RESPONSES TO THE EDITOR AND THE REVIEWERS

We would like to thank the Editor and the Reviewers for their careful assessment of the manuscript and for their comments, which we found very useful to improve the quality of the manuscript. We have addressed all the comments in the revised version to the best of our abilities, and you will find the point to point responses to your comments below. Your comments are in italics, our response follows after "**R:**".

To facilitate the second round of revisions, we highlighted the major changes using blue font in the manuscript. Although the manuscript has been thoroughly revised for clarity, consistence of notation, etc., minor typographical changes and rewording have not been highlighted.

Regards,
Bharath Keshavamurthy and Nicolò Michelusi

EDITOR

We thank the Editor for the timely completion of the review process. We have addressed all the Reviewers' comments to the best of our abilities, carrying out a major revision of the system model, along with a number of editorial changes and additional explanations. We believe that this version of the paper is much improved. As per the Reviewers' suggestions,

- We have addressed – both theoretically and empirically – our hypothesis that the correlation in LU occupancy behavior across frequency can be reversed, i.e., in time-step t, the occupancy of subcarrier k can depend interchangeably on the occupancy of either subcarrier $k-1$ or subcarrier $k+1$.
- In our numerical evaluations, we have incorporated a more realistic channel model, that includes random deployments of transmitters and receivers, probabilistic LoS and NLoS links, SINR variations due to stochastic variations of large- and small-scale channel conditions, and a rate adaptation scheme at the CRs to ensure maximum exploitation of the correctly inferred white-spaces.
- We have addressed the computational time complexity of the proposed neighbor discovery and channel access order allocation schemes by bench-marking their performance in highly-mobile DARPA SC2 scenarios – specifically, a disaster-relief scenario named "Payline", and a troop-deployment scenario named "Alleys of Austin".
- We have also performed computational time complexity analysis of our complete system (Monte-Carlo Fragmented PERSEUS with Hamming distance state filters), and bench-marked it against relevant works in the state-of-the-art to demonstrate superior scalability.
- Finally, we have addressed the suggested need for a Receiver Operating Characteristics (ROC) plot of the proposed system, along with evaluations against comparable works in the state-of-the-art.
- We have also changed the title from "Learning-based Spectrum Sensing in Cognitive Radio Networks via Approximate POMDPs" to "Learning-based Spectrum Sensing and Access in Cognitive Radios via Approximate POMDPs" to better reflect the contributions of the paper.

A detailed point-to-point response to the reviewers follows.

REVIEWER 1

1) **Comment**: *This work proposes a spectrum sensing and access strategy, wherein a cognitive radio learns a time-frequency correlation model defining the occupancy behavior of incumbents via the Baum-Welch algorithm, and concurrently devises an optimal spectrum sensing and access strategy that exploits this learned correlation model. This paper is in general well written. However, several major flaws exist, and they are listed below.*

R: We thank the reviewer for the valuable assessment. We have addressed the comments to the best of our abilities, and believe that this version of the paper is much improved. Please note that, in the revised manuscript, we refer to SU as CR (Cognitive Radio) and to PU as LU (Licensed User), as opposed to SU and PU used in the original submission. In addition, we have termed our proposed framework LESSA (LEarning-based Spectrum Sensing and Access) and its Multi-Agent version (MA-LESSA), to more directly refer to it throughout the manuscript. We have also changed the title from "Learning-based Spectrum Sensing in Cognitive Radio Networks via Approximate POMDPs" to "Learning-based Spectrum Sensing and Access in Cognitive Radios via Approximate POMDPs" to better reflect the contributions of the paper.

2) **Comment**: *As claimed, PUs typically occupy a set of adjacent subcarriers, which leads to frequency correlation. That is, the occupancy of frequency band k is related to the occupancy of the adjacent frequency bands $k-1$ and $k+1$. However, the authors assume that the occupancy of frequency band k only depends on the occupancy of frequency band $k-1$ in Eq. 6. Is that reasonable?*

R: We thank the reviewer for this valuable comment, and for giving us the opportunity to clarify the time-frequency correlation model. In the revised manuscript, we have added details to motivate the structure of this model, which we believe addresses this concern: from the revised manuscript (Section II-A on page 8),

