Adaptive Underwater Acoustic Communications with Limited Feedback: An AoI-Aware Hierarchical Bandit Approach

Fabio Busacca^o, Andrea Panebianco*§, Yin Sun§

^{o}University of Catania, Italy *University of Palermo, Italy §Auburn University, Alabama, USA

Abstract—Underwater Acoustic (UWA) networks are vital for remote sensing and ocean exploration but face inherent challenges such as limited bandwidth, long propagation delays, and highly dynamic channels. These constraints hinder real-time communication and degrade overall system performance. To address these challenges, this paper proposes a bilevel Multi-Armed Bandit (MAB) framework. At the fast inner level, a Contextual Delayed MAB (CD-MAB) jointly optimizes adaptive modulation and transmission power based on both channel state feedback and its Age of Information (AoI), thereby maximizing throughput. At the slower outer level, a Feedback Scheduling MAB dynamically adjusts the channel-state feedback interval according to throughput dynamics: stable throughput allows longer update intervals, while throughput drops trigger more frequent updates. This adaptive mechanism reduces feedback overhead and enhances responsiveness to varying network conditions. The proposed bilevel framework is computationally efficient and well-suited to resource-constrained UWA networks. Simulation results using the DESERT Underwater Network Simulator demonstrate throughput gains of up to 20.61% and energy savings of up to 36.60% compared with Deep Reinforcement Learning (DRL) baselines reported in the existing literature.

Index Terms—Underwater Communications, Adaptive Modulation, Power Control, Reinforcement Learning, Multi-Armed Bandit, Age of Information.

I. INTRODUCTION

UnderWater (UW) networks are attracting growing attention from academia and industry, enabled by advances in acoustic communication technologies [1], [2]. They support real-time data exchange in remote, harsh environments for applications such as environmental monitoring, exploration, and disaster response. However, UnderWater Acoustic (UWA) networks face intrinsic constraints—limited bandwidth, long propagation delays, and highly dynamic channels—that require adaptive protocols with low computational and signaling overhead. In particular, frequent feedback over costly acoustic channels can overwhelm the network, making intelligent scheduling critical to sustain throughput while limiting signaling and energy consumption.

A promising solution is Adaptive Modulation (AM), which dynamically selects Modulation Schemes (MSs) based on real-time channel conditions [3]. Adjusting transmission power (P) can further improve reliability. Yet, adapting these parameters in highly dynamic UW environments remains difficult, requiring fast responsiveness within the limited computational resources of UW devices.

Model-based approaches have long been applied to improve UW network performance [4], [5], predicting channel dynamics to guide decisions such as MS selection or data scheduling. Their effectiveness, however, is limited by the non-linear and unpredictable nature of UW environments, where factors like temperature, salinity, and water motion introduce uncertainties that are difficult to model.

To address these limitations, Deep Reinforcement Learning (DRL) has been explored for dynamic adaptation in UW environments. Although DRL provides accuracy and adaptability, it incurs high computational cost, slow convergence, and extensive training time, limiting its use in dynamic, resource-constrained UWA networks [6], [7]. In contrast, lightweight Multi-Armed Bandit (MAB) algorithms have recently attracted attention for their efficiency and ability to adapt to changing conditions without large datasets or heavy training [8]. They offer a computationally efficient alternative that balances performance and adaptability, making them well-suited for real-time UW applications.

This paper presents a bilevel MAB framework that jointly optimizes throughput and feedback overhead through adaptive transmission control and feedback scheduling. The main contributions are as follows:

