Neural Contextual Bandits without Regret

Parnian Kassraie, Andreas Krause

ETH zürich — Learning & Adaptive Systems

Overview

- Contextual Bandits are rich models for online decision-making problems where an agent sequentially interacts with an unknown, responsive environment and receives a reward.
- How can an agent leverage the experssive power of neural networks to learn the reward function and eventually converge to an optimal action selection policy?
- We propose algorithms that employ (convolutional) neural networks to estimate the reward, and provably attain sublinear regret.

Problem Setting

- At every step $1 \leq t \leq T$

$$y_t = f(\boldsymbol{x}_t) + \epsilon_t$$

Action	$\boldsymbol{a}_t \in \mathcal{A}$ one-hot vector of length $\lvert\mathcal{A}\rvert$
Context	$\boldsymbol{z}_t = (\boldsymbol{z}_{t,1}, \cdots, \boldsymbol{z}_{t,\lvert\mathcal{A}\rvert}) \in \mathbb{R}^{d \times \lvert\mathcal{A}\rvert}$
Noise	ϵ_t : σ^2 sub-Gaussian, i.i.d.
Input	$\boldsymbol{x}_t = \boldsymbol{z}_t \boldsymbol{a}_t \in \mathbb{R}^d$ $\quad \mathcal{X} \subseteq \mathbb{S}^{d-1}$
Reward	$f : \mathcal{X} \to \mathbb{R}$ $\quad f \in \mathcal{H}_{k_{\text{NN}}}$
	$\lVert f \rVert_{k_{\text{NN}}} \leq B$ $\quad k_{\text{NN}}$: the Neural Tangent Kernel

- Cumulative regret

$$R_T = \sum_{t=1}^{T} f(\boldsymbol{x}_t^*) - f(\boldsymbol{x}_t) \qquad \boldsymbol{x}_t^* = \underset{\boldsymbol{x}=\boldsymbol{z}_t\boldsymbol{a}, \boldsymbol{a}\in\mathcal{A}}{\arg\max}\, f(\boldsymbol{x})$$

- Learner's goal

$$R_T/T \to 0 \text{ as } T \to \infty$$

Neural Bandits

- We propose two algorithms, NN-UCB and CNN-UCB.
- At every step $1 \leq t \leq T$

Approximate the mean of reward with $\hat{\mu}_{t-1}(\boldsymbol{x}) = f^{(J)}(\boldsymbol{x}; \boldsymbol{\theta}_{t-1})$

the (C)NN trained for J steps of gradient descent on

$$\mathcal{L}(\theta) = \sum_{i=1}^{t-1} \left(f(\boldsymbol{x}_i; \boldsymbol{\theta}) - y_i\right)^2 + m\sigma^2 \lVert \boldsymbol{\theta} - \boldsymbol{\theta}^0 \rVert_2^2$$

Approximate the variance of reward with

$$\hat{\sigma}^2_{t-1}(\boldsymbol{x}) = \frac{\boldsymbol{g}^T(\boldsymbol{x})}{\sqrt{m}} \left(\sigma^2 \boldsymbol{I} + \sum_{i=1}^{t-1} \frac{\boldsymbol{g}^T(\boldsymbol{x}_i)\boldsymbol{g}(\boldsymbol{x}_i)}{m}\right)^{-1} \frac{\boldsymbol{g}(\boldsymbol{x})}{\sqrt{m}}$$

$$g(x) = \nabla_\theta f(x; \theta^{(0)})$$

Pick actions by maximizing the approximate Upper Confidence Bound

$$\boldsymbol{x}_t = \underset{\boldsymbol{x}=\boldsymbol{z}_t\boldsymbol{a},\, \boldsymbol{a}\in\mathcal{A}}{\arg\max}\; \hat{\mu}_{t-1}(\boldsymbol{x}) + \sqrt{\beta_t}\hat{\sigma}_{t-1}(\boldsymbol{x})$$

Information Gain

- It often appears in bandit regret bounds and quantifies the speed at which the agent learns about the reward function. It is the mutual information between the observed reward and true reward values. H denotes the entropy.

$$I(\boldsymbol{y}_T; \boldsymbol{f}_T) := H(\boldsymbol{y}_T) - H(\boldsymbol{y}_T | \boldsymbol{f}_T) = \frac{1}{2} \log\det(\boldsymbol{I} + \sigma^{-2}\boldsymbol{K}_T)$$

$$\boldsymbol{K}_T = [k_{\text{NN}}(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j \leq T}$$

- Its maximum depends only on the domain, the noise and the kernel function

$$\gamma_T = \max_{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T} I(\boldsymbol{y}_T; \boldsymbol{f}_T)$$

Theorem (Information Gain Bound, Informal)

Then the maximum information gain associated with the NTK of a fully-connected ReLU network is bounded by

$$\gamma_{T,NN} = \tilde{\mathcal{O}}\left(C_1(d, L) T^{\frac{d-1}{d}}\right)$$

Main Result

- We resolve the open problem of proving sublinear regret bounds for general context sequences, considering both fully-connected and convolutional nets.

