Learning Non-myopic Power Allocation in Constrained Scenarios

Arindam Chowdhury*, Santiago Paternain$^\#$, Gunjan Verma†, Ananthram Swami†, and Santiago Segarra*

*Dept. of ECE, Rice University $^\#$Dept. of ECSE, Rensselaer Polytechnic Institute
†DEVCOM Army Research Laboratory

***Abstract*—We propose a learning-based framework for efficient power allocation in ad hoc interference networks under episodic constraints. The problem of optimal power allocation – for maximizing a given network utility metric – under instantaneous constraints has recently gained significant popularity. Several learnable algorithms have been proposed to obtain fast, effective, and near-optimal performance. However, a more realistic scenario arises when the utility metric has to be optimized for an entire episode under time-coupled constraints. In this case, the instantaneous power needs to be regulated so that the given utility can be optimized over an entire sequence of wireless network realizations while satisfying the constraint at all times. Solving each instance independently will be myopic as the long-term constraint cannot modulate such a solution. Instead, we frame this as a constrained and sequential decision-making problem, and employ an actor-critic algorithm to obtain the constraint-aware power allocation at each step. We present experimental analyses to illustrate the effectiveness of our method in terms of superior episodic network-utility performance and its efficiency in terms of time and computational complexity.**

***Index Terms*—Non-myopic power allocation, episodic constraint, hierarchical model, GCNN, TD3, UWMMSE**

I. Introduction

Power allocation for interference management in wireless networks [1] is essential for satisfying high quality-of-service (QoS) requirements in modern communication systems [2]. Mathematically, it can be formulated as the problem of optimizing a certain system-level utility function (such as sum-rate or harmonic-rate) subject to instantaneous resource budget constraints. However, such a problem formulation is NP-hard [3], and a classical approximate solution involves reformulating it in the form of a surrogate tri-convex objective with instantaneous constraints and solving it via an iterative block-coordinate-descent based approach termed WMMSE [4]. Since the advent of deep learning, several *data-driven* alternatives have been proposed [5]–[11], that try to address multiple drawbacks in the WMMSE algorithm, including computational and time complexity [12]. However, none of these algorithms are capable of handling time-dependent constraints. For example, episodic sum-rate maximization under battery constraints is a crucial problem, especially in scenarios wherein the available battery is limited and channel conditions vary over time. The key challenge here is to selectively allocate power under favorable channel conditions only, such that the available battery can be preserved for a longer duration in an episode, resulting in a *non-myopic* (non-greedy) power allocation method.

The power allocation process at successive time steps, based on the current channel conditions and battery level, can be modeled as a sequential decision-making problem under constraints. Further, considering the fact that the instantaneous power allocation depends only on the last battery state and not the entire history, we frame this as a Markov decision process [13]. Recently, the paradigm of constrained reinforcement learning (CRL) [14]–[17] has become an extremely popular data-driven framework to tackle episodic optimization problems with time-coupled constraints. In this work, we combine the risk-aware reward function formulation [18] with constrained policy [17] to develop a hierarchical framework for learning *non-myopic power allocation* (NMPA) in wireless networks under battery constraints.

Notation: The entries of a matrix $\mathbf{X}$ and a vector $\mathbf{x}$ are denoted by $[\mathbf{X}]_{ij}$ and $[\mathbf{x}]_i$.The all-zeros and all-ones vectors are denoted by $\mathbf{0}$ and $\mathbf{1}$.$[\cdot]_+$ represents a $\max(\cdot, 0)$ operation. $\mathbb{E}(\cdot)$ is the expectation operator. The zero-mean normal distribution of variance σ^2 is denoted by $\mathcal{N}(0, \sigma^2)$.

