j'} \big\}.
\end{aligned} \tag{66}$$

We then calculate $O_{\theta,i_M} \cap O_{\theta,i'_M}$ according to (14) and (20) as

$$\begin{aligned}
&O_{\theta,i_M} \cap O_{\theta,i'_M} \\
&= \left[\max\left(\theta_{i_M+\Delta_1}, \theta_{i'_M+\Delta_1}\right), \min\left(\theta_{i_M-\Delta_1}, \theta_{i'_M-\Delta_1}\right)\right) \\
&= \begin{cases} \left[\theta_{i_M+\Delta_1}, \theta_{i'_M-\Delta_1}\right), & i_M < i'_M, \\ \left[\theta_{i'_M+\Delta_1}, \theta_{i_M-\Delta_1}\right), & i_M \geq i'_M. \end{cases}
\end{aligned} \tag{67}$$

From (67), $O_{\varphi,i_N} \cap O_{\varphi,i'_N} = \emptyset$ if and only if $\theta_{i_M+\Delta_1} \geq \theta_{i'_M-\Delta_1}$ or $\theta_{i'_M+\Delta_1} \geq \theta_{i_M-\Delta_1}$. According to the definition in

(14), the above two inequalities are equivalent to

$$\arccos\left(\frac{2\left(i_M+\Delta_1\right)}{M}-1\right)\geq\arccos\left(\frac{2\left(i'_M-\Delta_1\right)}{M}-1\right), \tag{68}$$

and

$$\arccos\left(\frac{2\left(i'_M+\Delta_1\right)}{M}-1\right)\geq\arccos\left(\frac{2\left(i_M-\Delta_1\right)}{M}-1\right), \tag{69}$$

respectively. By solving (68) and (69), we obtain that $O_{\varphi,i_N}\cap O_{\varphi,i'_N}=\emptyset$ if and only if $|i_M-i'_M|\geq 2\Delta_1$. Similarly, $O_{\varphi,i_N}\cap O_{\varphi,i'_N}=\emptyset$ and $O_{\tau,j}\cap O_{\tau,j'}=\emptyset$ if and only if $|i_N-i'_N|\geq 2\Delta_2$ and $|j-j'|\geq 2\Delta_3$, respectively. Therefore $\bar{O}_{i,j}\cap\bar{O}_{i',j'}=\emptyset$ if $|i_M-i'_M|\geq 2\Delta_1$ or $|i_N-i'_N|\geq 2\Delta_2$ or $|j-j'|\geq 2\Delta_3$, and (30) is obtained from (60).

When M, N, and N_c tend to infinity, let $\left[\mathbf{H}_k\left(t\right)\right]_{i,j}$ be replaced by $\left[\mathbf{H}_k^{\inf}\left(t\right)\right]_{i,j}$ according to Lemma 2. Since we have illustrated that (19) in Lemma 2 is a special case of (28) when M, N, and N_c tend to infinity, then similar to the derivation process of (60), (31) can be obtained immediately by substituting (5) and (19) into (29). This completes the proof.

Appendix D
Proof of Proposition 1

From (16) and (35), the estimated ADCRM is calculated as

$$\begin{aligned}\tilde{\mathbf{H}}_k\left(t\right)&=\frac{1}{\sqrt{MNN_c}}\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)^H\tilde{\mathbf{G}}_k\left(t\right)\mathbf{F}^*_{N_c\times N_g}\\&=\mathbf{H}_k\left(t\right)+\underbrace{\frac{1}{\sqrt{MNN_c}}\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)^H\mathbf{Z}_k\left(t\right)\mathbf{F}^*_{N_c\times N_g}}_{\triangleq\mathbf{N}_k(t)}.\end{aligned} \tag{70}$$

