Multi-source Transfer Learning for Signal Detection over a Fading Channel with Co-channel Interference

Ziyan Zheng*, Xinyi Tong*, Xinchun Yu*, Xiangxiang Xu†‡ and Shao-Lun Huang*§
* Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, China
† Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA

Abstract—**For signal detection tasks in wireless communications, most of the existing algorithms either ignore the co-channel interference or treat it as Gaussian noise, which may result in unsatisfactory accuracy when the interference is non-negligible and complex-distributed. When neural networks are motivated in the design of the detectors, difficulty arises in the training due to the fact that there are few accessible pilots in each packet. In this paper, we consider a data-driven detector based on multi-source transfer learning (MSTL) for signal detection in a fading channel with interference. The MSTL detector transfers channel knowledge of previous packets into the latest detection. In particular, we consider a linear combination of the pilots and historical symbols in the distribution space, and design the optimal combination coefficients based on the number of those symbols as well as distributions similarity. Numerical simulations on a Gauss-Markov flat Rayleigh fading channel with co-channel interference validate the advantages of our algorithms, compared with several existing training schemes including directly applying the fully connected deep neural network (FCDNN) and conventional linear minimum mean square error (LMMSE) detector.**

Index Terms—**Signal detection, fading channel, co-channel interference, multi-source transfer learning, neural network.**

I. Introduction

Signal detection is a classical hypothesis testing problem, which is applied in many areas, such as radar and wireless communication systems. Many works have been conducted on signal detection algorithms (e.g. [1], [2]), among which the maximum likelihood (ML) detector can achieve the optimality by minimizing the probability of detection error. However, because of the difficulty in the implementation of the ML detector in complicated scenarios, researchers have considered sub-optimal linear detectors, such as LMMSE detector and Kalman filter, which have lower computational complexity at the cost of performance degradation [3]. Unfortunately, such standard model-based approaches are inapplicable if (i) a precise channel model is unavailable, or (ii) the optimal receiver for the given transmission scheme and channel is prohibitively complicated or unknown [4]. Examples for both cases can be found in new communication set-ups (e.g., molecular channels lacking well-established models), links with strong nonlinearity (e.g., satellite channels with non-linear transceivers) and channels with non-negligible complex co-channel interference [5]–[7], where the latter one is mainly focused on in this paper. In practice, co-channel interference is inevitable in many wireless communication systems due to spectrum sharing and multiple access [8]. It is mainly caused by multiple radios transmitting on the same frequency band at the same time, and can probably degrade the communications performance. In most previous works, co-channel interference is usually either not studied or simply treated as Gaussian noise. Although it was proven in [9] that weak interference can be treated as Gaussian noise to achieve satisfactory performance, strong interference still requires sophisticated designs of the receiver to mitigate its influence in conventional approaches [10].

In the design of detectors, the absence of capability of capturing complex distributions of the received signals in traditional techniques has motivated deployment of machine learning methods, especially deep learning (e.g., [11]) which regards detection problems as classification tasks, where the received signals are input into the classifier, and the predicted transmitted messages are output. Benefiting from the multi-layer structure with non-linear operations, deep neural network (DNN) is powerfully capable of pattern recognition and feature extraction, resulting in high performance for analyzing complex-distributed data. In wireless communications, probing sequences or pilots known by the receiver are inserted into the transmitted signal packets to feed the detector, and can hence be utilized as training data to conduct the supervised learning. Layers of the DNN can be optimized off-line to provide accurate and cheap decision rules applied in real time. Several DNN-based signal detectors have been proposed to produce significant results. For example, a FCDNN was designed to estimate CSI in [11], and deep learning for decoding without CSI is developed in [12].

In general, the more complex the underlying non-Gaussian distribution is, the more samples are needed for training a DNN. In a time-varying channel with non-Gaussian-distributed co-channel interference, lack of pilots makes it difficult for DNN to capture the complex underlying distribution. Therefore, we consider training a detector based not only on the pilots, but also on the historical transmitted and received signals, without any direct estimation of CSI. Since the channel state is time-varying, only partial information of previous packets is helpful in the latest detection task, i.e., the joint distribution of transmitted and received symbols at different time slots may be different but related to some extent. To extract the useful information, transfer learning is served as a powerful tool to assist in overcoming such difficulties.

