Reinforcement Learning for Channel Estimation in Communications: A Comprehensive Review

Gemini

***Abstract*—Accurate channel state information (CSI) is crucial for modern wireless systems, but increasing complexity and dynamics pose significant challenges to traditional estimation techniques. Reinforcement Learning (RL) offers a promising solution by learning optimal strategies through environmental interaction without explicit models. This review examines RL applications in channel estimation, focusing on algorithms like Q-learning, DQN, DDPG, and PPO. We highlight key achievements in optimizing IRS configurations, estimating time-varying channels, facilitating MIMO systems, and enabling end-to-end communication. RL-based methods demonstrate superior performance over conventional approaches in terms of reduced estimation error and improved system-level metrics. While acknowledging challenges like computational complexity and generalization issues, we explore future directions including meta-RL, lightweight algorithms, and holistic optimization approaches. This review aims to advance understanding of RL's role in addressing channel estimation challenges for next-generation wireless networks.**

***Index Terms*—Reinforcement Learning, Channel Estimation, Wireless Communications, Deep Learning, Machine Learning, 5G, 6G, Intelligent Reflecting Surface (IRS), MIMO, Dynamic Channels, End-to-End Learning.**

I. Introduction

Modern wireless communications increasingly rely on accurate Channel State Information (CSI) as systems evolve toward 5G/6G with massive MIMO, mmWave/THz frequencies, and Intelligent Reflecting Surfaces (IRS). Traditional estimation methods face significant challenges in these complex scenarios: massive MIMO systems create high-dimensional channel matrices; vehicular and high-speed scenarios produce rapidly varying channels where quasi-static assumptions fail; mmWave sensitivity and IRS-created cascaded channels complicate modeling; and increasing pilots for better estimation reduces spectral efficiency. Reinforcement Learning (RL) offers a compelling solution by learning optimal policies through environmental interaction without explicit models or labeled data, making it particularly suitable for complex wireless channels. RL agents learn to adaptively select pilots, optimize IRS configurations, denoise estimates, track variations, and even optimize end-to-end communication parameters. Deep Reinforcement Learning (DRL), combining RL with neural networks, effectively handles the high-dimensional spaces typical in wireless communications. This review explores RL applications in channel estimation, examining methodologies, research achievements across diverse scenarios, performance improvements over conventional approaches, inherent challenges, and promising future directions.

II. Core Reinforcement Learning Methodologies for Channel Estimation

Several RL algorithms and frameworks are being explored for channel estimation. These can be broadly categorized based on how they learn and represent the policy and/or value function. Table **??** (placeholder for a table that could summarize these) would typically summarize these. For the purpose of this textual report, we discuss them below.

A. Value-Based Methods

Value-based methods learn a value function that estimates the expected return for each state or state-action pair. The policy is then derived implicitly by selecting actions that maximize this value function.

1) Q-Learning: Q-learning is a model-free, off-policy RL algorithm that aims to learn the optimal action-value function, $Q^*(s, a)$, which represents the maximum expected future reward achievable by taking action a in state s and following the optimal policy thereafter. The Q-values are typically stored in a table (Q-table) for discrete state and action spaces and updated iteratively using the Bellman equation: $Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)]$ where α is the learning rate and γ is the discount factor. In channel estimation, Q-learning has been applied to tasks like channel prediction in NOMA systems [41] and successive denoising in MIMO-OFDM systems [42] where the agent learns to select optimal actions (e.g., predicted channel coefficients, subcarriers to denoise) based on the current channel state representation.

2) Deep Q-Networks (DQN): For problems with large or continuous state spaces, Q-tables become impractical. DQN addresses this by using a DNN to approximate the Q-function, $Q(s, a; \theta)$, where θ represents the network weights. DQNs often employ techniques like experience replay (storing and randomly sampling past transitions to break correlations) and target networks (using a separate, periodically updated network to stabilize learning targets) to improve stability and performance. DQN and its variants (e.g., Double DQN - DDQN) are relevant for channel estimation in IoT backscatter communications [28] and can be applied to optimize parameters related to channel adaptation.

B. Policy Gradient and Actor-Critic Methods: Advanced Strategies

Policy gradient methods directly learn the policy function $\pi(a|s; \theta)$, which maps states to actions (or probabilities of actions). The policy parameters θ are updated by performing

gradient ascent on an objective function that measures the expected cumulative reward.

