Distributionally Robust Receive Combining

Shixiong Wang, Wei Dai, and Geoffrey Ye Li, *Fellow, IEEE*

Abstract—This article investigates signal estimation in wireless transmission (i.e., receive combining) from the perspective of statistical machine learning, where the transmit signals may be from an integrated sensing and communication system; that is, 1) signals may be not only discrete constellation points but also arbitrary complex values; 2) signals may be spatially correlated. Particular attention is paid to handling various uncertainties such as the uncertainty of the transmit signal covariance, the uncertainty of the channel matrix, the uncertainty of the channel noise covariance, the existence of channel impulse noises, the non-ideality of the power amplifiers, and the limited sample size of pilots. To proceed, a distributionally robust receive combining framework that is insensitive to the above uncertainties is proposed, which reveals that channel estimation is not a necessary operation. For optimal linear estimation, the proposed framework includes several existing combiners as special cases such as diagonal loading and eigenvalue thresholding. For optimal nonlinear estimation, estimators are limited in reproducing kernel Hilbert spaces and neural network function spaces, and corresponding uncertainty-aware solutions (e.g., kernelized diagonal loading) are derived. In addition, we prove that the ridge and kernel ridge regression methods in machine learning are distributionally robust against diagonal perturbation in feature covariance.

Index Terms—Wireless Transmission, Smart Antenna, Machine Learning, Robust Estimation, Robust Combining, Distributional Uncertainty, Channel Uncertainty, Limited Pilot.

I. Introduction

IN wireless transmission, detection and estimation of transmitted signals is of high importance, and combining at array receivers serves as a key signal-processing technique to suppress interference and environmental noises. The earliest beamforming solutions rely on the use of phase shifters (e.g., phased arrays) to steer and shape wave lobes, while advanced combining methods allow the employment of digital signal processing units, which introduce additional structural freedom (e.g., fully digital, hybrid, nonlinear, wideband) in combiner design and significant performance improvement in signal recovery [1]–[3].

In traditional communication systems, transmitted signals are discrete points from constellations. Therefore, signal recovery, commonly referred to as *signal detection*, can be cast into a classification problem from the perspective of statistical machine learning, and the number of candidate classes is determined by the number of points in the employed constellation. Research in this stream includes, e.g., [4]–[9] as well as references therein, and the performance measure for signal detection is usually the misclassification rate (i.e., symbol error

S. Wang, W. Dai, and G. Li are with the Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, United Kingdom (E-mail: s.wang@u.nus.edu; wei.dai1@imperial.ac.uk; geoffrey.li@imperial.ac.uk).

rate); representative algorithms encompass the maximum likelihood detector, the sphere decoding, etc. In another research stream, the signal recovery performance is evaluated using mean-squared errors (cf., signal-to-interference-plus-noise ratio), and the resultant signal recovery problem is commonly known as *signal estimation*, which can be considered as a regression problem from the perspective of statistical machine learning. By comparing the estimated symbols with the constellation points afterward, the detection of discrete symbols can be realized. For this case, till now, typical combining solutions include zero-forcing receivers, Wiener receivers (i.e., linear minimum mean-squared error receivers), Capon receivers (i.e., minimum variance distortionless response receivers), and nonlinear receivers such as neural-network receivers [10]–[12]. On the basis of these canonical approaches, variants such as robust beamformers working against the limited size of pilot samples and the uncertainty in steering vectors [13]–[18] have also been intensively reported; among these robust solutions, the diagonal loading method [19], [14, Eq. (11)] and the eigenvalue thresholding method [20], [14, Eq. (12)] are popular due to their excellent balance between practical performance and technical simplicity.

Different from traditional paradigms, in emerging communication systems, e.g., integrated sensing and communication (ISAC) systems, transmitted signals may be arbitrary complex values and spatially correlated [21]–[23]. As a result, mean-squared error is a preferred performance measure to investigate the *receive combining and estimation* problem of wireless signals, which is, therefore, the focus of this article.

Although a large body of problems have been attacked in the area, the following signal-processing problems of combining and estimation in wireless transmission remain unsolved.

1) What is the relation between the signal-model-based approaches (e.g., Wiener and Capon receivers) and the data-driven approaches (e.g., deep-learning receivers)? In other words, how can we build a mathematically unified modeling framework to interpret all the existing digital receive combiners?

2) In addition to the limited pilot size and the uncertainty in steering vectors, there exist other uncertainties in the signal model: the uncertainty of the transmit signal covariance, the uncertainty of the communication channel matrix, the uncertainty of the channel noise covariance, the presence of channel impulse noises (i.e., outliers), and the non-ideality of the power amplifiers. Therefore, how can we handle all these types of uncertainties in a unified solution framework?

3) Existing literature mainly studied the robustness theory of linear beamformers against limited pilot size and the uncertainty in steering vectors [13]–[18]. However, how can we develop the theory of robust nonlinear combiners

against all the aforementioned uncertainties?

To this end, this article designs a unified modeling and solution framework for receive combining of wireless signals, in consideration of the scarcity of the pilot data and the different uncertainties in the signal model.

A. Contributions

The contributions of this article can be summarized from the aspects of machine learning theory and wireless transmission theory.

In terms of machine learning theory, we give a justification of the popular ridge regression and kernel ridge regression (i.e., quadratic loss function plus squared-F-norm regularization) from the perspective of distributional robustness against diagonal perturbation in feature covariance, which enriches the theory of trustworthy machine learning; see Theorems 2 and 3, as well as Corollaries 3 and 5.

In terms of wireless transmission theory, the contributions are outlined below.

1) We build a fundamentally theoretical framework for receive combining from the perspective of statistical machine learning. In addition to the linear estimation methods, nonlinear approaches (i.e., nonlinear combining) are also discussed in reproducing kernel Hilbert spaces and neural network function spaces. In particular, we reveal that channel estimation is not a necessary operation in receive combining. For details, see Subsection III-A.

2) The presented framework is particularly developed from the perspective of distributional robustness which can therefore combat the limited size of pilot data and several types of uncertainties in the wireless signal model such as the uncertainty in the transmit power matrix, the uncertainty in the communication channel matrix, the existence of channel impulse noises (i.e., outliers), the uncertainty in the covariance matrix of channel noises, the non-ideality of the power amplifiers, etc. For details, see Subsection III-B, and the technical developments in Sections IV and V.

3) Existing methods such as diagonal loading and eigenvalue thresholding are proven to be distributionally robust against the limited pilot size and all the aforementioned uncertainties in the wireless signal model. Extensions of diagonal loading and eigenvalue thresholding are proposed as well. Moreover, the kernelized diagonal loading and the kernelized eigenvalue thresholding methods are put forward for nonlinear estimation cases. For details, see Corollary 1, Examples 4 and 5, and Subsections IV-B.

4) The distributionally robust receive combining and signal estimation problems across multiple frames, where channel conditions may change, are also investigated. For details, see Subsections IV-C and V-A2.

B. Notations

The N-dimensional real (coordinate) space and complex (coordinate) space are denoted as $\mathbb{R}^N$ and $\mathbb{C}^N$, respectively. Lowercase symbols (e.g., x) denote vectors (column by default) and uppercase ones (e.g., X) denote matrices. We use the Roman font for random quantities (e.g., $\mathbf{x}, \mathbf{X}$) and the italic font for deterministic quantities (e.g., x, X). Let $\operatorname{Re} X$ be the real part of a complex quantity X (a vector or matrix) and $\operatorname{Im} X$ be the imaginary part of X. For a vector $x \in \mathbb{C}^N$, let

$$\underline{x} := \left[\begin{array}{c} \operatorname{Re} x \\ \operatorname{Im} x \end{array} \right] \in \mathbb{R}^{2N}$$

be the real-space representation of x; for a matrix $H \in \mathbb{C}^{N \times M}$, let

$$\underline{H} := \left[\begin{array}{c} \operatorname{Re} H \\ \operatorname{Im} H \end{array} \right], \qquad \underline{\underline{H}} := \left[\begin{array}{cc} \operatorname{Re} H & -\operatorname{Im} H \\ \operatorname{Im} H & \operatorname{Re} H \end{array} \right]$$

be the real-space representations of H where $\underline{H} \in \mathbb{R}^{2N \times M}$ and $\underline{\underline{H}} \in \mathbb{R}^{2N \times 2M}$. The running index set induced by an integer N is defined as $[N] := \{1, 2, \ldots, N\}$. To concatenate matrices and vectors, MATLAB notations are used: i.e., $[A, \ B]$ for row stacking and $[A; \ B]$ for column stacking. We let $\Gamma_M := [I_M, \ J_M] \in \mathbb{C}^{M \times 2M}$ where I_M denotes the M-dimensional identity matrix, $J_M := j \cdot I_M$, and j denotes the imaginary unit. Let $\mathcal{N}(\mu, \Sigma)$ denote a real Gaussian distribution with mean μ and covariance Σ. We use $\mathcal{CN}(s, P, C)$ to denote a complex Gaussian distribution with mean s, covariance P, and pseudo-covariance C; if C is not specified, we imply $C = 0$.

II. PRELIMINARIES

We review two popular structured representation methods of nonlinear functions $\phi : \mathbb{R}^N \to \mathbb{R}^M$. More details can be seen in Appendix A.

A. Reproducing Kernel Hilbert Spaces

A reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ induced by the kernel function $\operatorname{ker} : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ and a collection of points $\{x_1, x_2, \ldots, x_L\} \subset \mathbb{R}^N$ is a set of functions from $\mathbb{R}^N$ to $\mathbb{R}$; L may be infinite. Every function $\phi : \mathbb{R}^N \to \mathbb{R}$ in the functional space $\mathcal{H}$ can be represented by a linear combination [24, p. 539; Chap. 14]

$$\phi(x) = \sum_{i=1}^{L} \omega_i \cdot \operatorname{ker}(x, x_i), \ \forall x \in \mathbb{R}^N \tag{1}$$

where $\{\omega_i\}_{i \in [L]}$ are the combination weights; $\omega_i \in \mathbb{R}$ for every $i \in [L]$. The matrix form of (1) for M-multiple functions are

$$\phi(x) := \left[\begin{array}{c} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_M(x) \end{array} \right] = W \cdot \varphi(x) := \left[\begin{array}{c} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_M \end{array} \right] \cdot \varphi(x), \tag{2}$$

where $\omega_1, \omega_2, \ldots, \omega_M \in \mathbb{R}^L$ are weight row-vectors for functions $\phi_1(x), \phi_2(x), \ldots, \phi_M(x)$, respectively, and

$$W := \left[\begin{array}{c} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_M \end{array} \right] \in \mathbb{R}^{M \times L}, \quad \varphi(x) := \left[\begin{array}{c} \operatorname{ker}(x, x_1) \\ \operatorname{ker}(x, x_2) \\ \vdots \\ \operatorname{ker}(x, x_L) \end{array} \right]. \tag{3}$$

Since a kernel function is pre-designed (i.e., fixed) for an RKHS $\mathcal{H}$, (2) suggests a W-linear representation of x-nonlinear functions $\phi(x)$ in $\mathcal{H}^M$. Note that there exists a

one-to-one correspondence between ϕ and $\boldsymbol{W}$: for every $\phi : \mathbb{R}^N \to \mathbb{R}^M$, there exists a $\boldsymbol{W} \in \mathbb{R}^{M \times L}$, and vice versa.

B. Neural Networks

Neural networks (NN) are another powerful tool to represent (i.e., approximate) nonlinear functions. A neural network function space (NNFS) $\mathcal{K}$ characterizes (or parameterizes) a set of multi-input multi-output functions. Typical choices are multi-layer feed-forward neural networks, recurrent neural networks, etc. For combining and estimation of wireless signals, the multi-layer feed-forward neural networks are standard [10]–[12]. Suppose that we have $R - 1$ hidden layers (so in total $R + 1$ layers including one input layer and one output layer) and each layer $r = 0, 1, \ldots, R$ contains T_r neurons. To represent a function $\phi : \mathbb{R}^N \to \mathbb{R}^M$, for the input layer $r = 0$ and output layer $r = R$, we have $T_0 = N$ and $T_R = M$, respectively. Let the output of the rth layer be $\boldsymbol{y}_r \in \mathbb{R}^{T_r}$. For every layer r, we have $\boldsymbol{y}_r = \boldsymbol{\sigma}_r(\boldsymbol{W}_r^\circ \cdot \boldsymbol{y}_{r-1} + \boldsymbol{b}_r)$ where $\boldsymbol{W}_r^\circ \in \mathbb{R}^{T_r \times T_{r-1}}$ is the weight matrix, $\boldsymbol{b}_r \in \mathbb{R}^{T_r}$ is the bias vector, and the multi-output function $\boldsymbol{\sigma}_r$ is the activation function which is entry-wise identical. Hence, every function $\phi : \mathbb{R}^N \to \mathbb{R}^M$ in a NNFS can be recursively expressed as [25, Chap. 5], [26]

$$
\begin{aligned}
\phi(\boldsymbol{x}) &= \boldsymbol{\sigma}_R(\boldsymbol{W}_R \cdot [\boldsymbol{y}_{R-1}(\boldsymbol{x}); \; 1]) \\
\boldsymbol{y}_r(\boldsymbol{x}) &= \boldsymbol{\sigma}_r(\boldsymbol{W}_r \cdot [\boldsymbol{y}_{r-1}(\boldsymbol{x}); \; 1]), \quad r \in [R-1] \\
\boldsymbol{y}_0(\boldsymbol{x}) &= \boldsymbol{x},
\end{aligned}
\tag{4}
$$

where $\boldsymbol{W}_r := [\boldsymbol{W}_r^\circ, \; \boldsymbol{b}_r]$ for $r \in [R]$. Note that the activation functions can vary from one layer to another.

III. PROBLEM FORMULATION

Consider a narrow-band wireless signal transmission model

$$
\mathbf{x} = \boldsymbol{H}\mathbf{s} + \mathbf{v}
\tag{5}
$$

where $\mathbf{x} \in \mathbb{C}^N$ is the received signal, $\mathbf{s} \in \mathbb{C}^M$ is the transmitted signal, $\boldsymbol{H} \in \mathbb{C}^{N \times M}$ is the channel matrix, and $\mathbf{v} \in \mathbb{C}^N$ is the zero-mean channel noise. The precoding operation (if exists) is integrated in $\boldsymbol{H}$. The transmitted symbols $\mathbf{s}$ have zero means, which may be not only discrete symbols from constellations such as quadrature amplitude modulation but also arbitrary values such as integrated sensing and communication signals. We consider L pilots $\mathbf{S} := (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L)$ in each frame, and the corresponding received symbols are $\mathbf{X} := (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L)$ under the noise $(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_L)$. We suppose that $\boldsymbol{R}_s := \mathbb{E}\mathbf{s}\mathbf{s}^{\mathsf{H}}$ and $\boldsymbol{R}_v := \mathbb{E}\mathbf{v}\mathbf{v}^{\mathsf{H}}$ may not be identity or diagonal matrices: i.e., the components of $\mathbf{s}$ can be correlated (e.g., in ISAC), so can be these of $\mathbf{v}$. Consider the real-space representation of the signal model (5) by stacking the real and imaginary components:

$$
\underline{\mathbf{x}} = \underline{\underline{\boldsymbol{H}}} \cdot \underline{\mathbf{s}} + \underline{\mathbf{v}},
\tag{6}
$$

where $\underline{\mathbf{x}} \in \mathbb{R}^{2N}$, $\underline{\underline{\boldsymbol{H}}} \in \mathbb{R}^{2N \times 2M}$, $\underline{\mathbf{s}} \in \mathbb{R}^{2M}$, and $\underline{\mathbf{v}} \in \mathbb{R}^{2N}$. The expressions of $\underline{\boldsymbol{R}_x} := \mathbb{E}\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{T}}$, $\underline{\boldsymbol{R}_s} := \mathbb{E}\underline{\mathbf{s}}\underline{\mathbf{s}}^{\mathsf{T}}$, $\underline{\boldsymbol{R}_{xs}} := \mathbb{E}\underline{\mathbf{x}}\underline{\mathbf{s}}^{\mathsf{T}}$, and $\underline{\boldsymbol{R}_v} := \mathbb{E}\underline{\mathbf{v}}\underline{\mathbf{v}}^{\mathsf{T}}$ can be readily obtained; see Appendix B. In some cases, signal estimation in real spaces can be technically simpler than that in complex spaces.

A. Optimal Estimation

1) Optimal Nonlinear Estimation (Receive Combining): To recover $\mathbf{s}$ using $\mathbf{x}$, we consider an estimator $\hat{\mathbf{s}} := \phi(\mathbf{x})$, called a receive combiner, at the receiver where $\phi : \mathbb{C}^N \to \mathbb{C}^M$ is a Borel-measurable function. Note that $\phi(\mathbf{x})$ may be nonlinear in general because the joint distribution of $(\mathbf{x}, \mathbf{s})$ is not necessarily Gaussian, for example, when the channel noise $\mathbf{v}$ is non-Gaussian or when the power amplifiers work in non-linear regions. The signal estimation problem at the receiver can be written as a statistical machine-learning problem under the joint data distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ of $(\mathbf{x}, \mathbf{s})$, that is,

$$
\min_{\phi \in \mathcal{B}_{\mathbb{C}^N \to \mathbb{C}^M}} \mathrm{Tr}\, \mathbb{E}_{\mathbf{x},\mathbf{s}}[\phi(\mathbf{x}) - \mathbf{s}][\phi(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}},
\tag{7}
$$

where $\mathcal{B}_{\mathbb{C}^N \to \mathbb{C}^M}$ contains all Borel-measurable estimators from $\mathbb{C}^N$ to $\mathbb{C}^M$. In what follows, we omit the notational dependence on $\mathbb{C}^N$ and $\mathbb{C}^M$, and use $\mathcal{B}$ as a shorthand. The optimal estimator, in the sense of minimum mean-squared error, is known as the conditional mean of $\mathbf{s}$ given $\mathbf{x}$, i.e.,

$$
\hat{\mathbf{s}} = \phi(\mathbf{x}) = \mathbb{E}(\mathbf{s}|\mathbf{x}).
\tag{8}
$$

Usually, it is computationally complicated to find the optimal $\phi(\cdot)$ from the whole space $\mathcal{B}$ of Borel-measurable functions, that is, to compute the conditional mean. Therefore, in practice, we may find the optimal approximation of $\phi(\cdot)$ in an RKHS $\mathcal{H}$ or a NNFS $\mathcal{K}$; note that $\mathcal{H}$ and $\mathcal{K}$ are two subspaces of $\mathcal{B}$. However, both $\mathcal{H}$ and $\mathcal{K}$ are sufficiently rich because they can be dense in the space of all continuous bounded functions.