- We propose a novel fully distributed bilevel MAB framework for UWA networks. The inner layer employs a Contextual Delayed MAB (CD-MAB) to jointly adapt modulation and power, maximizing throughput and reliability. The outer layer uses a Feedback Scheduling MAB to tune the feedback interval, extending it under stable performance and shortening it under degradation, thereby controlling context freshness, shaping the Age of Information (AoI), and reducing overhead. To the best of our knowledge, this is the first distributed bilevel MAB framework integrating adaptive modulation, power control, and AoI-aware feedback scheduling for UWA networks.
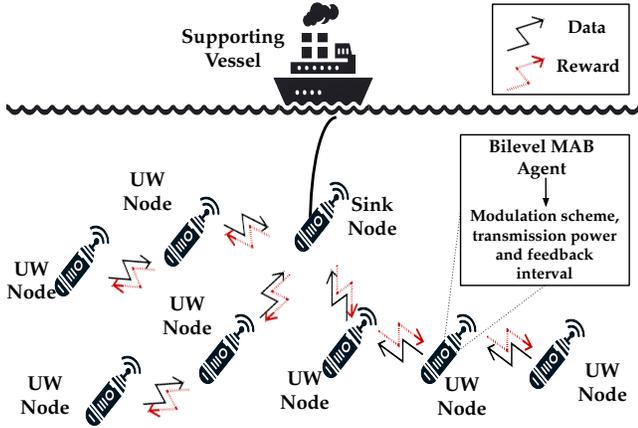
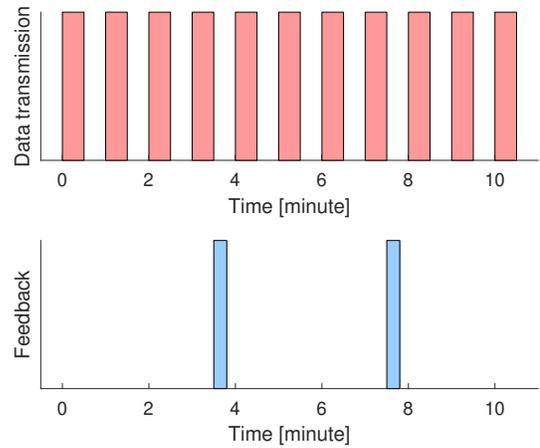Fig. 1: Underwater Acoustic Network architecture.

Fig. 2: Data is transmitted periodically, while feedback is not sent after every event. Our algorithm adapts the number of consecutive transmissions before the Rx sends feedback.

- We validate the proposed framework through extensive simulations using the DEsign, Simulate, Emulate and Realize Test-beds (DESERT) UW simulator [9], incorporating realistic environmental conditions and dynamic channel variations for a comprehensive assessment.
- We compare our framework with existing DRL-based schemes and show that it **achieves up to 20.61% higher throughput and 36.60% energy savings**, thanks to distributed design, joint modulation and power adaptation, and dynamic feedback scheduling. Unlike prior centralized DRL solutions [10], [11] with fixed MS, P, and feedback frequency, our benchmarks are drawn from terrestrial wireless networks, given the limited availability of effective RL methods specifically tailored to UWA environments.

II. REFERENCE SYSTEM

This section presents the reference scenario used to evaluate the proposed UWA network, modeled with DESERT.

The proposed architecture, shown in Fig. 1, includes a set of $\mathcal{U}$ Internet of Underwater Things (IoUT) nodes deployed in a 3D shallow-water environment. Each node $u \in \mathcal{U}$ communicates with a central *Sink* via single- or multi-hop acoustic routing. The Sink aggregates data from the nodes and forwards it via a wired link to a Supporting Vessel, which provides power and connectivity to the Sink. All nodes, except for leaf nodes, act as both transmitters (Txs) and receivers (Rxs) in a half-duplex fashion, with Tx–Rx interactions occurring at the link level, i.e., between directly connected neighbors.

The Tx periodically transmits sensor data packets embedding the selected P level, allowing the Rx to estimate the Signal-to-Noise Ratio (SNR) based solely on channel conditions, independent of the power adaptation policy. The Rx computes SNR and throughput from the received signal and periodically returns this information via dedicated single-hop feedback links. Feedback is sent only at the end of each feedback interval, summarizing performance and reporting the estimated SNR from the most recent channel observation. This aggregated approach reduces signaling overhead while still providing the Tx with sufficient information to update its transmission strategy.