‡ Work performed at Tsinghua-Berkeley Shenzhen Institute

§ Corresponding author

As an extension of conventional deep learning, MSTL has recently been an active research area. It focuses on applying the stored knowledge of several problems (sources) in solving a new problem (target). To solve the target task, we not only use a few accessible training samples of the target task, but also use the knowledge of source tasks to help improve the target performance. In our detection over a fading channel with co-channel interference, the usefulness of the historical signals is affected by several factors including (i) the similarity among the CSI of the varying channel, (ii) the number of available training samples in each packet, (iii) the complexity or dimensionality of the DNN model. Most of the existing transfer strategies are designed based on how similar the source and target tasks are [13], [14], without considering the impacts of the training sample sizes or the complexity of the models. In this paper, taking channel similarities and the sample sizes into account, we firstly introduce a practical mathematical framework following [15] to investigate the similarity among historical and the latest joint distributions of the transmitted and received signals, and establish a quantifiable measure for knowledge transfer, which then provides a theoretical guidance for the design of our detector. Performances are finally evaluated in a Gauss-Markov flat Rayleigh fading channel in the presence of co-channel interference.

The main contributions of this work are summarized as follows:

- We adopt a mathematical framework of MSTL and develop a detection algorithm for time-varying channels. Our MSTL-based detector involves pilots and much previous data in the training scheme to mitigate the effect of interference, without the direct estimation for CSI.
- Unlike most of the previous works that either ignored interference or treated interference as Gaussian noise, non-Gaussian-distributed co-channel interference is taken into consideration in the channel model.
- The proposed MSTL detector is implemented in the signals stream over a simulated flat fading channel in the presence of co-channel interference, in which we show that our algorithm outperforms both the traditional method and conventional neural networks without the design of knowledge transfer.

II. Model and Theory

This section formulates the problem and introduces the theoretical background for our detector.

A. Problem formulation

We model a time-varying fading channel with co-channel interference as

$$Y_t = H_t X_t + I_t + n_t, \tag{1}$$

where X_t is the transmitted symbol, Y_t the received symbol, H_t the time-varying channel state, I_t the co-channel interference caused by other signals transmitted in the same frequency band, and n_t the additive white Gaussian noise (AWGN) indexed with t. Here $t \in \mathbb{Z}$ implies the time in the data stream and represents the packet index. Note that each packet contains many symbols, the first several ones among which are pilots. The transmitted symbol X_t is from a discrete constellation set, and the dynamic model that H_t varies with respect to t is not specified. In addition, H_t, X_t, I_t and n_t are assumed to be independent from each other.

Our goal is to detect the transmitted signal X_t using the received signal Y_t subject to interference I_t, fading channel H_t and noise n_t at each latest time t. We assume that the detection task is currently implemented at the latest time labeled $T \in \mathbb{Z}$. Then our task is to detect transmitted signal X_T from the observed received signal Y_T. For previous time $t = 1, 2, ..., T-1$, N_t preserved samples extracted from each t are available for training, which we call source samples. N_T pilots at time T, which we call target samples, are accessible. Note that $N_T \ll N_t \, (\forall t < T)$ since previous detected signal pairs preserved can be utilized in the detection for the current packet.

B. Framework of Multi-source transfer learning

In this part, the problem above is embedded in the framework of MSTL. Firstly, the received symbol Y_t as a variable is assumed to be discrete, and will be extended to the continuous case later. The domains of X_t and Y_t are denoted as $\mathcal{X}$ and $\mathcal{Y}$, respectively. We let $\mathcal{P}$ denote the set of all distributions on $\mathcal{X} \times \mathcal{Y}$. For each time $t = 1, 2, ..., T$, we assume that N_t training samples $\left\{\left(x_t^{(i)}, y_t^{(i)}\right)\right\} (1 \le i \le N_t)$ are i.i.d. generated from some underlying joint distribution $P_{X_tY_t} \in \mathcal{P}$ with [1] $P_{X_tY_t}(x, y) > 0$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and the empirical distributions $\hat{P}_{X_tY_t} \in \mathcal{P}$ of the samples are defined as

$$\hat{P}_{X_tY_t}(x, y) \triangleq \frac{1}{N_t} \sum_{i=1}^{N_t} 1\left\{x_t^{(i)} = x, y_t^{(i)} = y\right\}, \tag{2}$$

where $1\{\cdot\}$ denotes the indicator function. Therefore $P_{X_tY_t}$ can be seen as the underlying joint distribution of transmitted signal x and received signal y at time t, while $\hat{P}_{X_tY_t}$'s are the learned model for $1 \le t \le T-1$. Since we only have few pilots that are accessible at time T, empirical distribution $\hat{P}_{X_TY_T}$ captured from pilots could be less possible to fit the true distribution $P_{X_TY_T}$, which makes it hard to detect the up-coming X_T from Y_T.

