A Learning-Based Two-Stage Spectrum Sharing Strategy With Multiple Primary Transmit Power Levels

Rui Zhang , Peng Cheng , *Member, IEEE*, Zhuo Chen , *Senior Member, IEEE*, Yonghui Li , *Fellow, IEEE*, and Branka Vucetic*, Fellow, IEEE*

Abstract—Multi-parameter cognition in a cognitive radio network (CRN) provides a more thorough understanding of the radio environments, and could potentially lead to far more intelligent and efficient spectrum usage for a secondary user. In this paper, we investigate the multi-parameter cognition problem for a CRN where the primary transmitter (PT) radiates multiple transmit power levels, and propose a learning-based two-stage spectrum sharing strategy. We first propose a data-driven/machine learning based multi-level spectrum sensing scheme, including the spectrum learning (Stage I) and prediction (the first part in Stage II). This fully blind sensing scheme does not require any prior knowledge of the PT power characteristics. Then, based on a novel normalized power level alignment metric, we propose two prediction-transmission structures, namely periodic and non-periodic, for spectrum access (the second part in Stage II), which enable the secondary transmitter (ST) to closely follow the PT power level variation. The periodic structure features a fixed prediction interval, while the non-periodic one dynamically determines the interval with a proposed reinforcement learning algorithm to further improve the alignment metric. Finally, we extend the prediction-transmission structure to an online scenario, where the number of PT power levels might change as a consequence of PT adapting to the environment fluctuation or quality of service variation. The simulation results demonstrate the effectiveness of the proposed strategy in various scenarios.

Index Terms—Cognitive radio, multiple primary transmit power levels, machine learning.

I. INTRODUCTION

A. Cognitive Radio

THE emerging new wireless technologies, such as 5G cellular networks and machine-to-machine enabled industrial Internet of Things, are fueling an ever-increasing demand for access to the radio frequency spectrum. Cognitive radio (CR), an intelligent wireless technology able to recognize the surrounding radio environments [1], creates a potential communication paradigm to achieve more efficient and flexible spectrum usage. A secondary user (SU) with CR capability monitors the spectrum utilization of a primary user (PU) and determines its access to such spectrum accordingly. Two fundamental challenges arise in the process: how to explore possible spectrum opportunities (spectrum sensing) and how to exploit such opportunities efficiently (spectrum access).

Spectrum sensing measures and percepts the surrounding radio spectrum state based on various signal processing methods, including matched filter detection [2], [3], cyclostationary detection [4], [5], and energy detection [6]–[10]. Matched filter can achieve the optimum performance, but the SU requires perfect knowledge of the PU signaling features *a priori*. Cyclostationary detection takes advantage of the signal cyclostationary feature to distinguish signals from the stationary noise. In contrast, energy detection carries out the hypothesis test to determine the PU spectrum state, based on the energy of the collected PU signals. It features a low computational complexity, and is widely adopted in the literature. Upon obtaining the radio spectrum state, spectrum access dynamically adjusts the resources available to the SU, including frequency band, transmission time and transmission power, and accesses the licensed spectrum by taking into account the interference to the PU. The SU can access the licensed spectrum either when the PU is idle (opportunistic access) [11], or concurrently with the PU following a power control strategy to constrain the interference to the PU (spectrum sharing) [12], [13]. It is clear that appropriately designed spectrum sharing achieves higher throughput for the SU compared with the opportunistic spectrum access.

It is worth noting that many contemporary wireless standards, such as IEEE 802.11 [14], GSM [15], and LTE [16], have specified multiple transmit power levels to dynamically adapt to the fast changing environment and varying quality of service (QoS). The majority of CR studies in the literature did not take this into account, and the SU usually adopts a binary approach in reporting the radio spectrum state as idle or busy. In fact, for this multiple power level scenario, multi-parameter cognition is required, and the binary approach may not represent the most efficient spectrum utilization for the SU. The question then arises: how to exploit the variation in the PU power levels to design an intelligent spectrum sharing strategy? Naturally the answer to this question shall consist of spectrum sensing and spectrum access, which will be elaborated on in the next two subsections.

Fig. 1. Spectrum learning: from hypothesis test to blind data driven.

Fig. 2. System model for a spectrum sharing CR network.

B. Multi-Level Spectrum Sensing

The priority in the design of an agile spectrum sensing method should aim to accurately map the sensing samples received by the secondary transmitter (ST) to the corresponding primary transmitter (PT) power levels. In a conventional *binary* case (PT is ON/OFF), there are two kinds of errors, and the goal of the spectrum sensing is to determine a detection threshold θ (the upper part of Fig. 1). For example, given the target probability of false alarm and the noise power, θ can be simply determined by the Neyman-Pearson criterion[17]. By contrast, the goal of the *multi-level* case is to jointly determine multiple thresholds $\{\theta_i\}_{i=1}^{L-1}$ to separate L different power levels through a multiple hypothesis test (the middle part of Fig. 1), which is far more complicated than the binary one. In essence, there are $L(L-1)$ kinds of errors, which are intertwined to exacerbate the complexity in threshold calculation. In [18] and [19], the authors proposed an energy detection based multiple hypothesis test to derive the decision thresholds for the multiple power level identification. The results were extended to the scenarios with noise uncertainty [20] and non-Gaussian transmission signals [21]. However, the calculation of the thresholds in [18]–[21] requires a large amount of prior knowledge at the ST, including the noise power and the PT transmit power mode (i.e., the number and exact value of different transmit powers, and the prior probability of each hypothesis). In practice, these parameters are unlikely to be available to the ST *a priori*.

In this paper, we aim to break the limits of the existing work, and achieve multi-level spectrum sensing with no or minimal prior information. We deviate from the above classical signal processing approaches, and machine learning arises as the tool of our choice for knowledge discovery to mine and extract the latent patterns reflecting the PT power level variation in the PT traffic data flow. On this basis, we propose a data-driven/machine learning based multi-level spectrum sensing scheme. It is fully blind in the sense that the ST does not require any prior knowledge of the noise power and the PT transmit power mode.
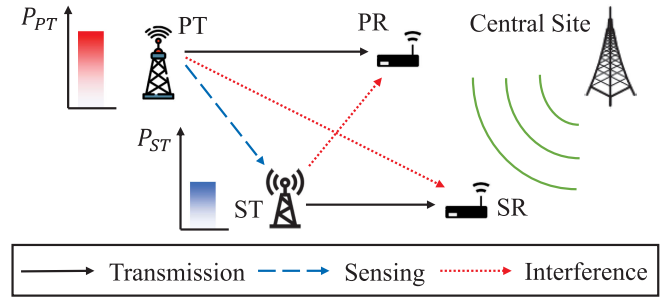
Specifically, the proposed spectrum sensing scheme spans across two stages as shown in Fig. 3. In Stage I (spectrum learning, a.k.a. the training phase in machine learning), the ST collects a multitude of signals that experience multiple PT transmit power level variation, and uses the Gaussian mixture models (GMMs) to capture the multi-level power characteristics inherent in the signals. Then, we introduce a Bayesian nonparametric method, referred to as conditionally conjugate Dirichlet process GMM (DPGMM), to automatically cluster the signals with the same PT transmit power level (the lower part of Fig. 1) and infer the model parameters (GMM parameters and PT power level duration distribution parameters). With the model parameters inferred in Stage I, the prediction part in Stage II (see Fig. 3) can easily identify the current PT power level through collecting PT signal samples. In this case, Stage I together with the prediction part in Stage II achieve fully blind multi-level spectrum sensing. Note that the second part in Stage II (ST transmission) will be detailed next.

C. Multi-Level Spectrum Access

With the big picture of multi-level PT radio environments learned in Section I-B, we next aim to establish a multi-level spectrum access strategy, which is the ultimate goal of spectrum sensing. To the best of our knowledge, this represents the first effort in such direction. For the SU, there is a fundamental tradeoff between two conflicting goals, namely, maximization of its own throughput and minimization of its interference to the PU. As typically there is no cooperation between the PU and SU, it is extremely difficult to optimize this tradeoff in practice. To provide a pragmatic solution to this dilemma, we first propose a new metric, referred to as the normalized power level alignment (NPLA), and it is defined as the time proportion that the ST matches its transmit power level to that of the PT.

To optimize the NPLA performance, we propose two prediction-transmission structures (periodic and non-periodic) in Stage II for spectrum access which enables the ST to closely follow the PT power level variation. As discussed before, the prediction part can identify the current PT power level. On this basis, the transmission part adjusts the ST power according to the required signal-to-interference-plus-noise ratio (SINR) at the primary receiver (PR). Specifically, the periodic structure features a fixed prediction interval, and is straightforward in implementation. By contrast, the non-periodic structure dynamically determines the interval, which can be formulated as a partially observable decision problem. This motivates us to develop a new algorithm based on reinforcement learning [22], exploiting the PT power level duration distribution. This
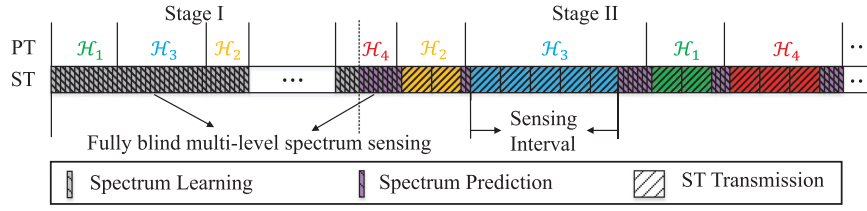
Fig. 3.   The proposed two-stage spectrum sharing strategy. The sensing slots of the ST in both stages have the same time duration T_{ss}. The sensing slots in Stage I are used for learning, while that in Stage II are for prediction. Here, the non-periodic structure is illustrated for Stage II.

structure further improves the NPLA performance. Finally, we extend the prediction-transmission structure to an online scenario, where the number of PT power levels might change as a consequence of PT adapting to the environment fluctuation or QoS variation.

D. Contribution

In a nutshell, we propose a learning-based two-stage spectrum sharing strategy for a CR network, enabling fully blind spectrum sensing when the PT power varies with time in multiple levels, and designs an adaptive spectrum access strategy for the NPLA optimization. The main contributions of this paper can be summarized as follows.

- We propose a novel data-driven/machine learning based multi-level blind spectrum sensing. The conditionally conjugate DPGMM with Bayesian inference is introduced to automatically cluster the signals and infer the model parameters, which is key to predict the PT power levels.
- We propose a new metric, NPLA, to strike an excellent tradeoff between the secondary network throughput and the interference to the primary network.
- To optimize the NPLA performance, we propose a prediction-transmission structure for spectrum access which enables the ST to closely follow the PT power level variation. Furthermore, the ST prediction interval is dynamically adjusted to achieve better performance.
- The spectrum access method is extended to the online scenario to accommodate a more realistic situation, where the number of PT power levels might change after the inital spectrum learning.

E. Related Work

Machine learning technology has recently played an important role in improving spectrum sensing. The work in [23] presented an adversarial machine learning approach to launch jamming attacks on CR and introduces a defense strategy. Several supervised and unsupervised machine learning algorithms for cooperative spectrum sensing (CSS) were investigated in [24]. In [25], the combination of infinite GMM and CSS was proposed to detect the primary user emulation attacks. In [26], a convolutional neural network-based CSS scheme was developed to detect multiple bands simultaneously. A mobile CSS framework was proposed in [27] for large-scale heterogeneous cognitive networks.

There are many efforts in sensing policy design for real-time decisions on which channel(s) to sense (dynamic multichannel selection). By contrast, our paper considers the single user and single channel case, and focuses on the design of multi-level

spectrum sensing to differentiate different PT power levels. On this basis, we also consider the policy design to dynamically adjust the prediction intervals to improve the NPLA performance.

The dynamic multichannel selection can be modelled as a partially observable Markov decision process (POMDP) [28]. The partial observation in [29]–[35] originates from each SU being unable to scan all the channels at any one time due to energy and hardware constraints. Therefore, a sensing policy needs to be developed to balance between utilizing a spectrum opportunity for immediate access and collecting spectrum occupancy statistics to track spectrum opportunity for future exploitation. As finding an optimal channel sensing policy in general is computationally prohibitive with the increased number of channels, several efforts endeavor to find the optimal/near-optimal policy with low computational cost. In [29]–[31], the dynamic multichannel access problem is modelled as a restless multi-armed bandit problem. The time horizon is divided by interleaving exploration and exploitation epochs with growing lengths, and the optimal policy can be translated into determining the length and allocation of each epoch. Recently, deep reinforcement learning (DRL) based channel selection [32]–[35] has attracted great attention, and it aims to handle the correlated channels with unknown channel dynamics. The essence of DRL is to provide a good approximation of the objective value (Q-value), facilitating the handling of the large state and action spaces.

