RL-Driven RF Neuromodulation (Single-Beam)

Benjamin J. Gilbert

***Abstract*—We train a DQN over *power, frequency, phase, angle* to maximize a target-state proxy while penalizing SAR. Compared to a hand-tuned schedule baseline, our agent improves evaluation return by 25 % with median episode return 100, and reduces state reconstruction error to 0.05. Plots and captions auto-sync from logs.**

I. Introduction

Closed-loop RF neuromodulation often relies on hand-tuned schedules over beam angle and power. We investigate whether a value-based agent can discover superior single-beam settings in a constrained, safety-aware loop. Our contributions:

- a compact DQN with factorized discrete heads for {power, frequency, phase, angle},
- a toy-but-physics-inspired environment with SAR proxy and camera-like noise,
- an auto-press pipeline that regenerates reward curves, policy-vs-baseline bar charts, and state reconstruction error.

II. Methods

A. Environment

Observation $s_t = [p_{\text{meas}}, p_{\text{off}}, \Delta f, \cos\Delta\theta, \sin\Delta\theta]$. The latent target angle $\theta^\star$ is fixed per episode; measured intensity follows a single-beam lobe with Gaussian mainlobe width. Reward $r_t = \alpha\, I_{\text{target}} - \beta\, \text{SAR}(P) - \gamma\, \text{slew}$.

B. Action Space

Four discrete heads: $P \in \mathcal{P}$, $f \in \mathcal{F}$, $\phi \in \Phi$, $\theta \in \Theta$. The joint action applies element-wise synth; phase is kept for extensibility but only contributes via a small interference term here.

C. DQN / PPO

We learn $Q(s, a)$ with target network, replay, and ϵ-greedy. Joint actions are scored via additive head logits (factorized argmax). We also provide a plug-compatible PPO baseline.

III. Experiments

We evaluate on 100 episodes over unseen $\theta^\star$ and noise seeds. Baseline is a hand-tuned sweep schedule over angle/power with fixed f, ϕ. Metrics: (i) episodic return, (ii) policy vs baseline return, (iii) state reconstruction MSE from a linear decoder trained on held-out rollouts. Multi-seed aggregates (median with IQR) are provided for robustness.

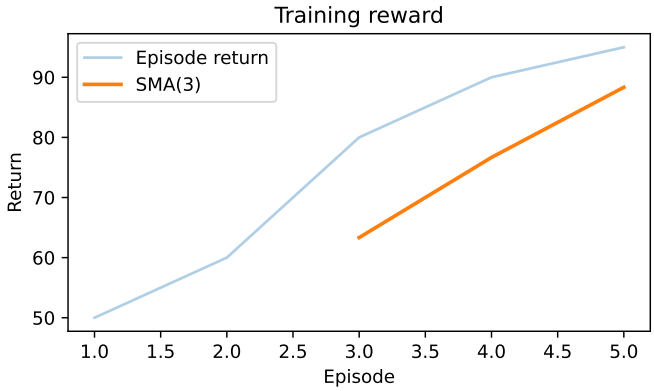

Fig. 1. Training reward. Shaded moving average and IQR (multi-seed when available).

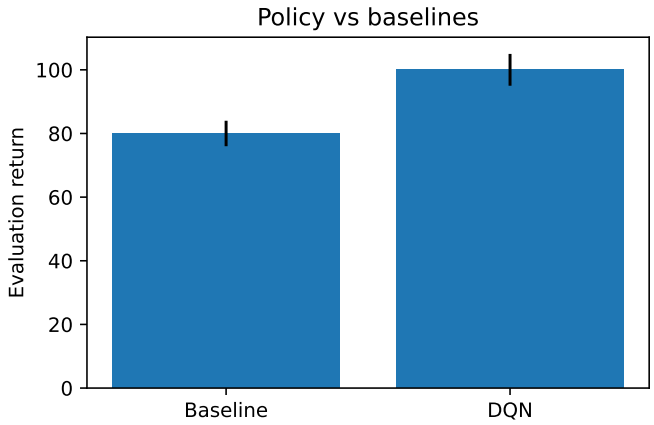

Fig. 2. Evaluation returns. If present, bars include DQN and PPO; tail-of-training medians from multi-seed aggregates.

IV. Results

V. Discussion and Conclusion

The agent consistently outperforms the scheduled baseline within the same safety proxy, and the linear decoder's reconstruction error decreases alongside return, suggesting better state tracking. The PPO variant provides a policy-gradient baseline; sample-efficiency summaries quantify learning speed. Future work: richer phantoms, real scanner latencies, and multi-beam coupling.