Based on channel sparsity, the power of ADCRM $P_k^{(\mathrm{AD})}$ can be approximated by the total power of the significant elements while the power of noise P_n can be approximated by the total power of the noise on the non-significant elements. Thus according to the definition of λ, it can be approximated by

$$\lambda\approx\frac{P_k^{(\mathrm{AD})}}{P_n+P_k^{(\mathrm{AD})}}, \tag{71}$$

where P_n is calculated as

$$\begin{aligned}P_n&=\sum_{i=0}^{N_a-1}\sum_{j=0}^{N_g-1}E\left\{\left|\left[\mathbf{N}_k\left(t\right)\right]_{i,j}\right|^2\right\}\\&=\mathrm{Tr}\left\{E\left\{\mathrm{vec}\left\{\mathbf{N}_k\left(t\right)\right\}\mathrm{vec}^H\left\{\mathbf{N}_k\left(t\right)\right\}\right\}\right\}\\&=\frac{1}{N_aN_c}\mathrm{Tr}\Big\{\left(\mathbf{F}_{N_c\times N_g}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)^HE\big\{\mathrm{vec}\left\{\mathbf{Z}_k\left(t\right)\right\}\\&\quad\cdot\mathrm{vec}^H\left\{\mathbf{Z}_k\left(t\right)\right\}\big\}\left(\mathbf{F}_{N_c\times N_g}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)\Big\}\\&=\frac{1}{N_aN_c}\mathrm{Tr}\Big\{\left(\mathbf{F}_{N_c\times N_g}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)^H\sigma_z^2\mathbf{I}_{N_aN_c}\\&\quad\cdot\left(\mathbf{F}_{N_c\times N_g}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)\Big\}=\frac{N_g\sigma_z^2}{N_c}.\end{aligned} \tag{72}$$

Then to calculate $P_k^{(\mathrm{AD})}$, we define the extended ADCRM as

$$\bar{\mathbf{H}}_k\left(t\right)\triangleq\frac{1}{\sqrt{MNN_c}}\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)^H\mathbf{G}_k\left(t\right)\mathbf{F}^*_{N_c}\in\mathbb{C}^{N_a\times N_g}, \tag{73}$$

with its (i,j)th element sharing the same representation of that in (17). Since the maximum delay $\tau_{\max}>N_gT_s$, then according to P4 in Lemma 1, $\left[\tilde{\mathbf{b}}(\tau)\right]_j=0$ for $j>N_g$ when N_c is sufficiently large, in that case we can approximate $\bar{\mathbf{H}}_k\left(t\right)$ as

$$\bar{\mathbf{H}}_k\left(t\right)=\left[\begin{array}{cc}\mathbf{H}_k\left(t\right)&\mathbf{0}_{N_a\times(N_c-N_g)}\end{array}\right]. \tag{74}$$

