Convergence Guarantees for Dynamical Neural Network Policy Learning

Michael G. Rawson
Department of Mathematics, University of Maryland at College Park
and the Pacific Northwest National Laboratory
Sept 13, 2022

Joint work with Radu Balan (UMD)

Norbert Wiener Center
for Harmonic Analysis and Applications

Contents

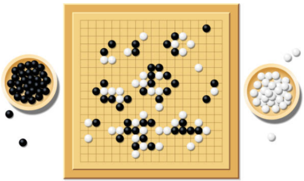
Policy Learning

Policy - Set of rules to choose an action from a set based on the state.

Policy Learning

Policy - Set of rules to choose an action from a set based on the state.

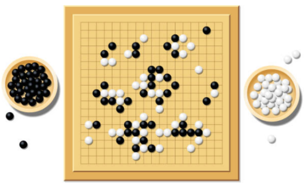
Finite (Actions and States):



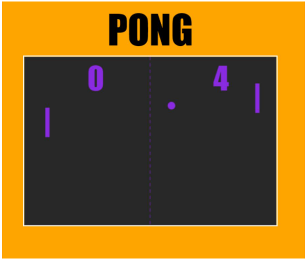
Policy Learning

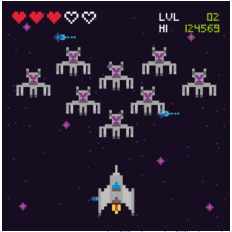
Policy - Set of rules to choose an action from a set based on the state.

Finite (Actions and States):



Infinite (Actions and States):





Active Signal Reconstruction. Reinforcement learning

State Machines have states S and transitions/actions A which take the system from state to state.

Active Signal Reconstruction. Reinforcement learning

State Machines have states S and transitions/actions A which take the system from state to state.
Reconstruct signal (reward) $R : S \times A \to \mathbb{R}$ where S are states and A are actions.

State Machines have states S and transitions/actions A which take the system from state to state.
Reconstruct signal (reward) $R : S \times A \to \mathbb{R}$ where S are states and A are actions.
The optimal policy, starting at state s_1, is

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{M} R(\gamma_{\pi}^{i-1}(s_1), \pi(\gamma_{\pi}^{i-1}(s_1)))$$

Active Signal Reconstruction. Reinforcement learning

State Machines have states S and transitions/actions A which take the system from state to state.
Reconstruct signal (reward) $R : S \times A \to \mathbb{R}$ where S are states and A are actions.
The optimal policy, starting at state s_1, is

$$\pi^* = \arg\max_\pi \sum_{i=1}^{M} R(\gamma_\pi^{i-1}(s_1), \pi(\gamma_\pi^{i-1}(s_1)))$$

where $\gamma_\pi^t(s)$ gives the state that follows from action $\pi(s)$ (policy π and state s) at time t, written $\gamma_\pi^{i-1}(s) = s'$ and γ^0 is identity.

Active Signal Reconstruction. Reinforcement learning

State Machines have states S and transitions/actions A which take the system from state to state.
Reconstruct signal (reward) $R : S \times A \to \mathbb{R}$ where S are states and A are actions.
The optimal policy, starting at state s_1, is

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{M} R(\gamma_{\pi}^{i-1}(s_1), \pi(\gamma_{\pi}^{i-1}(s_1)))$$

where $\gamma_{\pi}^{t}(s)$ gives the state that follows from action $\pi(s)$ (policy π and state s) at time t, written $\gamma_{\pi}^{i-1}(s) = s'$ and γ^0 is identity. If signal R_ω is a random variable on Ω,

$$\pi^* = \arg\max_{\pi} \sum_{i=1}^{M} \mathbb{E}_{\Omega}[R_\omega(\gamma_{\pi}^{i-1}(s_1), \pi(\gamma_{\pi}^{i-1}(s_1)))].$$

Active Signal Reconstruction

Common assumption which we use: $\gamma_\pi(s) \sim \text{distribution}(S)$ independent of policy π and state s (a.k.a. 'contextual bandit' or 'unconfoundedness') [Chen et al., 2020].

Active Signal Reconstruction

Common assumption which we use: $\gamma_\pi(s) \sim \text{distribution}(S)$ independent of policy π and state s (a.k.a. 'contextual bandit' or 'unconfoundedness') [Chen et al., 2020].

