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.

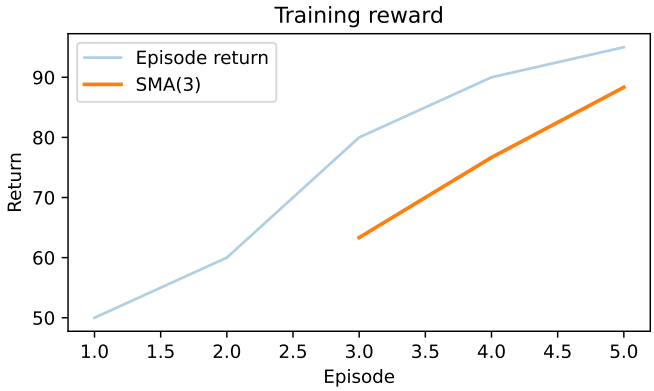

Fig. 1. Training reward. Shaded moving average and IQR (multi-seed when available).

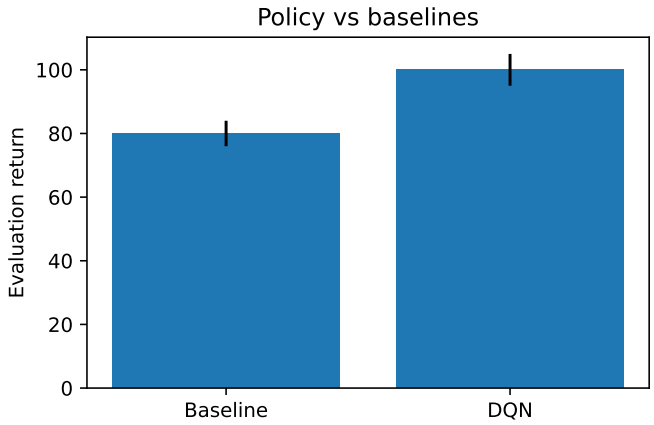

Fig. 2. Evaluation returns. If present, bars include DQN and PPO; tail-of-training medians from multi-seed aggregates.

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.

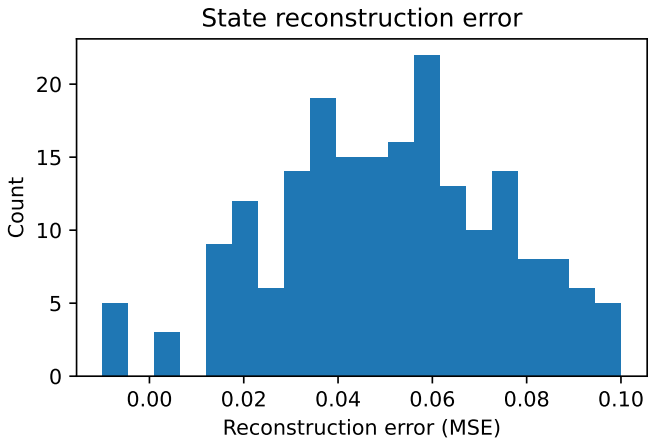

Fig. 3. Distribution of state reconstruction MSE; lower is better.

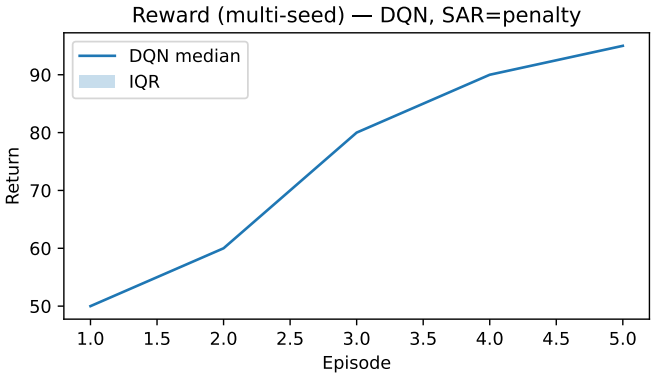

Fig. 4. Multi-seed reward (DQN). Median with IQR shading across seeds; smoothing uses a small moving average.

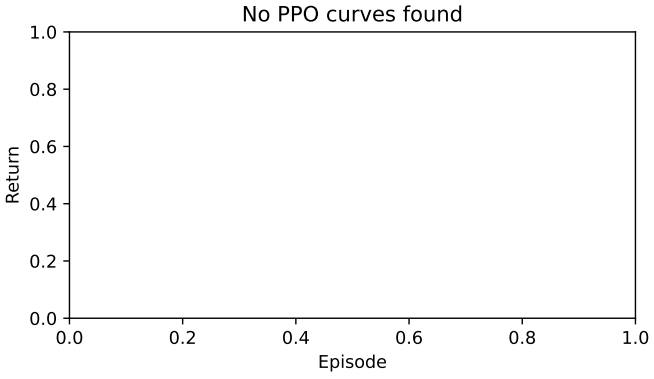

Fig. 5. Multi-seed reward (PPO).