II. System model and problem formulation

We start by laying out a procedure for constructing a finite, time-indexed set of wireless single-hop ad-hoc interference networks having M distinct single-antenna transceiver pairs. Time is slotted into uniform intervals indexed by t. At each time step t, a network realization is sampled randomly from a joint distribution over network topologies and fading conditions. In each sample, the transmitters are denoted by i and their unique intended receivers are denoted by $r(i)$ for $i \in \{1, \dots, M\}$. The transmitters are dropped uniformly at random as $\mathbf{l}_i \in [-S, S]^2$, S being user-defined, and their corresponding receivers are dropped uniformly at random as $\mathbf{o}_i \in [\mathbf{l}_i - \frac{R_i}{\sqrt{2}}, \mathbf{l}_i + \frac{R_i}{\sqrt{2}}]^2$, where R_i is the range of transmitter i. The i-th transmitter can only communicate with its corresponding receiver $r(i)$. It can, however, interfere with receivers $r(j)$, for all $j \neq i$, that lie within its range. We represent the set of all transmitters interfering with a receiver $r(i)$ as $\mathcal{E}_{r(i)}$. For

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-19-2-0269. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Emails: {ac131, segarra}@rice.edu, paters@rpi.edu, {gunjan.verma, ananthram.swami}.civ@army.mil.

simplicity, we assume that all transmitters have equal range i.e $R_i = R$ for all i. Further, we simulate fading effects in the sample by incorporating small-scale Rayleigh fading as well as large-scale path loss.

An arbitrary channel coefficient corresponding to the network realization at time t is given by $h(t)$. Now, denoting the signal transmitted by i at time t as $x_i(t) \in \mathbb{R}$, and assuming a linear transmission model, the received signal $y_i(t) \in \mathbb{R}$ at $r(i)$ is given by

$$y_i(t) = h_{ii}(t)x_i(t) + \sum_{j \in \mathcal{E}_{r(i)}} h_{ij}(t)x_j(t) + n_i(t), \tag{1}$$

where $h_{ii}(t) \in \mathbb{R}$ is the channel between the i-th transceiver pair at time t, $h_{ij}(t) \in \mathbb{R}$ for $j \in \mathcal{E}_i$ represents the interference between transmitter j and receiver $r(i)$, and $n_i(t) \sim \mathcal{N}(0, \sigma_N^2)$ represents the additive channel noise. The time-varying channel states $h_{ij}(t)$ are consolidated in a channel-state information (CSI) matrix $\mathbf{H}_t \in \mathbb{R}^{M \times M}$ where $[\mathbf{H}_t]_{ij} = h_{ij}(t)$. We define an *episode* of duration T as the sequence of i.i.d. CSI matrices $\{\mathbf{H}_1, \dots, \mathbf{H}_T\}$.

The instantaneous data rate c_i achievable at receiver $r(i)$ at time t is given by,

$$c_i(\mathbf{p}_t, \mathbf{H}_t) = \log_2 \left(1 + \frac{[\mathbf{H}_t]_{ii}^2 [\mathbf{p}_t]_i}{\sigma_N^2 + \sum_{j \in \mathcal{E}_{r(i)}} [\mathbf{H_t}]_{ij}^2 [\mathbf{p}_t]_j} \right), \tag{2}$$

where $\mathbf{p}_t \geq \mathbf{0}$ is the allocated power at time t. In the absence of episodic constraints, which are constraints coupled across time, optimal power allocation for a given CSI $\mathbf{H}_t$ at any time t, can be obtained by solving the following sum-rate maximization problem:

$$\max_{\mathbf{p}} \sum_{i=1}^{M} c_i(\mathbf{p}, \mathbf{H}_t) \tag{3}$$

$$\text{s.t. } \mathbf{p} \in [0, P_{\max}]^M \tag{4}$$

where $P_{\max} \in \mathbb{R}$ denotes the maximum available power at every transmitter.

Several solution models for (3)-(4) exist, including iterative block-coordinate descent based optimizers [4], neural network based approaches [5], [6] as well as a class of unfolding [19] based hybrid methods [7], [8] combining both. Recently, a GNN-based hybrid model UWMMSE [12] was proposed, that leverages the underlying graph structure of wireless networks captured implicitly in $\mathbf{H}$ to allocate power efficiently.