$$\begin{aligned}R^{\mathrm{AD}}_{i,j,i',j'}\left(\Delta t\right)&\overset{(a)}{=}E\Bigg\{\iiint\limits_{\bar{O}_{i,j}}c_k\left(\theta,\varphi,\tau\right)\left[\tilde{\mathbf{a}}^{(v)}\left(\theta\right)\right]_{i_M}\left[\tilde{\mathbf{a}}^{(h)}\left(\varphi;\theta\right)\right]_{i_N}\left[\tilde{\mathbf{b}}(\tau)\right]_j e^{\bar{\jmath}2\pi\nu_k(\theta,\varphi,\tau)(t+\Delta t)+\bar{\jmath}\phi_k(\theta,\varphi,\tau)}d\tau d\varphi d\theta\\&\quad\cdot\iiint\limits_{\bar{O}_{i',j'}}c_k^*\left(\theta',\varphi',\tau'\right)\left[\tilde{\mathbf{a}}^{(v)}\left(\theta'\right)\right]^*_{i'_M}\left[\tilde{\mathbf{a}}^{(h)}\left(\varphi';\theta'\right)\right]^*_{i'_N}\left[\tilde{\mathbf{b}}(\tau')\right]^*_j e^{-\bar{\jmath}2\pi\nu_k\left(\theta',\varphi',\tau'\right)t-\bar{\jmath}\phi_k\left(\theta',\varphi',\tau'\right)}d\tau' d\varphi' d\theta'\Bigg\}\\&\overset{(b)}{=}\iiint\limits_{\bar{O}_{i,j}}\iiint\limits_{\bar{O}_{i',j'}}E\left\{c_k\left(\theta,\varphi,\tau\right)c_k^*\left(\theta',\varphi',\tau'\right)\right\}E\left\{e^{\bar{\jmath}\phi_k(\theta,\varphi,\tau)-\bar{\jmath}\phi_k\left(\theta',\varphi',\tau'\right)}\right\}\\&\quad\cdot\left[\tilde{\mathbf{a}}\left(\theta,\varphi\right)\right]_i\left[\tilde{\mathbf{a}}\left(\theta',\varphi'\right)\right]^*_{i'}\left[\tilde{\mathbf{b}}(\tau)\right]_j\left[\tilde{\mathbf{b}}(\tau')\right]^*_{j'}e^{\bar{\jmath}2\pi\nu_k(\theta,\varphi,\tau)(t+\Delta t)-\bar{\jmath}2\pi\nu_k\left(\theta',\varphi',\tau'\right)t}d\tau' d\varphi' d\theta' d\tau d\varphi d\theta\\&\overset{(c)}{=}\iiint\limits_{\bar{O}_{i,j}}\iiint\limits_{\bar{O}_{i',j'}}S_k\left(\theta,\varphi,\tau\right)\delta\left(\theta-\theta'\right)\delta\left(\varphi-\varphi'\right)\delta\left(\tau-\tau'\right)\\&\quad\cdot\left[\tilde{\mathbf{a}}\left(\theta,\varphi\right)\right]_i\left[\tilde{\mathbf{a}}\left(\theta',\varphi'\right)\right]^*_{i'}\left[\tilde{\mathbf{b}}(\tau)\right]_j\left[\tilde{\mathbf{b}}(\tau')\right]^*_{j'}e^{\bar{\jmath}2\pi\nu_k(\theta,\varphi,\tau)(t+\Delta t)-\bar{\jmath}2\pi\nu_k\left(\theta',\varphi',\tau'\right)t}d\tau' d\varphi' d\theta' d\tau d\varphi d\theta\\&=\iiint\limits_{\bar{O}_{i,j}\cap\bar{O}_{i',j'}}S_k\left(\theta,\varphi,\tau\right)\left[\tilde{\mathbf{a}}\left(\theta,\varphi\right)\right]_i\left[\tilde{\mathbf{a}}\left(\theta,\varphi\right)\right]^*_{i'}\left[\tilde{\mathbf{b}}(\tau)\right]_j\left[\tilde{\mathbf{b}}(\tau)\right]^*_{j'}e^{\bar{\jmath}2\pi\nu_k(\theta,\varphi,\tau)\Delta t}d\tau d\varphi d\theta,\end{aligned} \tag{60}$$

Then the power of ADCRM $P_k^{(\mathrm{AD})}$ can be calculated as

$$\begin{aligned}
P_k^{(\mathrm{AD})} &= \sum_{i=0}^{N_a-1}\sum_{j=0}^{N_g-1} E\left\{\left|\left[\mathbf{H}_k\left(t\right)\right]_{i,j}\right|^2\right\} \\
&\approx \mathrm{Tr}\left\{E\left\{\mathrm{vec}\left\{\bar{\mathbf{H}}_k\left(t\right)\right\}\mathrm{vec}^H\left\{\bar{\mathbf{H}}_k\left(t\right)\right\}\right\}\right\} \\
&= \frac{1}{N_aN_c}\mathrm{Tr}\Big\{\left(\mathbf{F}_{N_c}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)^H E\big\{\mathrm{vec}\left\{\mathbf{G}_k\left(t\right)\right\} \\
&\quad\cdot\mathrm{vec}^H\left\{\mathbf{G}_k\left(t\right)\right\}\big\}\left(\mathbf{F}_{N_c}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)\Big\} \\
&= \frac{1}{N_aN_c}\mathrm{Tr}\Big\{E\left\{\mathrm{vec}\left\{\mathbf{G}_k\left(t\right)\right\}\mathrm{vec}^H\left\{\mathbf{G}_k\left(t\right)\right\}\right\} \\
&\quad\cdot\left(\mathbf{F}_{N_c}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)\left(\mathbf{F}_{N_c}\otimes\left(\mathbf{V}_M\otimes\mathbf{V}_N\right)\right)^H\Big\} \\
&= \frac{1}{N_aN_c}\mathrm{Tr}\left\{E\left\{\mathrm{vec}\left\{\mathbf{G}_k\left(t\right)\right\}\mathrm{vec}^H\left\{\mathbf{G}_k\left(t\right)\right\}\right\}\mathbf{I}_{N_aN_c}\right\} \\
&= \frac{1}{N_aN_c}\sum_{i=0}^{N_a-1}\sum_{j=0}^{N_c-1} E\left\{\left|\left[\mathbf{G}_k\left(t\right)\right]_{i,j}\right|^2\right\}.
\end{aligned} \tag{75}$$