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

TABLE I
CONSTRAINED SAR ABLATION (DQN).

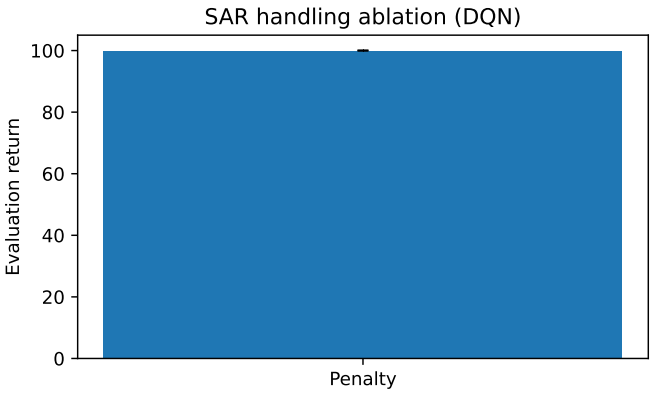

Fig. 6. SAR handling ablation (DQN).

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

Minimal reproducibility appendix stub (used when `sections/appendix.tex` *is missing).*

REFERENCES

[1] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, 2015.

[2] J. Schulman *et al.*, "Proximal policy optimization algorithms," in *ICML*, 2017.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, and et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.

[5] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2 ed., 2018.

[6] J. García and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, pp. 1437–1480, 2015.

[7] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *ICML*, 2017.

[8] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based safe policy optimization for continuous control," in *NeurIPS*, 2018.

[9] G. Dalal, K. Dvijotham, M. Vecerik, and et al., "Safe exploration in continuous action spaces," in *ICML*, 2018.

[10] A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," in *SafeML Workshop at ICLR*, 2019.

[11] H. L. Van Trees, *Optimum Array Processing*. Detection, Estimation, and Modulation Theory, Wiley, 2002.

[12] G. T. Clement and K. Hynynen, "A non-invasive method for focusing ultrasound through the human skull," *Physics in Medicine & Biology*, vol. 47, no. 8, pp. 1219–1236, 2002.

[13] A. Fomenko, C. Neudorfer, R. F. Dallapiazza, and et al., "Systematic review of low-intensity focused ultrasound neuromodulation and insights into its mechanisms," *Journal of Neural Engineering*, vol. 15, no. 5, p. 051004, 2018.

[14] W. Legon, T. Sato, A. Opitz, and et al., "Transcranial focused ultrasound modulates the activity of primary somatosensory cortex in humans," *Nature Neuroscience*, vol. 17, no. 2, pp. 322–329, 2014.

[15] R. Sitaram, T. Ros, L. Stoeckel, and et al., "Closed-loop brain training: the science of neurofeedback," *Nature Reviews Neuroscience*, vol. 18, no. 2, pp. 86–100, 2017.

[16] IEEE Standards Association, "Ieee standard for safety levels with respect to human exposure to electric, magnetic, and electromagnetic fields, 0 hz to 300 ghz," *IEEE Std C95.1-2019*, 2019.

[17] International Electrotechnical Commission, "IEC 60601-2-33: Medical electrical equipment — Particular requirements for the basic safety and essential performance of magnetic resonance equipment for medical diagnosis," 2015. +Amdts.

[18] C. M. Collins and Z. Wang, "Transmit array compression and radiofrequency power reduction using coil arrays for parallel transmission," *Magnetic Resonance in Medicine*, vol. 68, no. 4, pp. 1239–1246, 2012.

[19] B. van den Bergen, C. A. T. van den Berg, and et al., "Sar and temperature simulations in a human head within a 7 t mri birdcage coil," *Journal of Magnetic Resonance Imaging*, vol. 30, no. 2, pp. 194–203, 2009.

[20] A. Alkhateeb, S. Alex, P. Varkey, and et al., "Deep learning for mmwave beam and blockage prediction using sub-6 ghz channels," *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6306–6318, 2019.

[21] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," in *NeurIPS*, 2016.

[22] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[23] B. Efron, "Bootstrap methods: another look at the jackknife," *Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[24] A. Raffin, M. Plappert, and et al., "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research Open Source Software*, 2021. https://github.com/DLR-RM/stable-baselines3.