It is worth noting that the access policy design (prediction interval) in our paper is also formulated as a POMDP, but the nature of our formulation is fundamentally different from that in [29]–[35]. The partial observation in our work comes from imperfect multi-level sensing results and access feedback. To tackle this challenging POMDP, we reduce the infinite time horizon to a finite one, leading to a computationally tractable solution. Most importantly, we mathematically prove that such practice does not sacrifice the optimality in the utility.

F. Organization and Notation

The rest of the paper is organized as follows. In Section II, we discuss the system model for our proposed two-stage spectrum sharing strategy. In Section III, we introduce a Bayesian nonparametric method and its inference for the model parameters. The prediction-transmission structure with an online extension, which are adaptive to the PT power level variation, is presented in Section IV. Simulation results and discussion are presented in Section V followed by conclusions in Section VI.

$\mathcal{N}(\mu, S^{-1})$ denotes the Gaussian distribution with mean μ and precision S, $\mathcal{CN}(\mu, S^{-1})$ denotes the complex Gaussian distribution with mean μ and precision S, and $\mathcal{G}(a, b)$ denotes the Gamma distribution with shape parameter a and scale parameter b. $\Gamma(\cdot)$ denotes the Gamma function and $\Gamma(\cdot, \cdot)$ denotes

TABLE I
LIST OF SYMBOLS

Symbol	Definition
l, L	The index and total number of the actual PT transmit power levels.
k, K	The index and total number of the PT transmit power level estimated by the ST.
m, M	The index of the PT hypothesis and the total number of the PT hypotheses in Stage I.
n, N	The index of the ST action and the total number of sensing slots in Stage I.
$P_{PT,l}$	The PT transmission power value on the l-th level.
$\mathcal{H}_l$	The hypothesis that the PT transmits with $P_{PT,l}$.
$\hat{\mathcal{H}}_k$	The hypothesis determined by the ST that the PT transmits with $P_{PT,k}$.
X_n	The test statistic in the n-th single sensing slot.
N_s	The total number of samples collected by the ST in a single sensing slot.
T_{ss}, T_{st}, T_{po}	The duration of a ST sensing slot, a ST transmission block, and a PT hypothesis.
τ_s, τ_p	The discretized time of a ST transmission block and a PT hypothesis.
α, G_0	The concentration parameter and the base probability distribution of the Dirichlet process.
z_n	The latent variable indicating which component that X_n is associated with.
N_k	The total number of observations assigned to the k-th component.
μ_k, S_k, π_k	The value, precision, and mixing proportion of the k-th component in the GMM.
λ, R, W, β	The hyperparameters in the conditionally conjugate DPGMM.
H_{kj}	The probability that the PT is operating under $\mathcal{H}_k$ while the detection by the ST is in favor of $\mathcal{H}_j$.
C_{kj}	The probability that the PT transfers from the k-th transmit power level to the j-th level.
p_τ^k	The probability that the PT keeps operating with the k-th transmit power level at time τ.
a_τ	The action that the ST will take at time τ.
$\mathcal{T}_k$	The longest time that the SU should transmit when operating in the k-th transmit power level.
P_c	The probability of correct PT power level prediction in Stage II.
$U(\tau)$	The NPLA performance from time 0 to τ.

the incomplete Gamma function. $\lfloor \cdot \rfloor$ is the floor function. For convenience, we list most important symbols in Table I.

II. SYSTEM MODEL

We consider a spectrum sharing CR network in Fig. 2, with a primary network consisting of a PT and a PR, a secondary network consisting of a ST and a secondary receiver (SR), and a central site (broadcast tower). Transmission happens simultaneously in both networks sharing the same frequency band. Different from most CR networks considered in the literature, the PT operates with multiple (instead of binary: ON/OFF) power levels. The ST attempts to learn the model parameters, which will be defined later, and then optimize the spectrum access accordingly. The ultimate goal is to optimize the NPLA performance.

To achieve this target, we propose a novel two-stage spectrum sharing strategy, as illustrated in Fig. 3. Let $P_{PT,l}, l = 1, \ldots, L$, be the transmit powers of the PT, where $P_{PT,1} < P_{PT,2} < \cdots < P_{PT,L-1}$ and $P_{PT,L} = 0$ indicates an idle PT. For convenience, hypothesis $\mathcal{H}_l$ indicates that the PT transmits with power $P_{PT,l}$. It is assumed that $P_{PT,l}$ undergoes a slow change, as shown in the figure. We define the time duration of each hypothesis as a random variable T_{po}, which is usually much larger than that of the ST sensing slot T_{ss} and the ST transmission block T_{st}. In this paper, we consider a time discretized model, where T_{ss} is the minimum time unit. We define $\tau_s = T_{st}/T_{ss}$ as the fixed discretized time duration of the ST transmission block, and $\tau_p = T_{po}/T_{ss}$ as the varied discretized PT power level duration. As discussed before, the prior knowledge on the PT transmit power mode, defined as the number of transmit power levels L, the exact values $P_{PT,l}$, and the prior probability of each hypothesis $\Pr\{\mathcal{H}_l\}$, is unknown to the ST, which is fundamentally different from the assumptions in [18]–[21]. In the sequel, we describe in detail the operations of these two stages in Fig. 3.

A. Stage I

In this stage, the ST learns the multi-level power characteristics inherent in the received PT signals. The ST samples the received PT signals at a sampling frequency f_s and collects $N^{(s)}$ samples in each sensing slot with duration T_{ss} (for notation simplicity, we assume that $N^{(s)} = T_{ss} f_s$ is an integer). The ST observes N sensing slots in Stage I and collects a total of $NN^{(s)}$ samples. It is assumed that the learning period is reasonably large so that it covers all L hypotheses.[1] Thus, the i-th sample in the n-th sensing slot under hypothesis $\mathcal{H}_l$ can be given by

$$R_n[i] = \sqrt{P_{PT,l}} s_n[i] + u_n[i], \mathcal{H}_l,$$
$$i = 1, \ldots, N^{(s)}; n = 1, \ldots, N, \quad (1)$$

where $\sqrt{P_{PT,l}} s_n[i]$ is the received primary signal in the n-th sensing slot with average power $P_{PT,l}$, and $u_n[i] \sim \mathcal{CN}(0, \sigma_u^2)$ is the additive white Gaussian noise. Following [36], we assume that $s_n[i]$ is an independent and identically distributed (i.i.d.) random variable with mean 0 and variance 1. Following [36] and without loss of generality, we assume that $s_n[i]$ is complex PSK modulated signal.[2]

The test statistic in the n-th sensing slot can be calculated as

$$X_n = \frac{1}{N^{(s)}} \sum_{i=1}^{N^{(s)}} |R_n[i]|^2, \mathcal{H}_l. \quad (2)$$

When $N^{(s)}$ is large, according to the central limit theorem, the distribution of X_n under hypothesis $\mathcal{H}_l$ can be approximated by a Gaussian one, and we have

$$X_n \sim \mathcal{N}\left(\left(\gamma_{st}^l + 1\right) \sigma_u^2, \frac{1}{N^{(s)}} \left(2\gamma_{st}^l + 1\right) \sigma_u^4 \right), \mathcal{H}_l, \quad (3)$$

---

[1]There is a none-zero probability that some transmit power levels do not happen and are not observed by the ST during Stage I, even if the learning period is relatively large. In this case, these missed hypotheses can be viewed as small probability events. Consequently, they will have negligible impact on the performance of the proposed spectrum access method, and can be ignored.

[2]For other modulation schemes, the test statistic X_n still follows a Gaussian distribution [36]. Therefore, our proposed method is still valid for other modulation and/or adaptive modulation schemes.

where $\gamma_{st}^l = P_{PT,l}/\sigma_u^2$ is the received signal-to-noise ratio (SNR) at the ST. Considering all the hypotheses, we establish that X_n follows a mixed Gaussian distribution [37]

$$X_n \sim \sum_{l=1}^L \pi_l \mathcal{N}(\mu_l, S_l^{-1}), \qquad (4)$$

where $0 \leqslant \pi_l \leqslant 1$ is the mixing coefficients with $\sum_{l=1}^L \pi_l = 1$. Each Gaussian density $\mathcal{N}(\mu_l, S_l^{-1})$ is a component of the mixture with mean value $\mu_l = (\gamma_{st}^l + 1)\sigma_u^2$ and precision $S_l = ((2\gamma_{st}^l + 1)\sigma_u^4/N^{(s)})^{-1}$.

Given the observation set $\mathbf{X} = \{X_1, \ldots, X_N\}$, the proposed Bayesian nonparametric method aims to infer the GMM parameter set $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$, where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_L\}$ with $\theta_l = \{\mu_l, S_l\}$ and $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_L\}$. In other words, our method automatically clusters the signals with the same PT transmit power levels. In summary, Stage I establishes a big picture of the PT activities at the cost of an one-off overhead. After learning, the ST allocates the same number of power levels as that in the PT, with an initial ST power value $P_{ST,k}$ for each level k.

B. Stage II

In this stage, we propose two prediction-transmission structures (periodic and non-periodic) adapting to the PT power level variation for spectrum access. For the ease of exposition, Fig. 3 depicts the whole two-stage spectrum sharing strategy with the non-periodic structure for Stage II. With the aid of Fig. 3, the mechanisms of both structures can be exposed.

- For both periodic and non-periodic structures, Stage II consists of two alternating parts: prediction and transmission actions. In the prediction action (the sensing slots with purple color), the ST can easily identify the current PT power level l, which is jointly determined by the test statistic $X_n, n > N$, in (2) and the inferred GMM parameter set $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$. In the transmission action, the ST allocates its transmit power level k to match the latest identified PT power level l ($k = l$). Here, the corresponding $P_{ST,k}$ can be determined as follows. Assume that the required SINR for the PR is Γ_0 and the current received SINR is Γ_{PR}. A nearby monitoring station (not shown in Fig. 2) of the PR transmits Γ_{PR} to the central site, likely through optical fiber or microwave, which then broadcasts this information on a dedicated frequency. We assume that the ST is able to decode broadcast signals from the central site and communicate with the SR on different frequency bands. Through the broadcast nature, the ST obtains Γ_{PR}. If $\Gamma_{PR} < \Gamma_0$, the ST should reduce the transmit power and vice versa. In other words, $P_{ST,k}$ for each level k will gradually approach a desired power value. This guarantees that the PR is well protected, while the secondary network obtains the highest possible throughput.[3] On this basis, it is clear that the alignment between ST and PT power levels can optimize the trade-off between the interference to primary network and throughput of the secondary network.[4] In the case that the ST mismatches the PT power level variation, either the

PR is interfered below the required SINR or unnecessarily lower secondary network throughput is obtained.
- For the periodic structure, Stage II starts with a single sensing slot followed by a single transmission block. Then this pattern continues to be repeated (i.e. the prediction interval is fixed to be one transmission block).
- In the non-periodic structure, the intervals are dynamically determined to enhance the NPLA performance. Specifically, continuous sensing slots (zero prediction intervals) are used to track the PT power level variation, while continuous transmission blocks (long prediction intervals) are selected to avoid unnecessary prediction. If fact, zero prediction interval is used when the ST predicts that the current hypothesis is about to end, or it does not know when the current hypothesis begins (at the beginning of Stage II in Fig. 3). The non-periodic structure will be elaborated on in Section IV-C.

III. SPECTRUM LEARNING BASED ON BAYESIAN INFERENCE

In this section, we focus on Stage I, and introduce a Bayesian nonparametric method to infer the GMM parameter set $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$ based on the observation set $\mathbf{X}$. As L is unknown *a priori*, the traditional methods, such as the K-mean and expectation maximization, are inapplicable. This motivates us to resort to Dirichlet process Gaussian mixture model [40], which takes into account the Gaussian mixture property and is able to identify the unknown number of Gaussian components. For specific Bayesian inference, we choose the Markov chain Monte-Carlo based Gibbs sampling method [41] considering its simplicity.

In the following, we first review the preliminary knowledge on the Dirichlet process mixture model. On this basis, we introduce the DPGMM considering the specific distribution of the observation set $\mathbf{X}$. Furthermore, we modify the DPGMM to the conditionally conjugate case to simplify the inference process. Finally, we carry out Bayesian inference with Gibbs sampling method to infer $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$.

A. Dirichlet Process Mixture Model

The Dirichlet distribution is an extension of the Beta distribution for multivariate cases. It represents the probability of K events given that the k-th event x_k ($k = 1, \ldots, K$) has been observed $\alpha_k - 1$ times. The probability density function can be expressed as

$$\mathrm{Dir}(\alpha_1, \ldots, \alpha_K) = \frac{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}, \qquad (5)$$

where π_k is the probability of the k-th event x_k with $\sum_{k=1}^K \pi_k = 1$ and $\pi_k > 0$.