Actor-Critic methods combine the strengths of value-based and policy-based approaches. They consist of two components:

- **Actor:** Learns and implements the policy $\pi(a|s;\theta)$.
- **Critic:** Learns a value function (e.g., state-value function $V(s;\phi)$ or action-value function $Q(s,a;\phi)$) to evaluate the actions taken by the actor. The critic's evaluations are then used to guide the actor's policy updates.

1) Deep Deterministic Policy Gradient (DDPG): DDPG is an actor-critic, model-free algorithm designed for continuous action spaces. It learns a deterministic policy (Actor) and an action-value function (Critic). DDPG utilizes experience replay and target networks similar to DQN. It has been successfully applied to:

- Optimizing IRS configurations, where the action space (phase shifts) is continuous [7].
- End-to-end communication systems, where the transmitter (actor) learns to encode signals without explicit channel knowledge, using feedback from the receiver as a reward [11].

2) Proximal Policy Optimization (PPO): PPO is another popular actor-critic algorithm that aims for more stable and reliable policy updates than vanilla policy gradient methods. It achieves this by using a clipped surrogate objective function that limits the size of policy changes at each step, preventing performance collapse. PPO is often easier to implement and tune. While not always directly estimating channel coefficients, PPO is used for optimizing parameters in systems that rely on channel state, such as resource allocation in RIS-assisted MEC systems [16] and channel sampling patterns in ISAC [44]. It's also listed as a DRL algorithm for enhancing channel estimation in IoT backscatter systems by adapting policies [28].

3) Twin Delayed Deep Deterministic Policy Gradient (TD3): TD3 is an extension of DDPG that addresses the overestimation bias of the critic and improves stability. It employs three key techniques: clipped double Q-learning (uses two critic networks and takes the minimum of their Q-values), delayed policy updates (updates the actor less frequently than the critic), and target policy smoothing (adds noise to the target action). TD3 has been applied in RIS-assisted networks for optimizing the phase-shift matrix [20] and for signal detection in RIS-assisted ambient backscatter communication [20].

These algorithms provide a toolbox for tackling various aspects of channel estimation, from direct estimation and refinement to optimizing auxiliary systems that influence channel quality. The choice of algorithm often depends on the nature of the state and action spaces (discrete vs. continuous), the complexity of the problem, and the desired trade-off between sample efficiency and computational cost.

III. Reinforcement Learning in Action: Channel Estimation Across Diverse Communication Scenarios

RL techniques are being increasingly explored to address channel estimation challenges in a variety of modern wireless communication contexts.

A. Intelligent Reflecting Surfaces (IRS) / Reconfigurable Intelligent Surfaces (RIS)

IRS/RIS technology, which uses passive elements to reflect and steer incident signals, can create programmable wireless environments. However, optimizing the large number of phase shifts at the IRS elements, especially with limited or no dedicated sensing capabilities at the IRS, is a complex task that involves understanding and manipulating the cascaded channel (Base Station-IRS-User). RL, particularly DRL algorithms like DDPG [7], DQN, PPO, and TD3 [20], has shown significant promise in dynamically optimizing IRS phase shifts.

- **Problem Formulation:** The RL agent (often at the BS or a network controller) observes the state (e.g., received signal quality, estimated CSI, user locations) and takes actions (adjusting IRS phase shifts). The reward is typically based on metrics like sum-rate, Signal-to-Noise Ratio (SINR), coverage, or energy efficiency.
- **Achievements:** Studies show that RL can adaptively optimize IRS configurations in dynamic environments, improving signal quality, coverage, and overall system performance compared to traditional optimization or random phase shifts [7, 20]. RL can learn to overcome signal blockages and enhance weak signal paths.
- **Channel Estimation Aspect:** While RL primarily optimizes the IRS configuration, this process is intrinsically linked to the channel. The agent learns a policy that implicitly accounts for the channel characteristics to achieve its objective. Some research also looks at using RL to aid in the explicit estimation of the complex cascaded channels in IRS systems [7].