At the end of each interval, the Tx requests feedback from the Rx to update its strategy and remain synchronized. Upon receiving delayed feedback, the Tx node, which runs both the CD-MAB and the Feedback Scheduling MAB, updates its decisions. The CD-MAB selects the optimal Modulation Scheme (MS) (among BPSK, 8-PSK, and 16-PSK) and transmission power P (discretized transmission power, with three levels: low, medium, and high) to maximize throughput, while the Feedback Scheduling MAB selects the feedback interval from $\{4, 7, 10\}$ minutes, and hence balances throughput and feedback cost. More frequent feedback improves network tracking, indirectly reducing AoI, while longer intervals reduce signaling overhead at the cost of less timely channel information. This process is illustrated in Fig. 2. In realistic shallow-water scenarios, SNR typically ranges from 10 dB in poor conditions to 40 dB in optimal ones [12], [13]. To capture this variability with low overhead, we quantize the SNR into three non-uniform intervals, reflecting its nonlinear impact on communication performance. At low SNR (10–18 dB), even small variations significantly affect Bit Error Rate (BER) and throughput, requiring finer granularity. In the medium range (18–30 dB), performance improves more gradually, while at high SNR (above 30 dB), it saturates, making finer distinctions unnecessary.

- **Low Quality** [10, 18] dB: narrow range due to BER and packet loss being highly sensitive to minor SNR changes.
- **Medium Quality** (18, 30] dB: broader interval reflecting more gradual performance improvements.
- **High Quality** (30, 40] dB: performance saturates, further increases provide minimal additional gains.

III. AGE OF INFORMATION

In the proposed framework, the Age of Information (AoI) quantifies feedback freshness at the Tx. It measures the time elapsed since the last feedback packet, reflecting how outdated the channel knowledge of the Tx is [14], [15].

Let $k \in \mathbb{N}$ denote the index of feedback epochs, with t_k the time of the k-th feedback and $Q_k = t_k - t_{k-1}$ the interval duration. During Q_k, the AoI increases discretely by one unit per slot, and resets to zero at t_k. For any $t \in [t_{k-1}, t_k)$, the AoI evolves as:

$$\Delta(t) = t - t_{k-1}. \tag{1}$$

This time-evolving metric quantifies channel information staleness, guiding adaptive feedback decisions.

IV. OUR HIERARCHICAL BILEVEL MAB APPROACH

To jointly optimize throughput and feedback overhead, we formulate a bilevel MAB framework solving a unified problem. The inner Contextual Delayed Multi-Armed Bandit (CD-MAB) adapts modulation and power based on feedback, while the outer MAB adjusts the feedback interval to balance signaling cost and context freshness. Together, they coordinate to maximize long-term performance in dynamic UW settings.

A. Contextual Delayed Multi-Armed Bandit (Inner Loop)

The Tx uses delayed contextual feedback and an Upper Confidence Bound (UCB)-based policy [16], [17]. Within each feedback interval $[t_{k-1}, t_k)$, discretized into time slots t (each lasting 1 minute in wall-clock time), the CD-MAB selects an action $a_t \in \mathcal{A}$, where $\mathcal{A}$ is the set of feasible MS–P pairs. The context (X_t) includes the most recent SNR estimate ($\hat{\eta}$) and its associated AoI, indicating channel information freshness, so $X_t = (\hat{\eta}_{t-\Delta(t)}, \Delta(t))$.

Since the reward is observed only at the end of each feedback interval, the reward r_k received at t_k is the sum of throughput values from all actions taken during $[t_{k-1}, t_k)$. This design avoids slot-wise rewards, ensuring constant, minimal overhead. To assign credit to individual actions, we apply a Uniform Credit Assignment (UCA) strategy, distributing r_k equally over the interval. Each action a_t, executed under context X_t, is thus assigned a per-action reward:

$$g_t = \frac{r_k}{|\mathcal{T}_k|}, \quad \text{for } t \in [t_{k-1}, t_k), \tag{2}$$

where $|\mathcal{T}_k|$ is the number of actions taken during the interval.