In our application, the event x_k represents the k-th possible PT transmit power level, which can not be observed explicitly. Instead, the explicit observation is the test statistic X_n. As X_n is drawn from a distribution based on event x_k, we introduce the Dirichlet process (DP) to define the distribution of X_n. A random measure G is said to be a Dirichlet process distributed with a base probability distribution G_0 and a concentration parameter α, if we have

$$(G(A_1), \ldots, G(A_i)) \sim \mathrm{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_i)) \qquad (6)$$

---

[3]The similar idea of using broadcasting mechanism was also suggested by Federal Communications Commission [38, p. 6] and adopted in 4G LTE systems in the form of inter-cell interference overload indicator [39].

[4]Note that the channel state information for PT-ST and ST-PR is not required in our approach. In addition, the use of the received SINR has already included the impact of the channel fading.

for every finite measurable partition $\{A_1, \ldots, A_i\}$ of $\boldsymbol{\theta}$. It is written as $G \sim \mathrm{DP}(G_0, \alpha)$.

Next we model the observation set $\mathbf{X}$ using the parameter $\boldsymbol{\theta}$ based on the DP mixture model. A DP mixture model is suitable for the clustering purposes, where the number of Gaussian components is not known *a priori*. Here, X_n can be regarded as an independent draw from the distribution $F(\theta_n)$, where each θ_n is an i.i.d. draw from a DP G. Mathematically, the DP mixture model can be expressed as [42]

$$
\begin{aligned}
X_n | \theta_n &\sim F(\theta_n), \\
\theta_n | G &\sim G, \\
G | \{G_0, \alpha\} &\sim \mathrm{DP}(G_0, \alpha).
\end{aligned}
\tag{7}
$$

Note that the formulation (7) represents the most general case. In our case, two different observations X_n and $X_{n'}$ ($n \neq n'$) may follow the same distribution, and (7) can not explicitly reveal such property. Therefore, following [41], a latent variable z_n is introduced to explicitly indicate which transmit power level that X_n is associated with, and will be referred to as an indicator hereafter. Accordingly, an equivalent model can be obtained as

$$
\begin{aligned}
X_n | \{\boldsymbol{z}, \boldsymbol{\phi}\} &\sim F(\phi_{z_n}), \\
\phi_k | G_0 &\sim G_0, \\
p(z_n = k) &= \pi_k, \\
\boldsymbol{\pi} | \{\alpha, K\} &\sim \mathrm{Dir}(\alpha/K, \cdots, \alpha/K),
\end{aligned}
\tag{8}
$$

where $\boldsymbol{z} = \{z_1, \ldots, z_N\}$ is the set of indicators, $\boldsymbol{\phi} = \{\phi_1, \ldots, \phi_K\}$ is the set of unique values in $\boldsymbol{\theta}$, and $\theta_n = \phi_{z_n}$. Hereafter, K refers to the total number of Gaussian components, and each component consists of the observations that are determined by the ST as having the same transmit power level. Note that $\boldsymbol{\pi}$ is assumed to have a symmetric Dirichlet distribution, where all the concentration parameters are α/K. This assumption is widely adopted when there is no prior knowledge of the mixing proportions [41]. Let N_k denote the number of observations assigned to the k-th component, then N_k follows a multinomial distribution

$$
p(N_1, \ldots, N_K | \boldsymbol{\pi}) = \frac{N!}{N_1! \cdots N_K!} \prod_{k=1}^{K} \pi_k^{N_k},
\tag{9}
$$

where $\sum_{k=1}^{K} N_k = N$, and the distribution of the indicators is

$$
p(\boldsymbol{z} | \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{N_k}.
\tag{10}
$$

We can integrate out the mixing proportions of the product of $p(\boldsymbol{\pi} | \alpha)$ in (8) and $p(\boldsymbol{z} | \boldsymbol{\pi})$ in (10), and the prior on $\boldsymbol{z}$ in terms of α is expressed as [40]

$$
\begin{aligned}
p(\boldsymbol{z} | \alpha) &= \int p(\boldsymbol{z} | \boldsymbol{\pi}) p(\boldsymbol{\pi} | \alpha) \mathrm{d}\boldsymbol{\pi} \\
&= \int \frac{\Gamma(\alpha)}{\Gamma(\alpha/K)^K} \prod_{k=1}^{K} \pi_k^{N_k + \alpha/K - 1} \mathrm{d}\boldsymbol{\pi} \\
&= \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)} \prod_{k=1}^{K} \frac{\Gamma(N_k + \alpha/K)}{\Gamma(\alpha/K)}.
\end{aligned}
\tag{11}
$$

As all the observations are exchangeable, if we assume that $\boldsymbol{z}_{-n} = \{z_1, \ldots, z_{n-1}, z_{n+1}, \ldots, z_N\}$ has been obtained, the conditional distribution for the individual indicator can be given by

$$
p(z_n = k | \boldsymbol{z}_{-n}, \alpha) = \frac{p(z_n = k, \boldsymbol{z}_{-n} | \alpha)}{p(\boldsymbol{z}_{-n} | \alpha)} = \frac{N_{-n,k} + \alpha/K}{N - 1 + \alpha},
\tag{12}
$$

where $N_{-n,k}$ is the number of samples excluding X_n in the k-th component. Similarly, the prior distribution of θ_n over $\boldsymbol{\theta}_{-n} = \{\theta_1, \ldots, \theta_{n-1}, \theta_{n+1}, \ldots, \theta_N\}$ can be written as

$$
\theta_n | \boldsymbol{\theta}_{-n} \sim \frac{N_{-n,k}}{N-1+\alpha} \sum_{k=1}^{K} \delta(\phi_k) + \frac{\alpha}{N-1+\alpha} G_0.
\tag{13}
$$

B. Conditionally Conjugate Dirichlet Process Gaussian Mixture Model

Recall that the observation X_n in (4) follows a mixed Gaussian distribution given by

$$
X_n | \{\boldsymbol{\mu}, \boldsymbol{S}, \boldsymbol{\pi}\} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, S_k^{-1}),
\tag{14}
$$

where $k = \{1, \ldots, K\}$ replaces $l = \{1, \ldots, L\}$, as k denotes the index of inferred hypotheses while l is the index of the real hypothesis. Therefore, X_n can be modeled as a DPGMM and expressed as

$$
\begin{aligned}
X_n | \{\boldsymbol{z}, \boldsymbol{\phi}\} &\sim \mathcal{N}(\mu_{z_n}, S_{z_n}^{-1}), \\
(\mu_k, S_k^{-1}) &\sim G_0, \\
p(z_n = k) &= \pi_k, \\
\boldsymbol{\pi} | \{\alpha, K\} &\sim \mathrm{Dir}(\alpha/K, \cdots \alpha/K).
\end{aligned}
\tag{15}
$$

In (15), G_0 represents a prior guess of the distributions of μ_k and S_k^{-1} in the DPGMM. Its choice is usually guided by mathematical convenience, and the conjugate form is widely adopted. In our case, the distribution of G_0 specifies the prior on the mixed Gaussian distributions parameters $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$ and $\boldsymbol{S} = \{S_1, \ldots, S_K\}$, and it can be expressed in a conjugate form [37]

$$
\begin{aligned}
\mu_k | \{S_k, \xi, \rho\} &\sim \mathcal{N}(\xi, (\rho S_k)^{-1}), \\
S_k | \{\beta, W\} &\sim \mathcal{G}(\beta, W^{-1}),
\end{aligned}
\tag{16}
$$

where ξ, ρ, β and W are the hyperparameters for the DPGMM. It is clear that the prior distribution of μ_k is dependent on S_k. This undesirable property is inevitable due to the conjugacy requirement for G_0. To remove such dependency, we modify the original conjugate feature in the DPGMM and introduce the conditionally conjugate version of DPGMM to model $\mathbf{X}$. In a conditionally conjugate DPGMM, (16) can be rewritten as [37]

$$
\begin{aligned}
\mu_k | \{\lambda, R\} &\sim \mathcal{N}(\lambda, R^{-1}), \\
S_k | \{\beta, W\} &\sim \mathcal{G}(\beta, W^{-1}),
\end{aligned}
\tag{17}
$$

where R is the hyperparameter. To complete the conditionally conjugate DPGMM and capture the features inherent in $\mathbf{X}$, we need to give suitable values for the hyperparameters. However, the exact values are hard to know *a priori*, and small changes on them will dramatically affect the model performance. To achieve
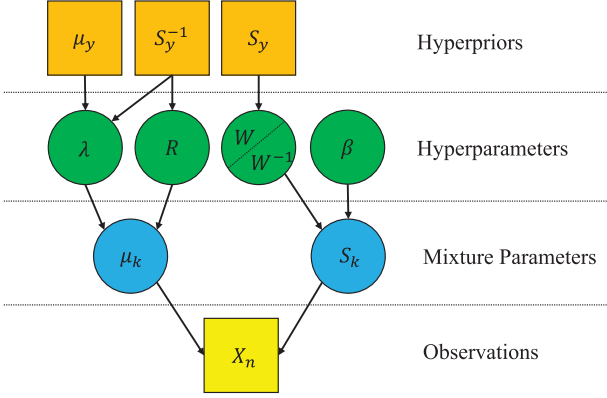
Fig. 4. Graphical representation of the conditionally conjugate DPGMM.

the robustness for the model, we impose vague priors for the hyperparameters following [40],

$$\lambda \sim \mathcal{N}(\mu_y, S_y^{-1}),$$
$$R \sim \mathcal{G}(1, S_y),$$
$$W \sim \mathcal{G}(1, S_y^{-1}),$$
$$\beta^{-1} \sim \mathcal{G}(1,1) \quad (18)$$

where the hyperpriors μ_y and S_y refer to the empirical mean and precision of $\mathbf{X}$, respectively. In theory, the prior should not depend on the observations. However, as shown in [40], the formulation for the priors in (18) is equivalent to normalizing observations, and a wide range of parameters in the priors lead to similar inference results.

The conditionally conjugate DPGMM in (18) can be graphically represented in Fig. 4.

C. Inference Using Gibbs Sampling

Given the conditionally conjugate DPGMM and observation $\mathbf{X}$, now we use the Markov chain Monte-Carlo algorithm based on Gibbs sampling to infer $\{\boldsymbol{\theta}, \boldsymbol{z}, L\}$. With the latent variable set $\boldsymbol{z}$, we can obtain the mixing proportion set $\boldsymbol{\pi}$. In the algorithm, we update the variables iteratively by sampling each variable from the posterior distribution conditioned on the others.

Given the likelihood of μ_k and S_k in (14) and their priors in (17), we can multiply the priors by the likelihood conditioned on $\boldsymbol{z}$ and obtain the conditional posterior distributions of μ_k and S_k, which can be given by

$$\mu_k | \{\boldsymbol{z}, S_k, \lambda, R\}$$
$$\sim \mathcal{N}\left(\frac{S_k \sum_{n=1}^N \mathbb{1}_{\mathbf{A}}(n)X_n + \lambda R}{N_k S_k + \lambda}, \frac{1}{N_k S_k + \lambda}\right),$$

$$S_k | \{\boldsymbol{z}, \beta, W\} \sim$$
$$\mathcal{G}\left(\beta + N_k, \left[\frac{1}{\beta + N_k}\left(W\beta + \sum_{n=1}^N \mathbb{1}_{\mathbf{A}}(n)(X_n - \mu_k)^2\right)\right]^{-1}\right),$$
$$(19)$$

where $\mathbf{A} = \{n | z_n = k\}$ and the indicator function $\mathbb{1}_{\mathbf{A}}(n) = 1$ when $n \in A$ and $\mathbb{1}_{\mathbf{A}}(n) = 0$ otherwise. Similarly, with the likelihood of the hyperparameters λ, R, β and W given in (17),

and their priors given in (18), the conditional posteriors of the hyperparameters can be written as

$$\lambda | \{\boldsymbol{\mu}, R\} \sim \mathcal{N}\left(\frac{\mu_y S_y + R \sum_{k=1}^K \mu_k}{S_y + KR}, \frac{1}{S_y + KR}\right),$$

$$R | \{\boldsymbol{\mu}, \lambda\} \sim \mathcal{G}\left(K+1, \left[\frac{1}{K+1}\left(S_y + \sum_{k=1}^K (\mu_k - \lambda)^2\right)\right]^{-1}\right),$$

$$W | \{\boldsymbol{S}, \beta\} \sim \mathcal{G}\left(K\beta+1, \left[\frac{1}{K\beta+1}\left(S_y + \beta \sum_{k=1}^K S_k\right)\right]^{-1}\right),$$

$$\beta | \{\boldsymbol{S}, W\} \propto \Gamma\left(\frac{\beta}{2}\right)^{-K} \exp\left(\frac{-1}{2\beta}\right)\left(\frac{\beta}{2}\right)^{(K\beta-3)/2}$$
$$\cdot \prod_{k=1}^K (S_k W)^{\beta/2} \exp\left(-\frac{\beta S_k W}{2}\right). \quad (20)$$

Note that the exact distribution of β is not given, but its samples can be obtained following [37]. In detail, we capture the log-concave property of $p(\log(\beta)|\boldsymbol{S}, W)$ and generate samples independently from the distribution of $\log(\beta)$. Then, we use the adaptive rejection sampling technique [43] to transform these samples and obtain the value of β.