B. Time-Varying and Doubly Dispersive Channels

In high-mobility scenarios (e.g., V2X, high-speed rail), wireless channels exhibit rapid variations in both time and frequency domains (doubly dispersive channels). Traditional channel estimation methods assuming quasi-static channels perform poorly. RL offers a framework for adaptive channel estimation and tracking in such dynamic environments.

- **Approach:** RL agents can be trained to select optimal detected data symbols for data-aided channel estimation, adapting to the channel's time-varying nature [8, 35]. The MDP formulation involves states representing historical channel information and detected symbols, actions related to selecting symbols for estimation, and rewards based on estimation accuracy (e.g., minimizing MSE). * Some works propose custom RL algorithms with state element refinement to explicitly capture channel variations [8].
- **Performance:** RL-aided estimators have demonstrated improved BER and NMSE compared to conventional pilot-aided methods and RL estimators designed for time-invariant channels, especially as channel variation increases [8].

C. MIMO and Massive MIMO Systems

The high dimensionality of channel matrices in MIMO and especially massive MIMO systems presents a significant esti-

mation challenge, increasing pilot overhead and computational complexity.

- **RL for Symbol Selection/Denoising:** Q-learning has been used for successive channel denoising in MIMO-OFDM systems by formulating the selection of sub-carriers for denoising as an MDP. The agent learns to identify and refine unreliable channel estimates based on channel curvature, approaching LMMSE performance without prior channel statistics [42]. Low-complexity RL algorithms have also been developed for selecting detected data symbols to aid channel estimation, reducing complexity while improving BER and NMSE [34].
- **Implicit Channel Handling in Resource Allocation:** While not always direct channel estimation, RL is used for resource allocation (e.g., beamforming, power control) in massive MIMO, where the agent learns to make decisions based on observed CSI or quality metrics, implicitly adapting to the high-dimensional channel environment.

D. Beyond Explicit Models: RL in End-to-End (E2E) Communication Systems

A more radical approach involves using DRL to learn the entire communication system (transmitter and receiver) in an end-to-end manner, without relying on traditional block-based designs or explicit channel models.

- **DDPG-E2E Approach:** The DDPG-E2E framework treats the transmitter as an RL agent (actor) that learns to encode messages into signals [11, 12]. The wireless channel is an unknown environment. The receiver decodes the message, and the end-to-end performance (e.g., BER, reconstruction loss) is used to generate a reward signal for the transmitter.
- **Implicit Channel Estimation:** The transmitter and receiver DNNs jointly learn to adapt their strategies to the unknown channel characteristics through the E2E training process. The channel is effectively treated as a "black box" that the system learns to communicate over.
- **Benefits:** This approach can potentially discover novel communication schemes optimized for specific unknown channels and can be more robust to channel model imperfections. It eliminates the need for a separate channel estimation module.
- **Performance:** DDPG-E2E has shown better BLER performance and convergence rates compared to state-of-the-art solutions over complex wireless channels (e.g., Rician, Rayleigh) [11].

E. Emerging Frontiers and Specific Scenarios

RL is also finding applications in channel estimation or related optimization for other emerging wireless scenarios:

1) Millimeter-Wave (mmWave) Communications: MmWave systems, while offering large bandwidth, suffer from high path loss and sensitivity to blockage. Beamforming is crucial, and accurate CSI is needed.

- **RL for Beamforming and ADC Optimization:** RL (e.g., policy gradient methods) is used to jointly optimize hybrid beamforming matrices and ADC threshold levels in mmWave MIMO systems with low-resolution ADCs. The RL agent learns to maximize achievable rates by adapting to channel statistics (implicitly handled via SNR variations during training) and can be robust to noisy CSI estimates [45].

2) Internet of Things (IoT) and Backscatter Communications: For low-power IoT devices and backscatter communication systems, efficient channel estimation and adaptation are critical.

- **DRL for Adaptive Policies:** A survey on DRL for backscatter communications highlights that DRL (including DQN, DDQN, DDPG, PPO) can enhance channel estimation by enabling devices to adapt their transmission parameters (modulation, power) and estimation policies based on changing SNR and interference, leading to more accurate channel estimates and improved system performance [28]. RL also optimizes resource allocation, packet scheduling, and radio access network selection, all of which rely on or interact with channel conditions.

3) Vehicular-to-Everything (V2X) Communications: V2X channels are characterized by high mobility, rapid variations, and non-stationarity.