The CD-MAB learns an optimal transmission policy π_{tx} that maps each context X_t to an action a_t, maximizing the expected throughput over a finite time horizon $\mathcal{T}$ (i.e., the total number of time slots observed across all intervals):

$$\max_{\pi_{\text{tx}}} \quad \mathbb{E}\left[\sum_{t=1}^{\mathcal{T}} g_t\right], \quad \text{with } a_t = \pi_{\text{tx}}(X_t). \tag{3}$$

The action selection follows the UCB criterion:

$$a_t = \arg\max_{a \in \mathcal{A}} \left(\hat{\mu}_t(a, X_t) + \sqrt{\frac{c \log n_t}{N_t(a, X_t)}} \right), \tag{4}$$

where $\hat{\mu}_t(a, X_t)$ is the empirical mean reward for action a under context X_t, c is the exploration–exploitation parameter,

n_t is the number of total decisions, and $N_t(a, X_t)$ is the number of times a was selected under that context.

Action value estimates are updated in two stages. First, an initial update is made upon action selection, assuming immediate feedback. Then, when the delayed reward r_k is received at the end of the interval, the estimate is corrected using the per-action reward g_t from the UCA scheme. The correction uses an incremental average:

$$\hat{\mu}_t(a_t, X_t) \leftarrow \hat{\mu}_t(a_t, X_t) + \frac{1}{N_t(a_t, X_t)} \left(g_t - \hat{\mu}_t(a_t, X_t) \right), \tag{5}$$

for all $t \in [t_{k-1}, t_k)$, where $N_t(a_t, X_t)$ denotes the cumulative number of times the pair (a_t, X_t) has been selected up to time t. This allows responsiveness despite delay and accurate long-term learning.

B. Feedback Scheduling MAB (Outer Loop)

In parallel with the CD-MAB, a second non-contextual stochastic bandit agent selects the feedback interval duration $Q_k \in \mathcal{Q}$ at each t_k, where $\mathcal{Q}$ defines the allowable durations between feedback transmissions. This choice directly affects control signaling frequency and context freshness. The Feedback Scheduling MAB aims to maximize throughput while minimizing the cost of frequent feedback, which interrupts transmission and consumes energy. Although not directly optimized, a lower AoI naturally enhances throughput.

To capture this trade-off, the reward is the cumulative throughput r_k during interval $[t_{k-1}, t_k]$ (with $Q_k = t_k - t_{k-1}$), penalized by the energy cost term proportional to feedback frequency.

$$r_{\text{fb},k} = \theta r_k - (1-\theta) C_{\text{fb}} \frac{1}{Q_k}, \tag{6}$$

where C_{fb} is the energy cost of a feedback packet, computed as the product of its transmission duration and transmission power P. The parameter $\theta \in [0, 1]$ tunes the trade-off between throughput and feedback transmission cost, and $\frac{1}{Q_k}$ reflects the average feedback rate. Since both r_k and Q_k vary at each round, $r_{\text{fb},k}$ is defined as a function of k. Smaller Q_k improves feedback freshness, boosting throughput at the cost of higher overhead. Longer intervals reduce signaling but risk outdated context. The agent learns to select shorter intervals when rapid adaptation is needed and longer ones when the channel is stable. The optimization objective is:
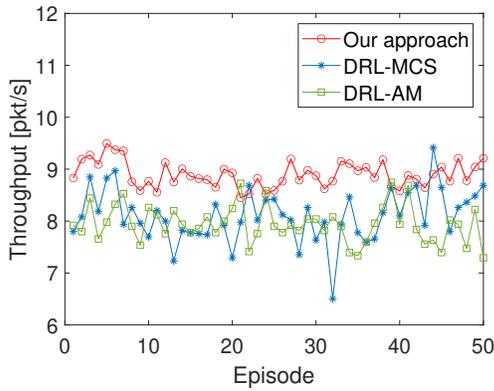
$$\max_{\pi_{\text{fb}}} \quad \sum_{k=1}^{K} \left(\theta r_k - (1-\theta) C_{\text{fb}} \frac{1}{Q_k} \right), \tag{7}$$

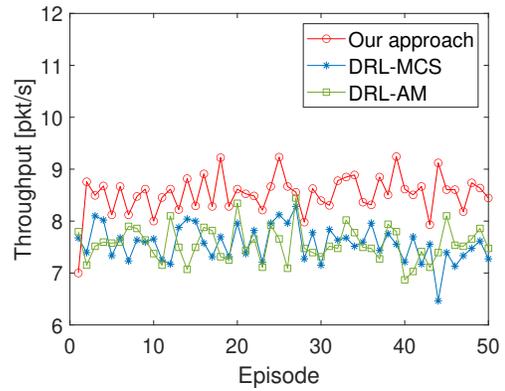where π_{fb} is the feedback scheduling policy.