Before introducing the conditional posterior for $\boldsymbol{z}$, we note that its prior in (12) only suits the case where K is a fixed finite parameter. To make the conditionally conjugate DPGMM applicable to the scenario with an infinite number of Gaussian components, we let $K \to \infty$ in (12) and the conditional prior reaches the following limits

$$p(z_n = k | \boldsymbol{z}_{-n}, \alpha) = \begin{cases} \dfrac{N_{-n,k}}{N - 1 + \alpha}, & N_{-n,k} > 0, \\[2mm] \dfrac{\alpha}{N - 1 + \alpha}, & N_{-n,k} = 0. \end{cases} \quad (21)$$

Now we combine the priors of $\boldsymbol{z}$ with its likelihood given in (12), and the conditional posterior can be given by (22). shown at the bottom of the next page.

Unfortunately, the integral in the second case of (22) is not analytically tractable. Therefore, the auxiliary variable sampling algorithm [41] is employed. In detail, we add k_0 auxiliary components in each sampling iteration to represent the effect of the auxiliary components. Note that the posterior probability that X_n belongs to the k-th component is proportional to $N_{-n,k}$, we use α/k_0 for the auxiliary components and rewrite (22) as

$$p(z_n = k' | \boldsymbol{z}_{-n}, \mu_{k'}, S_{k'}, \alpha)$$
$$= \begin{cases} \dfrac{qN_{-n,k'}}{N - 1 + \alpha} \mathcal{N}(X_n | \mu_{k'}, S_{k'}), & 1 \leqslant k' \leqslant K', \\[2mm] \dfrac{q\alpha/k_0}{N - 1 + \alpha} \mathcal{N}(X_n | \mu_{k'}, S_{k'}), & K' < k' \leqslant K' + k_0, \end{cases} \quad (23)$$

where k' is the index of the unique components in each iteration during the Gibbs sampling algorithm, q is the appropriate constant for normalization, and K' is the number of active components. To this end, we summarize the sampling algorithm in Algorithm 1.

Algorithm 1: Gibbs Sampling Algorithm.

Require:
  Initial observation set $\mathbf{X}$. Set a component which contains all X_n. Initialize the hyperparameters λ, R, β, and W, the hyperpriors μ_y and S_y, and the indicator set $\mathbf{z}$;

Ensure:
  The sets $\mathbf{z}$, $\boldsymbol{\mu}$, and $\boldsymbol{S}$.
1: Update $\boldsymbol{\mu}$ and $\boldsymbol{S}$ conditional on the indicator $\mathbf{z}$ and hyperparameters λ, R, β and W following (19);
2: Update the hyperparameters λ, R, β and W conditional on hyperpriors μ_y and S_y following (20);
3: **for** $n = 1, 2, \ldots, N$ **do**
4:   **if** $z_n \neq z_{n'}$ for all
     $n' \in \{1, \ldots, n-1, n+1, \ldots, N\}$ **then**
5:     Let $z_n = K' + 1$.
6:   **end if**
7:   Draw $\mu_{k'}$ and $S_{k'}$ for $k' \in \{K'+1, \ldots, K'+k_0\}$ following (17).
8:   Update the indicator z_n following (23).
9:   Discard the empty components.
10: **end for**

IV. PROPOSED PREDICTION-TRANSMISSION SPECTRUM ACCESS STRUCTURE

In this section, we propose two prediction-transmission structures (periodic and non-periodic) for spectrum access. We first introduce the functions of the prediction and transmission parts. Then, we present the details of how to determine the prediction intervals. As directly optimizing the NPLA performance is intractable, we propose to maximize an expected average utility by imposing reward (penalty) for power level match (mismatch). At last, we extend the prediction-transmission structure to an online scenario, which can handle the case where the number of the PT power levels L changes after Stage I.

A. Functions of the Prediction and Transmission Parts

In the prediction part, with the inferred GMM parameters $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$, the ST can easily identify the current PT power level by a single sensing slot with test statistic X_n, based on the following criterion

$$
\begin{aligned}
k &= \arg \max_{k \in \{1, \ldots, K\}} \Pr(\mathcal{H}_k | X_n) \\
  &= \arg \max_{k \in \{1, \ldots, K\}} \pi_k \mathcal{N}(X_n | \mu_k, S_k). \quad (24)
\end{aligned}
$$

In the transmission part, the ST allocates its transmit power level k to match the PT power level l, which means $k = l$. Note that K is an estimate of L. In the simulation, we find that the conditionally conjugate DPGMM is able to identify K ($K = L$)

with a high probability.[5] When the ST matches the PT power level, it will adjust its transmit power of each level, which has been explained in Section II-B.

B. Periodic Structure for the Prediction-Transmission Structure

In periodic structure, the ST periodically predicts the PT power level in a sensing slot with duration T_{ss}, and then transmits for a transmission block with duration $T_{st} = \tau_s T_{ss}$. This structure can be implemented in a straightforward way, and the prediction interval remains constant whether the SR decodes the signal successfully or not.

C. Non-Periodic Structure for the Prediction-Transmission Structure

For a non-periodic structure, an essential question is how to determine the prediction intervals. Basically, this needs to find out the distribution of the PT power level duration $T_{po} = \tau_p T_{ss}$, and the corresponding observation of each action. If the prediction action is taken, the observation will be the PT power level identified from the received test statistic X_n according to (24). If the transmission action is taken, the observation will be a positive or negative acknowledgment (ACK) received by the ST from the SR, which indicates whether the SR correctly receives the signal from the ST. Based on these observations, the ST can infer the PT transmit power level, and then dynamically adjust the prediction intervals. As the inference may not be correct all the time, the dynamic adjustment of intervals can be formulated as a partial observable decision problem [44]. To address this problem, we first estimate the distribution parameter of τ_p based on $\mathbf{z}$. Then, we develop a reinforcement learning algorithm that correlates the observations with the current PT power level identification.

1) PT Power Level Duration Distribution: Without loss of generality, the discretized PT power level duration τ_p of all hypotheses is assumed to follow a Poisson distribution with the same mean value ν.[6] Its cumulative distribution function can be given by

$$
F_\nu(\tau_p) = \frac{\Gamma(\tau_p + 1, \nu)}{\Gamma(\tau_p)}, \quad (25)
$$

where ν is the mean value of the Poisson distribution. In Stage I, we have attributed all signals X_n to different components by learning the GMM parameter set $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$, thus we can easily obtain the samples of the PT operation durations τ_p. Then, ν can be estimated through maximum likelihood estimation as

$$
\nu = \frac{1}{M} \sum_{m=1}^{M} \tau_p^m, \quad (26)
$$

---

[5]In the rare case of $K \neq L$, the NPLA performance will degrade and another round of learning will be necessary.

[6]If the time duration is correlated between adjacent hypotheses, the proposed algorithm also works, but at the cost of significantly increased computational complexity.

$$
p(z_n = k | \mathbf{z}_{-n}, \mu_k, S_k, \alpha) \propto
\begin{cases}
p(z_n = k | \mathbf{z}_{-n}, \alpha) p(X_n | \mu_k, S_k, \mathbf{z}_{-n}) \propto \dfrac{N_{-n,k} \mathcal{N}(X_n | \mu_k, S_k,)}{N - 1 + \alpha}, & N_{-n,k} > 0, \\
p(z_n = k | \mathbf{z}_{-n}, \alpha) \int p(X_n | \mu_k, S_k) p(\mu_k, S_k | \lambda, R, \beta, W) d\mu_k dS_k, & N_{-n,k} = 0.
\end{cases}
\quad (22)
$$

---

where τ_p^m is the discretized duration of the m-th PT hypothesis, and M is the number of PT hypotheses detected in Stage I.

We set a discrete time indicator $\tau = 0$ whenever the PT changes to a new power level. Under each hypothesis, if the PT has been keeping that power level until time point τ ($\tau \leqslant \tau_p$), then the probability that the PT will continue staying in the same power level during the following discretized time duration τ_0 can be expressed as

$$g_\tau(\tau_0) = \frac{1 - F_\nu(\tau + \tau_0)}{1 - F_\nu(\tau)}, \tau_0 \geqslant 1. \tag{27}$$

If the ST does not know when the current hypothesis begins, it will keep taking the prediction action until the next hypothesis. This is what occurs at the beginning of Stage II in Fig. 3.

2) PT Power Transition Probability: In Stage I, the observations in the conditionally conjugate DPGMM are indefinite exchangeable. Therefore, the ST can not infer the PT transmit power level transition probability directly. Alternatively, we take a pragmatic approach and use the mixing proportion set $\boldsymbol{\pi}$, obtained by counting the number of observations X_n in each power level, to represent the occupancy frequency of different power levels. Thus, we define an $L \times L$ transition probability matrix $\mathbf{C}$ for the PT transmit power levels. The element $C_{kj}, k, j \in \{1, \ldots, L\}$, of $\mathbf{C}$ refers to the probability that the PT transfers from the k-th to the j-th transmit power level, and is given by

$$C_{kj} = \begin{cases} \dfrac{\pi_j}{1 - \pi_k}, & k \neq j, \\ 0, & k = j. \end{cases} \tag{28}$$

If the PT does not transmit with the same power level under two consecutive hypotheses then $C_{kk} = 0$. We also define the vector $\boldsymbol{c}_k$ as the k-th row of matrix $\mathbf{C}$.

3) Estimation Probability Matrix of the PT Power Level: If the PT transmits with binary power levels, the prediction performance of the ST is dictated by the detection and false alarm probabilities. However, this is no longer the case when the PT has multiple transmit power levels. Instead, we define an $L \times L$ prediction probability matrix $\mathbf{H}$, with the element $H_{kj} = \Pr\{\hat{\mathcal{H}}_j | \mathcal{H}_k\}, k, j \in \{1, \ldots, L\}$, representing the probability that the PT is operating under hypothesis $\mathcal{H}_k$ while the detection by the ST is in favor of hypothesis $\mathcal{H}_j$. $\hat{\mathcal{H}}_j$ represents that the ST identifies the PT operating in the j-th transmit power level following (24). Thus the element H_{kk} refers to the detection probability for hypothesis $\mathcal{H}_k$. We also define vector $\boldsymbol{h}_j$ as the j-th column of matrix $\mathbf{H}$.

4) Benefits of the ST From Prediction and Transmission Actions: We denote the prediction action of the ST at time τ as $a_\tau = 0$, and its observation as $O_\tau^E \in \{\mathcal{H}_k\}$. We also denote the transmission action at time τ as $a_\tau = 1$, and its observation as $O_\tau^A \in \{\mathcal{A}(\text{positive ACK}), \mathcal{N}(\text{negative ACK})\}$. It is assumed that the positive/negative ACK is returned to the ST through a dedicated feedback channel. Let $p_{\tau+\tau_0}^k$ denote the conditional probability that the PT keeps operating with the k-th transmit power level at time $\tau + \tau_0$ given $p_0^k = 1$ and $\{a_0, \ldots, a_\tau, O_0, \ldots, O_\tau\}$, where $O_\tau \in \{O_\tau^E, O_\tau^A\}$. Based on Bayesian rule, the probabilities of PT staying in the k-th power level at time $\tau + \tau_0, \tau_0 \in \{1, \tau_s\}$, can be given as follows.

When $a_\tau = 0, \tau_0 = 1$, we have

$$p_{\tau+1}^k(O_{\tau+1}^E)$$
$$= \begin{cases} \dfrac{p_\tau^k g_\tau^E H_{kk}}{p_\tau^k g_\tau^E H_{kk} + (1 - p_\tau^k g_\tau^E) \boldsymbol{c}_k \boldsymbol{h}_k}, & O_{\tau+1}^E = \mathcal{H}_k, \\ \dfrac{p_\tau^k g_\tau^E H_{jk}}{p_\tau^k g_\tau^E H_{jk} + (1 - p_\tau^k g_\tau^E) \boldsymbol{c}_k \boldsymbol{h}_j}, & O_{\tau+1}^E = \mathcal{H}_j, j \neq k. \end{cases} \tag{29}$$

When $a_\tau = 1, \tau_0 = \tau_s$, we have

$$p_{\tau+\tau_s}^k(O_{\tau+\tau_s}^A)$$
$$= \begin{cases} \dfrac{p_\tau^k g_\tau^A}{p_\tau^k g_\tau^A + (1 - p_\tau^k g_\tau^A) \sum_{j=1}^{L} C_{kj} p_{kj}^{ACK}}, & O_{\tau+\tau_s}^A = \mathcal{A}, \\ 0, & O_{\tau+\tau_s}^A = \mathcal{N}, \end{cases} \tag{30}$$

where we assume the positive and negative ACKs from the SR can be received by the ST error free. Besides, p_{kj}^{ACK} denotes the probability that the SR decodes signals correctly when ST is on the k-th level and PT is one the j-th level. For simplicity, we set $p_{kj}^{ACK} = 1$ when $k > j$ and $p_{kj}^{ACK} = 0$ otherwise. In (29) and (30), $g_\tau^E = g_\tau(1)$ and $g_\tau^A = g_\tau(\tau_s)$, where the superscripts E and A represent the prediction and transmission, respectively. The expected utility that the ST obtains at time τ with the PT operating with the k-th transmit power level, which is $r(p_\tau^k, a_\tau, k)$, can be given by

$$r(p_\tau^k, a_\tau, k)$$
$$= \begin{cases} \left[p_\tau^k g_\tau^A D_k - (1 - p_\tau^k g_\tau^A) \sum_{j=1}^{L} C_{kj} Y_j \right] \tau_s, & a_\tau = 1, \\ 0, & a_\tau = 0. \end{cases} \tag{31}$$

Hereafter, k denotes the ST transmit power level, l is the PT transmit power level, and $K = L$. In (31), D_k is the reward that the ST will receive when the ST transmits with the k-th power level and $k = l$. Meanwhile, Y_j is the penalty that the ST will receive when the ST transmits with the j-th power and $j \neq l$.

