Generalized Linear Bandits with Limited Adaptivity

Ayush Sawarni[1], Nirjhar Das[1], Siddharth Barman[2], Gaurav Sinha[1]
[1]Microsoft Research India, [2]Indian Institute of Science

Generalized Linear Bandits

Generalized Linear Models: Random variable r has PDF with parameter z:

$$\mathbb{P}_z[r] = \exp(rz - b(z) + c(r))$$

$b(z)$ is convex and $\mu(z) := \dot{b}(z) = \mathbb{E}_z[r]$.

- We consider GLMs with $r \in [0, R]$ a.s.

At every round $t \in \{1, \dots, T\}$:
1. A context $\mathcal{X}_t = \{x_{1,t}, \dots, x_{K,t}\} \subset \mathbb{R}^d$ is presented
2. Learner plays *arm* $x_t \in \mathcal{X}_t$ according to some policy π_t
3. Learner observes *reward* r_t sampled from a GLM with parameter $x_t^\intercal \theta^*$
4. (Optional) Learner updates policy π_t to π_{t+1} using observation and history

Limited Adaptivity

Model M1: Learner can update policy only M (given) number of times. *Learner must declare before the start of bandit instance at which rounds it will update its policy.*

Model M2: Learner can update the policy for $\texttt{polylog}(T)$ times. *Learner can decide adaptively in which rounds it will update the policy.*

`B-GLinUCB` for M1

- Stochastic Contexts i.e., $\mathcal{X}_t \sim \mathcal{D}$
- Performance: Regret over T rounds given by-

$$\mathrm{R}_T = \mathbb{E}\Big[\sum_{t=1}^{T}\big(\max_{x\in\mathcal{X}_t}\mu(x^\intercal\theta^*) - \mu(x_t^\intercal\theta^*)\big)\Big]$$

- Non-linearity measures: For arm set $\mathcal{X}$, let $x^* = \operatorname{argmax}_{x\in\mathcal{X}} \mu(x^\intercal\theta^*)$. Define the quantities:

$$\kappa := \max_{\mathcal{X}\in\mathrm{supp}(\mathcal{D})}\max_{x\in\mathcal{X}} \frac{1}{\dot{\mu}(x^\intercal\theta^*)}$$

$$\frac{1}{\kappa^*} := \max_{\mathcal{X}\in\mathrm{supp}(\mathcal{D})} \dot{\mu}(x^{*\intercal}\theta^*)$$

$$\frac{1}{\widehat{\kappa}} := \mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\left[\dot{\mu}(x^{*\intercal}\theta^*)\right]$$

Optimal Design Policies

G-Optimal Design

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\Delta(\mathcal{X})$ be set of probability distributions supported on $\mathcal{X}$. For $\lambda \in \Delta(\mathcal{X})$, let $U(\lambda) = \mathbb{E}_{x\sim\lambda}[xx^\intercal]$. Define:

$$\pi_G = \operatorname*{argmin}_{\lambda\in\Delta(\mathcal{X})}\max_{x\in\mathcal{X}} \|x\|^2_{U(\lambda)^{-1}}$$

$$\pi_D = \operatorname*{argmax}_{\lambda\in\Delta(\mathcal{X})} \log\det\big(U(\lambda)\big)$$

Kiefer-Wolfowitz Theorem: $\pi_G = \pi_D$ and $\max_{x\in\mathcal{X}}\|x\|^2_{U(\pi_G)^{-1}} = d$.

Distributional Optimal Design [Ruan et al. (2021)]

Let $\mathcal{M} = \{(p_i, \mathbf{M}_i)\}_{i=1}^n$ where, $p_i \geq 0$ and $\sum_i p_i = 1$. For any $i \in [n]$, let $\pi_{\mathbf{M}_i} \in \Delta(\mathcal{X})$ defined as:

$$\pi_{\mathbf{M}_i}(x) = \frac{\|x\|^{2\alpha}_{\mathbf{M}_i}}{\sum_{y\in\mathcal{X}}\|y\|^{2\alpha}_{\mathbf{M}_i}} \quad \forall\, x \in \mathcal{X}$$

`Distributional Optimal Design` π for collection $\mathcal{M}$ is given as:

$$\pi(x) = \frac{1}{2}\pi_G(x) + \sum_{i=1}^{n}\frac{p_i}{2}\pi_{\mathbf{M}_i}(x),\ \forall x \in \mathcal{X}$$

Lemma: Let $\mathcal{X}_1, \dots, \mathcal{X}_s \overset{\text{i.i.d}}{\sim} \mathcal{D}$ and let $\mathcal{M}$ be constructed using Algorithm 2 of [1]. Further, define $\mathbf{W} = \mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\big[\mathbb{E}_{x\sim\pi}[xx^\intercal \mid \mathcal{X}]\big]$. Then, with high probability,

$$\mathbb{E}_{\mathcal{X}\sim\mathcal{D}}\big[\max_{x\in\mathcal{X}}\|x\|_{\mathbf{W}^{-1}}\big] \leq O(\sqrt{d\log d})$$

Algorithm

Batch lengths τ_k, $k \in [M]$ are calculated as:

$$\tau_1 := \left(\frac{\sqrt{\kappa}\, e^{3S} d^2\gamma^2}{S}\alpha\right)^{2/3},$$

$$\tau_2 := \alpha, \tau_k := \alpha\sqrt{\tau_{k-1}}, \text{ for } k \in [3, M]$$

where $\gamma := 30RS\sqrt{d\log T}$ ($\|\theta^*\| \leq S$) and $\alpha = T^{\frac{1}{2(1-2^{-M+1})}}$ if $M \leq \log\log T$ and $\alpha = 2\sqrt{T}$ else.