- **RL for Resource Management and Adaptation:** Most current RL research in V2X focuses on resource allocation, spectrum sharing, and adaptive beamforming/transmission strategies to cope with dynamic channel conditions [30, 31, 32]. For example, DRL is used to optimize beamforming and power allocation in ISAC for V2X, reducing reliance on extensive pilot signals by using sensing state information [32]. While not always estimating channel coefficients directly with RL, these systems learn to operate effectively in challenging V2X channels.

4) Underwater Acoustic Communications (UWA): UWA channels are notorious for long multipath delays, significant Doppler shifts, and limited bandwidth, making channel estimation extremely difficult.

- **RL for Adaptive Systems:** DRL (e.g., LSTM-DQN) has been applied to adaptive modulation in UWA communications to select appropriate modulation modes based on outdated or predicted CSI, outperforming traditional methods [39]. Some research mentions model-based RL frameworks to recursively estimate channel model parameters and track dynamics [39].

Across these diverse scenarios, RL provides a flexible and adaptive framework to tackle the intricacies of channel estimation, either directly or by optimizing systems to work efficiently with available channel information.

IV. Notable Achievements and Performance Benchmarks of RL in Channel Estimation

The application of Reinforcement Learning to channel estimation has yielded several notable achievements, often demonstrating superior performance compared to traditional methods or even other machine learning approaches in specific contexts.

Key performance improvements are typically measured in terms of:

- **Reduced Estimation Error:** RL-based methods have shown significant reductions in Normalized Mean Square Error (NMSE) or Mean Square Error (MSE) of channel estimates.
  - For IRS-assisted systems, DRL approaches combining CNNs and GRUs with DDPG for IRS phase shift optimization have achieved markedly smaller channel estimation errors (NMSE) for both direct and cascaded channels compared to traditional LS and LMMSE methods across public datasets and real test scenarios [7].
  - In MIMO-OFDM systems, Q-learning based successive denoising methods have approached the MSE performance of ideal LMMSE (which requires perfect channel statistics) and offered substantial gains (e.g., 6 dB) over LS estimation [42].
  - Low-complexity RL algorithms for selecting detected symbols in MIMO systems have significantly reduced NMSE compared to conventional LMMSE [34].
- **Improved System-Level Performance:**
  - **Bit Error Rate (BER) / Block Error Rate (BLER):** Lower channel estimation error translates to improved demodulation and decoding.
    - RL-aided channel estimators for time-varying MIMO systems have demonstrated better BLER than conventional pilot-aided methods and RL estimators designed for time-invariant channels [8, 35].
    - Low-complexity RL estimators for MIMO systems achieved BER improvements of approximately 0.7-1.2 dB at a BLER of 10^{-1} compared to conventional methods [34].
    - DDPG-E2E systems have shown better BLER performance and convergence rates over complex wireless channels compared to state-of-the-art E2E learning solutions [11].
  - **Achievable Rate / Sum-Rate / Throughput:** More accurate CSI enables more effective resource allocation, beamforming, and interference management.
    - Q-learning for channel prediction in MISO-NOMA systems has enhanced the sum rate compared to standard MMSE and DL-based LSTM procedures [41].
    - DRL-optimized IRS configurations have led to higher achievable system rates after beamforming co-optimization [7].
    - RL-based optimization of beamforming and ADC thresholds in mmWave MIMO systems has closely matched exhaustive search performance and significantly outperformed traditional baselines in terms of achievable rates [45].
- **Enhanced Robustness and Adaptability:**
  - RL agents can adapt to dynamic channel environments where traditional methods struggle. This is particularly evident in time-varying channels [8, 35] and systems with IRS where the environment can be actively shaped [7].
  - RL-based approaches can operate without perfect prior channel knowledge or extensive pre-labeled datasets, which is a significant advantage over some supervised DL techniques [42].
  - DDPG-E2E systems demonstrate the ability to jointly train transmitters and receivers over unknown channels, showcasing adaptability [11].
- **Reduced Overhead or Complexity (in some cases):**
  - By learning to select optimal data symbols or by enabling end-to-end learning, RL can potentially reduce the reliance on extensive pilot signaling.
  - Research into low-complexity RL algorithms aims to make these solutions practical for deployment by reducing computational demands compared to initial, more complex RL proposals [34]. For example, by using sub-blocks and backup samples, computational complexity and latency were significantly reduced without performance loss in some RL-aided channel estimators [35].