However, none of these methods is suited to handle episodic constraints, which are far more challenging. For example, under limited battery availability, the task of optimal power allocation would ideally involve transmitting optimally at time instants wherein the channels are "good", i.e, the achievable sum-rate is high, while intelligently preserving battery under poor channel conditions so that the overall episodic sum-rate can be maximized.

To formalize the aforementioned problem, let $\mathbf{b}_t \in \mathbb{R}^M$ denote the battery levels of all transmitters at time t. An initial battery budget is given by $\mathbf{b}_0$, where $\max_i [\mathbf{b}_0]_i = B_{\max}$ is user-defined and $\min_i [\mathbf{b}_0]_i \gg P_{\max}$. The battery dynamics can be modeled as

$$[\mathbf{b}_t]_i = [\mathbf{b}_{t-1}]_i - [\mathbf{p}_t]_i - \alpha \mathbb{1}([\mathbf{p}_t]_i > 0) \text{ for all } t, i. \tag{5}$$

Here, α represents a fixed cost of transmission. Note that the batteries for all transmitters evolve non-increasingly over time, effectively meaning that the spent batteries cannot be recharged within a given episode.

Further, by denoting a power allocation function as f, such that $\mathbf{p}_t = f(\mathbf{b}_{t-1}, \mathbf{H}_t)$, the problem of interest can be defined as:

$$\max_{f \in \mathcal{F}} \mathbb{E}_{\substack{\{\mathbf{H}_1, \dots, \mathbf{H}_T\} \sim \mathcal{H} \\ \mathbf{b}_0 \sim \mathcal{B}}} \left[\frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{M} c_i(\mathbf{p}_t, \mathbf{H}_t) \right] \tag{6}$$

$$\text{s.t. } \mathbf{p}_t \leq P_{\max}\mathbf{1} \text{ for all } t \tag{7}$$

$$\mathbf{b}_t \geq \mathbf{0} \text{ for all } t, \tag{8}$$

where $\mathcal{F}$ is a suitably chosen space of continuous functions, $\mathcal{H}$ is an unknown stationary distribution over CSI matrices and $\mathcal{B}$ is an unknown distribution over battery budgets.

A careful analysis of (5) and (8) reveals that the main purpose of the battery constraint is to restrict f from allocating power ever again within an episode to transmitters that fully deplete their battery, given the non-increasing battery dynamics. This implicitly forces f to utilize the available battery budget judiciously in a given episode. It is important to note here that solving the optimization problem (6)-(7), without satisfying (8), is equivalent to solving (3)-(4) over multiple time-steps $t = \{1, \dots, T\}$ independently. We define this method as the *myopic* (greedy) power allocation model. It is clearly ineffective as it becomes inapplicable once (8) is violated. For example, if batteries deplete at any $t < T$, then the achievable sum-rate for the remaining CSI matrices $\{\mathbf{H}_{t+1}, \dots, \mathbf{H}_T\}$ will be null. The main goal of this work, therefore, is the development of a *non-myopic* power allocation model to solve (6) under the time-coupling introduced by (5) and the corresponding constraint (8), while satisfying (7) at all t.

III. Proposed Method

The optimization problem (6-7) is non-convex and NP-hard [20], even for a deterministic episode (a fixed set of CSI matrices) without any episodic constraints. The stochasticity in (6) and the battery constraint (8) further add to the complexity of the problem. Moreover, the optimization in (6) is over an infinite-dimensional space $\mathcal{F}$ of continuous functions.