2) Optimal Linear Estimation (Receive Beamforming): If $\mathbf{x}$ and $\mathbf{s}$ are jointly Gaussian (e.g., when $\mathbf{s}$ and $\mathbf{v}$ are jointly Gaussian), the optimal estimator ϕ is linear in $\mathbf{x}$:

$$
\hat{\mathbf{s}} = \boldsymbol{W}\mathbf{x},
\tag{9}
$$

where $\boldsymbol{W} \in \mathbb{C}^{M \times N}$ is called a receive beamformer or a linear receive combiner. In this linear case, (7) reduces to the usual Wiener–Hopf beamforming problem

$$
\min_{\boldsymbol{W}} \mathrm{Tr}\, \mathbb{E}_{\mathbf{x},\mathbf{s}}[\boldsymbol{W}\mathbf{x} - \mathbf{s}][\boldsymbol{W}\mathbf{x} - \mathbf{s}]^{\mathsf{H}},
\tag{10}
$$

that is,

$$
\min_{\boldsymbol{W}} \mathrm{Tr}\, \big[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\boldsymbol{R}_{xs} - \boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \boldsymbol{R}_s\big],
\tag{11}
$$

where $\boldsymbol{R}_x := \mathbb{E}\mathbf{x}\mathbf{x}^{\mathsf{H}} \in \mathbb{C}^{N \times N}$ and $\boldsymbol{R}_{xs} := \mathbb{E}\mathbf{x}\mathbf{s}^{\mathsf{H}} \in \mathbb{C}^{N \times M}$. Since $\boldsymbol{R}_x = \boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v$ and $\boldsymbol{R}_{xs} = \boldsymbol{H}\boldsymbol{R}_s + \mathbb{E}\mathbf{v}\mathbf{s}^{\mathsf{H}} = \boldsymbol{H}\boldsymbol{R}_s$, the solution of (11), or (10), is

$$
\begin{aligned}
\boldsymbol{W}_{\text{Wiener}}^{\star} &= \boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1} \\
&= \boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}[\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v]^{-1},
\end{aligned}
\tag{12}
$$

which is known as the Wiener beamformer. With an additional constraint $\boldsymbol{W}\boldsymbol{H} = \boldsymbol{I}_M$ (i.e., distortionless response), (11) gives the Capon beamformer. Both the Wiener beamformer and the Capon beamformer maximize the output signal–to–interference-plus-noise ratio (SINR); hence, both are optimal in the sense of maximum output SINR.

No matter whether $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is Gaussian or not, (10) or (11) identifies the ***optimal linear estimator*** in the sense of minimum mean-squared error among all linear estimators.

3) Role of Channel Estimation: Eqs. (7) and (10) imply that channel estimation is not a necessary step in receive combining. The only necessary element, from the perspective of statistical machine learning, is the joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ of the received signal $\mathbf{x}$ and the transmitted signal $\mathbf{s}$. Therefore, the following two points can be highlighted.

a) If the joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is non-Gaussian, we just need to learn the mapping ϕ using (7).

b) If the joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is (or assumed to be) Gaussian, we just learn covariance matrices $\boldsymbol{R}_{xs}$ and $\boldsymbol{R}_x$; cf. (12); Gaussianity assumption of $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is beneficial in reducing computational burdens. If, further, the channel matrix $\boldsymbol{H}$ is known, $\boldsymbol{R}_{xs}$ and $\boldsymbol{R}_x$ can be expressed using $\boldsymbol{H}$.

B. Distributional Uncertainty and Distributional Robustness

For ease of conceptual illustration, we start with the following stationary-channel assumption in this subsection: The channel statistics remain unchanged within the communication frame so that the joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is fixed over time. That is, pilot data $\{(\boldsymbol{x}_1,\boldsymbol{s}_1),(\boldsymbol{x}_2,\boldsymbol{s}_2),\ldots,(\boldsymbol{x}_L,\boldsymbol{s}_L)\}$ and non-pilot communication data are drawn from the same unknown distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$. For the general case where the channel is not statistically stationary within a frame, see Appendix C; the statistical non-stationarity of $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ may be due to the time-selectivity of the transmit power matrix $\boldsymbol{R}_s$, of the channel matrix $\boldsymbol{H}$, and/or of the channel noise covariance $\boldsymbol{R}_v$.

1) Issue of Distributional Uncertainty: In practice, the true joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is unknown but can be estimated by the pilot data. Hence, the estimation of wireless signals is a data-driven statistical inference (i.e., statistical machine learning) problem. We let

$$\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}} := \frac{1}{L}\sum_{i=1}^{L}\delta_{(\boldsymbol{x}_i,\boldsymbol{s}_i)} \qquad (13)$$

denote the empirical distribution supported on the L collected data $\{(\boldsymbol{x}_i,\boldsymbol{s}_i)\}_{i\in[L]}$, where $\delta_{(\boldsymbol{x}_i,\boldsymbol{s}_i)}$ denotes the Dirac distribution (i.e., point-mass distribution) centered on $(\boldsymbol{x}_i,\boldsymbol{s}_i)$; note that $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ is a discrete distribution. If we use the estimated joint distribution $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ as a surrogate of the true joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$, (7) becomes the conventional empirical risk minimization (ERM)

$$\min_{\phi\in\mathcal{B}}\operatorname{Tr}\mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}}[\phi(\mathbf{x})-\mathbf{s}][\phi(\mathbf{x})-\mathbf{s}]^{\mathsf{H}}, \qquad (14)$$

i.e.,

$$\min_{\phi\in\mathcal{B}}\operatorname{Tr}\frac{1}{L}\sum_{i=1}^{L}[\phi(\boldsymbol{x}_i)-\boldsymbol{s}_i][\phi(\boldsymbol{x}_i)-\boldsymbol{s}_i]^{\mathsf{H}}. \qquad (15)$$

Likewise, (11) become the conventional beamforming problem

$$\min_{\boldsymbol{W}}\operatorname{Tr}\left[\boldsymbol{W}\hat{\boldsymbol{R}}_x\boldsymbol{W}^{\mathsf{H}}-\boldsymbol{W}\hat{\boldsymbol{R}}_{xs}-\hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}}+\hat{\boldsymbol{R}}_s\right], \qquad (16)$$

where $\hat{\boldsymbol{R}}_x$, $\hat{\boldsymbol{R}}_{xs}$, and $\hat{\boldsymbol{R}}_s$ are the training-sample-estimated (i.e., nominal) values of $\boldsymbol{R}_x$, $\boldsymbol{R}_{xs}$, and $\boldsymbol{R}_s$, respectively.

There exists the distributional difference between the sample-defined nominal distribution $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ and true data-generating distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ due to the limited size of the training data set (i.e., limited pilot length) and the time-selectivity of $\mathbb{P}_{\mathbf{x},\mathbf{s}}$. From the perspective of applied statistics and machine learning, the distributional difference between $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ and $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ (i.e., the distributional uncertainty of $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ compared to $\mathbb{P}_{\mathbf{x},\mathbf{s}}$) may cause significant performance degradation of (15) compared to (7), so is the performance deterioration of (16) compared to (11). For extensive reading on this point, see Appendix C. Therefore, to reduce the adverse effect introduced by the distributional uncertainty in $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$, a new surrogate of (7) rather than the sample-averaged approximation in (15) is expected.

2) Distributionally Robust Estimation: To combat the distributional uncertainty in $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$, we consider the distributionally robust counterpart of (7)

$$\min_{\phi\in\mathcal{B}}\max_{\mathbb{P}_{\mathbf{x},\mathbf{s}}\in\mathcal{U}_{\mathbf{x},\mathbf{s}}}\operatorname{Tr}\mathbb{E}_{\mathbf{x},\mathbf{s}}[\phi(\mathbf{x})-\mathbf{s}][\phi(\mathbf{x})-\mathbf{s}]^{\mathsf{H}}, \qquad (17)$$

where $\mathcal{U}_{\mathbf{x},\mathbf{s}}$, called a distributional uncertainty set, contains a collection of distributions that are close to the nominal distribution (i.e., the sample-estimated distribution) $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$;

$$\mathcal{U}_{\mathbf{x},\mathbf{s}} := \{\mathbb{P}_{\mathbf{x},\mathbf{s}}|\ d(\mathbb{P}_{\mathbf{x},\mathbf{s}},\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}})\le\epsilon\}, \qquad (18)$$

where $d(\cdot,\cdot)$ denotes a similarity measure (e.g., metric or divergence) between two distributions and $\epsilon\ge 0$ an uncertainty quantification level. Since $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ is discrete and $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is not, the Wasserstein distance [27, Def. 2] and the maximum mean discrepancy (MMD) distance [28, Def. 2.1] are the typical choices of $d(\cdot,\cdot)$ to construct $\mathcal{U}_{\mathbf{x},\mathbf{s}}$. When $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ and $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ are parametric distributions (e.g., Gaussian, exponential family), divergences such as the Kullback--Leibler (KL) divergence, or more general ϕ-divergence, are also applicable to particularize $d(\cdot,\cdot)$ because parameters can be estimated using samples. When $\epsilon=0$, (17) reduces to (15).

If $\mathcal{U}_{\mathbf{x},\mathbf{s}}$ contains (or is assumed, for computational simplicity, to contain) only Gaussian distributions, (17) particularizes to

$$\begin{aligned}\min_{\boldsymbol{W}}\max_{\boldsymbol{R}}\quad &\operatorname{Tr}\left[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}}-\boldsymbol{W}\boldsymbol{R}_{xs}-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}}+\boldsymbol{R}_s\right]\\ \text{s.t.}\quad &d_0(\boldsymbol{R},\ \hat{\boldsymbol{R}})\le\epsilon_0,\\ &\boldsymbol{R}\succeq\boldsymbol{0},\end{aligned}$$
$$(19)$$

where

$$\boldsymbol{R} := \begin{bmatrix}\boldsymbol{R}_x & \boldsymbol{R}_{xs}\\ \boldsymbol{R}_{xs}^{\mathsf{H}} & \boldsymbol{R}_s\end{bmatrix},\quad \hat{\boldsymbol{R}} := \begin{bmatrix}\hat{\boldsymbol{R}}_x & \hat{\boldsymbol{R}}_{xs}\\ \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}} & \hat{\boldsymbol{R}}_s\end{bmatrix}, \qquad (20)$$

because every zero-mean complex Gaussian distribution is uniquely characterized by its covariance and pseudo-covariance, but in receive beamforming, we do not consider pseudo-covariances; cf. (12); d_0 denotes the matrix similarity measures (e.g., matrix distances); $\epsilon_0\ge 0$ is the uncertainty quantification parameter. When $\epsilon_0=0$, (19) reduces to (16).

For additional discussions on the framework of distributionally robust estimation, see Appendix D.

IV. DISTRIBUTIONALLY ROBUST LINEAR ESTIMATION

Due to several practical benefits of linear estimation, for example, the simplicity of hardware structures, the clarity of physical meaning (i.e., constructive and destructive interference through beamforming), and the easiness of computations, investigating distributionally robust linear estimation problems is important. This section particularly studies Problem (19).

A. General Framework and Concrete Examples

The following lemma solves Problem (19).

Lemma 1: Suppose that the set $\{\boldsymbol{R}|\ d_0(\boldsymbol{R},\ \hat{\boldsymbol{R}}) \leq \epsilon_0\}$ is compact convex and $\boldsymbol{R}_x$ is invertible. Let $\boldsymbol{R}^\star$ solve the problem below:

$$\begin{aligned} \max_{\boldsymbol{R}} \quad & \mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs} + \boldsymbol{R}_s\right] \\ \mathrm{s.t.} \quad & d_0(\boldsymbol{R},\ \hat{\boldsymbol{R}}) \leq \epsilon_0, \\ & \boldsymbol{R} \succeq \boldsymbol{0}, \quad \boldsymbol{R}_x \succ \boldsymbol{0}. \end{aligned} \quad (21)$$

Construct $\boldsymbol{W}^\star$ using $\boldsymbol{R}^\star$ as follows:

$$\boldsymbol{W}^\star := \boldsymbol{R}_{xs}^{\star\mathsf{H}}\boldsymbol{R}_x^{\star-1}. \quad (22)$$

Then $(\boldsymbol{W}^\star, \boldsymbol{R}^\star)$ is a solution to Problem (19). On the other hand, if $(\boldsymbol{W}^\star, \boldsymbol{R}^\star)$ solves Problem (19), then $\boldsymbol{R}^\star$ is a solution to (21) and $(\boldsymbol{W}^\star, \boldsymbol{R}^\star)$ satisfies (22).

Proof: See Appendix E. $\qquad\square$

Let

$$f_1(\boldsymbol{R}) := \mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs} + \boldsymbol{R}_s\right] \quad (23)$$

denote the objective function of (21). When $\boldsymbol{R}_s$ and $\boldsymbol{R}_{xs}$ are fixed, we define

$$f_2(\boldsymbol{R}_x) := \mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs} + \boldsymbol{R}_s\right]. \quad (24)$$

The theorem below studies the properties of f_1 and f_2.

Theorem 1: Consider the definition of $\boldsymbol{R}$ in (20). The functions f_1 defined in (23) and f_2 defined in (24) are monotonically increasing in $\boldsymbol{R}$ and $\boldsymbol{R}_x$, respectively. To be specific, if $\boldsymbol{R}_1 \succeq \boldsymbol{R}_2 \succeq \boldsymbol{0}$, $\boldsymbol{R}_{1,x} \succ \boldsymbol{0}$, and $\boldsymbol{R}_{2,x} \succ \boldsymbol{0}$, we have $f_1(\boldsymbol{R}_1) \geq f_1(\boldsymbol{R}_2)$. In addition, if $\boldsymbol{R}_{1,x} \succeq \boldsymbol{R}_{2,x} \succ \boldsymbol{0}$, we have $f_2(\boldsymbol{R}_{1,x}) \geq f_2(\boldsymbol{R}_{2,x})$.

Proof: See Appendix F. $\qquad\square$

To concretely solve (21), we need to particularize d_0. This article investigates the following uncertainty sets.

Definition 1 (Additive Moment Uncertainty Set): The additive moment uncertainty set of $\boldsymbol{R}$ is constructed as

$$\{\boldsymbol{R}|\ \hat{\boldsymbol{R}} - \epsilon_0\boldsymbol{E} \preceq \boldsymbol{R} \preceq \hat{\boldsymbol{R}} + \epsilon_0\boldsymbol{E},\ \boldsymbol{R} \succeq \boldsymbol{0}\} \quad (25)$$

for some $\boldsymbol{E} \succeq \boldsymbol{0}$ and $\epsilon_0 \geq 0$. $\qquad\square$

Definition 1 is motivated by the fact that the difference $\boldsymbol{R} - \hat{\boldsymbol{R}}$ is bounded by some threshold matrix $\boldsymbol{E}$ and error quantification level ϵ_0: specifically, $-\epsilon_0\boldsymbol{E} \preceq \boldsymbol{R} - \hat{\boldsymbol{R}} \preceq \epsilon_0\boldsymbol{E}$. In practice, we can consider the threshold as an identity matrix because, for every non-identity $\boldsymbol{E} \succeq \boldsymbol{0}$, we have $\boldsymbol{E} \preceq \lambda_1\boldsymbol{I}_{N+M}$ where λ_1 is the largest eigenvalue of $\boldsymbol{E}$.

Definition 2 (Diagonal-Loading Uncertainty Set): The diagonal-loading uncertainty set of $\boldsymbol{R}$ is constructed as

$$\{\boldsymbol{R}|\ \hat{\boldsymbol{R}} - \epsilon_0\boldsymbol{I}_{N+M} \preceq \boldsymbol{R} \preceq \hat{\boldsymbol{R}} + \epsilon_0\boldsymbol{I}_{N+M},\ \boldsymbol{R} \succeq \boldsymbol{0}\} \quad (26)$$

for some $\epsilon_0 \geq 0$. $\qquad\square$

Due to the concentration property of the sample-covariance $\hat{\boldsymbol{R}}$ to the true covariance $\boldsymbol{R}$ when the true distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is fixed within a frame, finite values of ϵ_0 exist for every sample size L; NB: $\epsilon_0 \to 0$ as $L \to \infty$. However, given L, the smallest ϵ_0 cannot be practically calculated because it depends on the true but unknown $\mathbb{P}_{\mathbf{x},\mathbf{s}}$. If $\boldsymbol{E}$ is block-diagonal, the generalized diagonal-loading uncertainty set can be motivated.

Definition 3 (Generalized Diagonal-Loading Uncertainty Set): The generalized diagonal-loading uncertainty set of $\boldsymbol{R}$ is constructed by the following constraints: $\boldsymbol{R} \succeq \boldsymbol{0}$ and

$$\begin{aligned} & \begin{bmatrix} \hat{\boldsymbol{R}}_x & \hat{\boldsymbol{R}}_{xs} \\ \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}} & \hat{\boldsymbol{R}}_s \end{bmatrix} - \epsilon_0 \begin{bmatrix} \boldsymbol{F} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G} \end{bmatrix} \\ & \preceq \begin{bmatrix} \boldsymbol{R}_x & \boldsymbol{R}_{xs} \\ \boldsymbol{R}_{xs}^{\mathsf{H}} & \boldsymbol{R}_s \end{bmatrix} \\ & \preceq \begin{bmatrix} \hat{\boldsymbol{R}}_x & \hat{\boldsymbol{R}}_{xs} \\ \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}} & \hat{\boldsymbol{R}}_s \end{bmatrix} + \epsilon_0 \begin{bmatrix} \boldsymbol{F} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G} \end{bmatrix}, \end{aligned} \quad (27)$$

for some $\boldsymbol{F}, \boldsymbol{G} \succeq \boldsymbol{0}$ and $\epsilon_0 \geq 0$. $\qquad\square$

Definitions 1, 2, and 3 are introduced for the first time in this article. Another type of moment-based uncertainty set is popular in the literature, which we refer to as the multiplicative moment uncertainty set for differentiation.

Definition 4 (Multiplicative Moment Uncertainty Set [29]): The multiplicative moment uncertainty set of $\boldsymbol{R}$ is given as

$$\{\boldsymbol{R}|\ \theta_1\hat{\boldsymbol{R}} \preceq \boldsymbol{R} \preceq \theta_2\hat{\boldsymbol{R}}\} \quad (28)$$

for some $\theta_2 \geq 1 \geq \theta_1 \geq 0$. $\qquad\square$

The following corollary shows the distributionally robust linear beamformers associated with the various uncertainty sets in Definitions 1, 2, 3, and 4.