By substituting (72), (75) and (36) into (71), we obtain (37). This completes the proof.

REFERENCES

[1] H. Yin, H. Wang, Y. Liu, and D. Gesbert, "Addressing the curse of mobility in massive MIMO with Prony-based angular-delay domain channel predictions," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2903–2917, Dec. 2020.

[2] X. Li, S. Jin, H. A. Suraweera, J. Hou, and X. Gao, "Statistical 3-D beamforming for large-scale MIMO downlink systems over Rician fading channels," *IEEE Trans. Commun.*, vol. 64, no. 4, pp. 1529–1543, Apr. 2016.

[3] L. You, X. Gao, A. L. Swindlehurst, and W. Zhong, "Channel acquisition for massive MIMO-OFDM with adjustable phase shift pilots," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1461–1476, Mar. 2016.

[4] L. You, X. Gao, G. Y. Li, X. Xia, and N. Ma, "BDMA for millimeter-wave/terahertz massive MIMO transmission with per-beam synchronization," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1550–1563, Jul. 2017.

[5] Ye Li, N. Seshadri, and S. Ariyavisitakul, "Channel estimation for OFDM systems with transmitter diversity in mobile wireless channels," *IEEE J. Sel. Areas Commun.*, vol. 17, no. 3, pp. 461–471, Mar. 1999.

[6] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO: A survey," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 134–141, Sep. 2017.

[7] M. Alodeh, D. Spano, A. Kalantari, C. G. Tsinos, D. Christopoulos, S. Chatzinotas, and B. Ottersten, "Symbol-level and multicast precoding for multiuser multiantenna downlink: A state-of-the-art, classification, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1733–1757, May. 2018.

[8] R. Hadani, S. Rakib, M. Tsatsanis, A. Monk, A. J. Goldsmith, A. F. Molisch, and R. Calderbank, "Orthogonal time frequency space modulation," in *Proc. IEEE WCNC*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.

[9] A. Lu, X. Gao, W. Zhong, C. Xiao, and X. Meng, "Robust transmission for massive MIMO downlink with imperfect CSI," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5362–5376, Aug. 2019.

[10] L. Liu, H. Feng, T. Yang, and B. Hu, "MIMO-OFDM wireless channel prediction by exploiting spatial-temporal correlation," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 310–319, Jan. 2014.

[11] C. Lv, J. Lin, and Z. Yang, "Channel prediction for millimeter wave MIMO-OFDM communications in rapidly time-varying frequency-selective fading channels," *IEEE Access*, vol. 7, pp. 15 183–15 195, Jan. 2019.

[12] A. Duel-Hallen, "Fading channel prediction for mobile radio adaptive transmission systems," *Proc. IEEE*, vol. 95, no. 12, pp. 2299–2313, Dec. 2007.

[13] J. Yuan, H. Q. Ngo, and M. Matthaiou, "Machine learning-based channel prediction in massive MIMO with channel aging," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 2960–2973, May. 2020.

[14] I. C. Wong and B. L. Evans, "Joint channel estimation and prediction for OFDM systems," in *Proc. IEEE GLOBECOM*, vol. 4, St. Louis, MO, USA, Nov. 2005, pp. 5 pp.–2259.