The interval is selected using a standard UCB strategy.

$$Q_k = \arg\max_{Q \in \mathcal{Q}} \left(\hat{\mu}_k(Q) + \sqrt{\frac{c \log k}{N_k(Q)}} \right), \tag{8}$$

where $\hat{\mu}_k(Q)$ is the empirical average reward for action Q, c is the exploration–exploitation parameter, k is the current round index, and $N_k(Q)$ is the number of times Q has been selected.
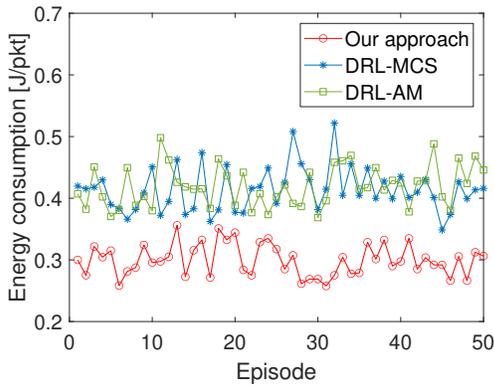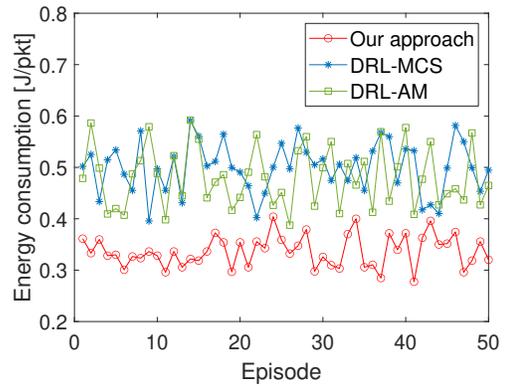
(a) 8-node scenario.

(b) 10-node scenario.

Fig. 3: Throughput comparison between our approach, DRL-MCS, and DRL-AM.

(a) 8-node scenario.

(b) 10-node scenario.

Fig. 4: Energy consumption comparison between our approach, DRL-MCS, and DRL-AM.

V. SIMULATION SCENARIO

To evaluate performance, we ran simulations with 4, 6, 8, and 10 UW nodes. Scalability is generally not a major concern in UWA networks, which rely on few nodes with long transmission ranges. The high deployment cost also limits large-scale setups. The topology ensured a maximum node distance of 357 meters, similar to the LOON testbed [18].

Simulation parameters reflect real-world UW devices, such as EvoLogics modems [19], using a 10.5 kHz carrier and 4.2 kHz bandwidth. Nodes transmitted 125-kilobyte data packets and control packets at 4800 bps. These settings capture UWA-specific challenges, including multipath, Doppler spread, and noise variability. The environment simulated wind speed $w = 50$ km/h, shipping factor $z = 0.5$, and spreading factor $s = 1.75$, introducing link quality variations from obstacles and wave motion. The framework continuously adapts to such dynamics. The learning process is organized into *episodes*, each consisting of a sequence of *decision epochs* where the agent selects an action. The episode performance corresponds to the average performance over all its decision epochs.

To assess the proposed bilevel MAB framework, which jointly selects Modulation Scheme (MS), transmission power

level P, and feedback interval Q_k, we compare it with two existing DRL solutions for adaptive modulation: the Deep Reinforcement Learning-based Adaptive Modulation (DRL-AM) algorithm [10] and the DRL-based intelligent Modulation and Coding Scheme selection (DRL-MCS) algorithm [11]. Note that, due to the lack of efficient RL solutions specifically designed for UW networks, we have considered the two aforementioned algorithms as they proved to be effective in challenging terrestrial networks characterized by the coexistence of multiple users and/or limited Channel State Information (CSI). Specifically:

- *DRL-AM* addresses adaptive modulation under outdated CSI using a DRL framework to dynamically adjust MSs. Its robustness to non-linear channel variations makes it relevant for UW environments.
- *DRL-MCS* targets cognitive networks with primary and secondary users sharing the same spectrum. It optimizes modulation and coding to minimize interference from secondary users and improve overall performance.