"To capture temporal correlation, we model the evolution of $\vec{B}(t)$ over time as a Markov process

$$\mathbb{P}(\vec{B}(t+1)|\vec{B}(j), \forall j \leq i) = \mathbb{P}(\vec{B}(t+1)|\vec{B}(t)) \triangleq P_B(\vec{B}(t+1)|\vec{B}(t)), \quad (32)$$

with one-step transition probability $P_B(\vec{B}(t+1)|\vec{B}(t))$. Using the chain rule of con-

ditional probability, we can further express it as

$$P_B(\vec{B}(t+1)|\vec{B}(t)) = \mathbb{P}(B_1(t+1)|\vec{B}(t)) \prod_{k=2}^{K} \mathbb{P}(B_k(t+1)|\vec{B}_{1:k-1}(t+1), \vec{B}(t)), \quad (33)$$

where $\vec{B}_{1:k-1}(t+1)$ is the spectrum occupancy state in subcarriers 1 to $k-1$ in time-slot $t+1$. We further assume a Markovian structure across frequency,

$$\mathbb{P}(B_k(t+1)|\vec{B}_{1:k-1}(t+1), \vec{B}(t)) = \mathbb{P}(B_k(t+1)|B_{k-1}(t+1), \vec{B}(t)),$$

i.e., $B_k(t+1)$ is independent of the spectrum occupancy in the non-adjacent subcarriers $\vec{B}_{1:k-2}(t+1)$, when conditioned on $\vec{B}(t)$ and on the spectrum occupancy in the adjacent subcarrier $B_{k-1}(t+1)$. This assumption reflects the intuition that the state of the adjacent subcarrier $k-1$ ($B_{k-1}(t+1)$) more directly affects the state of subcarrier k than non-adjacent subcarriers 1 to $k-2$. Replacing this expression into (33), we finally obtain

$$P_B(\vec{B}(t+1)|\vec{B}(t)) = \mathbb{P}(B_1(t+1)|\vec{B}(t)) \prod_{k=2}^{K} \mathbb{P}(B_k(t+1)|B_{k-1}(t+1), \vec{B}(t)). \quad (34)$$

We denote this model as *bottom-up frequency correlation*, since the state of subcarrier k depends on that of the adjacent lower subcarrier $k-1$, as opposed to the *top-down frequency correlation* where it depends on that of the adjacent upper subcarrier $k+1$. We remark that bottom-up and top-down frequency correlation models can be used interchangeably: in fact, replacing $\mathbb{P}(B_k(t+1)|B_{k-1}(t+1), \vec{B}(t)) = \mathbb{P}(B_{k-1}(t+1)|B_k(t+1), \vec{B}(t)) \frac{\mathbb{P}(B_k(t+1)|\vec{B}(t))}{\mathbb{P}(B_{k-1}(t+1)|\vec{B}(t))}$ (Bayes' rule) in (34) and simplifying, we obtain the top-down frequency correlation model

$$P_B(\vec{B}(t+1)|\vec{B}(t)) = \mathbb{P}(B_K(t+1)|\vec{B}(t)) \prod_{k=1}^{K-1} \mathbb{P}(B_k(t+1)|B_{k+1}(t+1), \vec{B}(t)), \quad (35)$$

so that the two models can be directly mapped to each other. We now embed the bottom-up frequency correlation of (34) into a parametric form. Expressing $\vec{B}(t) = [B_k(t), \vec{B}_{-k}(t)]$, where $\vec{B}_{-k}(t)$ is the spectrum occupancy state in subcarriers other than k in time-slot t, we define

$$q_w \triangleq \mathbb{P}(B_1(t+1) = 1|B_1(t)=w, \vec{B}_{-1}(t)) : \forall w \in \{0,1\}, \quad (36)$$

$$p_{u,v} \triangleq \mathbb{P}(B_k(t+1)=1|B_{k-1}(t+1)=u, B_k(t)=v, \vec{B}_{-k}(t)) : \forall u,v \in \{0,1\}, 2 \leq k \leq K. \quad (37)$$

Note that this parametric model assumes that $B_k(t+1)$ is independent of $\vec{B}_{-k}(t)$, when conditioned on $(B_{k-1}(t+1), B_k(t))$: intuitively, the state of subcarrier k in time-slot $t+1$ is most directly affected by the state on the same subcarrier in the previous time-slot t [...] "

In the revised manuscript, we have also performed a Bayesian Information Criterion (BIC) fit evaluation for the top-down and bottom-up correlation models using the DARPA SC2 [18] Active Incumbent [14] PSD data. This evaluation shows that the two models have similar BIC values: the bottom-up model has a BIC of 71.872; the top-down model, using a similar parameterization of (36), yields a similar BIC metric of 74.207 (the slightly different value is due to the parametric structure). Please refer to the discussion in Section II-B.