[15] H. P. Bui, Y. Ogawa, T. Nishimura, and T. Ohgane, "Performance evaluation of a multi-user MIMO system with prediction of time-varying indoor channels," *IEEE Trans. Antennas Propag.*, vol. 61, no. 1, pp. 371–379, Jan. 2013.

[16] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*, 1st ed. USA: Academic Press, Inc., 2015.

[17] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[18] W. Peng, W. Li, W. Wang, X. Wei, and T. Jiang, "Downlink channel prediction for time-varying FDD massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 1090–1102, Sep. 2019.

[19] C. Wu, X. Yi, W. Wang, Q. Huang, and X. Gao, "3D CNN-enabled positioning in 3D massive MIMO-OFDM systems," in *Proc. IEEE ICC*, Dublin, Ireland, Jun. 2020, pp. 1–6.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[21] Z. Qin, H. Ye, G. Y. Li, and B. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[22] H. Kim, S. Kim, H. Lee, C. Jang, Y. Choi, and J. Choi, "Massive MIMO channel prediction: Kalman filtering vs. machine learning," *IEEE Trans. Commun.*, pp. 1–1, Sep. 2020.

[23] W. Jiang, M. Strufe, and H. Dieter Schotten, "Long-range MIMO channel prediction using recurrent neural networks," in *Proc. IEEE CCNC*, Las Vegas, NV, USA, Jan. 2020, pp. 1–6.

[24] T. Ding and A. Hirose, "Fading channel prediction based on combination of complex-valued neural networks and chirp z-transform," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 9, pp. 1686–1695, Sep. 2014.

[25] W. Jiang and H. D. Schotten, "Neural network-based fading channel prediction: A comprehensive overview," *IEEE Access*, vol. 7, pp. 118 112–118 124, Aug. 2019.

[26] C. Luo, J. Ji, Q. Wang, X. Chen, and P. Li, "Channel state information prediction for 5G wireless communications: A deep learning approach," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 227–236, Jan. 2020.

[27] Y. Huangfu, J. Wang, R. Li, C. Xu, X. Wang, H. Zhang, and J. Wang, "Predicting the mumble of wireless channel with sequence-to-sequence models," in *Proc. IEEE PIMRC*, Istanbul, Turkey, Sep. 2019, pp. 1–7.

[28] A. Kulkarni, A. Seetharam, A. Ramesh, and J. D. Herath, "Deepchannel: Wireless channel quality prediction using deep learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 443–456, Jan. 2020.

[29] R. Shafin, L. Liu, Y. Li, A. Wang, and J. Zhang, "Angle and delay estimation for 3-D massive MIMO/FD-MIMO systems based on parametric channel modeling," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5370–5383, Aug. 2017.

[30] R. Shafin, L. Liu, J. Zhang, and Y. Wu, "DoA estimation and capacity analysis for 3-D millimeter wave massive-MIMO/fd-MIMO OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6963–6978, Oct. 2016.

[31] X. Sun, X. Gao, G. Y. Li, and W. Han, "Single-site localization based on a new type of fingerprint for massive MIMO-OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6134–6145, Jul. 2018.

[32] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[33] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005. [Online]. Available: https://doi.org/10.1017/CBO9780511807213

[34] L. You, X. Gao, X. Xia, N. Ma, and Y. Peng, "Pilot reuse for massive MIMO transmission over spatially correlated rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3352–3366, Jun. 2015.

[35] Ke Liu, V. Raghavan, and A. M. Sayeed, "Capacity scaling and spectral efficiency in wide-band correlated MIMO channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2504–2526, Oct. 2003.

[36] J. Ma, S. Zhang, H. Li, F. Gao, and S. Jin, "Sparse Bayesian learning for the time-varying massive MIMO channels: Acquisition and tracking," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1925–1938, Mar. 2019.

[37] K. E. Baddour and N. C. Beaulieu, "Autoregressive modeling for fading channel simulation," *IEEE Trans. Wireless Commun.*, vol. 4, no. 4, pp. 1650–1662, Jul. 2005.

[38] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Upper Saddle River, NJ, USA: Prentice-Hall, 2000.

[39] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J.