In order to utilize historical samples before time T, we consider the convex combination[2] learned from reserved source samples and pilots

$$Q_{X_TY_T}^{(\boldsymbol{w})} \triangleq \sum_{t=1}^{T} w_t \hat{P}_{X_tY_t}, \tag{3}$$

where the combining weight vector $\boldsymbol{w}$ is defined as

$$\boldsymbol{w} \in \{(w_1, w_2, ..., w_T) : \sum_{t=1}^{T} w_t = 1, w_t \ge 0\}. \tag{4}$$

[1] In practice, such joint distributions are usually modeled by some positive parameterized families, e.g., the softmax function. Therefore the positive assumption is without loss of generality.

[2] Such combination form follows [15].

Note that the combining vector $\boldsymbol{w}$ characterizes the knowledge transferred from each sources $1 \leq t \leq T-1$ to the target task $t = T$, and the designing of $\boldsymbol{w}$ should take the sample sizes and the task similarities into consideration.

Then, the performance of our model $Q^{(\boldsymbol{w})}_{X_T Y_T}$ is evaluated by the testing loss, which is measured by its empirical risk on the testing data of the latest task. The testing loss $L^{(\boldsymbol{w})}_{\text{test}}$ and the corresponding optimal combining coefficients $\boldsymbol{w}^*$ are defined as

$$L^{(\boldsymbol{w})}_{\text{test}} \triangleq \mathbb{E}\left[\chi^2\left(P_{X_T Y_T}, Q^{(\boldsymbol{w})}_{X_T Y_T}\right)\right], \tag{5}$$

$$\boldsymbol{w}^* \triangleq \underset{\boldsymbol{w}}{\arg\min}\ L^{(\boldsymbol{w})}_{\text{test}}. \tag{6}$$

Here the referenced χ^2-distance is applied as the measure, which is introduced as follows:

Definition 1. *For random variable X and Y, given a reference distribution R_{XY} for any distribution P_{XY} and Q_{XY}, the referenced χ^2-distance between them is defined as*

$$\chi^2_R\left(P_{XY}, Q_{XY}\right) \triangleq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{\left(P_{XY}\left(x,y\right) - Q_{XY}\left(x,y\right)\right)^2}{R_{XY}\left(x,y\right)}. \tag{7}$$

In particular, $\chi^2_R\left(P_{XY}, Q_{XY}\right)$ becomes Pearson χ^2-distance denoted by $\chi^2\left(P_{XY}, Q_{XY}\right)$ when $R_{XY} = P_{XY}$.

The following proposition reveals the factors affecting the performance of the knowledge transfer in detections.

Proposition 1 ([15], Theorem 3)**.** *The testing loss (5) for the target task is*

$$L^{(\boldsymbol{w})}_{test} = \chi^2\left(P_{X_T Y_T}, \sum_{t=1}^{T} w_t P_{X_t Y_t}\right) + \sum_{t=1}^{T} \frac{w_t^2}{N_t} V_t, \tag{8}$$

where V_t is defined as

$$V_t \triangleq \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{P_{X_t Y_t}\left(x,y\right)\left(1 - P_{X_t Y_t}\left(x,y\right)\right)}{P_{X_T Y_T}\left(x,y\right)}. \tag{9}$$

From (8) and the fact that $V_T = |\mathcal{X}||\mathcal{Y}| - 1$, the usefulness of previous signals is determined by the following three key factors: (i) the channel similarity measured by the χ^2-divergence $\chi^2\left(P_{X_T Y_T}, \sum_{t=1}^{T} w_t P_{X_t Y_t}\right)$, (ii) the number of pilots and previous signals for training, i.e., N_t for $t = 1, 2, ..., T$ and (iii) the complexity of the DNN characterized by the domain of the signals needed to capture, i.e., $|\mathcal{X}||\mathcal{Y}| - 1$ in V_T.

However, in our signal detection problem, the received signal Y_t may be continuous, which leads to the infinite cardinality $|\mathcal{Y}|$. In order to apply the framework, we adopt a parameterized representation for modeling features of the continuous-distributed received symbols.

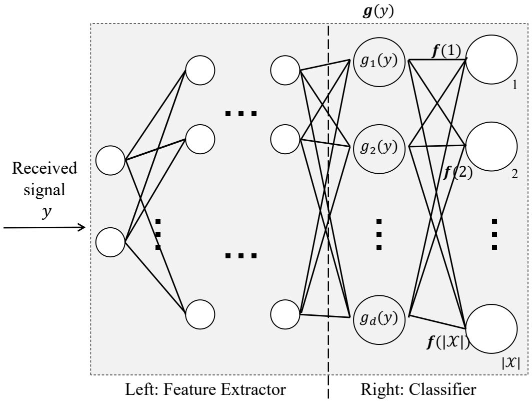

Fig. 1. A deep neural network for classification

As shown in Fig. 1, a DNN can be divided into two parts. The left part is the previous layers for extracting d-dimension features $\boldsymbol{g}(y) = [g_1(y), \cdots, g_d(y)]^T$ from the received symbol y, and the right part is the topmost layer for linear classification, with weights $\boldsymbol{f}(x) = [f_1(x), \cdots, f_d(x)]^T$ indexed by symbol x. When the features $\boldsymbol{g}(y)$ is given and fixed, the models learned from source and target tasks can be effectively represented by a finite collection of parameters, i.e., $\boldsymbol{f}(1), \cdots, \boldsymbol{f}(|\mathcal{X}|)$, where the class of discovered signal is represented from 1 to $|\mathcal{X}|$.