Now, for state set $S = \{\mathbb{s}\}$ (a.k.a. 'bandit'),

Active Signal Reconstruction

Common assumption which we use: $\gamma_\pi(s) \sim \text{distribution}(S)$ independent of policy π and state s (a.k.a. 'contextual bandit' or 'unconfoundedness') [Chen et al., 2020].
Now, for state set $S = \{\mathbb{s}\}$ (a.k.a. 'bandit'),

Algorithm 3: Epsilon Greedy Method [Sutton and Barto, 1998]

Parameters: $K > 1$, $c > 0$, $0 < d < 1$.
Initialization: $\epsilon_n := \min\{1, \frac{cK}{d^2 n}\}$ for $n = 1, 2, \ldots$

for $n = 1, 2, \ldots$ **do**
  i_n = the action with the highest current average reward
  if $\eta > \epsilon_n : \eta \sim Uniform([0, 1])$ **then**
    play i_n
  else
    play a uniform random action
  end
end

Contents

Related Methods

We fix dimension size and take limit in time, T.

Related Methods

We fix dimension size and take limit in time, T.

[Abbasi-yadkori et al., 2011] - Mean Regret: $\tilde{O}(1/\sqrt{T})$

Related Methods

We fix dimension size and take limit in time, T.

[Abbasi-yadkori et al., 2011] - Mean Regret: $\tilde{O}(1/\sqrt{T})$
Pros: Fast convergence rate, Simple Algorithm
Cons: Linear reward function of context

Related Methods

We fix dimension size and take limit in time, T.

[Abbasi-yadkori et al., 2011] - Mean Regret: $\tilde{O}(1/\sqrt{T})$
Pros: Fast convergence rate, Simple Algorithm
Cons: Linear reward function of context

[Zhou et al., 2020] - Mean Regret: $\tilde{O}(1/\sqrt{T})$

Related Methods

We fix dimension size and take limit in time, T.

[Abbasi-yadkori et al., 2011] - Mean Regret: $\tilde{O}(1/\sqrt{T})$
Pros: Fast convergence rate, Simple Algorithm
Cons: Linear reward function of context

[Zhou et al., 2020] - Mean Regret: $\tilde{O}(1/\sqrt{T})$ Pros: Fast convergence rate, General reward function
Cons: Expensive, Not almost surely (high probability)

Related Methods

We fix dimension size and take limit in time, T.

[Abbasi-yadkori et al., 2011] - Mean Regret: $\tilde{O}(1/\sqrt{T})$
Pros: Fast convergence rate, Simple Algorithm
Cons: Linear reward function of context

[Zhou et al., 2020] - Mean Regret: $\tilde{O}(1/\sqrt{T})$ Pros: Fast convergence rate, General reward function
Cons: Expensive, Not almost surely (high probability)

Rawson, Balan 2022 - Mean Regret: $\tilde{O}(1/\sqrt{\log T})$

Related Methods

We fix dimension size and take limit in time, T.

[Abbasi-yadkori et al., 2011] - Mean Regret: $\tilde{O}(1/\sqrt{T})$
Pros: Fast convergence rate, Simple Algorithm
Cons: Linear reward function of context

[Zhou et al., 2020] - Mean Regret: $\tilde{O}(1/\sqrt{T})$ Pros: Fast convergence rate, General reward function
Cons: Expensive, Not almost surely (high probability)

Rawson, Balan 2022 - Mean Regret: $\tilde{O}(1/\sqrt{\log T})$
Pros: General reward function, Almost surely, Simple algorithm, Fast computation
Cons: Slower convergence rate!

Contents

Active Signal Reconstruction

Input: $M \in \mathbb{N}$: total time steps, $m \in \mathbb{N}$: context dimension, $X \in \mathbb{R}^{M \times m}$ where state $X_t \in \mathbb{R}^m$ for time step t, $A = \{a_1, ..., a_K\}$: available actions, $\Phi : \mathbb{R}^m \to \mathbb{R}$: untrained neural network, function $Reward : \mathbb{N}_{[1,K]} \to \mathbb{R}$.

Active Signal Reconstruction

Input: $M \in \mathbb{N}$: total time steps, $m \in \mathbb{N}$: context dimension, $X \in \mathbb{R}^{M \times m}$ where state $X_t \in \mathbb{R}^m$ for time step t, $A = \{a_1, ..., a_K\}$: available actions, $\Phi : \mathbb{R}^m \to \mathbb{R}$: untrained neural network, function $\mathit{Reward} : \mathbb{N}_{[1,K]} \to \mathbb{R}$. **Output:** $D \in \mathbb{N}^M$: decision record, $R \in \mathbb{R}^M$ where R_t stores the reward from time step t.