Corollary 1 (of Theorem 1): Consider the moment-based uncertainty sets in Definitions 1, 2, 3, and 4. The distributionally robust linear beamforming (21) is analytically solved by the corresponding upper bounds of $\boldsymbol{R}$. To be specific,

C1) Under Definition 1, the additive-moment distributionally robust (DR-AM) beamformer is

$$\begin{aligned} \boldsymbol{W}_{\mathrm{DR\text{-}AM}}^\star &= (\hat{\boldsymbol{R}}_{xs} + \epsilon_0\boldsymbol{E}_{xs})^{\mathsf{H}}(\hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{E}_x)^{-1} \\ &= (\hat{\boldsymbol{H}}\hat{\boldsymbol{R}}_s + \epsilon_0\boldsymbol{E}_{xs})^{\mathsf{H}} \cdot \\ & \quad [\hat{\boldsymbol{H}}\hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}} + \hat{\boldsymbol{R}}_v + \epsilon_0\boldsymbol{E}_x]^{-1}, \end{aligned} \quad (29)$$

where $\hat{\boldsymbol{H}}$, $\hat{\boldsymbol{R}}_s$, and $\hat{\boldsymbol{R}}_v$ denote the estimates of $\boldsymbol{H}$, $\boldsymbol{R}_s$, and $\boldsymbol{R}_v$, respectively.

C2) Under Definition 2, the diagonal-loading distributionally robust (DR-DL) beamformer is

$$\begin{aligned} \boldsymbol{W}_{\mathrm{DR\text{-}DL}}^\star &= \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}[\hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{I}_N]^{-1} \\ &= \hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}}[\hat{\boldsymbol{H}}\hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}} + \hat{\boldsymbol{R}}_v + \epsilon_0\boldsymbol{I}_N]^{-1}, \end{aligned} \quad (30)$$

which is also known as the loaded sample matrix inversion method [19], [14, Eq. (11)] and widely-used in the practice of wireless communications.

C3) Under Definition 3, the generalized diagonal-loading distributionally robust beamformer (DR-GDL) is

$$\begin{aligned} \boldsymbol{W}_{\mathrm{DR\text{-}GDL}}^\star &= \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}[\hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{F}]^{-1} \\ &= \hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}}[\hat{\boldsymbol{H}}\hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}} + \hat{\boldsymbol{R}}_v + \epsilon_0\boldsymbol{F}]^{-1}. \end{aligned} \quad (31)$$

C4) Under Definition 4, the multiplicative-moment (MM) distributionally robust beamformer is identical to the Wiener beamformer (12) at nominal values:

$$\begin{aligned} \boldsymbol{W}_{\mathrm{DR\text{-}MM}}^\star &= \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\hat{\boldsymbol{R}}_x^{-1} \\ &= \hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}}[\hat{\boldsymbol{H}}\hat{\boldsymbol{R}}_s\hat{\boldsymbol{H}}^{\mathsf{H}} + \hat{\boldsymbol{R}}_v]^{-1}. \end{aligned} \quad (32)$$

The corresponding estimation errors are simple to obtain. $\quad\square$

Corollary 1 implies that, in the sense of the same induced robust beamformers, the diagonal-loading uncertainty set (26) and the generalized diagonal-loading uncertainty set (27) are technically equivalent to the following trimmed versions.

Definition 5 (Trimmed Diagonal-Loading Uncertainty Sets): By setting $\boldsymbol{G} := \boldsymbol{0}$ in (27), in terms of $\boldsymbol{R}_x$, (27) reduces to the trimmed generalized diagonal-loading uncertainty set:

$$\{\boldsymbol{R}_x| \ \hat{\boldsymbol{R}}_x - \epsilon_0 \boldsymbol{F} \preceq \boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_x + \epsilon_0 \boldsymbol{F}, \ \boldsymbol{R}_x \succeq \boldsymbol{0}\}. \qquad (33)$$

The trimmed diagonal-loading uncertainty set

$$\{\boldsymbol{R}_x| \ \hat{\boldsymbol{R}}_x - \epsilon_0 \boldsymbol{I}_N \preceq \boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_x + \epsilon_0 \boldsymbol{I}_N, \ \boldsymbol{R}_x \succeq \boldsymbol{0}\}, \qquad (34)$$

is obtained by letting $\boldsymbol{F} := \boldsymbol{I}_N$. $\qquad\square$

The robust beamformers corresponding to the trimmed uncertainty sets (33) and (34) remain the same as defined in (31) and (30), respectively; cf. Theorem 1.

As we can see from Corollary 1, the primary benefit of using the moment-based uncertainty sets is the computational simplicity due to the availability of closed-form solutions. If the uncertainty sets are constructed using the Wasserstein distance $\sqrt{\text{Tr}[\boldsymbol{R} + \hat{\boldsymbol{R}} - 2(\hat{\boldsymbol{R}}^{1/2}\boldsymbol{R}\hat{\boldsymbol{R}}^{1/2})^{1/2}]} \leq \epsilon_0$ or the KL divergence $\frac{1}{2}[\text{Tr}[\hat{\boldsymbol{R}}^{-1}\boldsymbol{R} - \boldsymbol{I}_{N+M}] - \ln\det(\hat{\boldsymbol{R}}^{-1}\boldsymbol{R})] \leq \epsilon_0$ between $\mathcal{CN}(\boldsymbol{0}, \boldsymbol{R})$ and $\mathcal{CN}(\boldsymbol{0}, \hat{\boldsymbol{R}})$, the induced distributionally robust linear beamforming problems have no closed-form solutions, and therefore, are computationally prohibitive in practice. In addition, Corollary 1 suggests that the distributionally robust beamformer under the multiplicative moment uncertainty set (28) is the same as the nominal beamformer $\hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\hat{\boldsymbol{R}}_x^{-1}$, which essentially does not introduce robustness in wireless signal estimation; this is another motivation why we construct new moment-based uncertainty sets in Definitions 1, 2, and 3. However, we can modify the multiplicative moment uncertainty set in Definition 4 to achieve robustness.

Definition 6 (Modified Multiplicative Moment Uncertainty Set): The modified multiplicative moment uncertainty set of $\boldsymbol{R}$ is defined by the following constraint:

$$\begin{bmatrix} \theta_1\hat{\boldsymbol{R}}_x & \hat{\boldsymbol{R}}_{xs} \\ \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}} & \theta_1\hat{\boldsymbol{R}}_s \end{bmatrix} \preceq \begin{bmatrix} \boldsymbol{R}_x & \boldsymbol{R}_{xs} \\ \boldsymbol{R}_{xs}^{\mathsf{H}} & \boldsymbol{R}_s \end{bmatrix} \preceq \begin{bmatrix} \theta_2\hat{\boldsymbol{R}}_x & \hat{\boldsymbol{R}}_{xs} \\ \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}} & \theta_2\hat{\boldsymbol{R}}_s \end{bmatrix} \qquad (35)$$

for some $\theta_2 \geq 1 \geq \theta_1 \geq 0$ such that the left-most matrix is positive semi-definite. $\qquad\square$

The robust beamformer under the modified multiplicative moment uncertainty set (35) is

$$\boldsymbol{W}_{\text{DR-MMM}}^{\star} = \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}} \cdot [\theta_2\hat{\boldsymbol{R}}_x]^{-1}. \qquad (36)$$

In terms of the uncertainties of $\boldsymbol{R}_s$ and $\boldsymbol{R}_v$, Problem (21) can be explicitly written as

$$\begin{aligned} \max_{\boldsymbol{R}_s, \boldsymbol{R}_v} \quad & \text{Tr}\left[\boldsymbol{R}_s - \boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v)^{-1}\boldsymbol{H}\boldsymbol{R}_s\right] \\ \text{s.t.} \quad & d_1(\boldsymbol{R}_s, \hat{\boldsymbol{R}}_s) \leq \epsilon_1, \\ & d_2(\boldsymbol{R}_v, \hat{\boldsymbol{R}}_v) \leq \epsilon_2, \\ & \boldsymbol{R}_s \succeq \boldsymbol{0}, \ \boldsymbol{R}_v \succeq \boldsymbol{0}, \end{aligned}$$
$$(37)$$

for some similarity measures d_1 and d_2 and nonnegative scalars ϵ_1 and ϵ_2. For every given $(\boldsymbol{R}_s, \boldsymbol{R}_v)$, the associated

beamformer is given in (12). When the uncertainty in the channel matrix must be investigated, we can consider

$$\begin{aligned} \max_{\boldsymbol{H}} \quad & \text{Tr}\left[\boldsymbol{R}_s - \boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v)^{-1}\boldsymbol{H}\boldsymbol{R}_s\right] \\ \text{s.t.} \quad & d_3(\boldsymbol{H}, \hat{\boldsymbol{H}}) \leq \epsilon_3, \end{aligned}$$
$$(38)$$

which is not a semi-definite program. In addition, the gradient of the objective function with respect to $\boldsymbol{H}$ is complicated to obtain. Hence, practically, we should avoid directly attacking Problem (38); this can be done by directly considering the uncertainties of $\boldsymbol{R}_x$ and $\boldsymbol{R}_{xs}$ (i.e., $\boldsymbol{R}$) because the uncertainties of $\boldsymbol{R}_s$, $\boldsymbol{R}_v$, and $\boldsymbol{H}$ can be reflected in the uncertainties of $\boldsymbol{R}_x$ and $\boldsymbol{R}_{xs}$; cf. $\boldsymbol{R}_x = \boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v$ and $\boldsymbol{R}_{xs} = \boldsymbol{H}\boldsymbol{R}_s$.

In addition to Corollary 1, below we provide other concrete examples to further showcase the usefulness and applications of the distributionally robust beamforming formulations (21) and (37), where the trimmed uncertainty sets are employed.

Example 1 (Distributionally Robust Capon Beamforming): We consider a distributionally robust Capon beamforming problem under the trimmed uncertainty set (34):

$$\begin{aligned} \min_{\boldsymbol{W}} \max_{\boldsymbol{R}_x} \quad & \text{Tr}\left[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}} - 2\boldsymbol{R}_s + \boldsymbol{R}_s\right] \\ \text{s.t.} \quad & \boldsymbol{W}\boldsymbol{H} = \boldsymbol{I}_M, \\ & \hat{\boldsymbol{R}}_x - \epsilon_0\boldsymbol{I}_N \preceq \boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{I}_N, \\ & \boldsymbol{R}_x \succeq \boldsymbol{0}, \end{aligned}$$

which is equivalent, in the sense of the same solutions, to

$$\begin{aligned} \min_{\boldsymbol{W}} \max_{\boldsymbol{R}_x} \quad & \text{Tr}\left[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}}\right] \\ \text{s.t.} \quad & \boldsymbol{W}\boldsymbol{H} = \boldsymbol{I}_M, \\ & \hat{\boldsymbol{R}}_x - \epsilon_0\boldsymbol{I}_N \preceq \boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{I}_N, \\ & \boldsymbol{R}_x \succeq \boldsymbol{0}. \end{aligned}$$

According to Theorem 1, the above display is equivalent to

$$\begin{aligned} \min_{\boldsymbol{W}} \quad & \text{Tr}\left[\boldsymbol{W}\hat{\boldsymbol{R}}_x\boldsymbol{W}^{\mathsf{H}}\right] + \epsilon_0 \cdot \text{Tr}\left[\boldsymbol{W}\boldsymbol{W}^{\mathsf{H}}\right] \\ \text{s.t.} \quad & \boldsymbol{W}\boldsymbol{H} = \boldsymbol{I}_M. \end{aligned}$$

The above formulation is the squared-F-norm–regularized Capon beamformer [14, Eq. (10)] whose solution is

$$\begin{aligned} \boldsymbol{W}_{\text{DR-Capon}}^{\star} = [\boldsymbol{H}^{\mathsf{H}}(\hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{I}_N)^{-1}\boldsymbol{H}]^{-1} \cdot \\ \boldsymbol{H}^{\mathsf{H}}(\hat{\boldsymbol{R}}_x + \epsilon_0\boldsymbol{I}_N)^{-1}, \end{aligned} \qquad (39)$$

which is the diagonal-loading Capon beamformer. $\qquad\square$

Example 2 (Eigenvalue Thresholding): Suppose that $\hat{\boldsymbol{R}}_x$ admits the eigenvalues of $\text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ in descending order and the eigenvectors in $\boldsymbol{Q}$ (columns). Let $0 \leq \mu \leq 1$ be a shrinking coefficient. If we assume $\boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_{x,\text{thr}}$ where

$$\hat{\boldsymbol{R}}_{x,\text{thr}} := \\ \boldsymbol{Q} \begin{bmatrix} \lambda_1 & & & \\ & \max\{\mu\lambda_1, \lambda_2\} & & \\ & & \ddots & \\ & & & \max\{\mu\lambda_1, \lambda_N\} \end{bmatrix} \boldsymbol{Q}^{-1},$$
$$(40)$$

we have the distributionally robust beamformer

$$\boldsymbol{W}_{\text{DR-ET}}^{\star} = \boldsymbol{R}_{xs}^{\mathsf{H}}\hat{\boldsymbol{R}}_{x,\text{thr}}^{-1}, \qquad (41)$$

which is known as the eigenvalue thresholding method [20], [14, Eq. (12)]. $\qquad\square$

Example 3 (Distributionally Robust Beamforming for Uncertain $\boldsymbol{R}_s$ and $\boldsymbol{R}_v$): Consider Problem (37). Since the objective of (37) is increasing in both $\boldsymbol{R}_s$ and $\boldsymbol{R}_v$,[1] if

$$\hat{\boldsymbol{R}}_s - \epsilon_1 \boldsymbol{I}_M \preceq \boldsymbol{R}_s \preceq \hat{\boldsymbol{R}}_s + \epsilon_1 \boldsymbol{I}_M,$$

we have a distributionally robust beamformer

$$\begin{aligned} \boldsymbol{W}_{\text{DR}}^\star &= (\hat{\boldsymbol{R}}_s + \epsilon_1 \boldsymbol{I}_M)\boldsymbol{H}^{\mathsf{H}}[\boldsymbol{H}(\hat{\boldsymbol{R}}_s + \epsilon_1 \boldsymbol{I}_M)\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v]^{-1} \\ &= (\hat{\boldsymbol{R}}_s + \epsilon_1 \boldsymbol{I}_M)\boldsymbol{H}^{\mathsf{H}}[\boldsymbol{H}\hat{\boldsymbol{R}}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v + \epsilon_1 \boldsymbol{H}\boldsymbol{H}^{\mathsf{H}}]^{-1}; \end{aligned} \tag{42}$$

if instead

$$\hat{\boldsymbol{R}}_s - \epsilon_1 \boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{H}^{\mathsf{H}})^{-2}\boldsymbol{H} \preceq \boldsymbol{R}_s \preceq \hat{\boldsymbol{R}}_s + \epsilon_1 \boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{H}^{\mathsf{H}})^{-2}\boldsymbol{H}, \tag{43}$$

we have

$$\begin{aligned} \boldsymbol{W}_{\text{DR}}^\star = [\hat{\boldsymbol{R}}_s\boldsymbol{H}^{\mathsf{H}} + \epsilon_1 \boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{H}^{\mathsf{H}})^{-1}] \cdot \\ [\boldsymbol{H}\hat{\boldsymbol{R}}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v + \epsilon_1 \boldsymbol{I}_N]^{-1}, \end{aligned} \tag{44}$$

which is a modified diagonal-loading beamformer. On the other hand, if

$$\hat{\boldsymbol{R}}_v - \epsilon_2 \boldsymbol{I}_N \preceq \boldsymbol{R}_v \preceq \hat{\boldsymbol{R}}_v + \epsilon_2 \boldsymbol{I}_N,$$

we have

$$\boldsymbol{W}_{\text{DR}}^\star = \boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}[\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \hat{\boldsymbol{R}}_v + \epsilon_2 \boldsymbol{I}_N]^{-1}, \tag{45}$$

which is also a diagonal-loading beamformer. $\square$

Motivated by Corollary 1 and Examples 1~3, as well as the trimmed uncertainty sets in Definition 5, we have the following important theorem, which justifies the popular ridge regression in machine learning.

Theorem 2 (Ridge Regression and Tikhonov Regularization): Consider a linear regression problem on $(\mathbf{x}, \mathbf{s})$, i.e.,

$$\mathbf{s} = \boldsymbol{W}\mathbf{x} + \mathbf{e},$$

where $\mathbf{e}$ denotes the error term and the distributionally robust estimator of $\boldsymbol{W}$, i.e.,

$$\min_{\boldsymbol{W} \in \mathbb{C}^{M \times N}} \max_{\mathbb{P}_{\mathbf{x},\mathbf{s}} \in \mathcal{U}_{\mathbf{x},\mathbf{s}}} \text{Tr}\,\mathbb{E}_{\mathbf{x},\mathbf{s}}[\boldsymbol{W}\mathbf{x} - \mathbf{s}][\boldsymbol{W}\mathbf{x} - \mathbf{s}]^{\mathsf{H}},$$

which can be particularized to (19). Supposing that the second-order moment of $\mathbf{x}$ is uncertain and quantified as

$$\hat{\boldsymbol{R}}_x - \epsilon_0 \boldsymbol{I}_N \preceq \boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_x + \epsilon_0 \boldsymbol{I}_N,$$

then the distributionally robust estimator of $\boldsymbol{W}$ becomes a ridge regression (i.e., squared-F-norm regularized) method

$$\min_{\boldsymbol{W}} \text{Tr}\left[\boldsymbol{W}\hat{\boldsymbol{R}}_x\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\hat{\boldsymbol{R}}_{xs} - \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \hat{\boldsymbol{R}}_s\right] + \epsilon_0 \text{Tr}\left[\boldsymbol{W}\boldsymbol{W}^{\mathsf{H}}\right].$$

The regularization term becomes $\text{Tr}\left[\boldsymbol{W}\boldsymbol{F}\boldsymbol{W}^{\mathsf{H}}\right]$, which is known as the Tikhonov regularizer, if

$$\hat{\boldsymbol{R}}_x - \epsilon_0 \boldsymbol{F} \preceq \boldsymbol{R}_x \preceq \hat{\boldsymbol{R}}_x + \epsilon_0 \boldsymbol{F}$$

for some $\boldsymbol{F} \succeq \mathbf{0}$.

Proof: This is due to Lemma 1 and Theorem 1. Just note that $\text{Tr}\left[\boldsymbol{W}(\hat{\boldsymbol{R}}_x + \epsilon_0 \boldsymbol{F})\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\hat{\boldsymbol{R}}_{xs} - \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \hat{\boldsymbol{R}}_s\right] = \text{Tr}\left[\boldsymbol{W}\hat{\boldsymbol{R}}_x\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\hat{\boldsymbol{R}}_{xs} - \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \hat{\boldsymbol{R}}_s\right] + \epsilon_0 \text{Tr}\left[\boldsymbol{W}\boldsymbol{F}\boldsymbol{W}^{\mathsf{H}}\right].$ This completes the proof. $\square$

---

[1]This claim can be routinely proven in analogy to Theorem 1 and a real-space case in [30, Thm. 1].