To address these challenges, we decouple the instantaneous power allocation objective from the battery constraint at all time steps. Since the instantaneous problem has already been solved effectively [7], [8], [12], we employ a suitable solver to construct a hierarchical model at two levels. The lower level generates an approximate solution $\bar{\mathbf{p}}_t$ of the optimal instantaneous power allocation at each time-step t based purely on the corresponding $\mathbf{H}_t$ without considering the state of the battery $\mathbf{b}_t$. The upper level then modulates $\bar{\mathbf{p}}_t$ such that the battery constraint is satisfied over the entire episode. Denoting

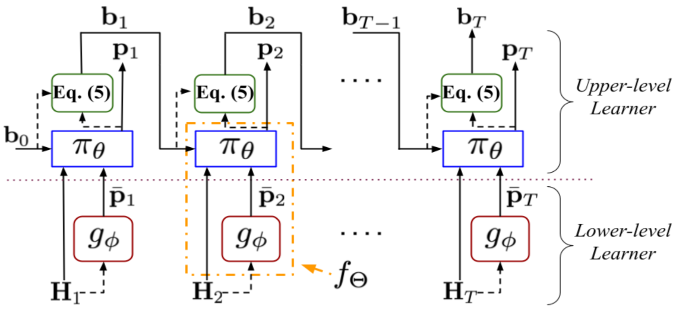


Fig. 1. Block diagram of the proposed model over T time-steps.

the lower-level solver as g and the upper-level solver as π, we propose the following solution model:

$$\mathbf{p}_t = \pi(\mathbf{b}_{t-1}, \mathbf{H}_t; \bar{\mathbf{p}}_t) \text{ for all } t \tag{9}$$

$$\text{where, } \bar{\mathbf{p}}_t = g(\mathbf{H}_t) \text{ for all } t, \tag{10}$$

and $\mathbf{b}_t$ evolves as per (5). Further, due to the relative ease of sampling from the unknown distributions $(\mathcal{H}, \mathcal{B})$, and inherent intractability of $\mathcal{F}$, we frame this as a *learning* problem defined on a parameterized function space completely characterized by the learnable parameters Θ. Any such function $f_\Theta(\mathbf{b}_{t-1}, \mathbf{H}_t)$ generates time-varying power allocations as $\mathbf{p}_t = f_\Theta(\mathbf{b}_{t-1}, \mathbf{H}_t) = \pi_\theta(\mathbf{b}_{t-1}, \mathbf{H}_t; g_\phi(\mathbf{H}_t))$ where $\Theta = \{\theta, \phi\}$. Figure 1 presents an illustration of the proposed model.

Owing to the action-feedback structure of the upper-level problem where the battery state at each time-step is affected by the choice of power allocation, we frame it as a sequential decision-making problem and employ CRL to train π_θ. On the other hand, g_ϕ is trained separately in an unsupervised setting prior to training π_θ and the learned weights ϕ are frozen such that g_ϕ essentially serves as a pre-trained feedforward model while training π_θ. More details on the choice of g_ϕ are provided in Remark 1.

Non-myopic Power Allocation (NMPA) formalizes the upper-level problem as a finite-horizon discounted Markov decision process (MDP) characterized by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma)$. The state space $\mathcal{S}$ and action space $\mathcal{A}$ are continuous. Tuples $(\mathbf{b}_{t-1}, \mathbf{H}_t)$ for all t constitute $\mathcal{S}$ while the corresponding actions $\mathbf{p}_t$ constitute $\mathcal{A}$. Since the current $\mathbf{p}_t$ depends only on the last battery state $\mathbf{b}_{t-1}$ and the current channel $\mathbf{H}_t$ but not on the entire history, the Markov assumption [13] holds. The environment is stochastic – on account of the randomness in the channel states at each time step – defined by a transition probability distribution $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. The initial state distribution ρ_0 consists of the tuple $(\mathcal{H}, \mathcal{B})$ and γ is the discount factor that controls the level of myopia in the system. A higher γ corresponds to a more non-myopic policy. A policy is a mapping from $\mathcal{S}$ to $\mathcal{A}$ that aims to optimize a certain objective. We redefine the upper-level learner π_θ as a parameterized deterministic policy given by

$$\mathbf{p}_t = \pi_\theta(\mathbf{b}_{t-1}, \mathbf{H}_t; \bar{\mathbf{p}}_t) = \mu_\theta(\mathbf{b}_{t-1}, \mathbf{H}_t) \odot \bar{\mathbf{p}}_t \tag{11}$$