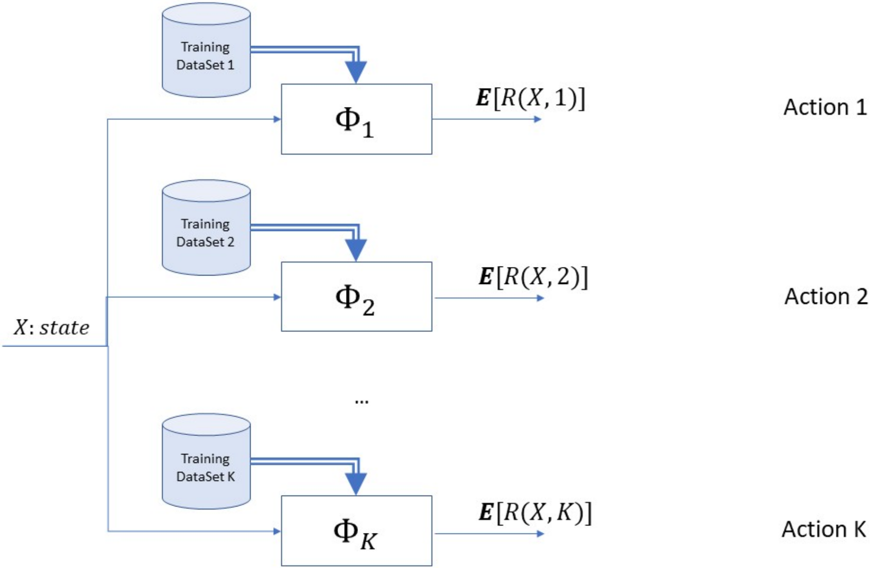
Active Signal Reconstruction

Input: $M \in \mathbb{N}$: total time steps, $m \in \mathbb{N}$: context dimension, $X \in \mathbb{R}^{M \times m}$ where state $X_t \in \mathbb{R}^m$ for time step t, $A = \{a_1, ..., a_K\}$: available actions, $\Phi : \mathbb{R}^m \to \mathbb{R}$: untrained neural network, function $Reward : \mathbb{N}_{[1,K]} \to \mathbb{R}$. **Output:** $D \in \mathbb{N}^M$: decision record, $R \in \mathbb{R}^M$ where R_t stores the reward from time step t.

Algorithm 6: Deep Epsilon Greedy

```
for t = 1, 2, ..., M do
    for j = 1 ... K do
        μ̂_{a_j} = Φ_{j,t}(X_t) (predict reward)
    end
    η ~ Uniform(0,1)
    ε_t = 1/t
    if η > ε_t then
        D_t = arg max_{1≤j≤K} μ̂_{a_j}
    else
        ρ ~ Uniform({1,...,K})
        D_t = A_ρ
    end
    R_t = Reward(D_t)
    for j = 1 ... K do
        S_j = {l : 1 ≤ l ≤ t, D_l = j}
        TrainNNet(Φ_{j,t−1}, input = X_{S_j}, output = R_{S_j})
    end
end
```

The Collection of Neural Networks



Contents

Active Signal Reconstruction

Theorem ([Györfi et al., 2002] Theorem 16.3)

Let Φ_n be a neural network with some number of parameters p and the parameters are optimized to minimize the penalized empirical risk of the training data, $S = \{(X_i, Y_i)\}_{i=1}^{n}$ where $X_i \in Sphere^m$ and Y_i almost surely bounded. Let the training data be of size n, and random variable $Y_i = R(x_i)$ depend on $x_i \in Sphere^m$. Then for n large enough,
$\mathbb{E}_S \int_{x \in Sphere} |\Phi_n(x) - \mathbb{E}(R(x))|^2 dP(x) \leq c\sqrt{\frac{\log(n)}{n}}$ *for some $c > 0$.*

Active Signal Reconstruction

Theorem ([Györfi et al., 2002] Theorem 16.3)

Let Φ_n be a neural network with some number of parameters p and the parameters are optimized to minimize the penalized empirical risk of the training data, $S = \{(X_i, Y_i)\}_{i=1}^{n}$ where $X_i \in Sphere^m$ and Y_i almost surely bounded. Let the training data be of size n, and random variable $Y_i = R(x_i)$ depend on $x_i \in Sphere^m$. Then for n large enough,
$\mathbb{E}_S \int_{x \in Sphere} |\Phi_n(x) - \mathbb{E}(R(x))|^2 dP(x) \leq c\sqrt{\frac{\log(n)}{n}}$ *for some $c > 0$.*

Assume there are K actions to play. Let random variable X be the state vector at some time step t and Y^j be the reward of action j at time step t both almost surely bounded. Let $\mu_j(X) := \mathbb{E}(Y^j|X)$.