As both DRL-based solutions are computationally intensive, we assume they operate at the Sink node rather than on resource-constrained IoUT devices. They adopt a centralized
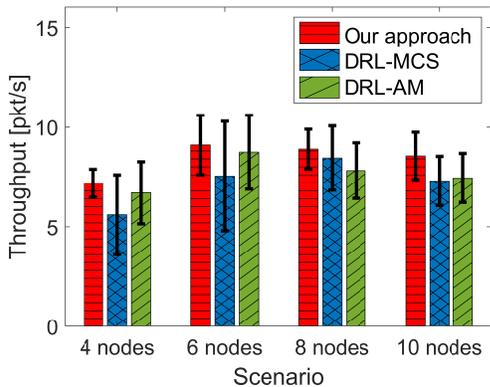
Fig. 5: Throughput in different scenarios.

Fig. 6: Energy efficiency in each considered scenario.

approach, applying the same MS to all nodes without per-node optimization. In addition, they do not adjust P nor control the feedback interval. As a result, each action requires immediate feedback, increasing network overhead. Simulations ran for a total duration of 6000 seconds. Both the bilevel MAB framework and the DRL baselines were trained on a high-performance system with an Nvidia GeForce RTX 4090 GPU. For MAB agents, c was set to 2, and $\theta = 0.7$.

VI. PERFORMANCE EVALUATION

This section analyzes the performance of our framework, comparing it against the DRL-based solutions under different network configurations.

A. Throughput Analysis

We evaluate our framework by analyzing training throughput across 4- to 10-node scenarios, as shown in Fig. 3, which details the 8- and 10-node cases. Our approach consistently outperforms DRL-based baselines in all scenarios, **achieving up to 20.61% higher throughput than DRL-MCS and 14.75% over DRL-AM**. This improvement is due to its ability to adapt MS, P, and feedback interval to local channel conditions in a fully distributed and low-variance fashion.

Unlike DRL-based solutions, which suffer from oscillatory behavior and centralized bottlenecks, our framework converges more rapidly and stably across different network sizes. DRL-AM and DRL-MCS apply uniform settings across nodes and rely on global coordination, limiting their adaptability in dense and heterogeneous environments.

Fig. 5 confirms this trend, showing that packet loss grows with network density. Our distributed scheme mitigates the issue by letting each node optimize transmissions via local feedback, ensuring high throughput and consistent performance even as the network scales.

B. Energy Consumption Analysis

We further assess our framework by analyzing energy consumption across all the network scenarios, as shown in Fig. 4, highlighting the 8- and 10-node cases. Our framework systematically surpasses DRL baselines, **achieving up to 36.60% lower energy consumption than DRL-MCS and 33.05% compared to DRL-AM**. This improvement
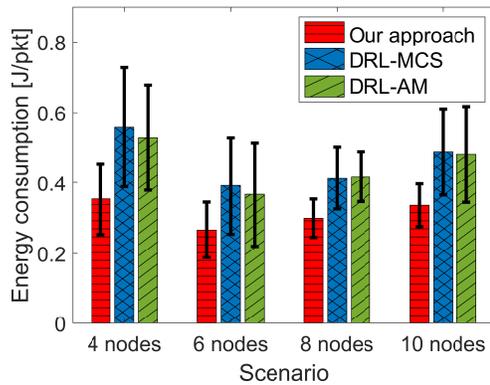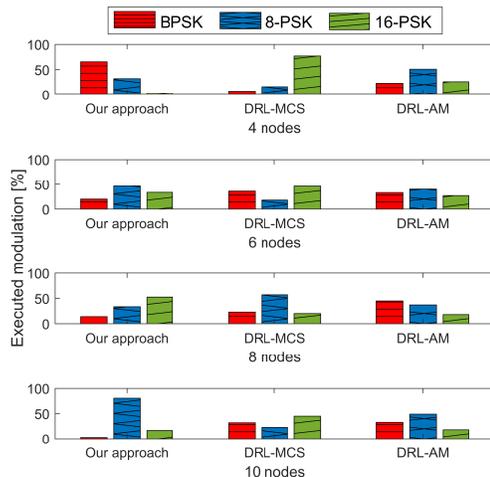