3) **Comment**: *What is the definition of* Γ *in Eq.* (8)*? I cannot find the explanations for this variable.*

R: We thank the reviewer for pointing out this typographical error in our original manuscript: Γ (the number of model parameters) was incorrectly identified as γ in the BIC evaluation description in Section II-B. In our revised manuscript, this has been corrected to accurately identify Γ as the number of model parameters in our BIC model fit validation.

4) **Comment**: *In Sec. IV, the SINRs of SU and PU under each scenario are assumed to be a constant regardless of the PU index, channel index, and time-slot index. This assumption may not be practical and should be addressed.*

R: We agree with the reviewer that this assumption may not be practical. We have therefore redone our numerical evaluations based on a more realistic channel model, that incorporates random deployments of transmitters and receivers, probabilistic LoS and NLoS links, SINR variations due to stochastic variations of large- and small-scale channel conditions, and a rate adaptation scheme at the CRs to ensure maximum exploitation of the correctly inferred white-spaces. We quote our modifications to the manuscript below (Section IV on page 18):

"LUs' and CR's transmitters and receivers are deployed randomly across an operational region: the three LUs' transmitters are placed at positions $(-225, 200)$m, $(225, 200)$m and $(0, -300)$m, at a height of 40m; the corresponding LU receivers are stationary nodes, placed randomly within a circular radius of 200m from the respective LU's transmitter. The CR's transmitter is located at position $(0, 0)$m at a height of 20m; the corresponding receiver is at ground level, and moves along a randomly generated

trajectory within a radius of 100m from the CR's transmitter. As described below, we have also facilitated rate adaptation at the CR based on the perceived SINR, which may vary with time and channel indices. The CR is capable of sensing κ=6 subcarriers per time-slot."

"For each transmitter (Tx, LU or CR) and receiver (Rx; the intended one, or unintended, thus experiencing interference) pairs, we model the channel on the kth subcarrier as $\hbar_k=\sqrt{\psi}\omega_k$, where ψ and $\omega_k\in\mathbb{C}$ encapsulate the large- and small-scale channel variations, respectively – with $\mathbb{E}[|\omega_k|^2]=1$. Given the angle of elevation between the Tx/Rx pair under consideration, $\chi\in(0,\frac{\pi}{2}]$, we generate the line-of-sight (LoS) condition with probability $P_{\text{LoS}}(\chi)=\frac{1}{1+z_1e^{-z_2(\chi-z_1)}}$ [39], so that the non-LoS (NLoS) condition occurs with probability $P_{\text{NLoS}}(\chi)=1-P_{\text{LoS}}(\chi)$, where z_1, z_2 are parameters specific to the propagation environment. Given the LoS or NLoS condition, and the distance d between Tx and Rx, we then generate the large- and small-scale channel conditions as follows. If the channel is LoS: the large-scale fading coefficient ψ is modeled as $\psi_{\text{LoS}}(d)=\psi_0 d^{-\mu_L}$, where ψ_0 is the reference pathloss at a distance of 1m from the Tx and $\mu_L\geq 2$ is the LoS pathloss exponent; the small-scale fading coefficient ω_k is modeled as a Rician with K-factor $\mathbb{K}(\chi)=f_1e^{f_2\chi}$, where f_1, f_2 are parameters specific to the propagation environment. Conversely, if the channel is NLoS: the large-scale fading coefficient ψ is modeled as $\psi_{\text{NLoS}}(d)=\iota\psi_0 d^{-\mu_N}$, where $\iota\in(0,1]$ is the additional NLoS attenuation and $\mu_N\geq\mu_L$ is the NLoS pathloss exponent; the small-scale fading coefficient ω_k is modeled as Rayleigh distributed (i.e., Rician with K-factor $\mathbb{K}(\chi)=0$). Throughout the simulation, we use μ_L=2.0, μ_N=2.8, ι=0.2, W=160 kHz, f_1=1.0, f_2=0.0512, z_1=9.12, and z_2 = 0.16 [40]."