Active Signal Reconstruction

We will use $*$ for an optimal action index, for example let $\mu_*(X)$ be the expectation of all optimal actions at X. Let $\Delta_j(X) := \max\{0, \mu_*(X) - \mu_j(X)\}$. Let $\epsilon_t = 1/t$. Let I_t be the action chosen at time t. Assume state X is sampled from an unknown distribution i.i.d. at each time step t.

Active Signal Reconstruction

We will use $*$ for an optimal action index, for example let $\mu_*(X)$ be the expectation of all optimal actions at X. Let $\Delta_j(X) := \max\{0, \mu_*(X) - \mu_j(X)\}$. Let $\epsilon_t = 1/t$. Let I_t be the action chosen at time t. Assume state X is sampled from an unknown distribution i.i.d. at each time step t.

Theorem ([Rawson and Balan, 2022])

Assume there is optimality gap δ with $0 < \delta \leq \Delta_j(X)$ for all j and X where j is suboptimal. Let C_i be the constant from above for neural network i and let n_i be the minimal value of the training data size such that neural net bounded. Then for every $t > t_0$ with probability at least $1 - K\exp(-3\log(t)/(28K))$,

$$\begin{aligned}\delta/(tK) &\leq \mathbb{E}_{X_t}\mathbb{E}_{I_t}\mathbb{E}_R\left[R_*(X_t) - R(X_t)\right] \\ &\leq \frac{\max_i \mathbb{E}_{X_t}\Delta_i(X_t)}{t} + K^{3/2}\frac{C_0}{\delta}\sqrt{\frac{\log(\log(t)) - \log(2K)}{\log(t)}}.\end{aligned}$$

Active Signal Reconstruction

Generalize ϵ_t by raising to the p power.

Active Signal Reconstruction

Generalize ϵ_t by raising to the p power.

Theorem ([Rawson and Balan, 2022])

Let $\epsilon_t = 1/t^p$ *where* $0 < p < 1$. *With the above assumptions, set* $C_0' = 8\sqrt{2(1-p)}\max_i C_i$ *and* $t_0 > (2(1-p)K\max\{e, \max_i n_i\})^{1/(1-p)}$. *Then for every* $t > t_0$ *with probability at least* $1 - K\exp\left(-(3\ t^{-p+1})/(28(-p+1)K)\right)$,

$$\delta/(Kt^p) \leq \mathbb{E}_{X_t}\mathbb{E}_{I_t}\mathbb{E}_R\left[R_*(X_t) - R(X_t)\right]$$

$$\leq \frac{\max_i \mathbb{E}_{X_t}\Delta_i(X_t)}{t^p} + K^{3/2}\frac{C_0'}{\delta}\sqrt{\frac{\log(t^{-p+1}) - \log(2(-p+1)K)}{t^{-p+1}}}.$$

Active Signal Reconstruction

Generalize ϵ_t by raising to the p power.

Theorem ([Rawson and Balan, 2022])

Let $\epsilon_t = 1/t^p$ where $0 < p < 1$. With the above assumptions, set $C_0' = 8\sqrt{2(1-p)}\max_i C_i$ and $t_0 > (2(1-p)K\max\{e, \max_i n_i\})^{1/(1-p)}$. Then for every $t > t_0$ with probability at least $1 - K\exp\left(-(3\ t^{-p+1})/(28(-p+1)K)\right)$,

$$\delta/(Kt^p) \leq \mathbb{E}_{X_t}\mathbb{E}_{I_t}\mathbb{E}_R\left[R_*(X_t) - R(X_t)\right]$$
$$\leq \frac{\max_i \mathbb{E}_{X_t}\Delta_i(X_t)}{t^p} + K^{3/2}\frac{C_0'}{\delta}\sqrt{\frac{\log(t^{-p+1}) - \log(2(-p+1)K)}{t^{-p+1}}}.$$

The expectations in above equations refer to the specific time step t. The probability refers to the stochastic policy's choices at previous time steps, 1 to $t-1$.