While quantitative comparisons vary widely depending on the specific scenario, RL algorithm, and baseline methods, the trend indicates that RL holds significant potential for pushing the boundaries of channel estimation performance, particularly in complex and dynamic next-generation wireless systems. The ability to learn and adapt from interaction provides a powerful alternative to model-based approaches that may falter when their assumptions are violated.

V. Overcoming Hurdles: Challenges and Limitations of RL Application in Channel Estimation

Despite the promising results, the practical application of Reinforcement Learning in channel estimation is not without its challenges and limitations. Addressing these is crucial for realizing the full potential of RL in real-world wireless systems.

- **Computational Complexity and Latency:**
  - **Training Complexity:** DRL algorithms, especially those involving deep neural networks (e.g., DQN, DDPG, PPO), can be computationally intensive to train, requiring significant data (interactions with the environment) and processing power. This can be a bottleneck for rapidly changing channel environments or resource-constrained devices [34, 35, 20].
  - **Inference Latency:** Even after training, the inference time of a complex DRL agent might be too high for real-time channel estimation, which often has stringent latency requirements.
  - **Q-table Size:** Traditional Q-learning is limited by the "curse of dimensionality," where the size of the Q-table grows exponentially with the state and action spaces, making it infeasible for many practical channel estimation problems [41]. DRL mitigates this but introduces neural network complexity.

- **Sample Efficiency and Convergence Speed:**
  - RL agents often require a large number of interactions with the environment (samples) to learn an effective policy. In wireless communication, this can translate to a long training time or the need for extensive pilot transmissions or feedback, potentially impacting spectral efficiency.
  - Convergence to an optimal or near-optimal policy can be slow, especially in complex environments with sparse rewards or long episodes.
- **Adaptation to Highly Dynamic and Non-Stationary Environments:**
  - While RL is designed for dynamic environments, extremely fast-changing or non-stationary channel conditions can still pose a challenge. A policy learned under one set of channel statistics might not perform well if the statistics change drastically and rapidly [8, 35].
  - Potential delays in dynamic decision-making by RL agents can affect real-time performance in fast-fading channels [7].
- **Reward Function Design:**
  - Designing an appropriate reward function that accurately reflects the channel estimation objective (e.g., minimizing NMSE, maximizing system throughput) and guides the agent effectively is non-trivial.
  - Sparse or delayed rewards can make learning difficult. For instance, the impact of a channel estimation action on the final BER might only be observable after many subsequent steps.
- **Generalization and Robustness:**
  - A policy trained in a specific simulation environment or under certain channel conditions might not generalize well to different, unseen real-world scenarios.
  - RL agents can be sensitive to hyperparameters, and finding the right set often requires extensive tuning.
  - Performance saturation at high SNRs has been observed, where models cannot achieve estimation errors infinitely close to zero [7].
- **State and Action Space Definition:**
  - Defining appropriate state and action spaces that capture relevant information without being excessively large is critical. Continuous or very large discrete action spaces (e.g., fine-grained IRS phase shifts) can be challenging for some RL algorithms, though methods like DDPG are designed for continuous actions.
- **Exploration vs. Exploitation Trade-off:**
  - The agent needs to balance exploring new actions to discover better policies and exploiting known good actions to maximize immediate rewards. Inefficient exploration can lead to suboptimal policies or slow convergence.
- **Challenges Specific to Certain Applications:**
  - **IRS Systems:** Continuously acquiring accurate CSI for effective IRS optimization in dynamic environments is a challenge. Optimal IRS placement and the power consumption of active/mobile RIS are also concerns [20].
  - **Data-Aided Estimation:** Using detected data symbols for estimation can lead to error propagation if symbols are detected incorrectly [8, 34]. RL helps in selective exploitation, but the risk remains.
  - **End-to-End Systems:** While promising, understanding the learned policies and ensuring interpretability in DDPG-E2E systems can be difficult. Training these systems can be computationally intensive [12].
- **Security and Vulnerability:**
  - Like other machine learning models, DRL-based channel estimation systems can be vulnerable to adversarial attacks, where malicious inputs are crafted to degrade estimation performance or compromise the system [29].