5) Prediction-Transmission Structure Optimization: An ST access policy $\epsilon = [d_0, \ldots, d_\tau, \ldots]$ maps the ST belief space $\{p_\tau^k, \tau \geqslant 0\}$ to the action space $\{a_\tau, \tau \geqslant 0\}$. Thus, the optimal prediction-transmission policy aims to maximize the expected average utility, which can be given by

$$\max_\epsilon \lim_{M' \to \infty} \frac{\sum_{m=M+1}^{M'} \sum_{\tau=0}^{\tau_p^m - 1} r(p_\tau^k, a_\tau, k) a_\tau / (M' - M)}{\sum_{m=M+1}^{M'} \tau_p^m / (M' - M)}. \tag{32}$$

Note that if $a_\tau = 1$ and $\tau_s \geqslant 2$, $a_{\tau+\tau_0}, \tau_0 = \{1, \ldots, \tau_s - 1\}$ is not defined in Section IV-C4. This is because the ST keeps transmitting between time τ and $\tau + \tau_s$, thus no action is taken during this period. Therefore, we assign $a_{\tau+\tau_0} = 0, \tau_0 = \{1, \ldots, \tau_s - 1\}$ if $a_\tau = 1$ and $\tau_s \geqslant 2$ in the action space $\{a_\tau, \tau \geqslant 0\}$. The total utility obtained by the ST during each PT hypothesis is i.i.d. Thus, by the law of large numbers, the maximization problem in (32) can be rewritten as

$$\max_\epsilon \mathrm{E}\left[\sum_{\tau=0}^{\tau_p - 1} r(p_\tau^k, a_\tau, k) a_\tau \right]. \tag{33}$$

In other words, instead of maximizing the average expected utility, we translate the problem to the maximization of the utility

in each PT hypothesis. For convenience, let $V_\epsilon(0, p_\tau^k, k)$ denote the expected utility that can be achieved in each PT hypothesis following policy ϵ, which is

$$V_\epsilon(\tau = 0, p_\tau^k = 1, k) = E_\epsilon\left[\sum_{\tau=0}^{\tau_p - 1} r(p_\tau^k, a_\tau, k)a_\tau\right]. \quad (34)$$

Then, we can define the maximum utility that can be achieved by the ST in each PT hypothesis as

$$V(\tau = 0, p_\tau^k = 1, k) = \sup_\epsilon V_\epsilon(\tau = 0, p_\tau^k = 1, k). \quad (35)$$

In (35), $V(\tau, p_\tau^k, k)$ is directly determined by the action taken at time τ, thus it can be expressed as

$$V(\tau, p_\tau^k, k) = \max\left\{E(\tau, p_\tau^k, k), A(\tau, p_\tau^k, k)\right\}, \quad (36)$$

where $E(\tau, p_\tau^k, k)$ and $A(\tau, p_\tau^k, k)$ are the expected utilities that can be obtained by the ST through prediction and transmission, respectively. We have

$$E(\tau, p_\tau^k, k) = \sum_{j=1}^{L} \Pr\left[O_{\tau+1}^E = \mathcal{H}_j\right] V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_j), k), \quad (37)$$

and

$$A(\tau, p_\tau^k, k) = \sum_{\mathcal{J} \in \{\mathcal{A}, \mathcal{N}\}} \Pr\left[O_{\tau+\tau_s}^A = \mathcal{J}\right]$$
$$\times V(\tau + \tau_s, p_{\tau+\tau_s}^k(\mathcal{J}), k) + r(p_\tau^k, 1, k)$$
$$= \Pr\left[O_{\tau+\tau_s}^A = \mathcal{A}\right] V(\tau + \tau_s, p_{\tau+\tau_s}^k(\mathcal{A}), k) + r(p_\tau^k, 1, k). \quad (38)$$

Lemma 1: $V(\tau, p_\tau^k, k)$ is a convex function of p_τ^k for given τ and k.

The proof of the above lemma can be easily obtained following [45, page 58–59], and is omitted here.

It is clear that $V(\tau, p_\tau^k, k)$ is derived backward in time domain in (36), (37) and (38). Thus it will be helpful for the derivation of the optimal policy if an upper bound of τ can be established, which is given by

$$\mathcal{T}_k = \min\left\{\tau' : g_\tau^A < \frac{\sum_{j=1}^{L} C_{kj}Y_j}{\sum_{j=1}^{L} C_{kj}Y_j + D_k}, \forall \tau > \tau'\right\}. \quad (39)$$

In each PT hypothesis with the k-th transmit power level, the transmission action will not be taken by the ST after $\mathcal{T}_k$. This is because, $\forall \tau > \mathcal{T}_k, r(1, 1, k) < 0$, which means the ST will certainly receive negative reward if it transmits after $\mathcal{T}_k$, even if the PT is estimated as idle at time τ. Therefore, $V(\tau, 1, k) = 0, \forall \tau > \mathcal{T}_k$. When it comes to the range of $\mathcal{T}_k$, as τ_p follows a Poisson distribution in (25), we find that $\mathcal{T}_k < \infty$ always holds according to [46].

Lemma 2: The optimal utility function $V(\tau, p_\tau^k, k)$ increases with p_τ^k for given τ and k.

Proof: The proof is given in Appendix. ∎

Lemma 3: $E(\tau, p_\tau^k, k)$ and $A(\tau, p_\tau^k, k)$ are convex functions increasing with p_τ^k for given τ and k.

Proof: We have proved that $E(\tau, p_\tau^k, k)$ and $A(\tau, p_\tau^k, k)$ increase with p_τ^k in Lemma 2. Next, to prove their convexity, we derive their second order derivatives with regard to p_τ^k. Combining (29) and (37), the second order derivative of $E(\tau, p_\tau^k, k)$ can
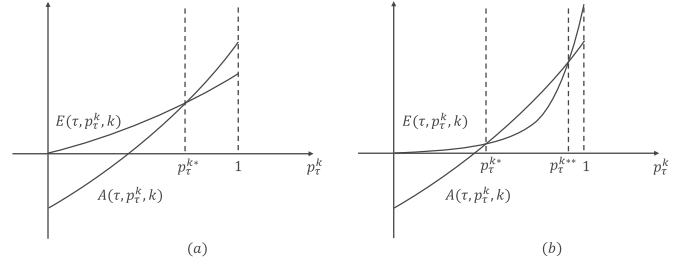

Fig. 5. The illustration of $A(\tau, p_\tau^k, k)$ and $E(\tau, p_\tau^k, k)$ versus p_τ^k, for (a) $p_\tau^{k*} = p_\tau^{k**}$ and (b) $p_\tau^{k*} < p_\tau^{k**}$.

be given by

$$\frac{dE^2(\tau, p_\tau^k, k)}{d^2 p_\tau^k} = \sum_{j=1}^{L} \frac{\left(g_\tau^E H_{jk} c_k h_j\right)^2 V''(\tau + 1, p_{\tau+1}^k(\mathcal{H}_j), k)}{\left[p_\tau^k g_\tau^E H_{jk} + (1 - p_\tau^k g_\tau^E)c_k h_j\right]^3}, \quad (40)$$

which is positive as $V(\tau, p_\tau^k, k)$ is convex. The second derivative of $A(\tau, p_\tau^k, k)$ can be proved positive similarly. Therefore, we complete the proof. ∎

We note that $E(\tau, 0, k) = 0$ in (37) and $A(\tau, 0, k) = -\sum_{j=1}^{L} C_{kj}Y_j\tau_s < 0$ in (38). Combining Lemma 3, we define the probability thresholds as

$$p_\tau^{k*} = \min\{p_\tau^k : E(\tau, p_\tau^k, k) \leqslant A(\tau, p_\tau^k, k)\}, \quad (41)$$

and

$$p_\tau^{k**} = \max\{p_\tau^k : E(\tau, p_\tau^k, k) \leqslant A(\tau, p_\tau^k, k)\}. \quad (42)$$

Then, we can give the following optimized protocol, where a_τ^* is the optimal action at time τ.

Theorem 1: If $p_\tau^{k*} = p_\tau^{k**}$, we have

$$a_\tau^* = \begin{cases} 0, & p_\tau^k \leqslant p_\tau^{k*}, \\ 1, & \text{others,} \end{cases} \quad (43)$$

and if $p_\tau^* < p_\tau^{**}$, we have

$$a_\tau^* = \begin{cases} 0, & p_\tau^k \leqslant p_\tau^{k*} \text{ or } p_\tau^k \geqslant p_\tau^{**}, \\ 1, & \text{others.} \end{cases} \quad (44)$$

The two cases are illustrated in Fig. 5. When $p_\tau^{k**} = p_\tau^{k*}$, $E(\tau, p_\tau^k, k)$ and $A(\tau, p_\tau^k, k)$ can be illustrated in Fig. 5(a) following Lemma 3. In this case, when $p_\tau^k \leqslant p_\tau^{k*}$, we get $E(\tau, p_\tau^k, k) \geqslant (\tau, p_\tau^k, k)$. The ST will choose the prediction action for larger expected utility ($a_\tau^* = 0$). When $p_\tau^k > p_\tau^{k*}$, we get $a_\tau^* = 1$. Note that the ST will choose the next action at the end of each transmission or prediction period. Fig. 5(b) follows a similar interpretation.

D. Prediction-Transmission in an Online Scenario

In Stage II, our proposed prediction-transmission structure can dynamically adjust the intervals between two sensing slots to increase the secondary network throughput. We note that this structure only works when the number of PT transmit power levels L remains unchanged. This is because the prediction-transmission structure is based on an one-off learning period in Stage I, and the learning results will become obsolete if L changes after Stage I. However, this change is likely to happen in the long run as a consequence of the PT adapting to the

environment fluctuation or QoS variation. This will render the proposed prediction-transmission structure less effective. To rectify this situation, we take advantage of the fact that the ST still receives the PT signals from time to time, reflecting the already changed PT power levels. We will explore these signals, and extend the spectrum access method to an online scenario to accommodate this highly dynamic PT power level characteristics.

In the online prediction-transmission structure, the ST not only identifies the PT power level in the sensing slot, but also stores the received PT signals (X_n^*). Then, it updates its observation set, starting from $\mathbf{X}$ from Stage I, through replacing the old signals with newly-received ones, while the size of $\mathbf{X}$ keeps the same. For example, after receiving the first signal X_1^* in Stage II, $\mathbf{X}$ will be updated as

$$\mathbf{X} = \{X_2, ..., X_N, X_1^*\}. \tag{45}$$

To make online update of $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$, we can revise Algorithm 1 by adding the following algorithm flow before line 1.

Algorithm 2: Algorithm Flow.

    for $n = 2, \ldots, N$ **do**
        Let $X_{n-1} = X_n$, $z_{n-1} = z_n$.
    end for
    Let $X_N = X_1^*$.
    Update the indicator z_n following (23).

V. NUMERICAL RESULTS

In this section, numerical results are provided to illustrate the advantages of the proposed two-stage spectrum sharing strategy. We first evaluate the performance of the prediction part in Stage II, in terms of different learning methods in Stage I including both parametric and nonparametric. Then, we numerically verify the theoretical results of the optimal thresholds in Theorem 1 in the prediction-transmission structure. Finally, we illustrate the superiority of the proposed spectrum sharing strategy on an overall basis in comparison with other methods, which are different from the proposed one in Stage I or II. In addition, we demonstrate the advantages of the proposed online prediction-transmission structure in handling the highly dynamic scenario where the number of PT power levels changes after Stage I.