where $\odot$ is an element-wise multiplication and $\mu_\theta : \mathcal{S} \to [0,1]^M$ is a 2-layered, 3-filter graph convolutional neural network (GCNN) [21] architecture used to learn a vector of scales as

$$\mu_\theta(\mathbf{b}_{t-1}, \mathbf{H}_t) = \text{sigmoid}\left(\sum_{v=0}^{2} \mathbf{H}_t^v \, \mathbf{Z} \, \theta_{1v}\right) \tag{12}$$

$$\text{where, } \mathbf{Z} = \text{leakyReLU}\left(\sum_{v=0}^{2} \mathbf{H}_t^v \, \mathbf{b}_{t-1} \, \theta_{0v}\right)$$

with $\theta = \{\theta_{0v}, \theta_{1v}\}_{v=0}^2$. The scale assigns one scalar weight per transmitter, which modulates $\bar{\mathbf{p}}_t$ based on the current battery condition. The primary motivation behind modeling the scale using a GCNN lies in the structure of $\mathbf{H}$, which can be interpreted as the weighted adjacency matrix of a directed graph with self-loops. The nodes of this graph are the transceivers. The self-loops represent transmission channels, and the remaining edges represent interference. Moreover, the transmitter-specific elements, such as battery and power allocation, can be considered as signals supported on the nodes of this graph. Due to this inherent graph structure, a GCNN can learn rich representations for each node by leveraging their local neighborhood information [10], [12]. Finally, note that the sigmoid non-linearity in (12) restricts the scale within the interval $[0, 1]$ element-wise. As, $\bar{\mathbf{p}}_t$ already satisfies (7) for all t, the product $\mathbf{p}_t$ satisfies (7) for all t by construction.

Further, we define a composite reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as the following:

$$r(\mathbf{b}_{t-1}, \mathbf{H}_t; \mathbf{p}_t) = \sum_{i=1}^{M} c_i(\hat{\mathbf{p}}_t, \mathbf{H}_t) - L\mathbb{1}([\mathbf{p}_t]_i > [\mathbf{b}_{t-1} - \alpha]_i) \tag{13}$$

where $\hat{\mathbf{p}}_t = \min(\mathbf{p}_t, [\mathbf{b}_{t-1} - \alpha]_+)$ and L controls the violation penalty at each transmitter. Note that the sum-rate component of the reward is computed based on transmission power, which depends on the available battery in addition to the power allocated by the policy, while the penalty component depends only on the allocated power. For example, if a certain transmitter has its battery level close to depletion and the policy allocates a high power to it at a given time step, it cannot transmit due to lack of battery and therefore cannot contribute to the overall sum-rate. However, a violation penalty is still attributed to the policy as a high power was allocated while the battery was low. Intuitively, such a scheme has the implicit effect of promoting non-myopic behavior by forcing the policy to be generally conservative in terms of power allocation within an episode and then specifically guiding it to allocate power optimally only when the achievable sum-rate is high.

The overall RL objective can now be defined as:

$$\max_\theta \, \mathbb{E}_{\mathcal{H},\mathcal{B}}\left[\sum_{t=1}^{T} \gamma^{t-1} r\Big(\mathbf{b}_{t-1}, \mathbf{H}_t; \pi_\theta(\mathbf{b}_{t-1}, \mathbf{H}_t; \bar{\mathbf{p}}_t)\Big)\right]. \tag{14}$$

We employ TD3 [22], a model-free off-policy deterministic policy gradient algorithm specialized for continuous action spaces, to solve (14). Its main advantage lies in employing two critic networks to mitigate the value-overestimation issue that plagues actor-critic algorithms. Similar to the policy network, we use GCNNs to model the critics to allow them to leverage the graph structure in $\mathbf{H}$.