We show that our algorithms converge to the optimal policy in polynomial time with high probability.

Theorem (Regret Bound, Neural, Informal)

Let $\delta \in (1, 0)$. Set $\beta_t = 2\log(2T|\mathcal{A}|/\delta)$. Choose the width s.t.

$$m \geq poly\left(T, L, |\mathcal{A}|, \sigma^{-2}, B, \lambda_0^{-1}, \log(1/\delta)\right)$$

(minimum eigenvalue of the kernel matrix)

and learning rate $\eta = C(LmT + m\sigma^2)^{-1}$ with some constant C. Then with probability greater than $1 - \delta$, SupNN-UCB satisfies

$$R(T) = \tilde{\mathcal{O}}(C_2(d, L) T^{\frac{2d-1}{2d}}).$$

- Comparison to prior work: [2] are the first to introduce neural contextual bandits. They provide a guarantee roughly of the form

$$R_T \leq \tilde{\mathcal{O}}\left(I(\boldsymbol{y}_T; \boldsymbol{f}_T)\sqrt{T}\right)$$

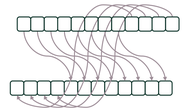
This bound is data-dependent via the Information Gain. The maximum information gain itself grows as $\gamma_{T,\text{NN}} = \tilde{\mathcal{O}}\left(T^{\frac{d-1}{d}}\right)$ for the NTK. Without further assumptions on the sequence of contexts, this bound may be vacuous.

Convolutional Guarantees

- Invariance Trick [1]

$$\boldsymbol{w} * \boldsymbol{x} = \sum_{l=1}^{d} \langle \boldsymbol{w}, c_l \cdot \boldsymbol{x} \rangle$$

$$c_l \cdot \boldsymbol{x} = (x_{l+1}, x_{l+2}, \cdots, x_d, x_1, \cdots, x_l)$$



2-Layer convolutional network is invariant to circular shifts.

$$f_{\text{CNN}}(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{1}{d}\sum_{i=1}^{m} v_i\, \sigma_{\text{relu}}(\boldsymbol{w}_i * \boldsymbol{x}) = \frac{1}{d}\sum_{l=1}^{d} f_{\text{NN}}(c_l \cdot \boldsymbol{x}; \boldsymbol{\theta})$$

$$f_{\text{CNN}}(c_l \cdot \boldsymbol{x}) \;=\; f_{\text{CNN}}(\boldsymbol{x})$$

- Connection between the NTK and the CNTK

$$k_{\text{CNN}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{d}\sum_{l=1}^{d} k_{\text{NN}}(\boldsymbol{x}, c_l \cdot \boldsymbol{x}')$$

When restricted to S^{d-1}, the NTK is described by $(\mu_k, \mathcal{F}_{d,k})_{k\geq 0}$ a sequence of eigenvalue eigenspace pairs. Here, the k-th eigenspace is spanned by degree-k polynomials. For the CNTK however, the k-th eigenspace is spanned by polynomials that are *invariant to circular shifts*. Searching through this shrunken RKHS gives us improved rates for CNN-UCB.

- Reward model assumption

$$f \in \mathcal{H}_{k_{\text{CNN}}} \qquad \lVert f \rVert_{k_{\text{CNN}}} \leq B$$

CNN-UCB benefits from a shift-invariant structure.
Our results suggest that for a high-dimensional input it outperforms NN-UCB.

- Guarantees for the Convolutional Neural Bandit

$$\gamma_{T,\text{CNN}} = \tilde{\mathcal{O}}\left(C_1(d, L)\left(\frac{T}{d}\right)^{\frac{d-1}{d}}\right)$$

Theorem (Regret Bound, Convolutional, Informal)

Let $\delta \in (1, 0)$. Set $\beta_t = 2\log(2T|\mathcal{A}|/\delta)$. For any T, there exists width m such that if $\eta = C(LmT + m\sigma^2)^{-1}$ with some constant C, then with probability greater than $1 - \delta$, SupCNN-UCB satisfies

$$R(T) = \tilde{\mathcal{O}}\left(\frac{C_2(d, L)}{d^{(d-1)/2d}} T^{\frac{2d-1}{2d}}\right).$$

References

[1] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3351–3418. PMLR, 15–19 Aug 2021.

[2] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with ucb-based exploration. In *International Conference on Machine Learning*, pages 11492–11502. PMLR, 2020.