Note that in Theorem 2, the second-order moment of $\mathbf{s}$ is not considered because it does not influence the optimal solution of $\boldsymbol{W}$: i.e., the optimal solution of $\boldsymbol{W}$ does not depend on the value of $\boldsymbol{R}_s$. Theorem 2 gives a new theoretical interpretation of the popular ridge regression in machine learning from the perspective of distributional robustness against second-moment uncertainties of the feature vector $\mathbf{x}$; another interpretation of ridge regression from the perspective of distributional robustness under martingale constraints is identified in [31, Ex. 3.3]. When the uncertainty is quantified by the Wasserstein distance, a similar result can be seen in [32, Prop. 3], [33, Prop. 2], which however is not a ridge regression formulation because in [32, Prop. 3] and [33, Prop. 2], the loss function is square-rooted and the norm regularizer is not squared; cf. also [27, Rem. 18 and 19]. The corollary below justifies the rationale of any norm-regularized method.

Corollary 2: The following squared-norm-regularized beamforming formulation can combat the distributional uncertainty:

$$\min_{\boldsymbol{W}} \text{Tr}\left[\boldsymbol{W}\hat{\boldsymbol{R}}_x\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\hat{\boldsymbol{R}}_{xs} - \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \hat{\boldsymbol{R}}_s\right] + \lambda\|\boldsymbol{W}\|^2, \tag{46}$$

where $\|\cdot\|$ denotes any matrix norm. This is because all norms on $\mathbb{C}^{M \times N}$ are equivalent; hence, there exists some $\lambda \geq 0$ such that $\lambda\|\boldsymbol{W}\|^2 \geq \epsilon_0\|\boldsymbol{W}\|_F^2 = \epsilon_0 \text{Tr}\left[\boldsymbol{W}\boldsymbol{W}^{\mathsf{H}}\right]$. As a result, (46) can upper bound the ridge cost in Theorem 2. $\square$

Motivated by Theorem 2, the following corollary is immediate, which gives another interpretation of ridge regression and Tikhonov regularization from the perspective of data augmentation through data perturbation (cf. noise injection in image [34] and speech [35] processing).

Corollary 3 (Data Augmentation for Linear Regression): Consider a linear regression problem on $(\mathbf{x}, \mathbf{s})$ with data perturbation vectors $(\boldsymbol{\Delta}_x, \boldsymbol{\Delta}_s)$

$$(\mathbf{s} + \boldsymbol{\Delta}_s) = \boldsymbol{W}(\mathbf{x} + \boldsymbol{\Delta}_x) + \mathbf{e},$$

and the distributionally robust estimator of $\boldsymbol{W}$

$$\min_{\boldsymbol{W} \in \mathbb{C}^{M \times N}} \max_{\mathbb{P}_{\boldsymbol{\Delta}_x,\boldsymbol{\Delta}_s} \in \mathcal{U}_{\boldsymbol{\Delta}_x,\boldsymbol{\Delta}_s}} \text{Tr}\,\mathbb{E}_{(\mathbf{x},\mathbf{s}) \sim \hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}} \mathbb{E}_{\boldsymbol{\Delta}_x,\boldsymbol{\Delta}_s}\Big\{ \\ [\boldsymbol{W}(\mathbf{x} + \boldsymbol{\Delta}_x) - (\mathbf{s} + \boldsymbol{\Delta}_s)][\boldsymbol{W}(\mathbf{x} + \boldsymbol{\Delta}_x) - (\mathbf{s} + \boldsymbol{\Delta}_s)]^{\mathsf{H}}\Big\}.$$

Suppose that $\boldsymbol{\Delta}_x$ is uncorrelated with $\mathbf{x}$, with $\mathbf{s}$, and with $\boldsymbol{\Delta}_s$; in addition, $\boldsymbol{\Delta}_s$ is uncorrelated with $\mathbf{x}$. If the second-order moment of $\boldsymbol{\Delta}_x$ is upper bounded as $\mathbb{E}\boldsymbol{\Delta}_x\boldsymbol{\Delta}_x^{\mathsf{H}} \preceq \epsilon_0 \boldsymbol{I}_N$, then the distributionally robust estimator of $\boldsymbol{W}$ becomes a ridge regression (i.e., squared-F-norm regularized) method

$$\min_{\boldsymbol{W}} \text{Tr}\left[\boldsymbol{W}\hat{\boldsymbol{R}}_x\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\hat{\boldsymbol{R}}_{xs} - \hat{\boldsymbol{R}}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \hat{\boldsymbol{R}}_s\right] + \epsilon_0 \text{Tr}\left[\boldsymbol{W}\boldsymbol{W}^{\mathsf{H}}\right].$$

The regularization term becomes $\text{Tr}\left[\boldsymbol{W}\boldsymbol{F}\boldsymbol{W}^{\mathsf{H}}\right]$, which is known as the Tikhonov regularizer, if $\mathbb{E}\boldsymbol{\Delta}_x\boldsymbol{\Delta}_x^{\mathsf{H}} \preceq \epsilon_0 \boldsymbol{F}$, for some $\boldsymbol{F} \succeq \mathbf{0}$. $\square$

The second-order moment of $\boldsymbol{\Delta}_s$ is not considered in Corollary 3 as it does not influence the optimal value of $\boldsymbol{W}$.

B. Complex Uncertainty Sets

Below we remark on more general construction methods for the uncertainty set of $\boldsymbol{R}$ using the Wasserstein distance and the F-norm, beyond the moment-based methods in Definitions

1∼6. However, note that such complicated approaches are computationally prohibitive in practice when N or M is large.

1) Wasserstein Distributionally Robust Beamforming: We start with the Wasserstein distance:

$$\max_{\boldsymbol{R}} \quad \mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs}+\boldsymbol{R}_s\right]$$
$$\text{s.t.} \quad \mathrm{Tr}\left[\boldsymbol{R}+\hat{\boldsymbol{R}}-2(\hat{\boldsymbol{R}}^{1/2}\boldsymbol{R}\hat{\boldsymbol{R}}^{1/2})^{1/2}\right]\le\epsilon_0^2 \quad (47)$$
$$\boldsymbol{R}\succeq\boldsymbol{0},\ \boldsymbol{R}_x\succ\boldsymbol{0}.$$

The first constraint in the above display is a particularization of the Wasserstein distance between $\mathcal{CN}(\boldsymbol{0},\boldsymbol{R})$ and $\mathcal{CN}(\boldsymbol{0},\hat{\boldsymbol{R}})$.

Problem (47) is a nonlinear positive semi-definite program (P-SDP). However, we can give it a linear reformulation.

Proposition 1: Problem (47) can be equivalently reformulated into a linear P-SDP

$$\max_{\boldsymbol{R},\boldsymbol{V},\boldsymbol{U}} \quad \mathrm{Tr}[\boldsymbol{R}_s-\boldsymbol{V}]$$
$$\text{s.t.} \quad \begin{bmatrix} \boldsymbol{V} & \boldsymbol{R}_{xs}^{\mathsf{H}} \\ \boldsymbol{R}_{xs} & \boldsymbol{R}_x \end{bmatrix}\succeq\boldsymbol{0}$$
$$\mathrm{Tr}\left[\boldsymbol{R}+\hat{\boldsymbol{R}}-2\boldsymbol{U}\right]\le\epsilon_0^2 \quad (48)$$
$$\begin{bmatrix} \hat{\boldsymbol{R}}^{1/2}\boldsymbol{R}\hat{\boldsymbol{R}}^{1/2} & \boldsymbol{U} \\ \boldsymbol{U} & \boldsymbol{I}_{N+M} \end{bmatrix}\succeq\boldsymbol{0}$$
$$\boldsymbol{R}\succeq\boldsymbol{0},\ \boldsymbol{R}_x\succ\boldsymbol{0},\ \boldsymbol{V}\succeq\boldsymbol{0},\ \boldsymbol{U}\succeq\boldsymbol{0}.$$

Proof: This is by applying the Schur complement. □

Complex-valued linear P-SDP can be solved using, e.g., the YALMIP solver.[2]

Suppose that $\boldsymbol{R}^{\star}$ solves (48). The corresponding Wasserstein distributionally robust beamformer is given as

$$\boldsymbol{W}_{\text{DR-Wasserstein}}^{\star}=\boldsymbol{R}_{xs}^{\star\mathsf{H}}\boldsymbol{R}_x^{\star-1}. \quad (49)$$

Next, we separately investigate the uncertainties in $\hat{\boldsymbol{R}}_s$ and $\hat{\boldsymbol{R}}_v$. From (37), we have

$$\max_{\boldsymbol{R}_s,\boldsymbol{R}_v} \quad \mathrm{Tr}\left[\boldsymbol{R}_s-\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}+\boldsymbol{R}_v)^{-1}\boldsymbol{H}\boldsymbol{R}_s\right]$$
$$\text{s.t.} \quad \mathrm{Tr}\left[\boldsymbol{R}_s+\hat{\boldsymbol{R}}_s-2(\hat{\boldsymbol{R}}_s^{1/2}\boldsymbol{R}_s\hat{\boldsymbol{R}}_s^{1/2})^{1/2}\right]\le\epsilon_1^2$$
$$\mathrm{Tr}\left[\boldsymbol{R}_v+\hat{\boldsymbol{R}}_v-2(\hat{\boldsymbol{R}}_v^{1/2}\boldsymbol{R}_v\hat{\boldsymbol{R}}_v^{1/2})^{1/2}\right]\le\epsilon_2^2$$
$$\boldsymbol{R}_s\succeq\boldsymbol{0},\ \boldsymbol{R}_v\succeq\boldsymbol{0}, \quad (50)$$

where we ignore the uncertainty of $\boldsymbol{H}$ for technical tractability. Problem (50) can be transformed into a linear P-SDP using a similar technique as in Proposition 1. One can just introduce an inequality $\boldsymbol{U}\succeq\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}(\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}+\boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{R}_s$ and the objective function will become $\mathrm{Tr}\left[\boldsymbol{R}_s-\boldsymbol{U}\right]$.

Suppose that $(\boldsymbol{R}_s^{\star},\boldsymbol{R}_v^{\star})$ solves (50). The corresponding Wasserstein distributionally robust beamformer is given as

$$\boldsymbol{W}_{\text{DR-Wasserstein-Individual}}^{\star}=\boldsymbol{R}_s^{\star}\boldsymbol{H}^{\mathsf{H}}[\boldsymbol{H}\boldsymbol{R}_s^{\star}\boldsymbol{H}^{\mathsf{H}}+\boldsymbol{R}_v^{\star}]^{-1}. \quad (51)$$

2) F-Norm Distributionally Robust Beamforming: Under the F-norm, we just need to replace the Wasserstein distance. To be specific, (47) becomes

$$\max_{\boldsymbol{R}} \quad \mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs}+\boldsymbol{R}_s\right]$$
$$\text{s.t.} \quad \mathrm{Tr}\left[(\boldsymbol{R}-\hat{\boldsymbol{R}})^{\mathsf{H}}(\boldsymbol{R}-\hat{\boldsymbol{R}})\right]\le\epsilon_0^2 \quad (52)$$
$$\boldsymbol{R}\succeq\boldsymbol{0},\ \boldsymbol{R}_x\succ\boldsymbol{0}.$$

The linear reformulation of the above display is given in the proposition below.

Proposition 2: The nonlinear P-SDP (52) can be equivalently reformulated into a linear P-SDP

$$\max_{\boldsymbol{R},\boldsymbol{V},\boldsymbol{U}} \quad \mathrm{Tr}[\boldsymbol{R}_s-\boldsymbol{V}]$$
$$\text{s.t.} \quad \begin{bmatrix} \boldsymbol{V} & \boldsymbol{R}_{xs}^{\mathsf{H}} \\ \boldsymbol{R}_{xs} & \boldsymbol{R}_x \end{bmatrix}\succeq\boldsymbol{0}$$
$$\mathrm{Tr}\left[\boldsymbol{U}\right]\le\epsilon_0^2, \quad (53)$$
$$\begin{bmatrix} \boldsymbol{U} & (\boldsymbol{R}-\hat{\boldsymbol{R}})^{\mathsf{H}} \\ (\boldsymbol{R}-\hat{\boldsymbol{R}}) & \boldsymbol{I}_{N+M} \end{bmatrix}\succeq\boldsymbol{0},$$
$$\boldsymbol{R}\succeq\boldsymbol{0},\ \boldsymbol{R}_x\succ\boldsymbol{0},\ \boldsymbol{V}\succeq\boldsymbol{0},\ \boldsymbol{U}\succeq\boldsymbol{0}.$$

Proof: This is by applying the Schur complement. □

C. Multi-Frame Case: Dynamic Channel Evolution

Each frame contains a pilot block used for beamformer design. Although the channel state information (CSI) may change from one frame to another, the CSI between the two consecutive frames is highly correlated. This correlation can benefit beamformer design across multiple frames. Suppose that $\{(\boldsymbol{s}_1,\boldsymbol{x}_1),(\boldsymbol{s}_2,\boldsymbol{x}_2),\ldots,(\boldsymbol{s}_L,\boldsymbol{x}_L)\}$ is the training data in the current frame and $\{(\boldsymbol{s}_1',\boldsymbol{x}_1'),(\boldsymbol{s}_2',\boldsymbol{x}_2'),\ldots,(\boldsymbol{s}_L',\boldsymbol{x}_L')\}$ is the history data in the immediately preceding frame. In such a case, the distributional difference between $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ and $\hat{\mathbb{P}}_{\mathbf{x}',\mathbf{s}'}$ is upper bounded, that is, $d(\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}},\ \hat{\mathbb{P}}_{\mathbf{x}',\mathbf{s}'})\le\epsilon'$ for some proper distance d and a real number $\epsilon'\ge0$ where $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}:=\frac{1}{L}\sum_{i=1}^{L}\delta_{(\boldsymbol{x}_i,\boldsymbol{s}_i)}$ and $\hat{\mathbb{P}}_{\mathbf{x}',\mathbf{s}'}:=\frac{1}{L}\sum_{i=1}^{L}\delta_{(\boldsymbol{x}_i',\boldsymbol{s}_i')}$.

Since a beamformer $\boldsymbol{W}=\mathcal{F}(\mathbb{P}_{\mathbf{x},\mathbf{s}})$ is a continuous functional $\mathcal{F}(\cdot)$ of data distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$, cf. (10), we have $\|\boldsymbol{W}-\boldsymbol{W}'\|_F=\|\mathcal{F}(\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}})-\mathcal{F}(\hat{\mathbb{P}}_{\mathbf{x}',\mathbf{s}'})\|_F\le C\cdot d(\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}},\ \hat{\mathbb{P}}_{\mathbf{x}',\mathbf{s}'})\le\epsilon$ for some positive constant $C\ge0$ and upper bound $\epsilon\ge0$ where $\boldsymbol{W}'$ is the beamformer associated with $\hat{\mathbb{P}}_{\mathbf{x}',\mathbf{s}'}$ in the previous frame. Therefore, the beamforming problem (11) becomes a constrained problem

$$\min_{\boldsymbol{W}} \quad \mathrm{Tr}\left[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}}-\boldsymbol{W}\boldsymbol{R}_{xs}-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}}+\boldsymbol{R}_s\right]$$
$$\text{s.t.} \quad \mathrm{Tr}[\boldsymbol{W}-\boldsymbol{W}'][\boldsymbol{W}-\boldsymbol{W}']^{\mathsf{H}}\le\epsilon^2.$$

By the Lagrange duality theory, it is equivalent to

$$\min_{\boldsymbol{W}}\mathrm{Tr}\left[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}}-\boldsymbol{W}\boldsymbol{R}_{xs}-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}}+\boldsymbol{R}_s\right]+$$
$$\lambda\cdot\mathrm{Tr}[\boldsymbol{W}-\boldsymbol{W}'][\boldsymbol{W}-\boldsymbol{W}']^{\mathsf{H}}$$
$$=\min_{\boldsymbol{W}}\mathrm{Tr}\left[\boldsymbol{W}(\boldsymbol{R}_x+\lambda\boldsymbol{I}_N)\boldsymbol{W}^{\mathsf{H}}-\boldsymbol{W}(\boldsymbol{R}_{xs}+\lambda\boldsymbol{W}'^{\mathsf{H}})-\right.$$
$$\left.(\boldsymbol{R}_{xs}+\lambda\boldsymbol{W}'^{\mathsf{H}})^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}}+(\boldsymbol{R}_s+\lambda\boldsymbol{W}'\boldsymbol{W}'^{\mathsf{H}})\right], \quad (54)$$

for some $\lambda\ge0$. As a result, we have the Wiener beamformer for the multi-frame case, where we can treat $\boldsymbol{W}'$ as a ***prior knowledge*** of $\boldsymbol{W}$.

Claim 1 (Multi-Frame Beamforming): The Wiener beamformer for the multi-frame case is given by

$$\boldsymbol{W}_{\text{Wiener-MF}}^{\star}=[\boldsymbol{R}_{xs}+\lambda\boldsymbol{W}'^{\mathsf{H}}]^{\mathsf{H}}[\boldsymbol{R}_x+\lambda\boldsymbol{I}_N]^{-1}$$
$$=[\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}+\lambda\boldsymbol{W}'][\boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}}+\boldsymbol{R}_v+\lambda\boldsymbol{I}_N]^{-1}, \quad (55)$$

where $\lambda\ge0$ is a tuning parameter to control the similarity between $\boldsymbol{W}$ and $\boldsymbol{W}'$. Specifically, if λ is large, $\boldsymbol{W}$ must be close to $\boldsymbol{W}'$; if λ is small, $\boldsymbol{W}$ can be far away from $\boldsymbol{W}'$. □

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2025.3582082

9

With the result in Claim 1, (21) becomes

$$
\max_{\boldsymbol{R}} \quad \mathrm{Tr}\big[-(\boldsymbol{R}_{xs}+\lambda\boldsymbol{W}'^{\mathsf{H}})^{\mathsf{H}}\cdot(\boldsymbol{R}_x+\lambda\boldsymbol{I}_N)^{-1}\cdot
$$
$$
(\boldsymbol{R}_{xs}+\lambda\boldsymbol{W}'^{\mathsf{H}})+(\boldsymbol{R}_s+\lambda\boldsymbol{W}'\boldsymbol{W}'^{\mathsf{H}})\big]
$$
$$
\text{s.t.} \quad d_0(\boldsymbol{R},\hat{\boldsymbol{R}})\le\epsilon_0,
$$
$$
\boldsymbol{R}\succeq\boldsymbol{0},
$$
(56)

whose objective function is monotonically increasing in $\boldsymbol{R}$.