Active Signal Reconstruction [Rawson and Balan, 2022]

Corollary

The Epsilon Greedy method with any predictor, neural network or otherwise, with convergence of $c\sqrt{\frac{\log(n)}{n}}$, *or better, will have regret converging to 0 almost surely.*

Active Signal Reconstruction [Rawson and Balan, 2022]

Corollary

The Epsilon Greedy method with any predictor, neural network or otherwise, with convergence of $c\sqrt{\frac{\log(n)}{n}}$*, or better, will have regret converging to 0 almost surely.*

Remark

With $\epsilon_t = 1/t^p$ with $p \leq 1$, enough samples will be taken to train an approximation to convergence. When $p > 1$, The number of samples is finite and the approximation will not converge in general. This is called a starvation scenario since the optimal action is not sampled sufficiently.

Active Signal Reconstruction [Rawson and Balan, 2022]

Corollary

The Epsilon Greedy method with any predictor, neural network or otherwise, with convergence of $c\sqrt{\frac{\log(n)}{n}}$*, or better, will have regret converging to 0 almost surely.*

Remark

With $\epsilon_t = 1/t^p$ with $p \leq 1$, enough samples will be taken to train an approximation to convergence. When $p > 1$, The number of samples is finite and the approximation will not converge in general. This is called a starvation scenario since the optimal action is not sampled sufficiently.

Corollary

The optimal p *for* $\epsilon_t = 1/t^p$ *with the fastest converging upper bound of above theorem for Deep Epsilon Greedy is* $p = 1/3$*.*

Contents

Active Signal Reconstruction: MNIST Experiments

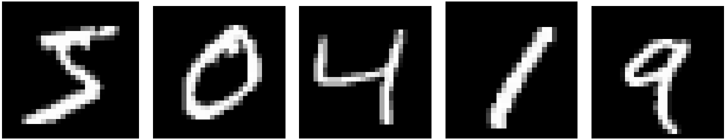
Find optimal policy to maximize the reward where
$R(a_i) = \text{digit}(image_i) + \text{Gaussian noise}$.

Active Signal Reconstruction: MNIST Experiments

Find optimal policy to maximize the reward where
$R(a_i) = \text{digit}(image_i) + \text{Gaussian noise}$.
Solution π^* selects a_i corresponding to $image_i$ with largest integer.

Active Signal Reconstruction: MNIST Experiments

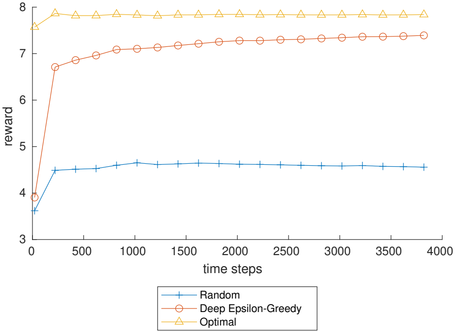
Find optimal policy to maximize the reward where
$R(a_i) = \text{digit}(image_i) + \text{Gaussian noise}$.
Solution π^* selects a_i corresponding to $image_i$ with largest integer.



(a) $image_1$ (b) $image_2$ (c) $image_3$ (d) $image_4$ (e) $image_5$

Figure: Example of MNIST images that form the random context (or state) vector.

Active Signal Reconstruction: MNIST Experiments



Active Signal Reconstruction: MNIST Experiments

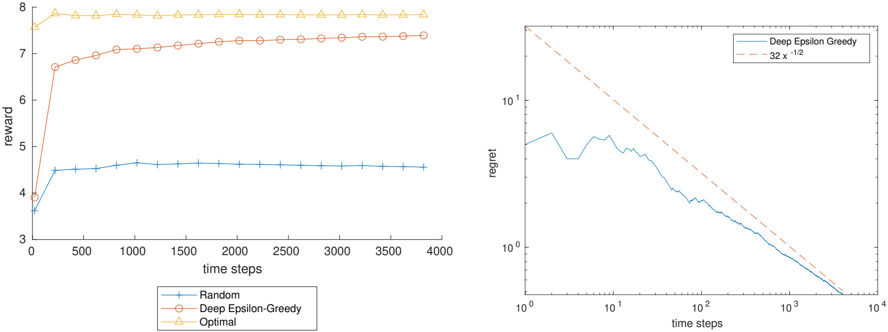
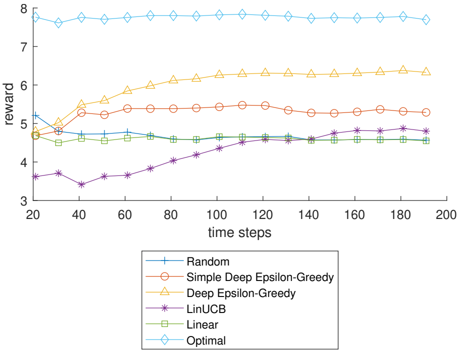