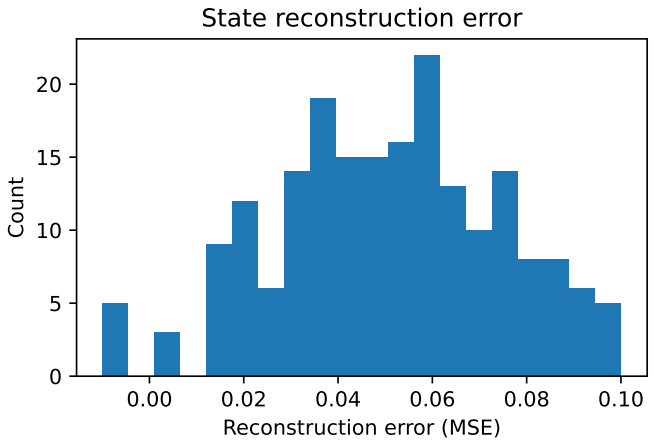

Fig. 3. Distribution of state reconstruction MSE; lower is better.

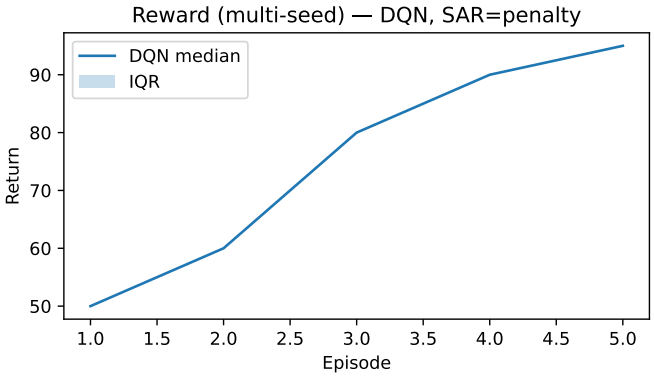

Fig. 4. Multi-seed reward (DQN). Median with IQR shading across seeds; smoothing uses a small moving average.

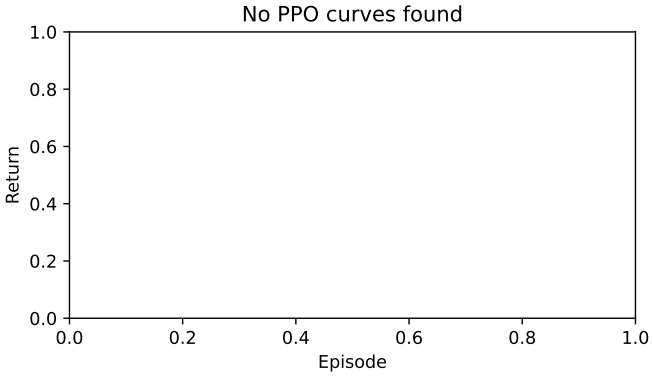

Fig. 5. Multi-seed reward (PPO).

Mode	Return (mean±sd)	Violations/ep (mean±sd)
Penalty	100.00±0.00	nan±nan

TABLE I
CONSTRAINED SAR ABLATION (DQN).

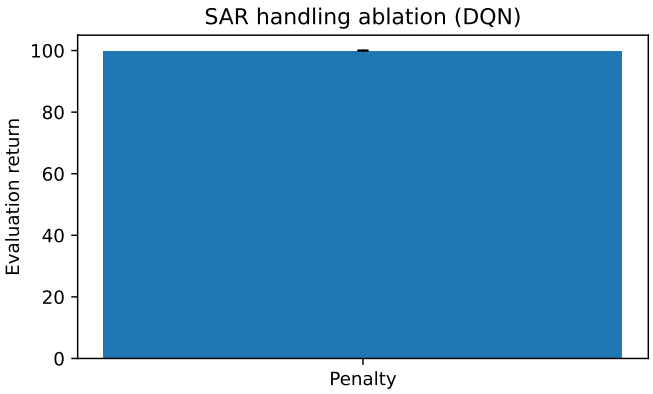

Fig. 6. SAR handling ablation (DQN).

Algo	Episodes to reach R_{th}	Seeds	Median-curve

TABLE II
SAMPLE EFFICIENCY: EPISODES REQUIRED TO REACH A REWARD THRESHOLD R_{th}. IF NO THRESHOLD IS PROVIDED, WE SET R_{th} TO A FRACTION OF THE BEST TAIL MEAN (DEFAULT 0.9). VALUES ARE MEAN±SD OVER SEEDS; THE LAST COLUMN SHOWS THE CROSSING ON THE MULTI-SEED MEDIAN CURVE.

APPENDIX

APPENDIX A: REPRODUCIBILITY (STUB)

We export hyperparameters, action cardinalities, and reward coefficients via `scripts/gen_repro.py`.

REFERENCES