The remaining distributional robustness modeling and analyses against the uncertainties in $\boldsymbol{R}$ are technically straightforward, and therefore, we omit them here. Upon using the diagonal-loading method on $\boldsymbol{R}$, a distributionally robust beamformer for the multi-frame case is

$$
\boldsymbol{W}^{\star}_{\text{DR-Wiener-MF}}=[\hat{\boldsymbol{R}}_{xs}+\lambda\boldsymbol{W}'^{\mathsf{H}}]^{\mathsf{H}}\cdot[\hat{\boldsymbol{R}}_x+\lambda\boldsymbol{I}_N+\epsilon_0\boldsymbol{I}_N]^{-1},
$$

where ϵ_0 is an uncertainty quantification parameter for $\boldsymbol{R}$.

V. DISTRIBUTIONALLY ROBUST NONLINEAR ESTIMATION

For the convenience of the technical treatment, we study the estimation problem in real spaces. Nonlinear estimators, which are suitable for non-Gaussian $\mathbb{P}_{\mathbf{x},\mathbf{s}}$, are to be limited in reproducing kernel Hilbert spaces and feedforward multi-layer neural network function spaces.

A. Reproducing Kernel Hilbert Spaces

1) General Framework and Concrete Examples: As a standard treatment in machine learning, we use the partial pilot data $\{\underline{\boldsymbol{x}}_1,\underline{\boldsymbol{x}}_2,\ldots,\underline{\boldsymbol{x}}_L\}$ to construct the reproducing kernel Hilbert spaces, and use the whole pilot data $\{(\underline{\boldsymbol{x}}_1,\underline{\boldsymbol{s}}_1),(\underline{\boldsymbol{x}}_2,\underline{\boldsymbol{s}}_2),\ldots,(\underline{\boldsymbol{x}}_L,\underline{\boldsymbol{s}}_L)\}$ to train the optimal estimator in an RKHS.

With the $\boldsymbol{W}$-linear representation of $\phi(\cdot)$ in (2), i.e., $\phi(\cdot)=\boldsymbol{W}\boldsymbol{\varphi}(\cdot)$, the distributionally robust estimation problem (17) becomes

$$
\min_{\boldsymbol{W}\in\mathbb{R}^{2M\times L}}\max_{\mathbb{P}_{\mathbf{x},\mathbf{s}}\in\mathcal{U}_{\mathbf{x},\mathbf{s}}}\mathrm{Tr}\,\mathbb{E}_{\mathbf{x},\mathbf{s}}[\boldsymbol{W}\cdot\boldsymbol{\varphi}(\underline{\mathbf{x}})-\underline{\mathbf{s}}][\boldsymbol{W}\cdot\boldsymbol{\varphi}(\underline{\mathbf{x}})-\underline{\mathbf{s}}]^{\mathsf{T}}.
$$
(57)

The proposition below reformulates and solves (57).

Proposition 3: Let $\boldsymbol{K}$ denote the *kernel matrix* associated with the kernel function $\ker(\cdot,\cdot)$ whose (i,j)-entry is defined as

$$
\boldsymbol{K}_{i,j}:=\ker(\underline{\boldsymbol{x}}_i,\underline{\boldsymbol{x}}_j),\quad\forall i,j\in[L].
$$

Let $\underline{\mathbf{z}}:=\boldsymbol{\varphi}(\underline{\mathbf{x}})$. Then, the distributionally robust $\underline{\mathbf{x}}$-nonlinear estimation problem (57) can be rewritten as a distributionally robust $\underline{\mathbf{z}}$-linear beamforming problem as

$$
\min_{\boldsymbol{W}}\max_{\boldsymbol{R}_{\underline{z}},\boldsymbol{R}_{\underline{z}s},\boldsymbol{R}_{\underline{s}}}\mathrm{Tr}\big[\boldsymbol{W}\boldsymbol{R}_{\underline{z}}\boldsymbol{W}^{\mathsf{T}}-\boldsymbol{W}\boldsymbol{R}_{\underline{z}s}-\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}\boldsymbol{W}^{\mathsf{T}}+\boldsymbol{R}_{\underline{s}}\big]
$$
$$
\text{s.t.}\quad d_0\left(\begin{bmatrix}\boldsymbol{R}_{\underline{z}}&\boldsymbol{R}_{\underline{z}s}\\\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}&\boldsymbol{R}_{\underline{s}}\end{bmatrix},\begin{bmatrix}\hat{\boldsymbol{R}}_{\underline{z}}&\hat{\boldsymbol{R}}_{\underline{z}s}\\\hat{\boldsymbol{R}}_{\underline{z}s}^{\mathsf{T}}&\hat{\boldsymbol{R}}_{\underline{s}}\end{bmatrix}\right)\le\epsilon_0,
$$
$$
\begin{bmatrix}\boldsymbol{R}_{\underline{z}}&\boldsymbol{R}_{\underline{z}s}\\\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}&\boldsymbol{R}_{\underline{s}}\end{bmatrix}\succeq\boldsymbol{0},
$$
(58)

where $\hat{\boldsymbol{R}}_{\underline{z}}=\frac{1}{L}\boldsymbol{K}^2$, $\hat{\boldsymbol{R}}_{\underline{z}s}=\frac{1}{L}\boldsymbol{K}\underline{\boldsymbol{S}}^{\mathsf{T}}$, $\hat{\boldsymbol{R}}_{\underline{s}}=\frac{1}{L}\underline{\boldsymbol{S}}\underline{\boldsymbol{S}}^{\mathsf{T}}$, and $\underline{\boldsymbol{S}}:=[\mathrm{Re}\,\boldsymbol{S};\ \mathrm{Im}\,\boldsymbol{S}]=[\underline{\boldsymbol{s}}_1,\underline{\boldsymbol{s}}_2,\ldots,\underline{\boldsymbol{s}}_L]$. In addition, the strong min-max property holds for (58): i.e., the order of $\min$ and $\max$ can be exchanged provided that the first constraint is compact

convex. As a result, given every pair of $(\boldsymbol{R}_{\underline{z}},\boldsymbol{R}_{\underline{z}s},\boldsymbol{R}_{\underline{s}})$, the optimal Wiener beamformer is

$$
\boldsymbol{W}^{\star}_{\text{RKHS}}=\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}\cdot\boldsymbol{R}_{\underline{z}}^{-1}
$$
(59)

which transforms (58) to

$$
\max_{\boldsymbol{R}_{\underline{z}},\boldsymbol{R}_{\underline{z}s},\boldsymbol{R}_{\underline{s}}}\mathrm{Tr}\big[-\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}\boldsymbol{R}_{\underline{z}}^{-1}\boldsymbol{R}_{\underline{z}s}+\boldsymbol{R}_{\underline{s}}\big]
$$
$$
\text{s.t.}\quad d_0\left(\begin{bmatrix}\boldsymbol{R}_{\underline{z}}&\boldsymbol{R}_{\underline{z}s}\\\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}&\boldsymbol{R}_{\underline{s}}\end{bmatrix},\begin{bmatrix}\hat{\boldsymbol{R}}_{\underline{z}}&\hat{\boldsymbol{R}}_{\underline{z}s}\\\hat{\boldsymbol{R}}_{\underline{z}s}^{\mathsf{T}}&\hat{\boldsymbol{R}}_{\underline{s}}\end{bmatrix}\right)\le\epsilon_0,
$$
$$
\begin{bmatrix}\boldsymbol{R}_{\underline{z}}&\boldsymbol{R}_{\underline{z}s}\\\boldsymbol{R}_{\underline{z}s}^{\mathsf{T}}&\boldsymbol{R}_{\underline{s}}\end{bmatrix}\succeq\boldsymbol{0},\quad\boldsymbol{R}_{\underline{z}}\succ\boldsymbol{0}.
$$
(60)

Proof: Treating $[\underline{\mathbf{z}};\underline{\mathbf{s}}]$ as, or approximating $[\underline{\mathbf{z}};\underline{\mathbf{s}}]$ using, a joint Gaussian random vector due to the linear estimation relation $\hat{\underline{\mathbf{s}}}=\boldsymbol{W}\underline{\mathbf{z}}$ in RKHS [cf. (57)], then the results in Lemma 1 apply. For details, see Appendix G. $\square$

In (58), d_0 defines a matrix similarity measure to quantify the uncertainty of the covariance matrix of $[\underline{\mathbf{z}};\underline{\mathbf{s}}]$, and $\epsilon_0\ge 0$ quantifies the uncertainty level. Proposition 3 reveals the benefit of the kernel trick (2), that is, the possibility to represent a nonlinear estimation problem as a linear one.

The claim below summarizes the solution of (17) in the RKHS induced by the kernel function $\ker(\cdot,\cdot)$.

Claim 2: Suppose that $(\boldsymbol{R}_{\underline{z}}^{\star},\boldsymbol{R}_{\underline{z}s}^{\star},\boldsymbol{R}_{\underline{s}}^{\star})$ solves (60). Then the optimal estimator of (17) in the RKHS induced by $\ker(\cdot,\cdot)$ is given by

$$
\phi^{\star}(\mathbf{x})=\boldsymbol{\Gamma}_M\cdot\boldsymbol{R}_{\underline{z}s}^{\star\mathsf{T}}\cdot\boldsymbol{R}_{\underline{z}}^{\star-1}\cdot\boldsymbol{\varphi}(\underline{\mathbf{x}}),
$$
(61)

where $\underline{\mathbf{x}}=[\mathrm{Re}\,\mathbf{x};\ \mathrm{Im}\,\mathbf{x}]$ is the real-space representation of $\mathbf{x}$, $\boldsymbol{\Gamma}_M:=[\boldsymbol{I}_M,\boldsymbol{J}_M]$ is defined in Subsection I-B, and

$$
\boldsymbol{\varphi}(\underline{\mathbf{x}}):=\begin{bmatrix}\ker(\underline{\mathbf{x}},\underline{\boldsymbol{x}}_1)\\\ker(\underline{\mathbf{x}},\underline{\boldsymbol{x}}_2)\\\vdots\\\ker(\underline{\mathbf{x}},\underline{\boldsymbol{x}}_L)\end{bmatrix}.
$$

In addition, the corresponding worst-case estimation error covariance is

$$
\boldsymbol{\Gamma}_M\cdot\big[-\boldsymbol{R}_{\underline{z}s}^{\star\mathsf{T}}\boldsymbol{R}_{\underline{z}}^{\star-1}\boldsymbol{R}_{\underline{z}s}^{\star}+\boldsymbol{R}_{\underline{s}}^{\star}\big]\cdot\boldsymbol{\Gamma}_M^{\mathsf{H}},
$$
(62)

which upper bounds the true estimation error covariance. $\square$

Concrete examples of Claim 2 are given as follows.

Example 4 (Kernelized Diagonal Loading): By using the trimmed diagonal-loading uncertainty set for $\boldsymbol{R}_{\underline{z}}$, i.e.,

$$
\hat{\boldsymbol{R}}_{\underline{z}}-\epsilon_0\boldsymbol{I}_L\preceq\boldsymbol{R}_{\underline{z}}\preceq\hat{\boldsymbol{R}}_{\underline{z}}+\epsilon_0\boldsymbol{I}_L,
$$

we have the kernelized diagonal loading method

$$
\phi^{\star}(\mathbf{x})=\boldsymbol{\Gamma}_M\cdot\frac{1}{L}\underline{\boldsymbol{S}}\boldsymbol{K}\cdot\left(\frac{1}{L}\boldsymbol{K}^2+\epsilon_0\boldsymbol{I}_L\right)^{-1}\cdot\boldsymbol{\varphi}(\underline{\mathbf{x}}),
$$
(63)

which is obtained at the upper bound of $\boldsymbol{R}_{\underline{z}}$. Furthermore, in this case, the distributionally robust formulation (57) is equivalent to a squared-F-norm-regularized formulation

$$
\min_{\boldsymbol{W}}\mathrm{Tr}\,\mathbb{E}_{(\underline{\mathbf{x}},\underline{\mathbf{s}})\sim\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}}[\boldsymbol{W}\cdot\boldsymbol{\varphi}(\underline{\mathbf{x}})-\underline{\mathbf{s}}][\boldsymbol{W}\cdot\boldsymbol{\varphi}(\underline{\mathbf{x}})-\underline{\mathbf{s}}]^{\mathsf{T}}+
$$
$$
\epsilon_0\cdot\mathrm{Tr}[\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}],
$$
(64)

which can be proven by replacing $\boldsymbol{R}_{\underline{z}}$ in (58) with its upper bound. $\square$

Example 5 (Kernelized Eigenvalue Thresholding): The kernelized eigenvalue thresholding method can be designed in analogy to Example 2. The two key steps are to obtain the eigenvalue decomposition of $\hat{\boldsymbol{R}}_{\underline{z}} = \boldsymbol{K}^2/L$ and then lift the eigenvalues; cf. (40). □

In addition, Example 4 motivates the following important theorem for statistical machine learning.

Theorem 3 (Kernel Ridge Regression and Kernel Tikhonov Regularization): Consider the nonlinear regression problem

$$\mathbf{s} = \phi(\mathbf{x}) + \mathbf{e},$$

and the distributionally robust estimator of $\phi(\underline{\mathbf{x}}) = \boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}})$ in the RKHS induced by the kernel function $\ker(\cdot, \cdot)$, i.e.,

$$\min_{\boldsymbol{W} \in \mathbb{R}^{2M \times L}} \max_{\mathbb{P}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}} \in \mathcal{U}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}}} \operatorname{Tr} \mathbb{E}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}} [\boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}][\boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}]^{\mathsf{T}}.$$

Supposing that only the second-order moment of $\underline{\mathbf{z}} := \boldsymbol{\varphi}(\underline{\mathbf{x}})$ is uncertain and quantified as

$$\hat{\boldsymbol{R}}_{\underline{z}} - \epsilon_0 \boldsymbol{I}_L \preceq \boldsymbol{R}_{\underline{z}} \preceq \hat{\boldsymbol{R}}_{\underline{z}} + \epsilon_0 \boldsymbol{I}_L,$$

then the distributionally robust estimator of $\boldsymbol{W}$ becomes a kernel ridge regression method (64). The regularization term in (64) becomes the Tikhonov regularizer $\operatorname{Tr}[\boldsymbol{W} \boldsymbol{F} \boldsymbol{W}^{\mathsf{T}}]$ if

$$\hat{\boldsymbol{R}}_{\underline{z}} - \epsilon_0 \boldsymbol{F} \preceq \boldsymbol{R}_{\underline{z}} \preceq \hat{\boldsymbol{R}}_{\underline{z}} + \epsilon_0 \boldsymbol{F}$$

for some $\boldsymbol{F} \succeq \mathbf{0}$.

Proof: See Example 4; cf. Theorem 2. □

Theorem 3 gives the kernel ridge regression an interpretation of distributional robustness. The usual choice of $\boldsymbol{F}$ in Theorem 3 is the L-divided kernel matrix $\boldsymbol{K}/L$; see, e.g., [36, Eq. (4)], [24, Eqs. (15.110) and (15.113)]. As a result, from (63), we have

$$\phi^{\star}(\mathbf{x}) = \boldsymbol{\Gamma}_M \cdot \underline{\boldsymbol{S}} \cdot (\boldsymbol{K} + \epsilon_0 \boldsymbol{I}_L)^{-1} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}), \qquad (65)$$

which is another type of kernel ridge regression (i.e., a new kernelized diagonal-loading method).

In analogy to Corollary 2, the following corollary motivated from (64) is immediate.

Corollary 4: The following squared-norm-regularized method in RKHSs can combat the distributional uncertainty:

$$\min_{\boldsymbol{W}} \operatorname{Tr} \mathbb{E}_{(\underline{\mathbf{x}}, \underline{\mathbf{s}}) \sim \hat{\mathbb{P}}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}}} [\boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}][\boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}]^{\mathsf{T}} + \\ \lambda \cdot \|\boldsymbol{W}\|^2, \qquad (66)$$

for any matrix norm $\|\cdot\|$; cf. Corollary 2. □

Moreover, in analogy to Corollary 3, the following corollary is immediate.

Corollary 5 (Data Augmentation for Kernel Regression): Consider the nonlinear regression problem in Theorem 3. Its data-perturbed counterpart can be constructed by taking into account the data perturbation vectors $(\boldsymbol{\Delta}_{\underline{s}}, \boldsymbol{\Delta}_{\underline{z}})$. Suppose that $\boldsymbol{\Delta}_{\underline{z}}$ is uncorrelated with $\underline{\mathbf{z}}$, with $\underline{\mathbf{s}}$, and with $\boldsymbol{\Delta}_{\underline{s}}$; in addition, $\boldsymbol{\Delta}_{\underline{s}}$ is uncorrelated with $\underline{\mathbf{z}}$. If the second-order moment of $\boldsymbol{\Delta}_{\underline{z}}$ is upper bounded by $\epsilon_0 \boldsymbol{I}_L$, then the distributionally robust estimator of $\boldsymbol{W}$ becomes a kernel ridge regression (i.e., squared-F-norm regularized) method (64). The regularization term becomes $\operatorname{Tr}[\boldsymbol{W} \boldsymbol{F} \boldsymbol{W}^{\mathsf{H}}]$, which is known as the Tikhonov regularizer, if the second-order moment of $\boldsymbol{\Delta}_{\underline{z}}$ is upper bounded by $\epsilon_0 \boldsymbol{F}$ for some $\boldsymbol{F} \succeq \mathbf{0}$. □

General uncertainty sets using the Wasserstein distance or the F-norm, beyond the diagonal ϵ_0-perturbation (cf. Example 4), can be straightforwardly employed and the distributional robustness modeling and analyses remain routine; cf. Subsection IV-B. Hence, we omit them here. However, such complicated approaches are computationally prohibitive in practice when L or M is large.

2) Multi-Frame Case: Dynamic Channel Evolution: As in (54), the multi-frame formulation in RKHSs is

$$\min_{\boldsymbol{W} \in \mathbb{R}^{2M \times L}} \operatorname{Tr} \mathbb{E}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}} [\boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}][\boldsymbol{W} \cdot \boldsymbol{\varphi}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}]^{\mathsf{T}} + \\ \lambda \cdot \operatorname{Tr}[\boldsymbol{W} - \boldsymbol{W}'][\boldsymbol{W} - \boldsymbol{W}']^{\mathsf{T}}, \qquad (67)$$

where $\boldsymbol{W}'$ denotes the beamformer in the immediately preceding frame and serves as a ***prior knowledge*** of $\boldsymbol{W}$.