Fig. 7: Modulation selection frequency for our approach, DRL-MCS and DRL-AM in each considered scenario.

results from the joint optimization of transmission settings and feedback frequency. Two factors drive this gain. First, by jointly selecting MS and P, each node transmits using only the energy needed for successful delivery. Second, by dynamically adjusting the feedback interval Q_k, the framework reduces control transmissions when frequent updates are unnecessary.

In contrast, prior DRL-based solutions lack both power control and feedback scheduling. Feedback is sent after every packet and P remains fixed, leading to excessive signaling and higher energy consumption. Fig. 6 confirms this behavior. By adapting P and Q_k to channel dynamics and throughput needs, the framework lowers both transmission and feedback energy, preserving performance.

C. Adaptive Modulation Strategy

Fig. 7 shows the modulation selection frequency across different scenarios for our framework and the DRL baselines.

Our approach dynamically selects the MS at each node by accounting for both network density and local channel conditions. In sparse scenarios (e.g., 4 nodes), lower-order modulations like BPSK are preferred for their robustness to channel variability. As density increases (e.g., 6–8 nodes), the

Scenario	4 min	7 min	10 min
4 nodes	48.40%	31.41%	20.19%
6 nodes	61.78%	28.11%	10.11%
8 nodes	57.83%	30.50%	11.67%
10 nodes	50.80%	33.93%	15.27%

TABLE I: Average distribution of feedback interval selections Q_k across different network scenarios.

agent shifts toward higher-order schemes such as 8-PSK and 16-PSK to exploit better connectivity and boost throughput.

In high-density cases (e.g., 10 nodes), higher-order modulations remain common, but with reduced variation. This reflects the need to balance spectral efficiency with resilience to interference in crowded acoustic environments.

In contrast, DRL-based schemes enforce uniform MS across nodes, ignoring channel diversity—hurting strong links and degrading weaker ones. This limited adaptability reduces efficiency and responsiveness, explaining the suboptimal use of 16-PSK even in sparse scenarios.

D. Feedback Interval Analysis

Table I reports the average distribution of feedback interval selections Q_k across different network sizes. As node density increases, the agent progressively favors shorter intervals to maintain context freshness.

In the 4-node case, the distribution is balanced, with 4-minute updates being most frequent (48.40%) but not dominant. The presence of 10-minute intervals (20.19%) suggests occasional feedback reduction under stable conditions. With 6 and 8 nodes, the preference for 4-minute intervals strengthens (61.78% and 57.83%), reflecting the need for more frequent updates. The persistent use of 7-minute intervals (28.11% and 30.50%) indicates a trade-off between feedback cost and freshness. In the 10-node case, the share of 10-minute intervals rises again (15.27%), likely to mitigate signaling overhead. The slight drop in 4-minute updates (50.80%) suggests that the agent relaxes feedback frequency as network load increases.

Overall, the agent adjusts Q_k to balance throughput and signaling cost, rather than aiming to minimize AoI directly.

VII. CONCLUSION

In this paper, we proposed a fully distributed bilevel MAB framework for UnderWater Acoustic (UWA) networks that jointly optimizes Modulation Schemes (MSs), transmission power (P), and feedback intervals. The first level uses a Contextual Delayed MAB (CD-MAB) to adapt transmission decisions from delayed feedback, while the second level leverages a Feedback Scheduling MAB to dynamically regulate the feedback interval, balancing throughput and signaling overhead. The fully distributed design ensures scalability and low overhead, suiting Internet of Underwater Things (IoUT) scenarios and enabling future extensions to non-stationary settings and physical-layer models. Simulations with the DESERT UW simulator show our framework outperforms DRL baselines, **achieving up to 20.61% higher throughput than DRL-MCS and 14.75% over DRL-AM, while cutting** energy use by up to **36.60% and 33.05%, respectively, with faster convergence and improved stability**.