In particular, we adopt the discriminative model in the factorization form[3]

$$\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}\left(x|y\right) \triangleq P_{X_T}\left(x\right)\left(1 + \boldsymbol{f}^T\left(x\right)\boldsymbol{g}\left(y\right)\right). \tag{10}$$

When we train the model for target time T, we learn the corresponding weights $\hat{\boldsymbol{f}}_t$ for each source time t and update the model $\tilde{P}^{(\hat{\boldsymbol{f}}_t,\boldsymbol{g})}_{X_T|Y_T}$ to fit the training samples , where the weight $\hat{\boldsymbol{f}}_t$ is defined as

$$\hat{\boldsymbol{f}}_t \triangleq \underset{\boldsymbol{f}}{\arg\min}\ \chi^2_{R_{XY}}\left(\hat{P}_{X_t Y_t}, P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}\right). \tag{11}$$

For convenience, we use a unified reference $R_{XY} \triangleq P_{X_T Y_T}$ here. The referenced χ^2-distance measures the fitness between the empirical distribution $\hat{P}_{X_t Y_t}$ and the joint distribution $P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}$ [4]. From (11) we know that $P_{Y_T}\hat{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}$ plays the role in the continuous case corresponding to $\hat{P}_{X_T Y_T}$ in the discrete case. Then we form the convex combination of discriminative models $\tilde{P}^{(\hat{\boldsymbol{f}}_t,\boldsymbol{g})}_{X_T|Y_T}$ as

$$Q^{(\boldsymbol{w})}_{X_T|Y_T} \triangleq \sum_{t=0}^{T} w_t \tilde{P}^{(\hat{\boldsymbol{f}}_t,\boldsymbol{g})}_{X_T|Y_T} = \tilde{P}^{(\hat{\boldsymbol{f}},\boldsymbol{g})}_{X_T|Y_T}, \tag{12}$$

with $\hat{\boldsymbol{f}} \triangleq \sum_{t=1}^{T} w_t \hat{\boldsymbol{f}}_t$. Then the testing loss corresponding to (5) becomes

$$L^{(\boldsymbol{w})}_{\text{test}} \triangleq \mathbb{E}\left[\chi^2_{R_{XY}}\left(P_{X_T Y_T}, P_{Y_T} Q^{(\boldsymbol{w})}_{X_T|Y_T}\right)\right], \tag{13}$$

[3] Such factorization form is similar to the theory proposed in [16], and is also applicated in natural language processing [17].

[4] The joint distribution $P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}$ is defined as
$\left[P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}\right](x,y) \triangleq P_{Y_T}\left(y\right)\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}\left(x|y\right)$ for all (x,y).
Note that $P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}$ corresponds to the optimal approximation of the target distribution $P_{X_T Y_T}$ when the discriminative model $\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}$ is fixed.

and the optimal combining coefficients are found by

$$\boldsymbol{w}^* \triangleq \arg\min_{\boldsymbol{w}} L_{\text{test}}^{(\boldsymbol{w})}. \tag{14}$$

The following characterization investigates the testing loss in this case.

Proposition 2 ([15], Theorem 4)**.** *The testing loss (13) associated with the model (12) is*

$$\begin{aligned} L_{\textit{test}}^{(\boldsymbol{w})} &= \chi^2_{R_{XY}}\left(P_{Y_T}\tilde{P}^{(\boldsymbol{f}_T,\boldsymbol{g})}_{X_T|Y_T}, \sum_{t=1}^{T} w_t \tilde{P}^{(\boldsymbol{f}_t,\boldsymbol{g})}_{X_T|Y_T}\right) \\ &+ \sum_{t=1}^{T} \frac{w_t^2}{N_t}\tilde{V}_t + \chi^2_{R_{XY}}\left(P_{X_TY_T}, P_{Y_T}\tilde{P}^{(\boldsymbol{f}_T,\boldsymbol{g})}_{X_T|Y_T}\right), \end{aligned} \tag{15}$$

where $\boldsymbol{f}_t \triangleq \arg\min_{\boldsymbol{f}} \chi^2_{R_{XY}}\left(P_{X_tY_t}, P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}\right)$ *and* $\tilde{V}_t$ *is a constant independent of* $\boldsymbol{w}$*.*

It could be observed in (15) that $\boldsymbol{w}^*$ can be efficiently computed by solving a non-negative quadratic programming problem.