Please refer to the description in Section IV. Furthermore, Fig. 5 illustrates the numerical evaluations of our system and those of comparable works in the state-of-the-art, under this enhanced channel model.

5) **Comment**: *In Sec. V, the update process for the aggregated-ranked list keeps repeating until a consensus is reached. The time consumed by this repetitive process should also be addressed.*

R: In our revised manuscript, we have incorporated a Big-O algorithmic computational complexity analysis of our neighbor discovery and channel access order allocation schemes

in multi-agent deployments, as well as numerical evaluations to bench-mark their performance in highly-mobile DARPA SC2 scenarios – specifically, a disaster-relief scenario named “Payline", and a troop-deployment scenario named “Alleys of Austin". We quote our modifications to the manuscript below:

"The computational time complexity of the RSSI thresholding scheme for neighbor discovery is $O(J_C^2)$." (Section V on page 25)

"The computational time-complexity of the quorum-based preferential ballot scheme for channel access rank allocation is $O(\tilde{T}J_C^2)$ [15], where $\tilde{T}$ corresponds to the number of iterations involved until a consensus is reached – which depends on the mobility patterns of these nodes along with the temporal evolution of the peer-to-peer link qualities." (Section V on page 26)

"To evaluate the proposed neighbor discovery (RSSI thresholding) and channel access rank allocation (quorum-based preferential ballot) heuristics from a computational time complexity perspective, we retrofit these schemes into the control channel design, collaboration, and data aggregation modules of the Purdue BAM! Wireless radio [25], and analyse their feasibility in emulations of highly mobile real-world scenarios – namely, military deployments in the Alleys of Austin scenario [32] (urban: $5\times$ $[9-$guardsmen $+$ $1-$UAV$]$) and disaster relief deployments in the Payline scenario [31] (urban: $5\times$ $[9-$EMTs $+$ $1-$HQ$]$). Along with the scenario-specific node mobility patterns (Fig. 8), the results of these emulations are shown in Fig. 9: neighbor discovery list changes are minimal in spite of node mobility (with an RSSI threshold of 22dB), and distributed convergence of channel access rank allocation across the ensemble is achieved within the first few iterations in any given 10s time-step." (Section V-B on page 28)

Please refer to the added discussion in Section V of the revised manuscript. Additionally, please refer to Fig. 8 and Fig. 9 in the revised manuscript and their accompanying discussions.

6) **Comment**: *One relevant issue for future bench-marking would be some information on the computation complexity of the proposed strategy.*

R: In our revised manuscript, we have included Big-O computational complexity analyses as well as numerical evaluations for our parameter estimation algorithm (Baum-Welch

[20]), our approximate POMDP value iteration algorithm (PERSEUS [21] with newly developed embedded heuristics including Fragmentation, Hamming distance state filter heuristics and Monte-Carlo techniques), and the concurrent combination of the two. We quote our modifications to the manuscript below.

"The computational time complexity of the Baum-Welch algorithm is $O(\tau K \tilde{T})$ [20], where $\tilde{T}$ is the number of iterations until convergence, which depends on the consistency of spectrum occupancy measurements, driven by our observation model and the CR's sensing limitations." (Section III-A on page 14)

"The computational time complexity of Algorithm 1 is $O(\tilde{T}|\tilde{\mathcal{B}}|^2 2^{2K'})$ [21], where $\tilde{T}$ denotes the number of iterations until convergence. Note here that incorporating Hamming distance state filters to alleviate the computational intractability inherent in PERSEUS belief update further mitigates the exponential dependence on the fragment size." (Section III-B on page 18)

"Fig. 4b plots the run time of LESSA as a function of the number of subcarriers, against TD-SARSA with LFA [7] and Adaptive DQN [13]. The algorithms are run on a $2\times$ 12-core Intel Xeon Gold 6126 @ 2.6 GHz compute node with 192 GB RAM [42]: as the number of subcarriers increases, our solution scales better yielding a more computationally tractable performance relative to the other two, owing to fragmentation and belief update simplification heuristics. Interestingly, for $K \geq 36$ with fixed fragment size of $K' = 6$, the run time of LESSA flattens out: this is because LESSA is carried out in parallel across all 6-subcarrier fragments." (Section IV on page 21)

Please refer to the added discussions in Section IV, along with the evaluations depicted in Fig. 4.