Claim 3 (Multi-Frame Estimation in RHKS): The solution to (67) is given by (cf. (59))

$$\begin{aligned} \boldsymbol{W}_{\text{RKHS-MF}}^{\star} &= [\boldsymbol{R}_{\underline{z}s} + \lambda \boldsymbol{W}'^{\mathsf{T}}]^{\mathsf{T}} [\boldsymbol{R}_{\underline{z}} + \lambda \boldsymbol{I}_L]^{-1} \\ &= \left(\tfrac{1}{L} \underline{\boldsymbol{S}} \boldsymbol{K} + \lambda \boldsymbol{W}'\right) \cdot \left(\tfrac{1}{L} \boldsymbol{K}^2 + \lambda \boldsymbol{I}_L\right)^{-1}, \end{aligned} \qquad (68)$$

where $\lambda \geq 0$ is a tuning parameter to control the similarity between $\boldsymbol{W}$ and $\boldsymbol{W}'$; cf. Claim 1. □

The remaining distributional robustness modeling and analyses on (67) against the uncertainties in $\hat{\boldsymbol{R}}_{\underline{z}}$, $\hat{\boldsymbol{R}}_{\underline{xz}}$, and $\hat{\boldsymbol{R}}_{\underline{s}}$ are technically straightforward; cf. Subsection IV-C. Therefore, we omit them here.

B. Neural Networks

With the $\boldsymbol{W}$-parameterization $\phi_{\boldsymbol{W}_{[R]}}(\underline{\mathbf{x}})$ of $\phi(\underline{\mathbf{x}})$ in feedforward multi-layer neural networks, i.e., (4), the distributionally robust estimation problem (17) becomes

$$\min_{\boldsymbol{W}_{[R]}} \max_{\mathbb{P}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}} \in \mathcal{U}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}}} \operatorname{Tr} \mathbb{E}_{\underline{\mathbf{x}}, \underline{\mathbf{s}}} [\phi_{\boldsymbol{W}_{[R]}}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}][\phi_{\boldsymbol{W}_{[R]}}(\underline{\mathbf{x}}) - \underline{\mathbf{s}}]^{\mathsf{T}}, \quad (69)$$

where $\boldsymbol{W}_{[R]} := \{\boldsymbol{W}_1, \boldsymbol{W}_2, \dots, \boldsymbol{W}_R\}$ and $\phi_{\boldsymbol{W}_{[R]}}(\underline{\mathbf{x}})$ is defined in (4). Problem (69) is highly nonlinear in both argument $\underline{\mathbf{x}}$ and parameter $\boldsymbol{W}_{[R]}$, which is different from the case in reproducing kernel Hilbert spaces where the $\boldsymbol{W}$-linearization features. Hence, problem (69) is too complicated to solve to global optimality. According to [27, Cor. 33], under several technical conditions (plus the boundedness of the feasible region of $\boldsymbol{W}_{[R]}$), (69) is upper bounded by a spectral-norm-regularized empirical risk minimization problem

$$\min_{\boldsymbol{W}_{[R]}} \frac{1}{L} \sum_{i=1}^{L} \operatorname{Tr}[\phi_{\boldsymbol{W}_{[R]}}(\underline{\boldsymbol{x}}_i) - \underline{\boldsymbol{s}}_i][\phi_{\boldsymbol{W}_{[R]}}(\underline{\boldsymbol{x}}_i) - \underline{\boldsymbol{s}}_i]^{\mathsf{T}} + \\ \lambda' \cdot \sum_{r=1}^{R} \|\boldsymbol{W}_r\|_2, \qquad (70)$$

for some regularization coefficient $\lambda' \geq 0$, where $\|\cdot\|_2$ denotes the spectral norm of a matrix (i.e., the induced 2-norm). Eq. (70) rigorously justifies the popular norm regularization method in training neural networks: By diminishing the upper bound (70) of (69), the true error in (69) can be controlled from above. The regularized ERM problem (70) is reminiscent of the ridge regression and the kernel ridge regression methods in Theorems 2 and 3 for distributional robustness in linear regression and RKHS linear regression, respectively. Supposing

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2025.3582082

11

that $\boldsymbol{W}_{[R]}^{\star}$ is an (approximated, or sub-optimal) solution[3] of (70), then the distributionally robust optimal estimator of the transmitted signal $\mathbf{s}$ can be obtained as

$$\hat{\mathbf{s}} = \boldsymbol{\Gamma}_M \cdot \boldsymbol{\phi}_{\boldsymbol{W}_{[R]}^{\star}}(\underline{\mathbf{x}}).$$

Therefore, in training a neural network for wireless signal estimation, it is recommended to apply the norm regularization methods. Since norms on real spaces are equivalent, (70) can be further upper bounded by

$$\min_{\boldsymbol{W}_{[R]}} \frac{1}{L} \sum_{i=1}^{L} \mathrm{Tr}[\boldsymbol{\phi}_{\boldsymbol{W}_{[R]}}(\underline{\boldsymbol{x}}_i) - \underline{\boldsymbol{s}}_i][\boldsymbol{\phi}_{\boldsymbol{W}_{[R]}}(\underline{\boldsymbol{x}}_i) - \underline{\boldsymbol{s}}_i]^{\mathsf{T}} + \\ \lambda \cdot \sum_{r=1}^{R} \|\boldsymbol{W}_r\|, \tag{71}$$

for any matrix norm $\|\cdot\|$ and some $\lambda \geq 0$; λ depends on λ' and $\|\cdot\|$. As a result, to achieve distributional robustness in training a neural network, any-norm-regularized learning method in (71) can be considered.

VI. Experiments

We consider a point-to-point multiple-input-multiple-output (MIMO) wireless communication problem where the transmitter is located at $[0,0]$ and the receiver is at $[500\text{m}, 450\text{m}]$. We randomly sample 25 points according to the uniform distribution on the square of $[0, 500\text{m}] \times [0, 500\text{m}]$ to denote the scatterers' positions; i.e., there exist 26 radio paths. All the source data and codes are available online at GitHub with thorough implementation comments: https://github.com/Spratm-Asleaf/DRRC. In this section, we only present major experimental setups and results; readers can use the shared source codes to explore (or verify) minor ones.

The following eleven methods are implemented in the experiments: 1) **Wiener**: Wiener beamformer (12), upper expression; 2) **Wiener-DL**: Wiener beamformer with diagonal loading (30), upper expression; 3) **Wiener-DR**: Distributionally robust Wiener beamformer (49) and (53); 4) **Wiener-CE**: Channel-estimation-based Wiener beamformer (12), lower expression; 5) **Wiener-CE-DL**: Channel-estimation-based Wiener beamformer with diagonal loading (30), lower expression; 6) **Wiener-CE-DR**: Distributionally robust channel-estimation-based Wiener beamformer (42) and (31); 7) **Capon**: Capon beamformer (39) for $\epsilon_0 = 0$; 8) **Capon-DL**: Capon beamformer with diagonal loading (39); 9) **ZF**: Zero-forcing beamformer where $\boldsymbol{W}_{\mathrm{ZF}} := (\hat{\boldsymbol{H}}^{\mathsf{H}}\hat{\boldsymbol{H}})^{-1}\hat{\boldsymbol{H}}^{\mathsf{H}}$ and $\hat{\boldsymbol{H}}$ denotes the estimated channel matrix; 10) **Kernel**: Kernel receiver (61) with $\epsilon_0 = 0$ in (60); and 11) **Kernel-DL**: Kernel receiver with diagonal loading (65). Note that the diagonal-loading-based methods are particular cases of distributionally robust combiners; see, e.g., Corollary 1 and Example 4. The deep-learning-based (DL-based) methods in Subsection V-B are not implemented in this section because they have been deeply studied in our previous publications, e.g., [10], [12]; we only comment on the advantages and disadvantages of DL-based methods compared with the listed eleven methods in Section VII (Conclusions).

[3]Neural networks are hard to be globally optimized.

When covariance matrix $\boldsymbol{R}_s$ of transmitted signal $\mathbf{s}$ is unknown for the receiver (e.g., in ISAC systems, $\boldsymbol{R}_s$ needs to vary from one frame to another for sensing), $\boldsymbol{R}_s$ is estimated by the sample covariance matrix $\hat{\boldsymbol{R}}_s = \boldsymbol{S}\boldsymbol{S}^{\mathsf{H}}/L$. The channel matrix $\boldsymbol{H}$ is estimated using the minimum mean-squared error method, i.e., $\hat{\boldsymbol{H}} = \boldsymbol{X}\boldsymbol{S}^{\mathsf{H}}(\boldsymbol{S}\boldsymbol{S}^{\mathsf{H}})^{-1}$. Covariance matrix $\boldsymbol{R}_v$ of channel noise $\mathbf{v}$ is estimated using the least-square method, i.e., $\hat{\boldsymbol{R}}_v = (\boldsymbol{X} - \hat{\boldsymbol{H}}\boldsymbol{S})(\boldsymbol{X} - \hat{\boldsymbol{H}}\boldsymbol{S})^{\mathsf{H}}/L$. The matrices $\hat{\boldsymbol{R}}_s$, $\hat{\boldsymbol{H}}$, and $\hat{\boldsymbol{R}}_v$ are therefore uncertain compared to their true (but unknown; possibly time-varying) values $\boldsymbol{R}_s$, $\boldsymbol{H}$, and $\boldsymbol{R}_v$, respectively. The matrices $\hat{\boldsymbol{R}}_s$, $\hat{\boldsymbol{H}}$, and $\hat{\boldsymbol{R}}_v$ are used in beamformers such as the channel-estimation-based Wiener beamformer (30), the Capon beamformer, and the zero-forcing beamformer.

The combiners are determined on the training data set (i.e., pilot data). The performance evaluation method of combiners is mean-squared estimation error (MSE) on the test data set (i.e., non-pilot communication data): to be specific, $\|\boldsymbol{S}_{\mathrm{test}} - \hat{\boldsymbol{S}}_{\mathrm{test}}\|_F^2 / (M \times L_{\mathrm{test}})$ where $\boldsymbol{S}_{\mathrm{test}} \in \mathbb{C}^{M \times L_{\mathrm{test}}}$ is the test data block, $\hat{\boldsymbol{S}}_{\mathrm{test}}$ is its estimate, and L_{test} is the length of non-pilot test data units. As data-driven machine learning methods, all parameters (e.g., uncertainty quantification coefficients ϵ's) of combiners can be tuned using the popular cross-validation (e.g., one-shot cross-validation) method. The parameters can also be empirically tuned to save training times because cross-validation imposes a significant computational burden. This article mainly uses the empirical tuning method (i.e., trial-and-error) to tune each combiner to achieve its best average performance. For each test case, the MSE performances are averaged on 250 Monte–Carlo episodes.

We consider an experimental scenario where impulse channel noises exist; i.e., the channel is non-Gaussian so linear beamformers are no longer sufficient. (Complementary experimental setups and results can be seen in online supplementary materials.) The detailed setups are as follows. The transmitter has four antennas (i.e., $M = 4$) with unit transmit power; without loss of generality, each antenna is assumed to emit continuous-valued complex Gaussian signals. The receiver has eight antennas (i.e., $N = 8$). The SNR is -10dB, which is a challenging situation. The channel has impulse noises: i.e., in L received signals (i.e., $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L]$) that are contaminated by usual complex Gaussian channel noises, 10% of them are also contaminated by uniform noises with the maximum amplitude of 1.5, which is a relatively large value compared to the amplitude of the usual Gaussian channel noises. We assume that a communication frame contains 500 non-pilot data units; i.e., $L_{\mathrm{test}} = 500$. The experimental results are shown in Tables I~VI, from which the following main points can be outlined.

1) A larger number of pilot data benefits the estimation performances of wireless signals.
2) The diagonal loading operation can significantly improve the estimation performances especially when the pilot data size is relatively small.
3) Since the signal model under impulse channel noises is no longer linear Gaussian, the optimal combiner in the MSE sense must be nonlinear. Therefore, the Kernel and the Kernel-DL methods have the potential to outper-

form other linear beamformers, i.e., to *suppress outliers*. However, in practice, the non-robust Kernel method may undergo numerical instability in calculating the inverse of the kernel matrix K. Therefore, its actual MSEs are not necessarily smaller than those of linear beamformers. Nevertheless, the robust Kernel-DL method consistently outperforms all other beamformers.

4) Distributionally robust combiners (including diagonal-loading ones) can combat the adverse effect introduced by the limited pilot size and several types of uncertainties in the signal model (e.g., outliers). To be specific, for example, all diagonal-loading combiners can outperform their original non-diagonal-loading counterparts; cf. the Wiener and the Wiener-DL methods, the Wiener-CE and the Wiener-CE-DL methods, the Capon and the Capon-DL methods, and the Kernel and the Kernel-DL methods. In addition, the Wiener-DR beamformer (53) using the F-norm uncertainty set has the potential to outperform the Wiener-DL beamformer (30) that employs the simple uncertainty set (26).

5) Although the Wiener-DR beamformer has the potential to work better than the Wiener-DL beamformer, it has a significant computational burden, which may not be suitable for timely use in practice especially when the computing resources are limited. Hence, the Wiener-DL beamformer is practically promising because it can provide an excellent balance between the computational burden and the actual performance.

Remarks on Parameter Tuning: From experiments, we find that the uncertainty quantification coefficients ϵ's (e.g., in diagonal loading) can be neither too large nor too small. When ϵ's are too large, the combiners become overly conservative, while when ϵ's are too small, the combiners cannot offer sufficient robustness against data scarcity and model uncertainties. In both cases of inappropriate ϵ's, the performances of combiners degrade significantly. Therefore, ϵ's must be carefully tuned in practice, and a rigorous method to tune ϵ's can be the cross-validation method on the training data set (i.e., the pilot data set). If practitioners just pursue satisfaction rather than optimality, the empirical tuning method is recommended to save training times.

TABLE I
EXPERIMENTAL RESULTS (PILOT SIZE = 10)

Combiner	MSE	Time	Combiner	MSE	Time
Wnr	3.30	1.49e-04	Wnr-DL	2.11	9.81e-06
Wnr-DR	1.97	3.16e+00	Wnr-CE	3.30	4.59e-05
Wnr-CE-DL	2.50	2.17e-05	Wnr-CE-DR	3.31	4.63e-05
Capon	5.44	4.42e-05	Capon-DL	4.52	2.50e-05
ZF	2.12	2.54e-05	Kernel	1.07	1.60e-04
Kernel-DL	0.80	5.59e-05			

Wnr: The abbreviation for Wiener.
Time: The training time averaged on 250 Monte–Carlo episodes.

VII. CONCLUSIONS

This article introduces a unified mathematical framework for receive combining of wireless signals from the perspective

TABLE II
EXPERIMENTAL RESULTS (PILOT SIZE = 15)

Combiner	MSE	Time	Combiner	MSE	Time
Wnr	1.38	1.65e-04	Wnr-DL	1.23	1.10e-05
Wnr-DR	1.07	3.21e+00	Wnr-CE	1.38	4.44e-05
Wnr-CE-DL	1.30	2.12e-05	Wnr-CE-DR	1.39	4.28e-05
Capon	4.48	4.31e-05	Capon-DL	4.34	2.42e-05
ZF	2.97	2.44e-05	Kernel	1.12	1.94e-04
Kernel-DL	0.70	9.23e-05			

TABLE III
EXPERIMENTAL RESULTS (PILOT SIZE = 20)

Combiner	MSE	Time	Combiner	MSE	Time
Wnr	1.12	1.86e-04	Wnr-DL	1.05	1.87e-05
Wnr-DR	0.93	7.19e+00	Wnr-CE	1.12	5.78e-05
Wnr-CE-DL	1.08	3.14e-05	Wnr-CE-DR	1.13	6.01e-05
Capon	5.01	5.93e-05	Capon-DL	4.94	3.81e-05
ZF	3.82	3.56e-05	Kernel	1.20	4.48e-04
Kernel-DL	0.66	3.11e-04			

TABLE IV
EXPERIMENTAL RESULTS (PILOT SIZE = 25)

Combiner	MSE	Time	Combiner	MSE	Time
Wnr	0.92	1.41e-04	Wnr-DL	0.88	1.11e-05
Wnr-DR	0.80	4.22e+00	Wnr-CE	0.92	5.02e-05
Wnr-CE-DL	0.90	2.44e-05	Wnr-CE-DR	0.92	4.78e-05
Capon	4.94	4.93e-05	Capon-DL	4.89	2.85e-05
ZF	4.06	2.72e-05	Kernel	1.14	4.26e-04
Kernel-DL	0.60	2.95e-04			

TABLE V
EXPERIMENTAL RESULTS (PILOT SIZE = 50)

Combiner	MSE	Time	Combiner	MSE	Time
Wnr	0.69	1.75e-04	Wnr-DL	0.68	1.85e-05
Wnr-DR	0.65	6.10e+00	Wnr-CE	0.69	5.81e-05
Wnr-CE-DL	0.68	3.03e-05	Wnr-CE-DR	0.70	5.90e-05
Capon	6.95	5.97e-05	Capon-DL	6.93	3.75e-05
ZF	6.36	3.38e-05	Kernel	0.92	1.81e-03
Kernel-DL	0.53	1.67e-03			

TABLE VI
EXPERIMENTAL RESULTS (PILOT SIZE = 100)

Combiner	MSE	Time	Combiner	MSE	Time
Wnr	0.57	3.41e-04	Wnr-DL	0.57	3.64e-05
Wnr-DR	0.55	4.96e+00	Wnr-CE	0.57	6.35e-05
Wnr-CE-DL	0.57	2.93e-05	Wnr-CE-DR	0.58	6.07e-05
Capon	9.89	6.88e-05	Capon-DL	9.88	3.99e-05
ZF	9.45	3.27e-05	Kernel	0.72	5.93e-03
Kernel-DL	0.49	5.83e-03			

of data-driven machine learning, which reveals that channel estimation is not a necessary operation. To combat the limited pilot size and several types of uncertainties in the signal model, the distributionally robust (DR) receive combining framework is then suggested. We prove that the diagonal-loading (DL) methods are distributionally robust against the scarcity of pilot data and the uncertainties in the signal model. In addition, we generalize the diagonal-loading methods to achieve better estimation performance (e.g., the DR Wiener beamformer using F-norm for uncertainty quantification), at the cost of significantly higher computational burdens. Experiments suggest that nonlinear combiners such as the Kernel and the Kernel-DL methods have the potential when the pilot size is small and/or the signal model is not linear Gaussian. Compared with the Kernel and the Kernel-DL combiners, neural-network-based solutions [10], [12] have a stronger expressive capability of nonlinearities, which however are unscalable in the numbers of transmit and receive antennas, and significantly more time-consuming in training and more troublesome in tuning hyper-parameters (e.g., the number of layers and the number of neurons in each layer) than the studied eleven combiners.