`B-GLinUCB`

1. τ_1 rounds, play arms using π_G and observe rewards.
2. Obtain $\widehat{\theta}_w$ via MLE.
3. For batches $k = 2, \dots, M$ do:
4. For τ_k rounds do:
5. Receive arm set $\mathcal{X}_t$.
6. Use previous estimates of θ^* to eliminate arms.
7. Scale the reduced arm set with a non-linearity factor.
8. Play an arm based on `Distributional Optimal Design` policy on the scaled arm set.
9. Estimate (via MLE) θ^* .
10. Construct a new `Distributional Optimal Design` policy.

Theorem: Regret of `B-GLinUCB` $\mathrm{R}_T \leq (\mathrm{R}_1 + \mathrm{R}_2)\log\log T$, where

$$\mathrm{R}_1 = O\left(RSd\left(\sqrt{\frac{d}{\widehat{\kappa}}} \wedge \sqrt{\frac{1}{\kappa^*}}\right)T^{\frac{1}{2(1-2^{1-M})}}\log T\right) \text{ and}$$

$$\mathrm{R}_2 = O\left(\kappa^{1/3}d^2e^{2S}(RS\log T)^{2/3}T^{\frac{1}{3(1-2^{1-M})}}\right).$$

Corollary: When $M \geq \log\log T$, `B-GLinUCB` achieves a regret bound of

$$\mathrm{R}_T \leq \widetilde{O}\Bigg(\left(\sqrt{\frac{d}{\widehat{\kappa}}} \wedge \sqrt{\frac{1}{\kappa^*}}\right) dRS\sqrt{T} + d^2e^{2S}(S^2R^2\kappa T)^{1/3}\Bigg)$$

`RS-GLinUCB` for M2

- Adversarial Contexts: $\mathcal{X}_t$ can be any subset of $\mathbb{R}^d$
- Performance: Regret over T rounds given by-

$$\mathrm{R}_T = \sum_{t=1}^{T}\big(\max_{x\in\mathcal{X}_t}\mu(x^\intercal\theta^*) - \mu(x_t^\intercal\theta^*)\big)$$

- Non-linearity measure: For adversarial context

$$\kappa := \max_{x\in\cup_{t=1}^{T}\mathcal{X}_t} \frac{1}{\dot{\mu}(x^\intercal\theta^*)}$$

Algorithm

Key Highlights

- Optimal Regret: Resolves conjecture in GLM Bandit by removing κ from $\sqrt{T}$-term
- Computationally Efficient: Update time is per round amortized $O(\texttt{poly}(d)\log T)$
- S-free Regret: Resolves conjecture of polynomial dependence on S in regret's leading term

Main Idea: Context-dependent switching criterion *in addition to* determinant-doubling trick

`RS-GLinUCB`

1. Initialize: $\mathbf{V} = \mathbf{H}_1 = \lambda\mathbf{I}$, $\mathcal{T}_o = \emptyset$, $\tau = 1$, $\lambda := d\log(T/\delta)/R^2$ and $\gamma := 25RS\sqrt{d\log\left(\frac{T}{\delta}\right)}$.
2. For rounds $t = 1, \dots, T$ do:
3. Observe arm set $\mathcal{X}_t$.
4. If $\max_{x\in\mathcal{X}_t}\|x\|^2_{\mathbf{V}^{-1}} \geq 1/(\gamma^2\kappa R^2)$ [`Criterion I`]
5. Select $x_t = \operatorname{argmax}_{x\in\mathcal{X}_t}\|x\|_{\mathbf{V}^{-1}}$ and observe r_t.
6. Update $\mathcal{T}_o \leftarrow \mathcal{T}_o \cup \{t\}$, $\mathbf{V} \leftarrow \mathbf{V} + x_tx_t^\intercal$ and $\mathbf{H}_{t+1} \leftarrow \mathbf{H}_t$.
7. Compute $\widehat{\theta}_o = \operatorname{argmin}_\theta \sum_{s\in\mathcal{T}_o}\ell(\theta, x_s, r_s) + \frac{\lambda}{2}\|\theta\|_2^2$.
8. Else
9. If $\det(\mathbf{H}_t) > 2\det(\mathbf{H}_\tau)$ [`Criterion II`]
10. Set $\tau = t$ and $\widetilde{\theta} \leftarrow \operatorname{argmin}_\theta \frac{\lambda}{2}\|\theta\|_2^2 + \sum_{s\in[t-1]\setminus\mathcal{T}_o}\ell(\theta, x_s, r_s)$
11. $\theta_\tau \leftarrow \texttt{Project}(\widetilde{\theta})$
12. Update $\mathcal{X}_t \leftarrow \mathcal{X}_t \setminus \{x \in \mathcal{X}_t : UCB_o(x) < \max_{z\in\mathcal{X}_t} LCB_o(z)\}$.
13. Select $x_t = \operatorname{argmax}_{x\in\mathcal{X}_t} UCB(x, \mathbf{H}_\tau, \theta_\tau)$ and observe reward r_t.
14. Update $\mathbf{H}_{t+1} \leftarrow \mathbf{H}_t + \frac{\dot{\mu}(x_t^\intercal\widehat{\theta}_w)}{e}x_tx_t^\intercal$.

Theorem: Given $\delta \in (0, 1)$, with probability $\geq 1 - \delta$, the regret of `RS-GLinUCB` satisfies

$$\mathrm{R}_T = O\left(d\sqrt{\sum_{t\in[T]}\dot{\mu}(x_t^{*\intercal}\theta^*)\log(RT/\delta)} + \kappa d^2R^5S^2\log^2(T/\delta)\right)$$

Lemma: `RS-GLinUCB`, during its entire execution, updates its policy at most $O(R^4S^2\,\kappa d^2\,\log^2(T/\delta))$ times.