Addressing these limitations through algorithmic improvements, more efficient learning paradigms (like meta-learning or transfer learning), and careful system design is key to the widespread adoption of RL for channel estimation.

VI. The Path Forward: Innovative Frontiers and Research Directions for RL in Channel Estimation

The application of Reinforcement Learning to channel estimation is a vibrant research area with numerous exciting avenues for innovation. Future efforts are likely to focus on enhancing the intelligence, efficiency, robustness, and practicality of RL-based solutions.

A. Advanced Learning Paradigms

1) Meta-Reinforcement Learning (Meta-RL): Enabling Rapid Adaptation: Wireless channels are often non-stationary and can change significantly based on the environment, user mobility, or frequency band. Meta-RL, or "learning to learn," aims to train agents that can quickly adapt to new channel environments or tasks with minimal additional training data.

- **Concept:** The agent is trained over a distribution of different (but related) channel estimation tasks, learning a meta-policy or a good initial set of parameters that can be rapidly fine-tuned for a new, unseen channel environment.
- **Applications in CE:** Model-Agnostic Meta-Learning (MAML) and its variants are being explored for time-varying OFDM channel estimation [52, 54], allowing networks to learn channel characteristics and quickly adapt to new tasks with few training samples. Meta-DRL (e.g., DQN + MAML) is also proposed for optimizing UAV operations which inherently depend on channel conditions [48].
- **Benefits:** Reduced sample complexity for new environments, faster convergence, and better generalization.

2) Transfer Learning: Leveraging Prior Knowledge: Transfer learning allows knowledge gained from a source task/environment to be applied to a different but related target task/environment.

- **Concept:** An RL agent pre-trained on a general set of channel conditions or a simulated environment can be fine-tuned for a specific, new deployment scenario, requiring less data and time than training from scratch.
- **Applications in CE:** Deep Transfer Learning (DTL) is used for downlink channel prediction in FDD massive MIMO systems, where models trained on previous environments are fine-tuned for new ones [49]. It's also proposed to accelerate the convergence of Heterogeneous Federated Learning (HFL) for channel estimation by enhancing local model parameter initialization [50].
- **Benefits:** Faster learning in new environments, improved performance with limited data.

3) Federated and Distributed Reinforcement Learning: Towards Privacy-Preserving and Scalable Channel Estimation: With growing concerns about data privacy and the distribution of intelligence to the edge, Federated Reinforcement Learning (FRL) and distributed RL are gaining traction.

- **Concept:** Multiple agents (e.g., user devices, edge nodes) collaboratively train a global RL model without sharing their raw local channel data. Each agent trains a local model based on its own channel observations and interactions, and then model updates (parameters or gradients) are aggregated centrally (or in a decentralized manner) to improve the global model.
- **Applications in CE:** FRL is explored for channel estimation in RIS-assisted cell-free MIMO systems, where DRL (e.g., Qmix) is used for coalition formation among users to improve accuracy and reduce overhead, with transfer learning accelerating convergence [50]. Distributed learning-based channel estimation models are seen as candidates for future wireless systems beyond 5G [4].
- **Benefits:** Enhanced data privacy, reduced communication overhead (compared to sending raw data), improved scalability, and robustness due to learning from diverse data sources.

B. Lightweight and Computationally Efficient DRL Algorithms

For practical deployment, especially on resource-constrained devices (e.g., IoT devices, mobile terminals), the computational complexity and energy consumption of DRL agents must be minimized.

- **Research Focus:** Developing lightweight DRL architectures and algorithms through techniques such as:
  - **Network Pruning and Quantization:** Reducing the size and complexity of the neural networks used in DRL.
  - **Knowledge Distillation:** Training smaller "student" networks to mimic the behavior of larger, pre-trained "teacher" networks.
  - **Efficient Exploration Strategies:** Reducing the number of samples needed for learning.
  - **Low-Complexity RL Algorithms:** Designing RL algorithms that are inherently less computationally demanding, such as the low-complexity RL for symbol selection in MIMO systems which divides data blocks into sub-blocks and uses partial soft information [34].
- **Goal:** To make DRL-based channel estimation feasible for on-device learning and real-time operation in a wider range of scenarios.