APPENDIX A
STRUCTURED REPRESENTATION OF NONLINEAR FUNCTIONS

In Section II, we have reviewed two popular frameworks for representing (nonlinear) functions: reproducing kernel Hilbert spaces (RKHS) and neural network function spaces (NNFS). Typical kernel functions $\ker(\cdot, \cdot)$ to define RKHSs include Gaussian kernel, Matern kernel, Linear kernel, Laplacian kernel, and Polynomial kernel. Mathematical details of these kernel functions can be found in [24, Subsec. 14.2], [27, Ex. 1]. Typical activation functions $\sigma(\cdot)$ to define NNFSs include Hyperbolic tangent (i.e, tanh) function, Softmax function, Sigmoid function, Rectified linear unit (ReLU) function, and Exponential linear unit (ELU) function. Mathematical details of these activation functions can be found in [27, Ex. 2].

APPENDIX B
DETAILS ON REAL-SPACE SIGNAL REPRESENTATION

Let $\boldsymbol{R}_x := \mathbb{E}\mathbf{x}\mathbf{x}^{\mathsf{H}}$, $\boldsymbol{C}_x := \mathbb{E}\mathbf{x}\mathbf{x}^{\mathsf{T}}$, $\boldsymbol{C}_s := \mathbb{E}\mathbf{s}\mathbf{s}^{\mathsf{T}}$, and $\boldsymbol{C}_v := \mathbb{E}\mathbf{v}\mathbf{v}^{\mathsf{T}} = \mathbf{0}$. We have

$$\boldsymbol{R}_{\underline{x}} := \mathbb{E}\underline{\mathbf{x}}\underline{\mathbf{x}}^{\mathsf{T}} = \frac{1}{2}\left[\begin{array}{cc} \mathrm{Re}(\boldsymbol{R}_x + \boldsymbol{C}_x) & \mathrm{Im}(-\boldsymbol{R}_x + \boldsymbol{C}_x) \\ \mathrm{Im}(\boldsymbol{R}_x + \boldsymbol{C}_x) & \mathrm{Re}(\boldsymbol{R}_x - \boldsymbol{C}_x) \end{array}\right].$$

$$\boldsymbol{R}_{\underline{s}} := \mathbb{E}\underline{\mathbf{s}}\underline{\mathbf{s}}^{\mathsf{T}} = \frac{1}{2}\left[\begin{array}{cc} \mathrm{Re}(\boldsymbol{R}_s + \boldsymbol{C}_s) & \mathrm{Im}(-\boldsymbol{R}_s + \boldsymbol{C}_s) \\ \mathrm{Im}(\boldsymbol{R}_s + \boldsymbol{C}_s) & \mathrm{Re}(\boldsymbol{R}_s - \boldsymbol{C}_s) \end{array}\right],$$

and

$$\boldsymbol{R}_{\underline{v}} := \mathbb{E}\underline{\mathbf{v}}\underline{\mathbf{v}}^{\mathsf{T}} = \frac{1}{2}\left[\begin{array}{cc} \mathrm{Re}\,\boldsymbol{R}_v & \mathrm{Im}-\boldsymbol{R}_v \\ \mathrm{Im}\,\boldsymbol{R}_v & \mathrm{Re}\,\boldsymbol{R}_v \end{array}\right].$$

Note that the following identities hold: $\boldsymbol{R}_x = \boldsymbol{H}\boldsymbol{R}_s\boldsymbol{H}^{\mathsf{H}} + \boldsymbol{R}_v$, $\boldsymbol{C}_x = \boldsymbol{H}\boldsymbol{C}_s\boldsymbol{H}^{\mathsf{T}}$, $\boldsymbol{R}_{\underline{x}} = \underline{\boldsymbol{H}} \cdot \boldsymbol{R}_{\underline{s}} \cdot \underline{\boldsymbol{H}}^{\mathsf{T}} + \boldsymbol{R}_{\underline{v}}$, and $\boldsymbol{R}_{\underline{xs}} = \underline{\boldsymbol{H}} \cdot \boldsymbol{R}_{\underline{s}}$.

APPENDIX C
EXTENSIVE READING ON DISTRIBUTIONAL UNCERTAINTY

A. Generalization Error and Distributional Robustness

We use (7) and (15) as examples to illustrate the concepts. Supposing that $\phi^{\star}$ solves the true problem (7) and $\phi^{\star}_{\mathrm{ERM}}$ solves the surrogate problem (15), we have

$$\begin{aligned} \min_{\phi} \mathrm{Tr}\, &\mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}}[\phi(\mathbf{x}) - \mathbf{s}][\phi(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} \\ &= \mathrm{Tr}\, \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}}[\phi^{\star}(\mathbf{x}) - \mathbf{s}][\phi^{\star}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} \\ &\leq \mathrm{Tr}\, \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}}[\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}][\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}}. \end{aligned} \quad (72)$$

To clarify further, the testing error in the last line (evaluated at the true distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$) of the learned estimator $\phi^{\star}_{\mathrm{ERM}}$ may be (much) larger than the optimal error in the first two lines, although $\phi^{\star}_{\mathrm{ERM}}$ has the smallest training error (evaluated at the nominal distribution $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$), i.e.,

$$\begin{aligned} \min_{\phi} \mathrm{Tr}\, &\mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}}[\phi(\mathbf{x}) - \mathbf{s}][\phi(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} \\ &= \min_{\phi} \mathrm{Tr}\, \frac{1}{L}\sum_{i=1}^{L}[\phi(\boldsymbol{x}_i) - \boldsymbol{s}_i][\phi(\boldsymbol{x}_i) - \boldsymbol{s}_i]^{\mathsf{H}} \\ &= \mathrm{Tr}\, \frac{1}{L}\sum_{i=1}^{L}[\phi^{\star}_{\mathrm{ERM}}(\boldsymbol{x}_i) - \boldsymbol{s}_i][\phi^{\star}_{\mathrm{ERM}}(\boldsymbol{x}_i) - \boldsymbol{s}_i]^{\mathsf{H}} \\ &\leq \mathrm{Tr}\, \frac{1}{L}\sum_{i=1}^{L}[\phi^{\star}(\boldsymbol{x}_i) - \boldsymbol{s}_i][\phi^{\star}(\boldsymbol{x}_i) - \boldsymbol{s}_i]^{\mathsf{H}}. \end{aligned} \quad (73)$$

In the terminologies of machine learning, the difference between the testing error and the training error, i.e.,

$$\begin{aligned} \mathrm{Tr}\, &\mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}}[\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}][\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} - \\ &\mathrm{Tr}\, \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}}[\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}][\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} \\ &= \mathrm{Tr}\, \mathbb{E}_{\mathbf{x},\mathbf{s}}[\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}][\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} - \\ &\mathrm{Tr}\, \frac{1}{L}\sum_{i=1}^{L}[\phi^{\star}_{\mathrm{ERM}}(\boldsymbol{x}_i) - \boldsymbol{s}_i][\phi^{\star}_{\mathrm{ERM}}(\boldsymbol{x}_i) - \boldsymbol{s}_i]^{\mathsf{H}} \end{aligned}$$

is called the *generalization error* of $\phi^{\star}_{\mathrm{ERM}}$; the difference between the testing error and the optimal error, i.e.,

$$\begin{aligned} &\mathrm{Tr}\, \mathbb{E}_{\mathbf{x},\mathbf{s}}[\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}][\phi^{\star}_{\mathrm{ERM}}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} - \\ &\mathrm{Tr}\, \mathbb{E}_{\mathbf{x},\mathbf{s}}[\phi^{\star}(\mathbf{x}) - \mathbf{s}][\phi^{\star}(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}} \end{aligned}$$

is called the *excess risk* of $\phi^{\star}_{\mathrm{ERM}}$. In machine learning practice, we want to reduce both the generalization error and the excess risk. Most attention in the literature has been particularly paid to reducing generalization errors. Specifically, an upper bound of the true cost $\mathrm{Tr}\, \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}}[\phi(\mathbf{x}) - \mathbf{s}][\phi(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}}$ is first found and then minimize the upper bound: by minimizing the upper bound, the true cost can also be reduced.

Fact 1: Suppose that the true distribution $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$ of $(\mathbf{x},\mathbf{s})$ is included in $\mathcal{U}_{\mathbf{x},\mathbf{s}}$; for notational clarity, we hereafter distinguish $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$ from $\mathbb{P}_{\mathbf{x},\mathbf{s}}$. The true objective function evaluated at $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$, i.e.,

$$\mathrm{Tr}\, \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{0,\mathbf{x},\mathbf{s}}}[\phi(\mathbf{x}) - \mathbf{s}][\phi(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}}, \quad \forall \phi \in \mathcal{B}, \quad (74)$$

is upper bounded by the worst-case objective function of (17), i.e.,

$$\max_{\mathbb{P}_{\mathbf{x},\mathbf{s}}\in\mathcal{U}_{\mathbf{x},\mathbf{s}}} \mathrm{Tr}\, \mathbb{E}_{(\mathbf{x},\mathbf{s})\sim\mathbb{P}_{\mathbf{x},\mathbf{s}}}[\phi(\mathbf{x}) - \mathbf{s}][\phi(\mathbf{x}) - \mathbf{s}]^{\mathsf{H}}, \quad \forall \phi \in \mathcal{B}. \quad (75)$$

Therefore, by diminishing the upper bound in (75), the true estimation error evaluated at $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$ can also be reduced. However, the conventional empirical estimation error evaluated at $\hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}$ cannot upper bound the true estimation error (74).

This performance guarantee is the benefit of considering the distributionally robust method (17). Due to the weak convergence property of the empirical distribution to the true data-generating distribution, that is, $d(\mathbb{P}_{0,\mathbf{x},\mathbf{s}}, \ \hat{\mathbb{P}}_{\mathbf{x},\mathbf{s}}) \to 0$ as the sample size $L \to \infty$, there exists ϵ in (18) for every L, such that $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$ is included in $\mathcal{U}_{\mathbf{x},\mathbf{s}}$ in $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}^L$-probability ($L$-fold product measure of $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$). □

B. Non-Stationary Channel Statistics

In the main body of the article (see also Fact 1), we assume that the true data-generating distribution $\mathbb{P}_{0,\mathbf{x},\mathbf{s}}$ is time-invariant within a frame. In real-world operations, however, this assumption might be untenable.

As shown in Fig. 1, the frame contains eight data units; we suppose that the first four units are pilot symbols and the rest four units are communication-data symbols.
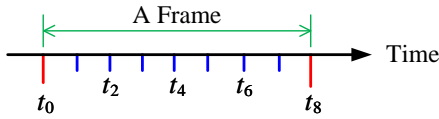

Fig. 1. True data-generating distributions might be time-varying in a frame.

Let $\mathbb{P}_{0,\mathbf{x},\mathbf{s},i}$ denote the true data-generating distribution at time point t_i where $i = 1, 2, \ldots, 8$. Specifically, we have $(\mathbf{x}_i, \mathbf{s}_i) \sim \mathbb{P}_{0,\mathbf{x},\mathbf{s},i}$ for every i. Therefore, the pilot data set (i.e., the training data set) $\{(\boldsymbol{x}_1, \boldsymbol{s}_1), (\boldsymbol{x}_2, \boldsymbol{s}_2), (\boldsymbol{x}_3, \boldsymbol{s}_3), (\boldsymbol{x}_4, \boldsymbol{s}_4)\}$ can be seen as realizations of the mean distribution $\mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}}$ of underlying true training-data distributions where $\mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}} = \sum_{i=1}^4 h_i \mathbb{P}_{0,\mathbf{x},\mathbf{s},i}$, which is a mixture distribution with mixing weights $0 \le h_1, h_2, h_3, h_4 \le 1$; $\sum_{i=1}^4 h_i = 1$. Similarly, the communication data set (i.e., the testing data set) $\{(\boldsymbol{x}_5, \boldsymbol{s}_5), (\boldsymbol{x}_6, \boldsymbol{s}_6), (\boldsymbol{x}_7, \boldsymbol{s}_7), (\boldsymbol{x}_8, \boldsymbol{s}_8)\}$ can be seen as realizations of the mean $\mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}}$ of the underlying true testing-data distributions where $\mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}} = \sum_{i=5}^8 h_i \mathbb{P}_{0,\mathbf{x},\mathbf{s},i}$, with mixing weights $0 \le h_5, h_6, h_7, h_8 \le 1$; $\sum_{i=5}^8 h_i = 1$.

Suppose that

$$d(\hat{\mathbb{P}}_{\text{train},\mathbf{x},\mathbf{s}}, \ \mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}}) \le \epsilon_1,$$

where $\hat{\mathbb{P}}_{\text{train},\mathbf{x},\mathbf{s}} := \frac{1}{4} \sum_{i=1}^4 \delta_{(\boldsymbol{x}_i, \boldsymbol{s}_i)}$ is the data-driven estimate of $\mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}}$ and

$$d(\mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}}, \ \mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}}) \le \epsilon_2,$$

for some $\epsilon_1, \epsilon_2 \ge 0$. We have the uncertainty quantification

$$d(\mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}}, \ \hat{\mathbb{P}}_{\text{train},\mathbf{x},\mathbf{s}}) \le \epsilon := \epsilon_1 + \epsilon_2.$$

Therefore, the distributionally robust modeling and solution framework is still valid to hedge against the distributional uncertainty in the nominal distribution $\hat{\mathbb{P}}_{\text{train},\mathbf{x},\mathbf{s}}$ compared to the underlying true distribution $\mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}}$. When $\mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}} = \mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}}$, as assumed in the main body of the article, we have $\epsilon_1 \to 0$ and $\epsilon \to \epsilon_2 = 0$ as the pilot size tends to infinity; however, when $\mathbb{P}_{\text{train},0,\mathbf{x},\mathbf{s}} \ne \mathbb{P}_{\text{test},0,\mathbf{x},\mathbf{s}}$, the radius $\epsilon \to \epsilon_2 \ne 0$ although $\epsilon_1 \to 0$.

Another justification for the DRO method is as follows. Suppose that there exists $\epsilon \ge 0$ such that

$$d(\mathbb{P}_{0,\mathbf{x},\mathbf{s},i}, \ \hat{\mathbb{P}}_{\text{train},\mathbf{x},\mathbf{s}}) \le \epsilon, \quad \forall i \in \{1, 2, \ldots, 8\}.$$

It means that, at every snapshot in the frame, the true data-generating distribution is included in the uncertainty set. Hence, the DRO cost can still upper bound the true cost even though the true distribution is time-varying; cf. Fact 1.

APPENDIX D
ADDITIONAL DISCUSSIONS ON DISTRIBUTIONALLY ROBUST ESTIMATION

To develop this article, the typical minimum mean-squared error (MSE) criterion is employed; see (7) and (10). Accordingly, the distributionally robust receive combining framework in this article is exemplified using the MSE cost function. The cost function for wireless signal estimation, however, can be any Borel-measurable function $h : \mathbb{C}^M \times \mathbb{C}^M \to \mathbb{R}_+$. As a result, the optimal estimation problem under the distribution $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ is given by

$$\min_{\phi \in \mathcal{B}_{\mathbb{C}^N \to \mathbb{C}^M}} \mathbb{E}_{\mathbf{x},\mathbf{s}} h[\phi(\mathbf{x}), \mathbf{s}]. \tag{76}$$

Specific examples of h in wireless communications can be, e.g., mean absolute error, Huber's cost function [37], [38] where h is no longer quadratic as in (7) and (10). Accordingly, when the distributional uncertainty exists in $\mathbb{P}_{\mathbf{x},\mathbf{s}}$, the distributionally robust receive combining framework becomes

$$\min_{\phi \in \mathcal{B}_{\mathbb{C}^N \to \mathbb{C}^M}} \max_{\mathbb{P}_{\mathbf{x},\mathbf{s}} \in \mathcal{U}_{\mathbf{x},\mathbf{s}}} \mathbb{E}_{\mathbf{x},\mathbf{s}} h[\phi(\mathbf{x}), \mathbf{s}]. \tag{77}$$

Problem (77) is generally challenging to solve because it is an infinite-dimensional program. Therefore, in practice, we can limit the feasible region of ϕ to a parameterized subspace of $\mathcal{B}_{\mathbb{C}^N \to \mathbb{C}^M}$, for example, a reproducing kernel Hilbert space $\mathcal{H}$ or a neural network function space $\mathcal{K}$; see Section II. Consequently, Problem (77) is approximated by the following finite-dimensional (in terms of $\boldsymbol{W}$) program

$$\min_{\boldsymbol{W}} \max_{\mathbb{P}_{\mathbf{x},\mathbf{s}} \in \mathcal{U}_{\mathbf{x},\mathbf{s}}} \mathbb{E}_{\mathbf{x},\mathbf{s}} h[\phi_{\boldsymbol{W}}(\mathbf{x}), \mathbf{s}], \tag{78}$$

where $\boldsymbol{W}$ parameterizes ϕ and lies in real or complex coordinate spaces; note that both $\mathcal{H}$ and $\mathcal{K}$ can be dense in $\mathcal{B}$. Under the MSE cost function, (78) is particularized in (19) for linear function spaces, in (57) for reproducing kernel Hilbert spaces, and in (69) for neural network function spaces, which build this article in a technically tractable manner.

The distributionally robust receive combining problem (77) under generic cost functions h and generic feasible regions of ϕ can be technically challenging. Even for the simplified problem (78), the solution method can be quite complex, and closed-form solutions cannot be generally guaranteed; see, e.g., [39], [40]. The complication further arises when the distributional uncertainty sets $\mathcal{U}_{\mathbf{x},\mathbf{s}}$ for $\mathbb{P}_{\mathbf{x},\mathbf{s}}$ are complicated; see, e.g., [32]. Therefore, this article serves as the starting point of distributionally robust receive combining, in which **closed-form solutions** are largely ensured by leveraging

F1) the MSE cost function as in (7) and (10);

F2) the linear function spaces as in (19) and reproducing kernel Hilbert spaces as in (57);

F3) the second-moment-based uncertainty sets in Definitions 1, 2, 3, and 4; see also Corollary 1, Claim 2, and Example 4.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2025.3582082

15

Note that even under the features F1) and F2), the closed-form solutions cannot be guaranteed. For example, if Wasserstein or F-norm uncertainty sets are used, the associated distributionally robust receive combining problems can be computationally heavy; see (47) and (52) as well as Propositions 1 and 2. However, for emerging high-performance computing devices, the computational burden may be no longer an issue in the future. Hence, advanced distributionally robust receive combining formulations based on (77) and (78) are still attractive for future-generation communication systems. This article seeks to provide a foundation for this direction.