As a direct consequence of using GCNNs to model the policy, the trained model can also be deployed in a fully distributed fashion with each transmitter provided with a copy of the trained policy weights along with some additional feedback links [4], [12]. However, training has to be centralized under the current formulation. Further, the computational complexity of the NMPA framework at inference is given as $\mathcal{O}(M^2D)$, equivalent to that of a GCNN [21], [23], where M is the number of transmitters and D is the hidden dimension.

Remark 1 (Choice of lower-level learner) We choose the UWMMSE [12] model to serve as g_ϕ. UWMMSE has been shown to achieve superior sum-rate performance on instantaneous CSI matrices with reasonable robustness while incurring very low time and computational cost [9], [12], [24]. Further, UWMMSE enforces (4) explicitly as a non-linearity in its architecture. The parameters ϕ are trained using gradient feedback by maximizing the sum-rate objective (3) over a sufficiently large set of CSI matrices $\{\mathbf{H}_k\}_{k=1}^N$ sampled from $\mathcal{H}$. It is important to note here that while we present a specific choice of g_ϕ in this work for the sake of completeness, the proposed NMPA framework is compatible with any algorithm, learnable or otherwise, that provides a solution to (3)-(4).

IV. Numerical Experiments

We now empirically evaluate the performance of the proposed model in terms of average episodic sum-rate and average episodic violations per transmitter under a variable battery budget. We further demonstrate the generalization performance of the model across various episode lengths. Maximum instantaneous power is set to $P_{\max} = 1$ unit and maximum battery is set at $B_{\max} = 20$ units. The violation control parameter is fixed at $L = 1$. Fixed cost α is set to $0.5P_{\max}$. To model g_ϕ, we use a 4-layered UWMMSE [12] model. It is trained for a maximum of 10,000 epochs with early stopping. The batch size is set at 32. The hidden layer dimension for all GCNNs in policy π_θ and corresponding critics is set to 32. For training π_θ, we use a replay buffer of size $100k$ transitions with an off-policy batch size of 32. The learning rate for the actor network is set to 5×10^{-4} while that for the critic networks is 1×10^{-3}. The update ratio for target networks is set to 1×10^{-3}. Training is performed for a maximum of 10,000 episodes of fixed length 100, with early stopping. Inference performances are averaged over 10 independently sampled episodes. All computations are performed on an AMD Ryzen Threadripper 3970X 32-Core CPU with 128GB RAM.[1]

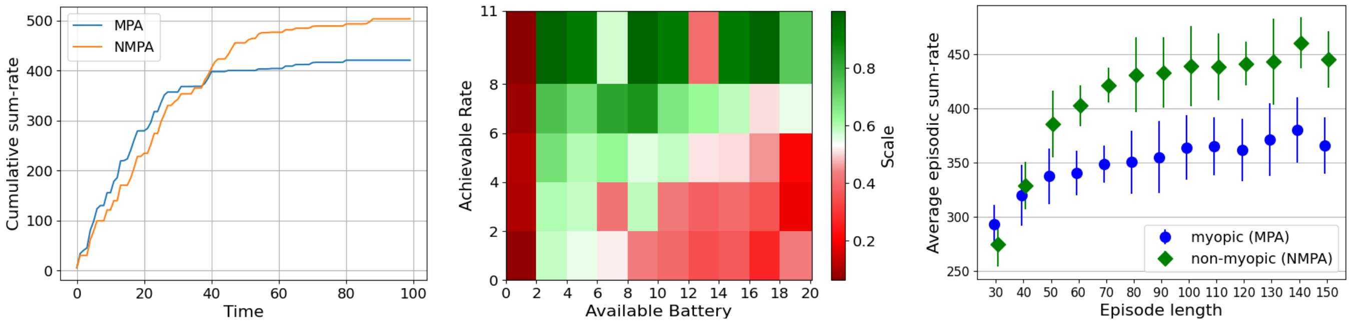
Dataset. We set the wireless network size $M = 10$, spatial size $S = 60$ units, and transmitter range $R = 20$ units for all experiments. The path loss between transmitter i and receiver $r(j)$ is computed as a function of their corresponding physical distance $d_{ij}(t)$. Incorporating small-scale Rayleigh fading, the elements of the channel matrix $[\mathbf{H}_t]_{ij}$ are given by

$$[\mathbf{H}_t]_{ij} = \frac{1}{1+d_{ij}^2(t)}\left|\frac{\mathcal{N}(0,1)}{\sqrt{2}} + \mathrm{i}\frac{\mathcal{N}(0,1)}{\sqrt{2}}\right| \quad \text{for all } i,j. \tag{15}$$