III. Algorithm for MSTL Detector

In this part, we develop a signal detection algorithm for application based on the theory above. According to Proposition 2, our algorithm optimizes the extracted features $\boldsymbol{g}$, the weights $\boldsymbol{f}$ and the combining coefficients $\boldsymbol{w}$ to achieve better performance.

For given $\boldsymbol{f}$, $\boldsymbol{g}$ and $\boldsymbol{w}$, the loss function is proposed as

$$L^{(\boldsymbol{w},\boldsymbol{f},\boldsymbol{g})} \triangleq \sum_{t=1}^{T} w_t \chi^2_{R_{XY}}\left(\hat{P}_{X_tY_t}, P_{Y_T}\tilde{P}^{(\boldsymbol{f},\boldsymbol{g})}_{X_T|Y_T}\right). \tag{16}$$

The following proposition shows that the estimated $\hat{\boldsymbol{f}}$ can be computed by directly minimizing the loss without evaluating each $\hat{\boldsymbol{f}}_t$ individually.

Proposition 3 ([15], Proposition 5)**.** *The* $\hat{\boldsymbol{f}}$ *defined in (12) satisfies*

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f}'} \ L^{(\boldsymbol{w},\boldsymbol{f}',\boldsymbol{g})}. \tag{17}$$

Then, with training samples from time 1 to $T-1$ and pilots obtained from time T, our algorithm consists of two different kinds of optimizations: (i) the optimization of $\boldsymbol{w}$ for given $(\boldsymbol{f},\boldsymbol{g})$ to minimize the testing loss $L_{\text{test}}^{(\boldsymbol{w})}$ defined by (15), via dealing with a non-negative quadratic programming problem; and (ii) the optimization of $(\boldsymbol{f},\boldsymbol{g})$ for given $\boldsymbol{w}$ to minimize the loss $L^{(\boldsymbol{w},\boldsymbol{f},\boldsymbol{g})}$ defined in (16), via training the DNN. We summarize the procedures as Algorithm 1.

With the trained DNN by Algorithm 1, for an up-coming observed received signal y at time T, the predicted signal $\hat{x}$ could be determined by the MAP (maximum a posterior) decision rule

$$\begin{aligned} \hat{x}(y) &= \arg\max_{x\in\mathcal{X}} \ \tilde{P}^{(\boldsymbol{f}^*,\boldsymbol{g}^*)}_{X_T|Y_T}(x|y) \\ &= \arg\max_{x\in\mathcal{X}} \ P_{Y_T}(y)\left(1 + \boldsymbol{f}^{*T}(x)\,\boldsymbol{g}^*(y)\right). \end{aligned} \tag{18}$$

Algorithm 1 Algorithm for training the MSTL detector

```
Input:
    Historical samples $\left\{\left(x_t^{(i)} y_t^{(i)}\right)\right\}_{i=1}^{N_t}$ $(t = 1, 2, ..., T-1)$
    Pilots $\left\{\left(x_T^{(i)} y_T^{(i)}\right)\right\}_{i=1}^{N_T}$
    Randomly initialize $\boldsymbol{w}^*$ and DNN parameters
Output:
repeat
    $(\boldsymbol{f}^*, \boldsymbol{g}^*) \leftarrow \arg\min_{\boldsymbol{f},\boldsymbol{g}} L^{(\boldsymbol{w}^*,\boldsymbol{f},\boldsymbol{g})}$
    $\boldsymbol{w}^* \leftarrow \arg\min_{\boldsymbol{w}} L_{\text{test}}^{(\boldsymbol{w})}$
until $\boldsymbol{w}^*$ converges
$(\boldsymbol{f}^*, \boldsymbol{g}^*) \leftarrow \arg\min_{\boldsymbol{f},\boldsymbol{g}} L^{(\boldsymbol{w}^*,\boldsymbol{f},\boldsymbol{g})}$
return $\boldsymbol{f}^*, \boldsymbol{g}^*$;
```

IV. Simulation

In this section, we conduct our multi-source transfer learning algorithm on the simulated time-varying flat Rayleigh fading channel.

A. Environment design

Following the channel model in the works [7], [18], the distributions of the variables in (1) are determined as following.

1. The dynamics of the time-varying channel state H_t are modeled by a first-order Gauss-Markov process[5]

$$H_t = aH_{t-1} + u_t, \quad u_t \sim \mathcal{CN}\left(0, \left(1-a^2\right)\sigma_H^2\right), \tag{19}$$

where u_t is the white Gaussian driving noise. Parameter $a \in [0,1]$ is the fading correlation coefficient that describes the degree of time variation. Small a corresponds to fast fading, and large a represents slow fading. The value of a could be affected by Doppler spread and the transmission bandwidth [22], and is assumed to be unknown for our detection.