C. Robust Reinforcement Learning Agents

Real-world wireless channels can be adversarial, with unpredictable interference, jamming, or sudden changes.

- **Concept:** Designing RL agents that are robust to uncertainties, disturbances, and even adversarial attacks on the learning process or the input data (e.g., pilot signals). This involves techniques from robust optimization and game theory.
- **Applications in CE:** RL combined with IRS can inherently improve robustness by adaptively reconfiguring the environment [7]. Secure channel estimation using AI/DL, considering adversarial attacks, is an active area [29].
- **Goal:** To ensure reliable performance of RL-based channel estimators even in challenging and potentially hostile wireless environments.

D. Explainable AI (XAI) for RL-based Channel Estimation

As DRL models become more complex ("black boxes"), understanding why an agent makes a particular decision becomes crucial for debugging, trust, and certification.

- **Concept:** Developing XAI techniques tailored for DRL in the context of channel estimation to provide insights into the agent's learned policy and decision-making process.
- **Benefits:** Increased trustworthiness, easier debugging of unexpected behaviors, and better insights into the underlying channel physics that the agent might have learned.

E. Holistic Optimization: Joint Channel Estimation and Cross-Layer Design

Channel estimation is not an isolated task; its quality directly impacts higher-layer operations like resource allocation, beamforming, scheduling, and even application-level QoS.

- **Concept:** Using RL to perform joint optimization of channel estimation with other communication tasks in a holistic manner. The RL agent could learn a policy that considers the end-to-end performance or cross-layer objectives.
- **Applications:**
  - End-to-end communication systems where DDPG jointly optimizes the transmitter and receiver [11].
  - RL agents that adapt channel estimation strategies (e.g., pilot density, estimation algorithm) based on current network load, QoS requirements, or application needs.
  - Joint optimization of IRS phase shifts (affecting the channel) and resource allocation.
- **Benefits:** Potentially superior overall system performance compared to optimizing each module separately.

The pursuit of these innovative directions promises to further solidify the role of RL as a key enabler for intelligent

and adaptive channel estimation in the complex landscape of future wireless communication systems.

VII. Conclusion: The Transformative Impact of Reinforcement Learning on Channel Estimation

The journey of wireless communications into an era of unprecedented complexity, dynamism, and diverse service demands necessitates a paradigm shift in how we approach fundamental challenges like channel estimation. Reinforcement Learning, particularly when augmented by the representational power of Deep Learning, has emerged as a highly promising and transformative approach. This review has charted the landscape of RL applications in channel estimation, highlighting its core methodologies, diverse applications, notable achievements, and the hurdles that still need to be overcome.

RL's inherent ability to learn from interaction, adapt to changing environments without explicit models, and optimize for long-term goals makes it uniquely suited for the intricacies of modern wireless channels. We have seen its successful application in:

- Actively shaping the wireless environment through intelligent control of IRS/RIS.
- Enhancing the accuracy and robustness of channel estimates in time-varying, doubly dispersive, and massive MIMO scenarios.
- Enabling novel end-to-end communication paradigms that implicitly handle channel effects.
- Optimizing system parameters in emerging domains like mmWave, IoT, V2X, and potentially underwater communications, all ofwhich rely on accurate channel understanding.

Performance gains in terms of reduced estimation error (NMSE), improved system metrics (BER, sum-rate), and enhanced adaptability are well-documented across various studies.

However, the path to widespread practical deployment is paved with challenges. Computational and sample complexity, convergence speed, generalization to truly novel environments, robust reward engineering, and ensuring security remain active areas of research. The "black-box" nature of some DRL models also calls for greater emphasis on explainability.

The future innovative frontiers are rich with potential. Advanced learning paradigms like meta-RL and transfer learning promise to accelerate adaptation and reduce data dependency. The development of lightweight, computationally efficient DRL algorithms is crucial for on-device intelligence. Building robust RL agents capable of withstanding adversarial conditions and unpredictable dynamics will be key for mission-critical applications. Furthermore, holistic optimization frameworks, where RL jointly tackles channel estimation alongside other cross-layer communication tasks, offer a pathway to globally optimized system performance.