In the simulation, we set the power level $L = 4$ and the probability of each hypothesis $\Pr(\mathcal{H}_l) = 0.25$. It is assumed that the noise variance $\sigma_u^2 = 1$, the PT transmit powers $P_{PT,1} : P_{PT,2} : P_{PT,3} = 1 : 2 : 3$, and $P_{PT,4} = 0$. Recall that $\gamma_{st}^l = P_{PT,l}/\sigma_u^2$ is the received SNR at the ST, thus we have $\gamma_{st}^1 : \gamma_{st}^2 : \gamma_{st}^3 = 1 : 2 : 3$ and $\gamma_{st}^4 = 0$. The average SNR at the ST is defined as $\gamma_{st} = (1/L)\sum_{l=1}^{L} \gamma_{st}^l$. In addition, we set the time duration of a sensing slot $T_{ss} = 2$ ms, the reward $D_k = 1$, and penalty $Y_l = 1$ in (31).

A. Performance of the PT Power Level Identification

For our proposed strategy, in the prediction part of Stage II, the ST can easily identify the current PT power level based on the inferred $\{\boldsymbol{\theta}, \boldsymbol{\pi}, L\}$ in Stage I and X_n, $n > N$. Note that in addition to the proposed conditionally conjugate DPGMM, there are other learning methods to obtain this parameter set, such as expectation maximization GMM (EMGMM) [47], mean shift [48], kernel density estimation (KDE), and density based spatial clustering of application with noise (DBSCAN). Among them,

the EMGMM is a parametric clustering method requiring the prior knowledge of L, while the others belong to a nonparametric class without the need for such prior knowledge. Taking into account the multiple power levels, we define the probability of correct PT power level prediction as

$$P_c = \sum_{l=1}^{L} \Pr\{\hat{\mathcal{H}}_l | \mathcal{H}_l\} \Pr\{\mathcal{H}_l\}. \tag{46}$$

In Fig. 6, we illustrate P_c for five different learning methods, with respect to different γ_{st} (Fig. 6(a)) and $N^{(s)}$ (Fig. 6(b)). It is shown in Fig. 6(a) that, in general, P_c increases with γ_{st} for all methods. This is because the gap between the adjacent transmit powers increases with γ_{st}, rendering them more distinguishable from the perspective of machine learning. In Fig. 6(b), P_c improves with increasing $N^{(s)}$ for all methods, because a larger $N^{(s)}$ results in a smaller variance of each Gaussian distribution in the mixture model. In Fig. 6, without the prior knowledge of L, the proposed DPGMM significantly outperforms mean shift and DBSCAN, particularly for small γ_{st}. The DPGMM also achieves larger P_c than KDE. Meanwhile, despite the lack of the prior knowledge of L, DPGMM is only slightly inferior to EMGMM, and the gap becomes negligible for increased γ_{st} or $N^{(s)}$.

B. Theoretical Verification of the Prediction-Transmission Structure

In this subsection, we first illustrate the optimized thresholds through numerical search. Note that only the case in Fig. 5(a) will occur in our scenario. Then, we compare the maximized utility $V(0, 1, k)$ in (35) with numerical results.

In Fig. 7(a), we illustrate p_τ^{k*} in (43) for each PT transmit power levels identified by the ST. It is shown that p_τ^{k*} is time varying, and is larger for higher PT transmit power level, which indicates that the ST favors a prediction action for a higher transmit power level k. However, it is seen that p_τ^{k*} becomes 1 when the time exceeds $\mathcal{T}_k$, which means that the ST will not transmit after $\mathcal{T}_k$ regardless of the value of p_τ^k. This observation is consistent with (39). In Fig. 7(b), we compare the maximized utility $V(0, p_\tau^{k*}, k)$ in (35) for different k. It is observed that the numerical results match the theoretical ones well, which verifies the effectiveness of the proposed reinforcement learning method. It is seen that $V(0, 1, k)$ decreases with τ_s for all k. This is because the increase in τ_s renders the prediction of l in the next transmission block less accurate. Thus, the ST is more likely to carry out prediction when the current PT power level duration is close to the end. Naturally, this decreases the utility.

C. Advantages of the Conditionally Conjugate DPGMM in Stage I

In this subsection, we illustrate the impact of different learning methods in Stage I on the NPLA performance of the proposed spectrum sharing strategy. The NPLA performance from the beginning to time point τ_t is defined as

$$U(\tau_t) = \frac{\tau_s}{\tau_t} \sum_{\tau_0=0}^{\tau_t} a_{\tau_0 - \tau_0^*} \psi\left(\sum_{\tau_1=0}^{\tau_s-1} |k_{\tau_0+\tau_1} - l_{\tau_0+\tau_1}|\right), \tag{47}$$

where τ_0^* is the time point of last PT power level change, k_τ and l_τ denote the ST and PT transmit power levels at time τ, respectively. We have $\psi(x) = 1$ when $x = 0$, and $\psi(x) = 0$
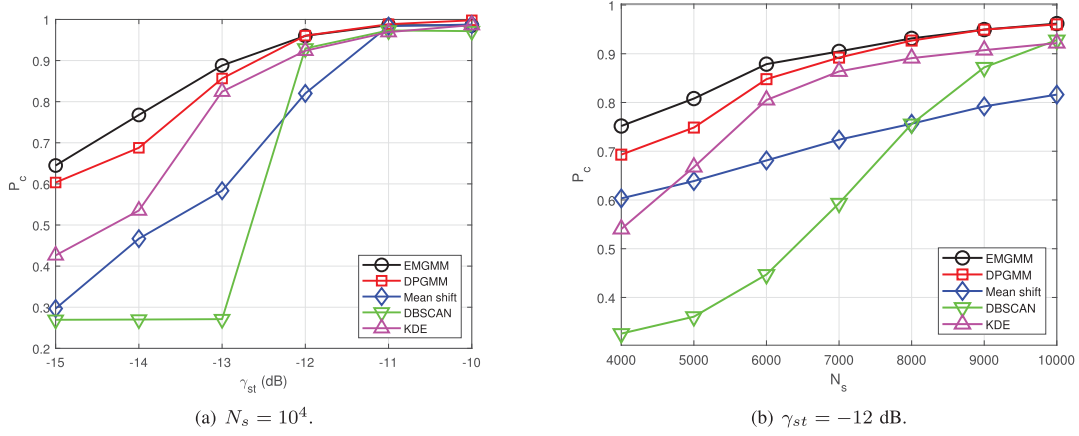
(a) $N_s = 10^4$.

(b) $\gamma_{st} = -12$ dB.

Fig. 6. The probability of correct PT power level prediction in Stage II (P_c) versus γ_{st} and $N^{(s)}$.

(a) The optimized threshold p_τ^{k*} versus τ with $\tau_s = 4$.

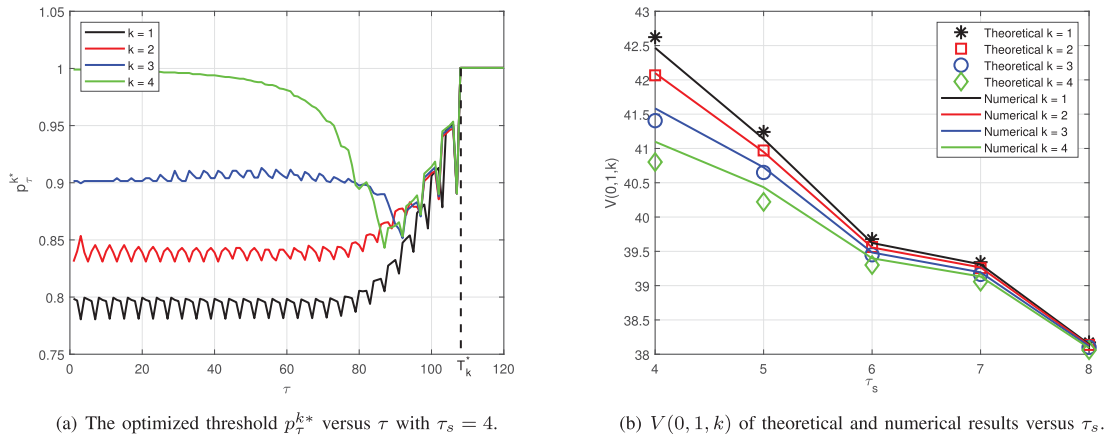(b) $V(0, 1, k)$ of theoretical and numerical results versus τ_s.

Fig. 7. Theoretical verification of the prediction-transmission structure with $f_s = 5$ MHz, $N^{(s)} = 10^4$, $\gamma_{st} = -12$ dB, and $\nu = 100$.

(a) Periodic structure for the prediction-transmission structure.

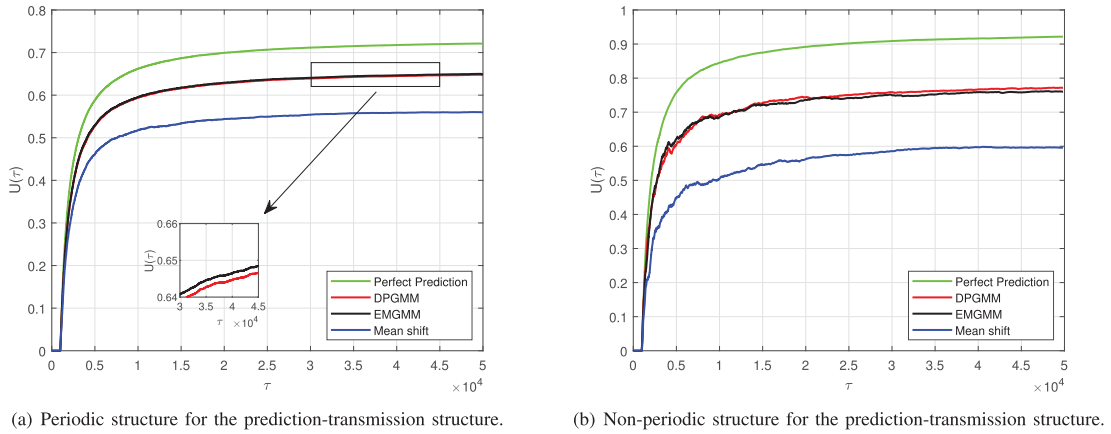(b) Non-periodic structure for the prediction-transmission structure.

Fig. 8. The NPLA performance $U(\tau_t)$ for the spectrum sharing strategies using different clustering methods in Stage I with $f_s = 2.5$ MHz, $N^{(s)} = 5 \times 10^3$, $\tau_s = 4$, $\gamma_{st} = -12$ dB, and $\nu = 50$.

otherwise. Basically, a larger $U(\tau_t)$ leads to a better tradeoff between the secondary network throughput and the interference to the primary network.

In Fig. 8, we compare $U(\tau_t)$ for different learning algorithms, considering both periodic (Fig. 8(a)) and non-periodic structures (Fig. 8(b)) in the prediction part of Stage II. In addition to

the three learning methods in Fig. 6 (conditionally conjugate DPGMM, EMGMM and mean shift), we include for reference an upper bound with perfect prediction. The same N, T_{ss} and τ_s are assumed for all the methods.

In Fig. 8(a), it is shown that $U(\tau_t)$ remains 0 when $\tau \leqslant 10^3$ due to the learning period (i.e., no transmission), and begins
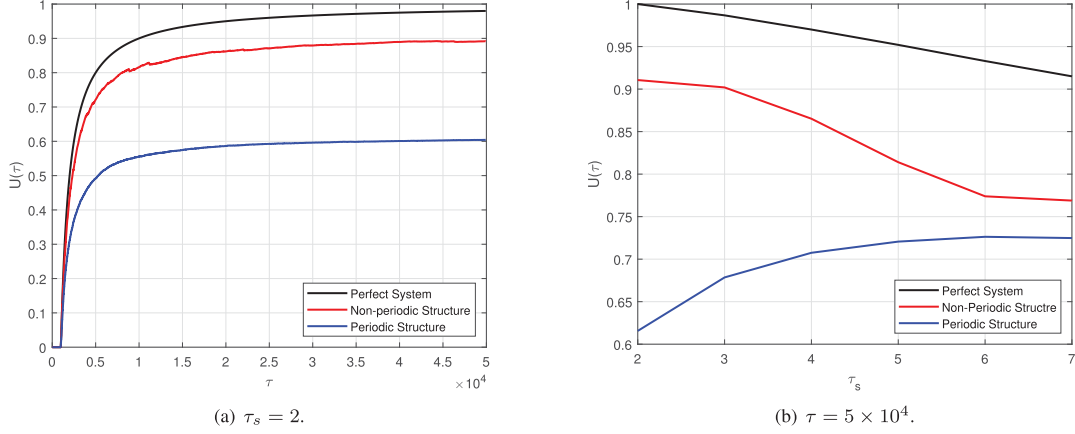
Fig. 9. The NPLA performance $U(\tau_t)$ for the spectrum sharing strategies using different structures in Stage II with $f_s = 5$ MHz, $N^{(s)} = 10^4$, $\gamma_{st} = -12$ dB, and $\nu = 50$.

to increase when the ST goes into Stage II $(\tau_t > 10^3)$. $U(\tau_t)$ refers to the average utility from the beginning of Stage I to time τ_t, and will keep stable in the long term. Thus it first increases significantly, and then asymptotically approaches a constant. With reference to Fig. 6(b), we can see that the learning method with better P_c corresponds to higher $U(\tau_t)$. This clearly collaborates the significance of the prediction accuracy. Similar conclusions can be drawn in Fig. 8(b) for the non-periodic structure.