Since our aim is to emphasize the selectivity of our model in transmitting under “good” versus “poor” channel conditions, we manually construct a set of two topologies, low-interference (good) – wherein intra-transceiver spacing is small while inter-transceiver spacing is large – and high interference (poor), which is the reverse. At each time-step t, we sample a topology uniformly at random from this set and combine the corresponding path losses with fading effects in (15). The battery budget is sampled from an M-dimensional uniform distribution $\mathcal{U}^M_{[0.5B_{\max},B_{\max}]}$.

Episodic Behavior. To evaluate the effectiveness of the proposed NMPA framework, we compare its achieved episodic sum-rate with that of the myopic power allocation (MPA) model, i.e., UWMMSE with battery-agnostic power allocation. As shown in Figure 2(a), the cumulative episodic sum-rate achieved by NMPA at the end of a randomly sampled episode of length 100 is significantly higher than that of MPA. The average improvement in episodic sum-rate achieved by NMPA over MPA, computed over 10 randomly sampled episodes, is $\sim 15-20\%$. Moreover, this performance improvement is achieved by incurring merely $0.0005 - 0.0008$ average violations per transmitter, computed for a 10-transceiver network over 10 independent episodes of length 100. Clearly, the proposed model learns not to allocate power once the battery is (close to being) depleted. And, in doing that, it learns to spend battery intelligently only when the achievable sum-rate is high (good channel conditions). This behavior can be observed in Figure 2(a), where the NMPA curve flattens at several time steps initially, for which the MPA model shows an increment, indicating that NMPA chooses not to transmit for those channels, presumably due to their relatively poor channel conditions. By contrast, MPA spends battery greedily and runs out of battery faster, not being able to leverage good channels occurring later in the episode. This is further emphasized in Figure 2(b), wherein the scale values $\mu_\theta(\cdot)$ for NMPA are high for transmitters that have better channel conditions (high achievable rate) when the available battery is non-zero. It can be clearly observed that the scales are minimized when the batteries are fully depleted or close to depletion. The inference time for the proposed hierarchical framework is 7 ms, wherein the NMPA model takes 3 ms while the remaining time is consumed by the lower-level model. This fast power allocation is well-suited for rapidly varying channel conditions.

[1] Code to replicate the numerical experiments presented here can be found at https://github.com/archo48/nmpa.git.



(a) (b) (c)

Fig. 2. (a) Episodic performance of NMPA and MPA in terms of cumulative sum-rate over a randomly sampled episode of length 100. (b) 2D histogram of proportional scale values $\mu_\theta(\cdot)$ generated by NMPA for $10k$ transmitters under multiple configurations of available battery and achievable rate. (c) Generalization performance of NMPA across variable episode lengths.

Generalization performance. We evaluate the generalization behavior of the NMPA framework by varying the inference episode length in the range $[30, 150]$ while the model is trained on a fixed episode length of 100. We observe that the average episodic sum-rate achieved by NMPA is consistently higher than that of MPA for episode lengths both greater and smaller than 100. This is illustrated in Figure 2(c). It is important to note that the length of the episode is not known to the model apriori. Clearly, the model learns to be agnostic to the episode lengths and makes decisions purely based on the immediate state of the system.