APPENDIX E
PROOF OF LEMMA 1

Proof: The objective function of Problem (19) equals to

$$\left\langle \begin{bmatrix} \boldsymbol{W}^{\mathsf{H}}\boldsymbol{W} & -\boldsymbol{W}^{\mathsf{H}} \\ -\boldsymbol{W} & \boldsymbol{I}_M \end{bmatrix}, \begin{bmatrix} \boldsymbol{R}_x & \boldsymbol{R}_{xs} \\ \boldsymbol{R}_{xs}^{\mathsf{H}} & \boldsymbol{R}_s \end{bmatrix} \right\rangle, \quad (79)$$

where $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \mathrm{Tr}\,\boldsymbol{A}^{\mathsf{H}}\boldsymbol{B}$ for two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Therefore, the objective function of (19) is convex in $\boldsymbol{W}$ and linear (thus concave) in the matrix variable $\boldsymbol{R}$. Hence, due to Sion's minimax theorem [41, Corollary 3.3], Problem (19) is equivalent to

$$\begin{aligned} \max_{\boldsymbol{R}} \min_{\boldsymbol{W}} \quad & \mathrm{Tr}\left[\boldsymbol{W}\boldsymbol{R}_x\boldsymbol{W}^{\mathsf{H}} - \boldsymbol{W}\boldsymbol{R}_{xs} - \boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{W}^{\mathsf{H}} + \boldsymbol{R}_s \right] \\ \text{s.t.} \quad & d_0(\boldsymbol{R},\ \hat{\boldsymbol{R}}) \le \epsilon_0, \\ & \boldsymbol{R} \succeq \boldsymbol{0}. \end{aligned}$$
$$(80)$$

Note that the feasible region of $\boldsymbol{R}$ is compact convex, and that of $\boldsymbol{W}$ (i.e., $\mathbb{C}^{M\times N}$) is convex.

For every given $\boldsymbol{R}$, the inner minimization sub-problem of (80) is solved by the Wiener beamformer $\boldsymbol{W}_{\mathrm{Wiener}}^{\star} = \boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}$, which transforms (80) to (21). This completes the proof. □

APPENDIX F
PROOF OF THEOREM 1

Proof: Consider the following optimization problem

$$\begin{aligned} \max_{\boldsymbol{R}} \quad & \mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs} + \boldsymbol{R}_s \right] \\ \text{s.t.} \quad & \boldsymbol{R} \succeq \boldsymbol{R}_2, \\ & \boldsymbol{R}_x \succ \boldsymbol{0}, \end{aligned}$$
$$(81)$$

which, due to Lemma 1, is equivalent [in the sense of the same optimal objective value and maximizer(s) $\boldsymbol{R}^{\star}$] to

$$\begin{aligned} \min_{\boldsymbol{W}} \max_{\boldsymbol{R}} \quad & \left\langle \begin{bmatrix} \boldsymbol{W}^{\mathsf{H}}\boldsymbol{W} & -\boldsymbol{W}^{\mathsf{H}} \\ -\boldsymbol{W} & \boldsymbol{I}_M \end{bmatrix}, \begin{bmatrix} \boldsymbol{R}_x & \boldsymbol{R}_{xs} \\ \boldsymbol{R}_{xs}^{\mathsf{H}} & \boldsymbol{R}_s \end{bmatrix} \right\rangle \\ \text{s.t.} \quad & \boldsymbol{R} \succeq \boldsymbol{R}_2, \\ & \boldsymbol{R}_x \succ \boldsymbol{0}. \end{aligned}$$
$$(82)$$

Note that $\begin{bmatrix} \boldsymbol{W}^{\mathsf{H}}\boldsymbol{W} & -\boldsymbol{W}^{\mathsf{H}} \\ -\boldsymbol{W} & \boldsymbol{I}_M \end{bmatrix} \succeq \boldsymbol{0}$, because for all $\boldsymbol{x} \in \mathbb{C}^N$ and $\boldsymbol{y} \in \mathbb{C}^M$, we have

$$[\boldsymbol{x}^{\mathsf{H}},\ \boldsymbol{y}^{\mathsf{H}}] \begin{bmatrix} \boldsymbol{W}^{\mathsf{H}}\boldsymbol{W} & -\boldsymbol{W}^{\mathsf{H}} \\ -\boldsymbol{W} & \boldsymbol{I}_M \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}\|_2^2 \ge 0.$$

Therefore, for every given $\boldsymbol{W}$, the objective function of (82) is increasing in $\boldsymbol{R}$. As a result, the objective value of (81) is lower-bounded at $\boldsymbol{R}_2$: To be specific, $\forall \boldsymbol{R} \succeq \boldsymbol{R}_2$, we have

$$\mathrm{Tr}\left[-\boldsymbol{R}_{xs}^{\mathsf{H}}\boldsymbol{R}_x^{-1}\boldsymbol{R}_{xs} + \boldsymbol{R}_s \right] \ge \mathrm{Tr}\left[-\boldsymbol{R}_{2,xs}^{\mathsf{H}}\boldsymbol{R}_{2,x}^{-1}\boldsymbol{R}_{2,xs} + \boldsymbol{R}_{2,s} \right],$$

i.e., $f_1(\boldsymbol{R}) \ge f_1(\boldsymbol{R}_2)$, which proves the first part.

On the other hand, if $\boldsymbol{R}_{1,x} \succeq \boldsymbol{R}_{2,x} \succ \boldsymbol{0}$, we have $\boldsymbol{R}_{2,x}^{-1} \succeq \boldsymbol{R}_{1,x}^{-1}$. As a result, $f_2(\boldsymbol{R}_{1,x}) - f_2(\boldsymbol{R}_{2,x}) = \mathrm{Tr}\left[\boldsymbol{R}_{xs}^{\mathsf{H}}(\boldsymbol{R}_{2,x}^{-1} - \boldsymbol{R}_{1,x}^{-1})\boldsymbol{R}_{xs} \right] \ge 0$, completing the proof. □

APPENDIX G
PROOF OF PROPOSITION 3

Proof: Letting $\underline{\boldsymbol{z}} := \boldsymbol{\varphi}(\underline{\boldsymbol{x}})$, (57) can be rewritten as

$$\min_{\boldsymbol{W} \in \mathbb{R}^{2M\times L}} \max_{\mathbb{P}_{\underline{\boldsymbol{z}},\underline{\boldsymbol{s}}} \in \mathcal{U}_{\underline{\boldsymbol{z}},\underline{\boldsymbol{s}}}} \mathrm{Tr}\,\mathbb{E}_{\underline{\boldsymbol{z}},\underline{\boldsymbol{s}}}[\boldsymbol{W}\underline{\boldsymbol{z}} - \underline{\boldsymbol{s}}][\boldsymbol{W}\underline{\boldsymbol{z}} - \underline{\boldsymbol{s}}]^{\mathsf{T}}. \quad (83)$$

Tantamount to the distributionally robust beamforming problem (19), Problem (83) reduces to (58) where

$$\hat{\boldsymbol{R}}_{\underline{z}} := \frac{1}{L}\sum_{i=1}^{L} \underline{\boldsymbol{z}}_i\underline{\boldsymbol{z}}_i^{\mathsf{T}} = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{\varphi}(\underline{\boldsymbol{x}}_i)\boldsymbol{\varphi}^{\mathsf{T}}(\underline{\boldsymbol{x}}_i) = \frac{1}{L}\boldsymbol{K}^2,$$

$$\hat{\boldsymbol{R}}_{\underline{z}\underline{s}} := \frac{1}{L}\sum_{i=1}^{L} \underline{\boldsymbol{z}}_i\underline{\boldsymbol{s}}_i^{\mathsf{T}} = \frac{1}{L}\sum_{i=1}^{L} \boldsymbol{\varphi}(\underline{\boldsymbol{x}}_i)\cdot\underline{\boldsymbol{s}}_i^{\mathsf{T}} = \frac{1}{L}\boldsymbol{K}\underline{\boldsymbol{S}}^{\mathsf{T}},$$

$$\hat{\boldsymbol{R}}_{\underline{s}} := \frac{1}{L}\sum_{i=1}^{L} \underline{\boldsymbol{s}}_i\underline{\boldsymbol{s}}_i^{\mathsf{T}} = \frac{1}{L}\sum_{i=1}^{L} \underline{\boldsymbol{s}}_i\cdot\underline{\boldsymbol{s}}_i^{\mathsf{T}} = \frac{1}{L}\underline{\boldsymbol{S}}\underline{\boldsymbol{S}}^{\mathsf{T}},$$

and

$$\boldsymbol{K} := [\boldsymbol{\varphi}(\underline{\boldsymbol{x}}_1), \boldsymbol{\varphi}(\underline{\boldsymbol{x}}_2), \ldots, \boldsymbol{\varphi}(\underline{\boldsymbol{x}}_L)] \in \mathbb{R}^{L\times L}.$$

The rest claims are due to Lemma 1; NB: $\boldsymbol{K}$ is invertible. □

REFERENCES

[1] T. Lo, H. Leung, and J. Litva, "Nonlinear beamforming," *Electronics Letters*, vol. 4, no. 27, pp. 350–352, 1991.

[2] S. Yang and L. Hanzo, "Fifty years of MIMO detection: The road to large-scale MIMOs," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 1941–1988, 2015.

[3] A. M. Elbir, K. V. Mishra, S. A. Vorobyov, and R. W. Heath, "Twenty-five years of advances in beamforming: From convex and nonconvex optimization to learning techniques," *IEEE Signal Processing Mag.*, vol. 40, no. 4, pp. 118–131, 2023.

[4] S. Chen, S. Tan, L. Xu, and L. Hanzo, "Adaptive minimum error-rate filtering design: A review," *Signal Processing*, vol. 88, no. 7, pp. 1671–1697, 2008.

[5] S. Chen, A. Wolfgang, C. J. Harris, and L. Hanzo, "Symmetric RBF classifier for nonlinear detection in multiple-antenna-aided systems," *IEEE Trans. Neural Networks*, vol. 19, no. 5, pp. 737–745, 2008.

[6] A. Navia-Vazquez, M. Martinez-Ramon, L. E. Garcia-Munoz, and C. G. Christodoulou, "Approximate kernel orthogonalization for antenna array processing," *IEEE Trans. Antennas Propagat.*, vol. 58, no. 12, pp. 3942–3950, 2010.

[7] M. Neinavaie, M. Derakhtian, and S. A. Vorobyov, "Lossless dimension reduction for integer least squares with application to sphere decoding," *IEEE Trans. Signal Processing*, vol. 68, pp. 6547–6561, 2020.

[8] J. Liao, J. Zhao, F. Gao, and G. Y. Li, "Deep learning aided low complex breadth-first tree search for MIMO detection," *IEEE Trans. Wireless Commun.*, 2023.

[9] D. A. Awan, R. L. Cavalcante, M. Yukawa, and S. Stanczak, "Robust online multiuser detection: A hybrid model-data driven approach," *IEEE Trans. Signal Processing*, 2023.

[10] H. Ye, G. Y. Li, and B.-H. Juang, "Power of deep learning for channel estimation and signal detection in OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 114–117, 2017.

[11] H. He, C.-K. Wen, S. Jin, and G. Y. Li, "Model-driven deep learning for MIMO detection," *IEEE Trans. Signal Processing*, vol. 68, pp. 1702–1715, 2020.

[12] N. Van Huynh and G. Y. Li, "Transfer learning for signal detection in wireless networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2325–2329, 2022.

This article has been accepted for publication in IEEE Transactions on Signal Processing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TSP.2025.3582082

16

[13] J. Li, P. Stoica, and Z. Wang, "On robust Capon beamforming and diagonal loading," *IEEE Trans. Signal Processing*, vol. 51, no. 7, pp. 1702–1715, 2003.

[14] R. G. Lorenz and S. P. Boyd, "Robust minimum variance beamforming," *IEEE Trans. Signal Processing*, vol. 53, no. 5, pp. 1684–1696, 2005.

[15] X. Zhang, Y. Li, N. Ge, and J. Lu, "Robust minimum variance beamforming under distributional uncertainty," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 2514–2518.

[16] B. Li, Y. Rong, J. Sun, and K. L. Teo, "A distributionally robust minimum variance beamformer design," *IEEE Signal Processing Lett.*, vol. 25, no. 1, pp. 105–109, 2017.

[17] Y. Huang, W. Yang, and S. A. Vorobyov, "Robust adaptive beamforming maximizing the worst-case SINR over distributional uncertainty sets for random inc matrix and signal steering vector," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4918–4922.

[18] Y. Huang, H. Fu, S. A. Vorobyov, and Z.-Q. Luo, "Robust adaptive beamforming via worst-case SINR maximization with nonconvex uncertainty sets," *IEEE Trans. Signal Processing*, vol. 71, pp. 218–232, 2023.

[19] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 35, no. 10, pp. 1365–1376, 1987.

[20] K. Harmanci, J. Tabrikian, and J. L. Krolik, "Relationships between adaptive minimum variance beamforming and optimal source localization," *IEEE Trans. Signal Processing*, vol. 48, no. 1, pp. 1–12, 2000.

[21] F. Liu, L. Zhou, C. Masouros, A. Li, W. Luo, and A. Petropulu, "Toward dual-functional radar-communication systems: Optimal waveform design," *IEEE Trans. Signal Processing*, vol. 66, no. 16, pp. 4264–4279, 2018.

[22] J. A. Zhang, F. Liu, C. Masouros, R. W. Heath, Z. Feng, L. Zheng, and A. Petropulu, "An overview of signal processing techniques for joint communication and radar sensing," *IEEE J. Select. Topics Signal Processing*, vol. 15, no. 6, pp. 1295–1315, 2021.

[23] Y. Xiong, F. Liu, Y. Cui, W. Yuan, T. X. Han, and G. Caire, "On the fundamental tradeoff of integrated sensing and communications under Gaussian channels," *IEEE Trans. Inform. Theory*, 2023.

[24] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[25] C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4, no. 4.

[26] G. Li and J. Ding, "Towards understanding variation-constrained deep neural networks," *IEEE Trans. Signal Processing*, vol. 71, pp. 631–640, 2023.

[27] S. Shafieezadeh-Abadeh, D. Kuhn, and P. M. Esfahani, "Regularization via mass transportation," *Journal of Machine Learning Research*, vol. 20, no. 103, pp. 1–68, 2019.

[28] M. Staib and S. Jegelka, "Distributionally robust optimization and generalization in kernel methods," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[29] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations Research*, vol. 58, no. 3, pp. 595–612, 2010.

[30] S. Wang, "Distributionally robust state estimation for jump linear systems," *IEEE Trans. Signal Processing*, 2023.

[31] J. Li, S. Lin, J. Blanchet, and V. A. Nguyen, "Tikhonov regularization is optimal transport robust under martingale constraints," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 677–17 689, 2022.

[32] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*. Informs, 2019, pp. 130–166.

[33] J. Blanchet, Y. Kang, and K. Murthy, "Robust Wasserstein profile inference and applications to machine learning," *Journal of Applied Probability*, vol. 56, no. 3, pp. 830–857, 2019.

[34] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, 2019.

[35] G. Saon, Z. Tüske, K. Audhkhasi, and B. Kingsbury, "Sequence noise injected training for end-to-end speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6261–6265.

[36] K. Vu, J. C. Snyder, L. Li, M. Rupp, B. F. Chen, T. Khelif, K.-R. Müller, and K. Burke, "Understanding kernel ridge regression: Common behaviors from simple functions to density functionals," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1115–1128, 2015.
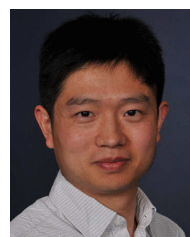
[37] X. Wang and H. V. Poor, "Robust adaptive array for wireless communications," *IEEE J. Select. Areas in Commun.*, vol. 16, no. 8, pp. 1352–1366, 1998.

[38] V. Katkovnik, M.-S. Lee, and Y.-H. Kim, "Performance study of the minimax robust phased array for wireless communications," *IEEE Trans. Wireless Commun.*, vol. 54, no. 4, pp. 608–613, 2006.

[39] H. Rahimian and S. Mehrotra, "Frameworks and results in distributionally robust optimization," *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022.

[40] D. Kuhn, S. Shafiee, and W. Wiesemann, "Distributionally robust optimization," *Acta Numerica*, 2024.

[41] M. Sion, "On general minimax theorems." *Pacific Journal of Mathematics*, vol. 8, no. 1, pp. 171 – 176, 1958.

Shixiong Wang (Member, IEEE) received the B.Eng. degree in detection, guidance, and control technology, and the M.Eng. degree in systems and control engineering from the School of Electronics and Information, Northwestern Polytechnical University, China, in 2016 and 2018, respectively. He received his Ph.D. degree from the Department of Industrial Systems Engineering and Management, National University of Singapore, Singapore, in 2022.

He is currently a Postdoctoral Research Associate with the Intelligent Transmission and Processing Laboratory, Imperial College London, London, United Kingdom, from May 2023. He was a Postdoctoral Research Fellow with the Institute of Data Science, National University of Singapore, Singapore, from March 2022 to March 2023.

His research interest includes statistics and optimization theories with applications in signal processing (especially optimal estimation theory), machine learning (especially generalization error theory), and control technology.

Wei Dai (Member, IEEE) received the Ph.D. degree from the University of Colorado Boulder, Boulder, Colorado, in 2007. He is currently a Senior Lecturer (Associate Professor) in the Department of Electrical and Electronic Engineering, Imperial College London, London, UK. From 2007 to 2011, he was a Postdoctoral Research Associate with the University of Illinois Urbana-Champaign, Champaign, IL, USA. His research interests include electromagnetic sensing, biomedical imaging, wireless communications, and information theory.

Geoffrey Ye Li is currently a Chair Professor at Imperial College London, UK. Before joining Imperial in 2020, he was a Professor at Georgia Institute of Technology for 20 years and a Principal Technical Staff Member with AT&T Labs – Research (previous Bell Labs) for five years. He made fundamental contributions to orthogonal frequency division multiplexing (OFDM) for wireless communications, established a framework on resource cooperation in wireless networks, and introduced deep learning to communications. In these areas, he has published over 700 journal and conference papers in addition to over 40 granted patents. His publications have been cited around 80,000 times with an H-index over 130. He has been listed as a Highly Cited Researcher by Clarivate/Web of Science almost every year.

Dr. Geoffrey Ye Li was elected to Fellow of the Royal Academic of Engineering (FREng), IEEE Fellow, and IET Fellow for his contributions to signal processing for wireless communications. He received 2024 IEEE Eric E. Sumner Award, 2019 IEEE ComSoc Edwin Howard Armstrong Achievement Award, and several other awards from IEEE Signal Processing, Vehicular Technology, and Communications Societies.