REVIEWER 2

1) **Comment**: *This paper studied the problem of learning-based spectrum sensing using approximate POMDPs. Some issues in this paper should be revised and addressed.*

   R: We thank the reviewer for the valuable assessment. We have addressed the comments to the best of our abilities, and believe that this version of the paper is much improved. Please note that, in the revised manuscript, we refer to SU as CR (Cognitive Radio) and to PU as LU (Licensed User), as opposed to SU and PU used in the original submission. In addition, we have termed our proposed framework LESSA (LEarning-based Spectrum Sensing and Access) and its Multi-Agent version (MA-LESSA), to more directly refer to it throughout the manuscript. We have also changed the title from "Learning-based Spectrum Sensing in Cognitive Radio Networks via Approximate POMDPs" to "Learning-based Spectrum Sensing and Access in Cognitive Radios via Approximate POMDPs" to better reflect the contributions of the paper.

2) **Comment**: *The authors claimed that the proposed scheme can achieve an optimal sensing performance, which is a very strong statement, requiring a detailed analysis and proof.*

   R: We agree with the reviewer that claiming optimality of the sensing performance would be an overly strong statement, and was an oversight on our part in the original submission. In fact, since PERSEUS is an *approximate*, randomized, point-based POMDP value-iteration algorithm, the sensing & access policy obtained through it *cannot* be claimed to be optimal. We have reworded any reference to the policy generated by our algorithm as "approximately optimal" instead of "optimal" throughout the manuscript. We have corrected this claim and provided additional comments in Sections I and IV.

   Regarding the claim of "approximate optimality" of our proposed framework, we would like to remark that an *optimal POMDP* policy would be optimal in the sense of maximizing the discounted reward defined in (19). However, the computation of such policy is intractable (see explanations provided after (19)), reason why we investigate approximate techniques in this paper. To evaluate the goodness of such approximation (quoted from the manuscript, Section IV on page 20):

   "In Fig. 4a, we also plot the convergence of the PERSEUS algorithm, using a discount factor of γ=0.9 and a termination threshold of ϵ=10^{-5}, λ=1, for both cases in which it is run concurrently with the Baum-Welch algorithm (red curve) or after its convergence

(green curve). To evaluate convergence, we use a normalized loss metric, defined as the loss of utility (defined in (13)) with respect to an oracle that performs spectrum access based on knowledge of the current occupancy state $\vec{B}(t)$, defined mathematically as

$$1 - \frac{R(\vec{\phi}^*(\hat{\beta}_t), \vec{B}(t))}{\max_{\vec{\phi}\in\{0,1\}^K} R(\vec{\phi}, \vec{B}(t))},$$

averaged out over time and multiple realizations, where $\vec{\phi}^*(\hat{\beta}_t)$ is the spectrum access decision defined in (15), based on the posterior belief $\hat{\beta}_t$ generated in our POMDP formulation, and $\max_{\vec{\phi}\in\{0,1\}^K} R(\vec{\phi}, \vec{B}(t))$ is the Oracle reward, which uses knowledge of $\vec{B}(t)$. We find that the concurrent method converges in approximately half the time of the non-concurrent one. Remarkably, the normalized loss is around 5% after convergence, i.e., LESSA performs on par with an Oracle with knowledge of the current spectrum occupancy – a significant result considering the lack of *a priori* knowledge of the underlying Markov transition model and the noisy and constrained spectrum sensing environment. In addition, since the Oracle performs better than the *optimal POMDP* policy, this result demonstrates that solving for the *optimal POMDP* policy (a computationally intractable task) does not yield a tangible boost in spectrum white-space detection, thereby legitimizing the validity of the approximate POMDP approach proposed in this paper."