REFERENCES

[1] I. F. Akyildiz, D. Pompili, and T. Melodia, "Underwater acoustic sensor networks: research challenges," *Ad hoc networks*, vol. 3, no. 3, pp. 257–279, 2005.

[2] M. Murad, A. A. Sheikh, M. A. Manzoor, E. Felemban, and S. Qaisar, "A survey on current underwater acoustic sensor network applications," *International Journal of Computer Theory and Engineering*, vol. 7, no. 1, pp. 51–56, 2015.

[3] F. Busacca, L. Galluccio, S. Palazzo, A. Panebianco, Z. Qi, and D. Pompili, "Adaptive versus predictive techniques in underwater acoustic communication networks," *Computer Networks*, vol. 252, p. 110679, 2024.

[4] V. Sadhu, Z. Li, Z. Qi, and D. Pompili, "High-resolution data acquisition and joint source-channel coding in underwater iot," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14 003–14 013, 2023.

[5] F. Busacca, L. Galluccio, S. Palazzo, and A. Panebianco, "A comparative analysis of predictive channel models for real shallow water environments," *Computer Networks*, vol. 250, p. 110557, 2024.

[6] Y. Zhang, Z. Zhang, L. Chen, and X. Wang, "Reinforcement learning-based opportunistic routing protocol for underwater acoustic sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 3, pp. 2756–2770, 2021.

[7] Y. Wang, W. Li, and Q. Huang, "Reinforcement learning-based underwater acoustic channel tracking for correlated time-varying channels," in *IEEE OCEANS: San Diego–Porto*, 2021, pp. 1–5.

[8] F. Busacca, L. Galluccio, S. Palazzo, A. Panebianco, and R. Raftopoulos, "AMUSE: a Multi-Armed Bandit Framework for Energy-Efficient Modulation Adaptation in Underwater Acoustic Networks," *IEEE Open Journal of the Communications Society*, 2025.

[9] F. Campagnaro, R. Francescon, F. Guerra, P. Favaro, P. Casari, R. Diamant, and M. Zorzi, "The desert underwater framework v2: Improved capabilities and extension tools," in *IEEE Third Underwater Communications and Networking Conference (UComms)*, 2016, pp. 1–5.

[10] S. Mashhadi, N. Ghiasi, S. Farahmand, and S. M. Razavizadeh, "Deep reinforcement learning based adaptive modulation with outdated CSI," *IEEE Communications Letters*, vol. 25, no. 10, pp. 3291–3295, 2021.

[11] L. Zhang, J. Tan, Y.-C. Liang, G. Feng, and D. Niyato, "Deep reinforcement learning-based modulation and coding scheme selection in cognitive heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 6, pp. 3281–3294, 2019.

[12] J. J. Zhang, A. Papandreou-Suppappola, B. Gottin, and C. Ioana, "Time-frequency characterization and receiver waveform design for shallow water environments," *IEEE Transactions on Signal Processing*, vol. 57, no. 8, pp. 2973–2985, 2009.

[13] T. Yang, "Properties of underwater acoustic communication channels in shallow water," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 129–145, 2012.

[14] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 5, pp. 1183–1210, 2021.

[15] M. K. C. Shisher and Y. Sun, "On the monotonicity of information aging," in *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2024, pp. 01–06.

[16] V. Kuleshov and D. Precup, "Algorithms for multi-armed bandit problems," 2014.

[17] X. Chen and I.-H. Hou, "Contextual restless multi-armed bandits with application to demand response decision-making," pp. 2652–2657, 2024.

[18] J. Alves, J. Potter, P. Guerrini, G. Zappa, and K. LePage, "The LOON in 2014: Test bed description," in *IEEE Underwater Communications and Networking (UComms)*, 2014, pp. 1–4.

[19] EvoLogics GmbH, https://www.evologics.com.