V. Conclusions

We proposed a constrained-reinforcement-learning based non-myopic power allocation method for episodic power allocation under battery constraints. We employ GCNNs to leverage the underlying graph structure in wireless networks. The proposed framework is fast, effective, and generalizes across multiple episode lengths. It is also compatible with a large class of instantaneous power allocation algorithms. Future work includes power allocation under multiple time-coupled constraints, distributed training, and applying this framework to mobile wireless networks.

References

[1] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] L. Liang, H. Ye, G. Yu, and G. Y. Li, “Deep-learning-based wireless resource allocation with application to vehicular networks,” *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, 2019.

[3] Z.-Q. Luo and S. Zhang, “Dynamic Spectrum Management: Complexity and Duality,” *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, 2008.

[4] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, “An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel,” *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, 2011.

[5] M. Eisen, C. Zhang, L. F. Chamon, D. D. Lee, and A. Ribeiro, “Learning optimal resource allocations in wireless systems,” *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2775–2790, 2019.

[6] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, “A graph neural network approach for scalable wireless power control,” in *2019 IEEE Globecom Workshops*, 2019, pp. 1–6.

[7] Q. Hu, Y. Cai, Q. Shi, K. Xu, G. Yu, and Z. Ding, “Iterative algorithm induced deep-unfolding neural networks: Precoding design for multiuser MIMO systems,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1394–1410, 2020.

[8] L. Pellaco, M. Bengtsson, and J. Jaldén, “Deep weighted mmse downlink beamforming,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4915–4919.

[9] A. Chowdhury, G. Verma, A. Swami, and S. Segarra, “Deep Graph Unfolding for Beamforming in MU-MIMO Interference Networks,” *IEEE Trans. Wireless Commun. (early access)*, pp. 1–1, 2023.

[10] L. Schynol and M. Pesavento, “Coordinated sum-rate maximization in multicell MU-MIMO with deep unrolling,” *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1120–1134, 2023.

[11] B. Li, G. Verma, and S. Segarra, “Graph-based algorithm unfolding for energy-aware power allocation in wireless networks,” *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1359–1373, 2023.

[12] A. Chowdhury, G. Verma, C. Rao, A. Swami, and S. Segarra, “Unfolding WMMSE using graph neural networks for efficient power allocation,” *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 6004–6017, 2021.

[13] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[14] E. Altman, *Constrained Markov decision processes*. Routledge, 2021.

[15] S. Junges, N. Jansen, C. Dehnert, U. Topcu, and J.-P. Katoen, “Safety-constrained reinforcement learning for MDPs,” in *Intl. Conf. on tools and algo. for the constr. and analy. of sys.* Springer, 2016, pp. 130–146.

[16] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, “Constrained reinforcement learning has zero duality gap,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, “Safe policies for reinforcement learning via primal-dual methods,” *IEEE Transactions on Automatic Control*, vol. 68, no. 3, pp. 1321–1336, 2022.

[18] P. Geibel and F. Wysotzki, “Risk-sensitive reinforcement learning applied to control under constraints,” *Journal of Artificial Intelligence Research*, vol. 24, pp. 81–108, 2005.

[19] V. Monga, Y. Li, and Y. C. Eldar, “Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing,” *IEEE Signal Process. Mag.*, vol. 38, no. 2, pp. 18–44, 2021.

[20] Z.-Q. Luo and S. Zhang, “Dynamic spectrum management: Complexity and duality,” *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 1, pp. 57–73, 2008.

[21] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro, “Convolutional neural network architectures for signals supported on graphs,” *IEEE Trans. Signal Process.*, vol. 67, no. 4, pp. 1034–1049, 2018.

[22] S. Fujimoto, H. Hoof, and D. Meger, “Addressing function approximation error in actor-critic methods,” in *Intl. Conf. Mach. Learn. (ICML)*. PMLR, 2018, pp. 1587–1596.

[23] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *Intl. Conf. Learn. Repres. (ICLR)*, 2017.

[24] A. Chowdhury, F. Gama, and S. Segarra, “Stability analysis of unfolded wmmse for power allocation,” in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, 2022, pp. 5298–5302.