D. Advantages of the Prediction-Transmission Structure in Stage II

In Fig. 9, we compare $U(\tau_t)$ of the proposed spectrum sharing strategy with periodic and non-periodic structures in Stage II for different τ_t and τ_s. Note that conditionally conjugate DPGMM is used in Stage I for all the cases. As an upper bound, we include a perfect system where the ST can always accurately track the PT power level variation.

Similar to Fig. 8, it is shown in Fig. 9(a) that, for each curve, $U(\tau_t)$ remains 0 when $\tau_t \leqslant 10^3$ and asymptotically approaches a constant as τ_t increases. Fig. 9(a) shows that the non-periodic structure outperforms the periodic one, which verifies the benefit of dynamically adjusting the prediction intervals.

Fig. 9(b) shows $U(\tau_t)$ of the three different structures versus τ_s when $\tau_t = 5 \times 10^4$. We find that when τ_s increases, $U(\tau_t)$ of the periodic structure increases. This is because a larger τ_s results in a smaller time proportion of prediction, and it will benefit the periodic strategy. By contrast, for non-periodic structure and the perfect system, $U(\tau_t)$ decreases with the increase of τ_s, as the ST is less flexible in transmission block allocation when τ_s increases.

E. Performance of the Two-Stage Spectrum Sharing Strategy With Online Function

In Fig. 10, we compare $U(\tau_t)$ of the proposed spectrum sharing strategies with and without the online function when the PT power level changes after Stage I. In this scenario, L increases to 5 when $\tau_t = 10^4$, where $\gamma_{st}^4 = 4\gamma_{st}^1$, $\gamma_{st}^5 = 0$, and other γ_{st}^l remain the same. When $\tau_t \leqslant 10^4$, the online strategy
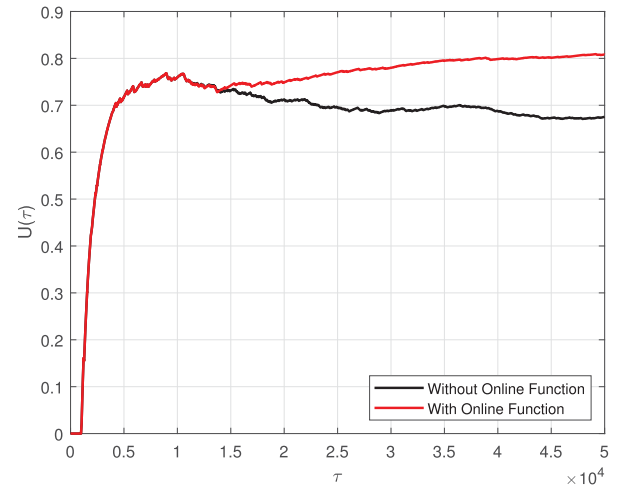

Fig. 10. $U(\tau_t)$ for the spectrum sharing strategies with and without online function in an online scenario with $f_s = 5$ MHz, $N^{(s)} = 10^4$, $\tau_s = 4$, $\gamma_{st} = -12$ dB, and $\nu = 50$.

has the same $U(\tau_t)$ performance as the regular strategy without the online function. When $\tau_t \geqslant 10^4$, $U(\tau_t)$ of the regular strategy decreases slowly. This is because the learning results becomes less accurate with an updated L, thus the existing strategy becomes less effective. In contrast, $U(\tau_t)$ of the online strategy first decreases when $10^4 \leqslant \tau_t \leqslant 1.4 \times 10^4$ due to the time consumption when updating $\mathbf{X}$. When $\tau_t > 1.4 \times 10^4$, it is shown that $U(\tau_t)$ can continue to increase based on the updated knowledge of L.

In summary, the simulation results provided in the above subsections verify that the learning method is beneficial to the design of spectrum sharing with multiple PT power levels. With the advent of 5G and industrial Internet of Things in the years to come, the proposed spectrum sharing provides a promising solution to the spectrum scarcity problem.

F. Discussion and Extension

1) Rayleigh Fading Channel: For the convenience of the exposition of our proposed method, we have not considered

fading. However, the extension to fading scenario is straightforward. Specifically, for Rayleigh fading, R_n in (1) can be written as [36]

$$R_n[i] = \sqrt{P_{PT,l}} h_n s_n[i] + u_n[i], \quad \mathcal{H}_l, \tag{48}$$

where $h_n \sim \mathcal{CN}(0,1)$ is the fading channel coefficient. Accordingly, the test statistic X_n still follows a Gaussian distribution as is the non-fading case, but with a different mean $(|h_n|^2 \gamma_{st}^l + 1)\sigma_u^2$ and variance $\left(2|h_n|^2 \gamma_{st}^l + 1\right) \sigma_u^4 / N^{(s)}$. Clearly, when h_n experiences deep fading, the mean $(|h_n|^2 \gamma_{st}^l + 1)\sigma_u^2$ will become very small and our solution is less capable of distinguishing the PT power levels. To solve this issue, motivated by [36], we can introduce time diversity by combining test statistics covering N_p consecutive time slots ($N_p \geqslant 2$), i.e.,

$$Y_{n^*} = \sum_{n_p=1}^{N_p} w_{n_p} X_{(n^*-1)N_p+n_p}, \tag{49}$$

where w_{n_p} is the weighting factor associated with the n_p-th sensing slot and Y_{n^*} follows a Gaussian distribution with mean $\sum_{n_p=1}^{N_p} w_{n_p}(|h_{n_p}|^2 \gamma_{st}^l + 1)\sigma_n^2$ and variance $\sum_{n_p=1}^{N_p} w_{n_p}^2 \left(2|h_{n_p}|^2 \gamma_{st}^l + 1\right) \sigma_u^4 / N^{(s)}$.

2) Two Stages Switch: In the multi-armed bandit problem, dynamic switching between exploration and exploitation epochs [29]–[31] achieves a regret with logarithmic order. However, the design of the two stages in our work are decoupled. That is, Stage I is a one-off design, which solely determines the accuracy in the model parameters estimation. Consequently, simply switching between Stage I and Stage II brings no benefits to the NPLA performance improvement, unless some information in Stage II (e.g., the correctness of the prediction) is fed to Stage I to improve the accuracy in the model parameters estimation. How to incorporate the information in Stage II into Stage I is an interesting problem to be considered in our future work.

3) Practical Implementation: Note that energy detection with classical signal processing could potentially suffer from hidden node problem, and can not perform well at very low SNRs. From machine learning perspective, it is possible to combine signal strength with other features to dramatically improve the performance. For example, our recent results show that much more accurate sensing performance at very low SNRs (e.g., -20 dB) can be achieved by integrating cyclic-prefix induced correlation as another feature with signal energy feature. Therefore, the data-driven method may provide a potential solution to the hidden node problem. In Section II, we made the assumption that the test statistic X_n follows a mixed Gaussian distribution. In a real scenario, such assumption might not hold as the real radio propagation environment is far more complicated. In this case, as suggested in [27], a kernel based process may be applied, which could be considered as our future work.

VI. CONCLUSIONS

In this paper, we considered the challenging problem of spectrum sharing where the PT transmits with multiple power levels. We endeavored to provide a practical solution with minimal or no prior information on the PT power profiles. We drew on the recent development in machine learning, and proposed a learning based two-stage spectrum sharing strategy. In Stage I, a conditionally conjugate DPGMM was proposed to capture the PT power level variation. Then, a Bayesian inference method

was designed for model parameters inference to mathematically establish the PT power profile. Basically, Stage I depicts a big picture of the multi-level radio environments. Based on this knowledge, in Stage II, we designed prediction-transmission structures to enable the ST transmit power level to closely match that of the PT, minimizing the interference to the primary network. To achieve this, we relied on a new metric, NPLA, to characterize the extent of matching between PT and ST transmit power levels. To accommodate a more realistic scenario where the number of PT power levels might dynamically change after the learning, we developed an online function for the strategy. Finally, we verified the effectiveness of the proposed strategy with numerical results.

The significance of our work for future impact are two-fold. On one hand, this data-driven based method may open a door to a new modality of spectrum sensing. On the other hand, our work may find its potential applications from industrial aspects. For example, the recent Ofcom [49] development on TV white space is considering both geolocation database and signal strength-based spectrum sensing.

APPENDIX
PROOF OF LEMMA 2

We get the result by induction on τ. In Lemma 1, we give the $V(\tau, p_\tau^k, k)$ being a convex function of p_τ^k. Thus $V(\tau, p_\tau^k, k) \leqslant p_\tau^k V(\tau, 1, k) = 0, \forall \tau \geqslant \mathcal{T}_k$, which means that $V(\tau, p_\tau^k, k)$ increases in p_τ^k for given k with $\tau \geqslant \mathcal{T}_k$. Then, we prove the statement for $\tau < \mathcal{T}_k$.

Let the statement holds for $V(\tau, p_\tau^k, k), \tau \geqslant \mathcal{T}_k - \upsilon$, then Lemma 2 can be proved if both $E(\tau, p_\tau^k, k)$ and $A(\tau, p_\tau^k, k)$ in (36) increase in p_τ^k. Thus, we give the first derivative of $E(\tau, p_\tau^k, k)$, which is expressed as

$$\frac{dE(\tau, p_\tau^k, k)}{dp_\tau^k} = \sum_{j=1}^{L} g_\tau^S (H_{jk} - \boldsymbol{c}_k \boldsymbol{h}_j) V(\tau+1, p_{\tau+1}^k(\mathcal{H}_j), k)$$
$$+ \sum_{j=1}^{L} \frac{g_\tau^S H_{jk} \boldsymbol{c}_k \boldsymbol{h}_j V'(\tau+1, p_{\tau+1}^k(\mathcal{H}_j), k)}{p_\tau^k g_\tau^S H_{jk} + (1 - p_\tau^k g_\tau) \boldsymbol{c}_k \boldsymbol{h}_j}. \tag{50}$$

As the discretized sensing time equals to 1, we have $V'(\tau+1, p_{\tau+1}^k(\mathcal{H}_j), k) > 0$ for $\tau \geqslant \mathcal{T}_k - \upsilon - 1$, thus the second term in (50) holds positive. In order to prove the first term in (50) positive, we define

$$\mathcal{C} = \sum_{j=1}^{L} (H_{jk} - \boldsymbol{c}_k \boldsymbol{h}_j) V(\tau+1, p_{\tau+1}^k(\mathcal{H}_j), k)$$
$$= (H_{kk} - \boldsymbol{c}_k \boldsymbol{h}_k) V(\tau+1, p_{\tau+1}^k(\mathcal{H}_k), k)$$
$$+ \sum_{j=1,j\neq k}^{L} (H_{jk} - \boldsymbol{c}_k \boldsymbol{h}_j) V(\tau+1, p_{\tau+1}^k(\mathcal{H}_j), k). \tag{51}$$

In (51), we define

$$\mathcal{D}(k,j) = H_{jk} - \boldsymbol{c}_k \boldsymbol{h}_j < H_{jk} - H_{jj} C_{jk}. \tag{52}$$

In this paper, we consider such a case that the PT transfers to different states with similar probabilities ($C_{jk} \approx 1/(L-1)$,

$j \neq k$). Besides, the probability of the ST correctly clustering the signals is much larger than incorrectly clustering ($H_{jj} \gg LH_{jk}, j \neq k$), which can be verified by the numerical results. Thus $\mathcal{D}(k, j) < 0$. Besides, we define

$$
\begin{aligned}
\mathcal{F} &= p_{\tau+1}^k(\mathcal{H}_k) - p_{\tau+1}^k(\mathcal{H}_j) \\
&= \frac{p_\tau^k g_\tau^S (1 - p_\tau^k g_\tau^S)(H_{kk} \boldsymbol{c}_k \boldsymbol{h}_j - H_{jk} \boldsymbol{c}_k \boldsymbol{h}_k)}{[p_\tau^k g_\tau^S H_{kk} + (1 - p_\tau^k g_\tau^S) \boldsymbol{c}_k \boldsymbol{h}_k][p_\tau^k g_\tau^S H_{jk} + (1 - p_\tau^k g_\tau^S) \boldsymbol{c}_k \boldsymbol{h}_j]} \\
&= \frac{p_\tau^k g_\tau^S (1 - p_\tau^k g_\tau^S) \sum_{i=1}^{L} (H_{kk} H_{ji} - H_{jk} H_{ki}) C_{ik}}{[p_\tau^k g_\tau^S H_{kk} + (1 - p_\tau^k g_\tau^S) \boldsymbol{c}_k \boldsymbol{h}_k][p_\tau^k g_\tau^S H_{jk} + (1 - p_\tau^k g_\tau^S) \boldsymbol{c}_k \boldsymbol{h}_j]}.
\end{aligned}
\tag{53}
$$

Similarly, we can get $\mathcal{F} > 0$ for $j \neq k$. Thus, $V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_k), k) > V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_j), j)$ for $j \neq k$ and $\tau \geqslant \mathcal{T}_k - \upsilon$.