3) **Comment**: *It seems that the authors proposed their methods by borrowing some ideas or methods from other references. Did the authors make some innovations when applying other methods into their work? If the authors just put the existing methods into their scheme, the contribution of the proposed method is limited. If the authors made some changes, a more detailed explanation is required.*

R: In this response and in the revised manuscript, we provided more details on the main novelties of our proposed method for spectrum sensing and access, and added Table I to clarify the contributions with respect to the state-of-the-art on spectrum sensing and access. In terms of algorithmic solutions, although PERSEUS has been proposed in [21] as a randomized point-based value iteration algorithm to solve *generic* POMDP models, we have optimized it and designed custom heuristics to alleviate the computational challenges of our specific spectrum sensing and access problem. To the best of our knowledge, these are novel contributions. We quote relevant excerpts from our revised manuscript below.

"In Section III-B, we concurrently leverage the learned model in a randomized PBVI algorithm known as PERSEUS [21] to devise an approximately optimal spectrum sensing and access policy; we alleviate its computational complexity by introducing fragmentation, belief update simplification heuristics via Hamming distance state filters and Monte-Carlo based methods." (Section I on page 5)

These methods are further explained in Section III-B on page 15.

Additionally, in the multi-agent extension, we have proposed novel neighbor discovery and channel access order allocation schemes. We quote relevant excerpts from our revised manuscript below.

"...adapting the Multi-band Directional Neighbor Discovery scheme described in [22] to distributed multi-agent CR deployments, we propose a novel neighbor discovery heuristic centered around RSSI thresholding; inspired by cluster fallback mechanisms (leader selection, broker failover, master-slave auto-configuration, and data replication) involved in Message Oriented Middleware [23], we propose a channel access rank allocation technique centered around a quorum-based preferential ballot algorithm." (Section I on page 5)

In terms of novelties with respect to the state-of-the-art, further details have been added in the introductory section and Table I has been added to clarify the novelties. To summarize:

- The works [3]–[6], [15] assume independence in LU occupancy behavior across both time and frequency, which is impractical. The systems described in [7]–[10] leverage LU occupancy correlation in time only, and therefore exhibit worse white-space detection performance than LESSA. Additionally, in [7], LU occupancy is estimated directly via energy detection thereby ignoring errors in state estimation; [8] assumes non-adaptive sensing, vs our adaptive POMDP-based scheme; [9] assumes *a priori* knowledge of the correlation model, vs our scheme which learns it from observations; and [10] makes independence assumptions during projection approximation to a lower-dimensional space.
- Spectrum sensing and access strategies exploiting the time-frequency correlation structure in LU occupancies are studied in [12], [13], [16], [17] – however, [12] learns this correlation structure offline using pre-loaded databases (vs our concurrent learning and adaptation strategy); [13] proposes a POMDP formulation wherein the uncertainty involved is assumed to be only due to sensing restrictions by the CR, while making no

claims about system operations in noisy settings; and [16], [17] propose black-box DL models that involve tedious data collection and pre-processing tasks.

- Contrasting MA-LESSA relative to solutions in the multi-agent state-of-the-art, we find that [5] focuses primarily on data aggregation strategies within the ensemble; [7] fails to detail neighbor discovery and channel access order allocation schemes, which we do; and [15] details schemes that employ ACKs as a feedback mechanism to gauge the utility of an access decision, which introduces unnecessary lag into the model – instead, we utilize a threshold-based decision heuristic, the simplicity of which is demonstrated by implementing our solution on a testbed of ESP32 radios.
- Finally, unlike all of the above works, LESSA provides for a mechanism to regulate the trade-off between CR throughput and LU interference, through a weighted POMDP metric: a crucial feature in real-world deployments.

We have highlighted these contributions in more detail in our revised manuscript in order to convey them better to the reader. Please refer to Section I and Table I for details.

4) **Comment**: *The authors claimed that their proposed work can achieve better performance than Q-learning which is a very basic tool for optimization. How about the comparison with deep Q learning or other deep reinforcement learning networks?*