2. The modulated signal $X_t = K_t + jQ_t$ can be described by the in-phase and quadrature components as

$$A\cos\left(2\pi v l + \phi_t\right) = \operatorname{Re}\left[\left(K_t + jQ_t\right)e^{j2\pi v l}\right], \tag{20}$$

where ϕ_t is the initial phase at time t[6], A and v are the amplitude and frequency respectively. K_t is the in-phase component and Q_t is the quadrature component at time t. (K_t, Q_t) can also be treated as the coordinates of a point on the constellation diagram determined by both phase and amplitude to give the modulation scheme. In our simulation, we consider Quadratic Phase shift keying (QPSK) as modulation scheme for the transmitted signal. In QPSK modulation, the transmitter transmits cosinusoids with a orthogonal phase, which we set $\phi_t = \frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}$ here to represent bit pairs 00, 01, 11 and 10 with a constant amplitude $A = \sqrt{2}$, resulting in points $(K_t, Q_t) \in \{(1,1), (-1,1), (-1,-1), (1,-1)\}$ on the constellation.

[5] The Gauss-Markov model is widely adopted as a simple but effective model to characterize the fading process [19]–[22].

[6] Here, t refers to the time varying the channel state, while l refers to the abscissa of the cosine wave transmitted.

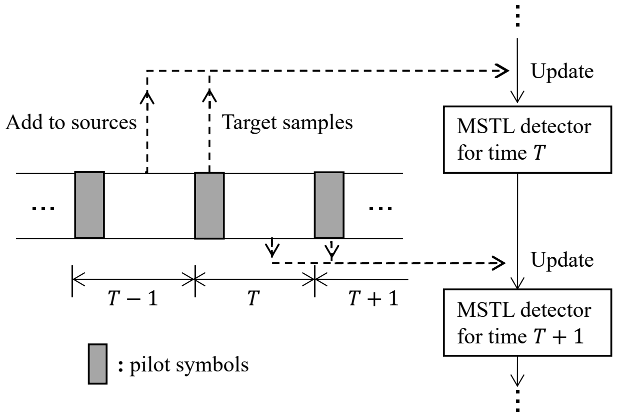

Fig. 2. The update process of the detector

3. Co-channel interference I_t is caused by other signals transmitted at the same time and in the same frequency band. Here we model I_t as QPSK signals with stronger power, i.e., $I_t \in \{q+jq, -q+jq, -q-jq, q-jq\}$ for some constant q, which is similar to the design of interference in [23].

4. Additive white Gaussian noise n_t has zero mean and covariance σ_n^2 for both real and imaginary components.

B. Implementation and Results

The simulation is conducted on binary bits stream with periodic placements of pilots. The binary bits are firstly encoded by a convolutional code with rate $\frac{1}{3}$ and modulated into QPSK symbols. which are then transmitted following the channel model (1). The received symbols are demodulated by the detector and then decoded to recover the original bits stream. The accuracy is performed through bit error rate (BER). Considering the limited memory space of the detector and the effectiveness of knowledge transfer, we simply include samples from the past time of length 9, i.e. samples from time $T-9$ to time $T-1$, as source samples for training at time T, and pilots from time T are trained as target samples. When the detection tasks on time T are completed, the sliding windows move forward, involving several samples from time T in the sources and dismissing the ones from time $T-9$, and use pilots from time $T+1$ as new target samples. The update process of the detector is described in Fig. 2. The performance of our MSTL detector and those for comparison is evaluated on the average of the moving-forward sliding windows.

Assume that the sliding windows are currently detecting the signals at time T. The following detectors are considered as comparison:

FCDNN-1 Conventional FCDNN based on L_2 loss[7], using N_T pilots for training.

FCDNN-2 Conventional FCDNN based on L_2 loss, using all the N_T pilots and $\sum_{t=T-9}^{T-1} N_t$ reserved samples from time $T-9$ to $T-1$ for training. Note that those samples are consistently extracted for our MSTL detector.