Figure: Deep Epsilon Greedy method convergence of regret to 0 at rate $x^{-1/2}$. Plotting normalized reward of optimal method minus normalized reward of Deep Epsilon Greedy method. No noise added to MNIST dataset. Single run with 1000 neurons in the fully connected, final layer.

Active Signal Reconstruction: MNIST Experiments



Active Signal Reconstruction: MNIST Experiments

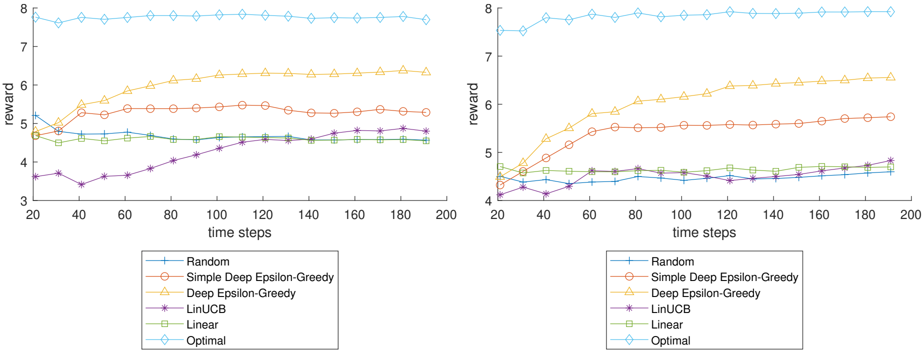


Figure: Left: Low Noise with no Gaussian noise added to reward. Right: High Noise with Gaussian noise, sigma = 1, added to reward. Left and Right: Mean reward normalized (divide by time step) plotted over time steps for each method. Task is to choose the largest MNIST image (digit) of 5 random images. Mean is over 12 independent runs.

Contents

Conclusion

- Deep Epsilon Greedy Method is simplest, fastest method.

Conclusion

- Deep Epsilon Greedy Method is simplest, fastest method.
- Showed convergence of Deep Epsilon Greedy Method.

Conclusion

- Deep Epsilon Greedy Method is simplest, fastest method.
- Showed convergence of Deep Epsilon Greedy Method.
- Showed that $\epsilon_t = t^{-1/3}$ minimizes error bound.

Conclusion

- Deep Epsilon Greedy Method is simplest, fastest method.
- Showed convergence of Deep Epsilon Greedy Method.
- Showed that $\epsilon_t = t^{-1/3}$ minimizes error bound.
- Flexible convergence accommodates various learning methods.

Conclusion

- Deep Epsilon Greedy Method is simplest, fastest method.
- Showed convergence of Deep Epsilon Greedy Method.
- Showed that $\epsilon_t = t^{-1/3}$ minimizes error bound.
- Flexible convergence accommodates various learning methods.
- Confirmed theory on real-world MNIST dataset.

References I

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011).
Improved algorithms for linear stochastic bandits.
In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Chen, M., Liu, H., Liao, W., and Zhao, T. (2020).
Doubly robust off-policy learning on low-dimensional manifolds by deep neural networks.
Submitted to Operations Research, under revision.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002).
A Distribution-Free Theory of Nonparametric Regression.
Springer Series in Statistics. Springer New York.

Rawson, M. and Balan, R. (2022).
Convergence Guarantees for Deep Epsilon Greedy Policy Learning.
arXiv:2112.03376.

References II

Sutton, R. S. and Barto, A. G. (1998).
Reinforcement Learning: An Introduction.
Cambridge, MA.

Zhou, D., Li, L., and Gu, Q. (2020).
Neural contextual bandits with UCB-based exploration.
In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 11492–11502.

Thank You!

Questions?