Therefore, we can reach the following result.

$$
\begin{aligned}
\mathcal{C} &> (H_{kk} - \boldsymbol{c}_k \boldsymbol{h}_k) V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_k), k) \\
&\quad + \sum_{j=1, j \neq k}^{L} (H_{jk} - \boldsymbol{c}_k \boldsymbol{h}_j) V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_k), k) \\
&= \left(\sum_{j=1}^{L} H_{jk} - \sum_{j=1}^{L} \sum_{i=1}^{L} H_{ji} C_{ik} \right) V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_k), k) \\
&= \left(1 - \sum_{i=1}^{L} C_{ik} \right) V(\tau + 1, p_{\tau+1}^k(\mathcal{H}_k), k) = 0.
\end{aligned}
\tag{54}
$$

Therefore, the first term in (50) is proved positive. Thus, $L(\tau, p_\tau^k, k)$ increases in p_τ^k for $\tau = \mathcal{T}_k - \upsilon - 1$. Similarly, the first derivative of $A(\tau, p_\tau^k, k)$ can be given by

$$
\begin{aligned}
\frac{dA(\tau, p_\tau^k, k)}{dp_\tau^k} &= g_\tau^T \left(1 - \sum_{j=1}^{L} C_{kj} \right) V \left(\tau + \tau_s, p_{\tau+\tau_s}^k(\mathcal{A}), k \right) \\
&\quad + \frac{g_\tau^T \sum_{j=1}^{L} C_{kj} V'(\tau + \tau_s, p_{\tau+\tau_s}^k(\mathcal{A}), k)}{p_\tau^k g_\tau^T + (1 - p_\tau^k g_\tau^T) \sum_{j=1}^{L} C_{kj}} \\
&\quad + \left(D_k + \sum_{j=1}^{L} C_{kj} Y_j \right) g_\tau^T \tau_s.
\end{aligned}
\tag{55}
$$

According to (28) and (39), we can get $\sum_{j=1}^{L} C_{kj} = 1$ and $D_k + \sum_{j=1}^{L} C_{kj} Y_j > 0$, which means (55) is positive and $A(\tau, p_\tau^k, k)$ increases in p_τ^k for $\tau = \mathcal{T}_k - \upsilon - 1$. In summary, both $E(\tau, p_\tau^k, k)$ and $A(\tau, p_\tau^k, k)$ in (36) increase in p_τ^k for $\tau = \mathcal{T}_k - \upsilon - 1$ and we complete the proof.
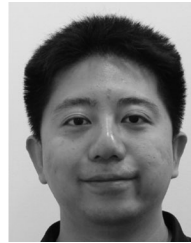
REFERENCES

[1] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[2] D. Cabric, S. M. Mishra, and R. W. Brodersen, "Implementation issues in spectrum sensing for cognitive radios," in *Conf. Rec. 38th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2004, vol. 1, pp. 772–776.

[3] A. I. Perez-Neira, M. A. Lagunas, M. A. Rojas, and P. Stoica, "Correlation matching approach for spectrum sensing in open spectrum communications," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4823–4836, Dec. 2009.

[4] P. D. Sutton, K. E. Nolan, and L. E. Doyle, "Cyclostationary signatures in practical cognitive radio applications," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 1, pp. 13–24, Jan. 2008.

[5] V. R. Cadambe and S. A. Jafar, "Interference alignment and degrees of freedom of the k-user interference channel," *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[6] T. Yucek and H. Arslan, "A survey of spectrum sensing algorithms for cognitive radio applications," *IEEE Commun. Surv. Tuts.*, vol. 11, no. 1, pp. 116–130, Jan.–Mar. 2009.

[7] S. Atapattu, C. Tellambura, and H. Jiang, "Energy detection based cooperative spectrum sensing in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 4, pp. 1232–1241, Apr. 2011.

[8] Y. Zou, Y. D. Yao, and B. Zheng, "Cooperative relay techniques for cognitive radio systems: Spectrum sensing and secondary user transmissions," *IEEE Commun. Mag.*, vol. 50, no. 4, pp. 98–103, Apr. 2012.

[9] W. Zhang, R. K. Mallik, and K. B. Letaief, "Optimization of cooperative spectrum sensing with energy detection in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5761–5766, Dec. 2009.

[10] M. E. Ahmed, J. B. Song, Z. Han, and D. Y. Suh, "Sensing-transmission edifice using Bayesian nonparametric traffic clustering in cognitive radio networks," *IEEE Trans. Mobile Comput.*, vol. 13, no. 9, pp. 2141–2155, Sep. 2014.

[11] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2053–2071, May 2008.

[12] A. Ghasemi and E. S. Sousa, "Fundamental limits of spectrum-sharing in fading environments," *IEEE Trans. Wireless Commun.*, vol. 6, no. 2, pp. 649–658, Feb. 2007.

[13] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, May 2007.

[14] I. Committee *et al.*, "Wireless LAN medium access control (MAC) and physical layer (PHY) specifications," IEEE Std 802-11-1997.

[15] M. Mouly and M.-B. Pautet, *The GSM System for Mobile Communications*. Palaiseau, France: Telecom Publishing, 1992.

[16] J. Remy and C. Letamendia, *LTE Standards*, ser. ISTE. New York, NY, USA: Wiley, 2014.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[18] F. Gao, J. Li, T. Jiang, and W. Chen, "Sensing and recognition when primary user has multiple transmit power levels," *IEEE Trans. Signal Process.*, vol. 63, no. 10, pp. 2704–2717, May 2015.

[19] Z. Li, S. Cheng, F. Gao, and Y. C. Liang, "Sequential detection for cognitive radio with multiple primary transmit power levels," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2769–2780, Jul. 2017.

[20] F. Gao, C. Qian, H. Qian, and T. Zhang, "Sensing and recognition for multiple-primary-power-level scenario with noise uncertainty," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2289–2300, Mar. 2017.

[21] D. Wang, N. Zhang, Z. Li, F. Gao, and X. Shen, "Leveraging high order cumulants for spectrum sensing and power recognition in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1298–1310, Feb. 2018.

[22] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1238–1274, 2013.

[23] Y. Shi, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, Z. Lu, and J. H. Li, "Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2018, pp. 1–6.

[24] K. M. Thilina, K. W. Choi, N. Saquib, and E. Hossain, "Machine learning techniques for cooperative spectrum sensing in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2209–2221, Nov. 2013.

[25] N. T. Nguyen, R. Zheng, and Z. Han, "On identifying primary user emulation attacks in cognitive radio systems using nonparametric Bayesian classification," *IEEE Trans. Signal Process.*, vol. 60, no. 3, pp. 1432–1445, Mar. 2012.

[26] W. Lee, M. Kim, and D. Cho, "Deep cooperative sensing: Cooperative spectrum sensing based on convolutional neural networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3005–3009, Mar. 2019.

[27] Y. Xu, P. Cheng, Z. Chen, Y. Li, and B. Vucetic, "Mobile collaborative spectrum sensing for heterogeneous networks: A Bayesian machine learning approach," *IEEE Trans. Signal Process.*, vol. 66, no. 21, pp. 5634–5647, Nov. 2018.

[28] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[29] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.

[30] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.

[31] T. Gafni and K. Cohen, "Learning in restless multi-armed bandits using adaptive arm sequencing rules," in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 2018, pp. 1206–1210.

[32] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.

[33] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[34] U. Challita, L. Dong, and W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4674–4689, Jul. 2018.

[35] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7.

[36] Y. C. Liang, Y. Zeng, E. C. Y. Peh, and A. T. Hoang, "Sensing-throughput tradeoff for cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 4, pp. 1326–1337, Apr. 2008.

[37] C. E. Rasmussen, "The infinite Gaussian mixture model," in *Proc. Advances Neural Inf. Process. Syst.*, 2000, pp. 554–560.

[38] FCC Notice of Inquiry and Notice of Proposed Rule Making, "In the matter of establishment of an interference temperature metric to quantify and manage interference and to expand available unlicensed operation in certain fixed, mobile and satellite frequency bands," *ET Docket*, no. 03–237, Nov. 2003.

[39] D. Astély, E. Dahlman, A. Furuskär, Y. Jading, M. Lindström, and S. Parkvall, "LTE: The evolution of mobile broadband," *IEEE Commun. Mag.*, vol. 47, no. 4, pp. 44–51, Apr. 2009.

[40] D. Görür and C. Edward Rasmussen, "Dirichlet process Gaussian mixture models: Choice of the base distribution," *J. Comput. Sci. Technol.*, vol. 25, no. 4, pp. 653–664, Jul. 2010.

[41] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *J. Comput. Graph. Statist.*, vol. 9, no. 2, pp. 249–265, 2000.

[42] Y. W. Teh, "Dirichlet process," in *Encyclopedia of Machine Learning*. New York, NY, USA: Springer, 2011, pp. 280–287.

[43] W. R. Gilks and P. Wild, "Adaptive rejection sampling for Gibbs sampling," *J. Appl. Statist.*, vol. 41, no. 2, pp. 337–348, 1992.

[44] G. E. Monahan, "A survey of partially observable Markov decision processes: Theory, models, and algorithms," *Manag. Sci.*, vol. 28, no. 1, pp. 1–16, 1982.

[45] S. M. Ross, *Introduction to Stochastic Dynamic Programming*. Orlando, FL, USA: Academic, 1983.

[46] S. Huang, X. Liu, and Z. Ding, "Optimal sensing-transmission structure for dynamic spectrum access," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2295–2303.

[47] J. A. Bilmes *et al.*, "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models," *Int. Comput. Sci. Inst.*, vol. 4, no. 510, p. 126, 1998.

[48] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 603–619, May 2002.

[49] Ofcom, "Implementing TV white spaces." Feb. 2015.

Peng Cheng (M'12) received the B.S. and M.S degrees (with great Hons.) in communication and information systems from the University of Electronic Science and Technology of China, Chengdu, China, in 2006 and 2009, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2013. From 2014 to 2017, he was a Postdoctoral Research Scientist with Commonwealth Scientific and Industrial Research Organization, Sydney, Australia, and he is currently with the University of Sydney, Australia. He has authored and coauthored more than 60 peer-reviewed research papers in leading international journals and conferences. His current research interests include machine learning, network MIMO, OFDM, millimeter-wave communications, and compressive sensing theory. He is a recipient of the Australian DECRA Fellowship in 2019.
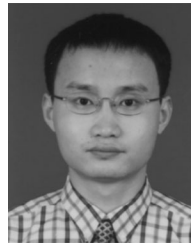
Zhuo Chen (S'01–M'05–SM'18) received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, in 1997 and the M.S. and Ph.D. degrees from the School of Electrical and Information Engineering, the University of Sydney, Sydney, Australia, in 2001 and 2004, respectively. He was a Senior Research Scientist with Commonwealth Scientific and Industrial Research Organization, Sydney, Australia. His research interests include wireless communications, machine learning, and wireless sensor networks.

Yonghui Li (M'04–SM'09–F'19) received the Ph.D. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in November 2002. From 1999 to 2003, he was affiliated with the Linkair Communication Inc, where he held a position of Project Manager with responsibility for the design of physical layer solutions for the LAS-CDMA system. Since 2003, he has been with the Centre of Excellence in Telecommunications, the University of Sydney, Australia. He is currently a Professor with the School of Electrical and Information Engineering, University of Sydney. His current research interests are in the area of wireless communications, with a particular focus on MIMO, millimeter wave communications, machine to machine communications, coding techniques, and cooperative communications.

Prof. Li is the recipient of the Australian Queen Elizabeth II Fellowship in 2008 and the Australian Future Fellowship in 2012. He holds a number of patents granted and pending in these fields. He is currently an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He was also the Guest Editor for the IEEE Journal on Selected Areas in Communications Special issue on Millimeter Wave Communications for Future Mobile Networks. He received the best paper awards from IEEE International Conference on Communications 2014, IEEE PIMRC 2017, and IEEE Wireless Days Conferences 2014.

Rui Zhang received the B.S. degree in information engineering from the Southeast University, Nanjing, China, in 2016. He is currently working toward the Ph.D. degree in telecommunication at The University of Sydney, Sydney, Australia. He holds a postgraduate research scholarship. His research interest includes cognitive radio networks and machine learning in wireless communications.

Branka Vucetic (F'03) is an ARC Laureate Fellow and Director of the Centre of Excellence for IoT and Telecommunications with the University of Sydney. Her current research interests include wireless networks and the Internet of Things. In the area of wireless networks, she works on communication system design for millimeter wave frequency bands. In the area of the Internet of Things, her works on providing wireless connectivity for mission critical applications. She is a Fellow of the Australian Academy of Technological Sciences and Engineering and the Australian Academy of Science.