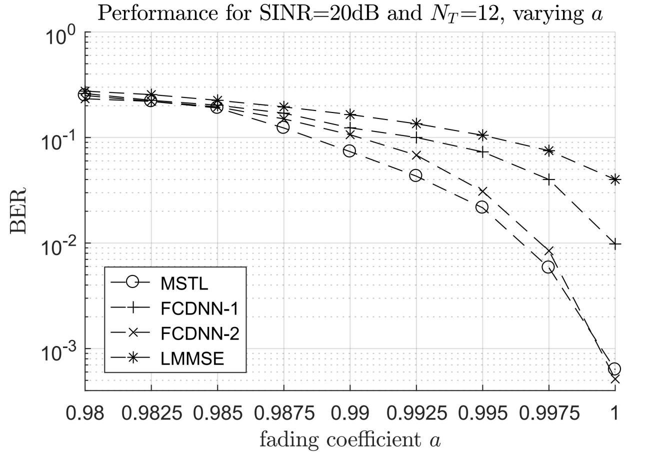

Fig. 3. Performance of the detectors with different fading coefficient a when SINR= 20dB and $N_T = 12$.

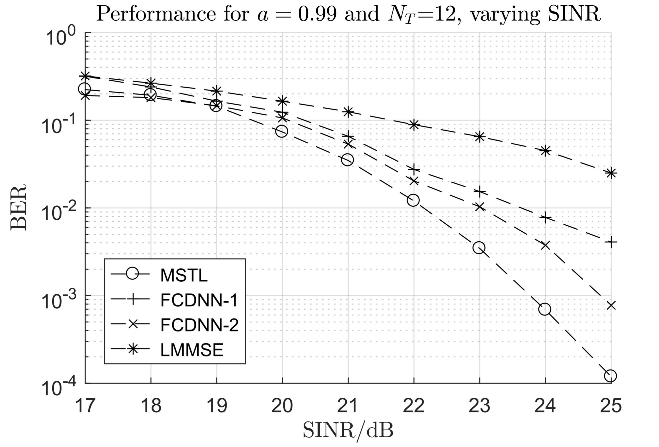

Fig. 4. Performance of the detectors with different SINR when $a = 20$dB and $N_T = 12$.

LMMSE LMMSE detector is one of the most widely-used and straightforward detector [25]. We use LMMSE detector for comparison, representing a class of detectors estimating the channel state but hardly recognizing the non-Gaussian noise.

To measure the interference level of the system, the Signal to Interference plus Noise Ratio (SINR) at some time t is defined as

$$\mathrm{SINR} \triangleq 10 \lg \frac{\sigma_H^2 |X_t|^2}{\sigma_n^2}, \tag{21}$$

and $q \propto \sigma_n$ for the consistency of power of the interference and noise. For each source, 200 pairs of transmitted and received signals are used for our training, i.e. $N_t = 200$ for $t = T-1, T-2, ..., T-9$. To validate the accuracy of detectors, we vary (i) the fading coefficient a, (ii) the number of pilots N_T, and (iii) SINR respectively. In each simulation, one of the factor is varying and the other two factors remain constant.

It is observed in Fig. 3 that the largest gain obtained using MSTL detector occurs when the fading coefficient a is in the range from 0.9875 to 0.9975. Relatively fast fading (small a) increases the difficulties in knowledge transfer, while the

[7] L_2 loss gives better accuracy than cross-entropy loss in our problem, similar to the result in [24].

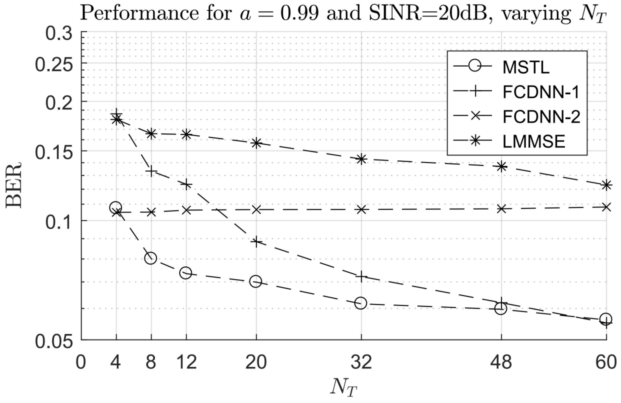

Fig. 5. Performance of the detectors with different number of pilots N_T when $a = 0.99$ and SINR= 20dB.

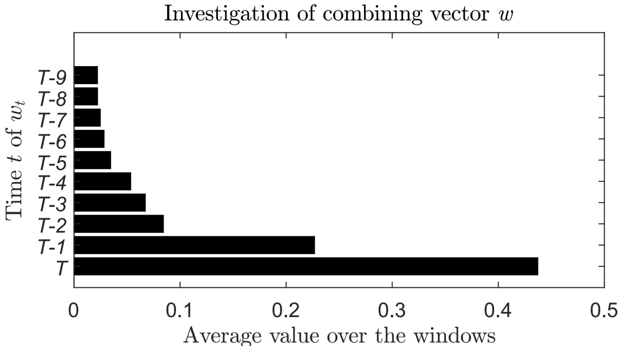

Fig. 6. The value of the combining vector $\boldsymbol{w}$ averaged on the sliding windows.

invariant channel state (i.e., $a = 1$) is naturally favourable for FCDNN-2, in which the invariance of the channel state is regarded as prior knowledge. In addition, Fig. 4 plots BER of the detectors in the variance of SINR, which shows that the BER of the MSTL detector is significantly lower than the other three schemes in the channel with high SINR. Fig. 5 exhibits that the MSTL detector outperforms the other algorithms in a large range of pilots length, while considerable pilots may be sufficient for FCDNN-1 to capture the target distribution simply using pilots for training. Finally, Fig. 6 shows the monotonous similarity of the sources and reveals the validity for the proposed combination in our MSTL framework.