In conclusion, while challenges persist, the trajectory of RL in the domain of channel estimation is clearly upward. Its principles of adaptive learning and optimization are fundamentally aligned with the requirements of next-generation wireless systems. As research continues to refine algorithms, address current limitations, and explore new synergistic combinations with other AI techniques, RL is poised to play an increasingly pivotal role in making wireless communication systems more intelligent, efficient, and resilient. The ongoing exploration in this field will undoubtedly unlock new capabilities and redefine the boundaries of what is achievable in channel estimation and beyond.

References

[1] V. W. Wong, R. Schober, D. W. K. Ng, and L. C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[2] Y. Chen, X. Chen, and S. Gu, "Distributed learning based channel estimation in wireless networks beyond 5G," *PeerJ Comput. Sci.*, vol. 10, p. e2852, 2024.

[3] L. Zhang, X. Wang, and J. Ding, "Research on channel estimation based on joint perception and deep enhancement learning in complex communication scenarios," *PeerJ Comput. Sci.*, vol. 10, p. e2852, 2024.

[4] Y. Han, S. Hong, J. Lee, and J. Kim, "Reinforcement learning-aided channel estimator in time-varying MIMO systems," *Sensors*, vol. 23, no. 12, p. 5689, 2023.

[5] F. Martinez, A. Liu, and M. Wei, "DDPG-E2E: A novel policy gradient approach for end-to-end communication systems," *arXiv:2404.06257v2*, 2024.

[6] Y. Wang, Z. Zhang, and J. Chen, "Proximal policy optimization for energy-efficient MEC systems with STAR-RIS assistance," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, pp. A-3-3, 2024.

[7] J. Huang, C. Liu, and Y. Shi, "A survey on reinforcement learning for reconfigurable intelligent surfaces in wireless communications," *Sensors*, vol. 23, no. 5, p. 2554, 2023.

[8] S. Li, L. Wang, and M. Chen, "Deep reinforcement learning for backscatter communications: Augmenting intelligence in future internet of things," *arXiv:2309.12507*, 2023.

[9] R. Zhao and D. Kumar, "Secure channel estimation using norm estimation model for 5G next generation wireless networks," *Comput. Mater. Contin.*, vol. 82, no. 1, p. 59225, 2025.

[10] K. Zhang, T. Wang, and P. Liu, "Spectrum sharing using deep reinforcement learning in vehicular networks," *arXiv:2410.12521*, 2024.

[11] M. Sun and Z. Wu, "Energy-efficient and intelligent ISAC in V2X networks with spiking neural networks-driven DRL," *arXiv:2501.01038v1*, 2025.

[12] J. Lee, S. Park, and H. Kim, "A low-complexity algorithm for a reinforcement learning-based channel estimator for MIMO systems," *IEEE Access*, vol. 10, pp. 66984-66995, 2022.

[13] L. Wang and T. Zhang, "Deep reinforcement learning-based adaptive modulation for underwater acoustic communication with outdated channel state information," *Remote Sens.*, vol. 14, no. 16, p. 3947, 2022.

[14] S. Chen and R. Li, "A study on the impact of integrating reinforcement learning for channel prediction and power allocation scheme in MISO-NOMA system," *Appl. Sci.*, vol. 13, no. 3, p. 1450, 2023.

[15] G. Liu, X. Chen, and Z. Ding, "Channel estimation via successive denoising in MIMO OFDM systems: A reinforcement learning approach," *arXiv:2101.10300*, 2021.

[16] C. Zhang and D. Wang, "Using deep reinforcement learning to enhance channel sampling patterns in integrated sensing and communication," *arXiv:2412.03157v1*, 2024.

[17] P. Yang and R. Chen, "Deep reinforcement learning for MIMO communication with low-resolution ADCs," *arXiv:2504.18957*, 2025.

[18] K. Li, J. Wang, and M. Zhang, "Deep transfer learning based downlink channel prediction for FDD massive MIMO systems," *arXiv:1912.12265*, 2019.

[19] Q. Wang and Y. Liu, "Coalition formation for heterogeneous federated learning enabled channel estimation in RIS-assisted cell-free MIMO," *arXiv:2502.05538v1*, 2025.

[20] Z. Liu and X. Wang, "Meta-learning based time-varying channel estimation method," *J. Syst. Eng. Electron.*, vol. 34, no. 3, pp. 739-749, 2023.