R: We would like to remark that, indeed, the original manuscript *does* compare our solution against a state-of-the-art Adaptive Deep Q-Network (DQN) strategy originally proposed in [13]. We regret that it was not sufficiently emphasized in the original manuscript. Specifically, this state-of-the-art scheme uses experiential replay (Memory Size C=10^6), 2048 input neurons, 4096 neurons with ReLU activation functions in each of the 2 hidden layers, a Mean-Squared Error cost function with an Adam Optimizer, a fixed exploration factor ϵ=0.1, a learning rate of α=10^{-4}, a batch size of W=32, and a sensing restriction of 6. The numerical comparisons, shown in Fig. 5, demonstrate that our solution offers a 9% improvement over this strategy [13]. In addition, the adaptive DQN formulation in [13] fails to provide a mechanism to manage the trade-off between CRs' network throughput and LUs' interference, hence it achieves a single point in the trade-off region of Fig. 5; in contrast, we enable this feature through a weighted reward metric in our approximate POMDP model. In the revised manuscript, we have stressed this comparison (in addition to that against Temporal Difference (TD) SARSA with Linear Function Approximation (LFA)

[7]) further in the Abstract and in the Introduction.

Since computational complexity is another key factor, in the revised manuscript we have added an evaluation of the runtime of TD-SARSA with LFA and adaptive DQN, which shows computational gains of our framework over adaptive DQN (in addition to the throughput gains already mentioned). The results are shown in Fig. 4b, as a function of the number of subcarriers. Quoted from the manuscript (Section IV on page 21):

"The algorithms are run on a $2\times$ 12-core Intel Xeon Gold 6126 @ 2.6GHz compute node with 192GB RAM [42]: as the number of subcarriers increases, our solution scales better yielding a more computationally tractable performance relative to the other two, owing to fragmentation and belief update simplification heuristics. Interestingly, for $K \geq 36$ with fixed fragment size of $K' = 6$, the run time of LESSA flattens out: this is because LESSA is carried out in parallel across all 6-subcarrier fragments."

5) **Comment**: *There are only five simulation results in the paper. More simulation figures are required to comprehensively demonstrate the performance of the proposed scheme. Moreover, it is better for the authors to use the commonly accepted sensing performance metrics, e.g.,* P_A *or* P_D.

R: We thank the reviewer for this suggestion. In the revised manuscript, we have added more simulation figures to demonstrate more comprehensively the performance of the proposed scheme against the state-of-the-art, and used a more realistic channel model in the numerical evaluations, that includes random deployments of transmitters and receivers, probabilistic LoS and NLoS links, SINR variations due to stochastic variations of large- and small-scale channel conditions, and a rate adaptation scheme at the CRs to ensure maximum exploitation of the correctly inferred white-spaces. Please refer to the description of the new channel model provided in Section IV. Fig. 5 illustrates the numerical evaluations of our system and those of comparable works in the state-of-the-art, under this enhanced channel model. In addition to the evaluations already included in the original manuscript (revised based on the more realistic channel model), we conducted additional evaluations as a part of this revision, as itemized below:

- In Fig. 4b, we have added a time-complexity evaluation of our proposed method against Adaptive DQN [13] and TD-SARSA with LFA [7], as discussed in our previous response. Our solution scales better yielding a more computationally tractable performance relative

to the other two.

- Fig. 6b plots the Receiver Operating Characteristics (ROC) of our framework (“Missed Detection Probability" versus “False Alarm Probability") – in addition to those corresponding to comparable works in the state-of-the-art (Neyman-Pearson Detection [6] and TD-SARSA with LFA [7]); in line with the results of Fig. 5, our proposed framework exhibits a more favorable trade-off; and
- The two plots ((a) and (b)) in Fig. 9 plot the performance of the proposed neighbor discovery (RSSI thresholding) and channel access rank allocation (quorum-based preferential ballot) heuristics from a computational time complexity perspective. Quoted from the manuscript (Section V-B on page 28):

  "[...] we retrofit these schemes into the control channel design, collaboration, and data aggregation modules of the Purdue BAM! Wireless radio [25], and analyse their feasibility in emulations of highly mobile real-world scenarios – namely, military deployments in the Alleys of Austin scenario [32] (urban: $5\times$ $[9-\text{guardsmen}+1-\text{UAV}]$) and disaster relief deployments in the Payline scenario [31] (urban: $5\times$ $[9-\text{EMTs}+1-\text{HQ}]$). Along with the scenario-specific node mobility patterns (Fig. 8), the results of these emulations are shown in Fig. 9: neighbor discovery list changes are minimal in spite of node mobility (with an RSSI threshold of 22dB), and distributed convergence of channel access rank allocation across the ensemble is achieved within the first few iterations in any given 10s time-step."