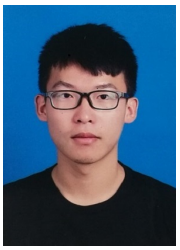
Pal, "Deep complex networks," in *ICLR*, 2018. [Online]. Available: https://openreview.net/forum?id=H1T2hmZAb

[40] S. Jaeckel, L. Raschkowski, K. Borner, and L. Thiele, "QuaDRiGa: A 3-D multi-cell channel model with time evolution for enabling virtual field trials," *IEEE Trans. Antennas Propag.*, vol. 62, no. 6, pp. 3242–3256, Jun. 2014.

[41] *Study on channel model for frequencies from 0.5 to 100 GHz, Version 15.0.0*, document 3GPP T.R. 38.901 , Jun. 2018.

Chi Wu (Student Member, IEEE) received the B.E. degree in information science and engineering from Southeast University, Nanjing, China, in 2018. He is currently pursuing the M.E. degree with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. From August 2019 to November 2019, he was a visiting student at the Department of Electrical Engineering and Electronics, University of Liverpool, U.K. His research interests include massive MIMO wireless localization, channel prediction, and their integration with machine learning.

Xinping Yi (Member, IEEE) received the Ph.D. degree in electronics and communications from Télécom ParisTech, Paris, France, in 2015. He is currently a Lecturer (Assistant Professor) with the Department of Electrical Engineering and Electronics, University of Liverpool, U.K. Prior to Liverpool, he was a Research Associate with Technische Universität Berlin, Berlin, Germany, from 2014 to 2017, a Research Assistant with EURECOM, Sophia Antipolis, France, from 2011 to 2014, and a Research Engineer with Huawei Technologies, Shenzhen, China, from 2009 to 2011. His main research interests include information theory, graph theory, and machine learning, and their applications in wireless communications and artificial intelligence.



Yiming Zhu (Student Member, IEEE) received the B.E. degree in communication engineering from Nanjing University of Science and Technology, Nanjing, China, in 2020. He is currently working towards the M.E. degree with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China. His research interests include massive MIMO wireless channel estimation and prediction.

Wenjin Wang (Member, IEEE) received the Ph.D. degree in communication and information systems from Southeast University, Nanjing, China, in 2011. From 2010 to 2014, he was with the School of System Engineering, University of Reading, Reading, U.K. He is currently an Associate Professor with the National Mobile Communications Research Laboratory, Southeast University, Nanjing. His research interests include advanced signal processing for future wireless communications and satellite communications. He was a recipient of the Best Paper Award at IEEE WCSP 2009 and the first grade Technological Invention Award of the State Education Ministry of China, in 2009.

Li You (Member, IEEE) received the B.E. and M.E. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree from Southeast University, Nanjing, in 2016, all in electrical engineering.

From 2014 to 2015, he conducted Visiting Research at the Center for Pervasive Communications and Computing, University of California Irvine, Irvine, CA, USA. Since 2016, he has been with the Faculty of the National Mobile Communications Research Laboratory, Southeast University. His research interests lie in the general areas of communications, signal processing, and information theory with the current emphasis on massive MIMO communications.

Xiqi Gao (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Southeast University, Nanjing, China, in 1997.

He joined the Department of Radio Engineering, Southeast University, in April 1992. Since May 2001, he has been a professor of information systems and communications. From September 1999 to August 2000, he was a visiting scholar at Massachusetts Institute of Technology, Cambridge, and Boston University, Boston, MA. From August 2007 to July 2008, he visited the Darmstadt University of Technology, Darmstadt, Germany, as a Humboldt scholar. His current research interests include broadband multi-carrier communications, MIMO wireless communications, channel estimation and turbo equalization, and multi-rate signal processing for wireless communications. From 2007 to 2012, he served as an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. From 2009 to 2013, he served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. From 2015 to 2017, he served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Dr. Gao received the Science and Technology Awards of the State Education Ministry of China in 1998, 2006 and 2009, the National Technological Invention Award of China in 2011, and the 2011 IEEE Communications Society Stephen O. Rice Prize Paper Award in the field of communications theory.