In summary, simulation results demonstrate that significant performance gain can be obtained by deploying the MSTL algorithm in a variety of scenarios.

Acknowledgment

The research of Shao-Lun Huang is supported in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146, National Key RD Program of China under Grant 2021YFA0715202 and High-end Foreign Expert Talent Introduction Plan under Grant G2021032013L.

References

[1] Z. Guo and P. Nilsson, "Algorithm and implementation of the k-best sphere decoding for mimo detection," *IEEE Journal on selected areas in communications*, vol. 24, no. 3, pp. 491–503, 2006.

[2] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser mimo-ofdm systems using approximate message passing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 902–915, 2014.

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

[4] S. Park, H. Jang, O. Simeone, and J. Kang, "Learning to demodulate from few pilots via offline and online meta-learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 226–239, 2020.

[5] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 648–664, 2018.

[6] S. Bouchired, D. Roviras, and F. Castanié, "Equalisation of satellite mobile channels with neural network techniques," *Space Communications*, vol. 15, no. 4, pp. 209–220, 1998.

[7] C. Liu, Y. Chen, and S.-H. Yang, "Signal detection with co-channel interference using deep learning," *Physical Communication*, vol. 47, p. 101343, 2021.

[8] S. Catreux, P. F. Driessen, and L. J. Greenstein, "Attainable throughput of an interference-limited multiple-input multiple-output (mimo) cellular system," *IEEE transactions on communications*, vol. 49, no. 8, pp. 1307–1311, 2001.

[9] V. S. Annapureddy and V. V. Veeravalli, "Gaussian interference networks: Sum capacity in the low-interference regime and new outer bounds on the capacity region," *IEEE Transactions on Information Theory*, vol. 55, no. 7, pp. 3032–3050, 2009.

[10] A. Carleial, "A case where interference does not reduce capacity (corresp.)," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 569–570, 1975.

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

[12] N. Farsad and A. Goldsmith, "Neural network detection of data sequences in communication systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 21, pp. 5663–5678, 2018.

[13] S. Thrun, L. K. Saul, and B. Schölkopf, *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, vol. 16. MIT press, 2004.

[14] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*, pp. 270–279, Springer, 2018.

[15] X. Tong, X. Xu, S.-L. Huang, and L. Zheng, "A mathematical framework for quantifying transferability in multi-source transfer learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[16] S. Rendle, "Factorization machines," in *2010 IEEE International conference on data mining*, pp. 995–1000, IEEE, 2010.

[17] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," *Advances in neural information processing systems*, vol. 27, pp. 2177–2185, 2014.

[18] M. Dong, L. Tong, and B. M. Sadler, "Optimal insertion of pilot symbols for transmissions over time-varying flat fading channels," *IEEE Transactions on Signal Processing*, vol. 52, no. 5, pp. 1403–1418, 2004.

[19] G. A. Arredondo, W. H. Chriss, and E. H. Walker, "A multipath fading simulator for mobile radio," *IEEE Transactions on Vehicular Technology*, vol. 22, no. 4, pp. 241–244, 1973.

[20] R. A. Iltis, "Joint estimation of pn code delay and multipath using the extended kalman filter," *IEEE Transactions on communications*, vol. 38, no. 10, pp. 1677–1685, 1990.

[21] M. Stojanovic, J. G. Proakis, and J. A. Catipovic, "Analysis of the impact of channel estimation errors on the performance of a decision-feedback equalizer in fading multipath channels," *IEEE Transactions on Communications*, vol. 43, no. 2/3/4, pp. 877–886, 1995.

[22] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Transactions on Information theory*, vol. 46, no. 3, pp. 933–946, 2000.

[23] D. Qiu, *Representation and transfer learning using information-theoretic approximations*. PhD thesis, Massachusetts Institute of Technology, 2020.

[24] H. Kim, Y. Jiang, R. Rana, S. Kannan, S. Oh, and P. Viswanath, "Communication algorithms via deep learning," *arXiv preprint arXiv:1805.09317*, 2018.

[25] S. Verdu *et al.*, *Multiuser detection*. Cambridge university press, 1998.