ACHIEVING LIMITED ADAPTIVITY FOR MULTINOMIAL LOGISTIC BANDITS

Sukruta Prakash Midigeshi [1] Tanmay Goyal [1] Gaurav Sinha [1]

[1]Microsoft Research India

Summary and Contributions

- Batched Algorithm for Multinomial Logistic Bandits: `B-MNL-CB`
  - **Setting**: Stochastic Contextual
  - **Regret**: $\tilde{O}(\sqrt{T})$
  - **Policy Updates**: $\Omega(\log\log T)$
- Rarely-Switching Algorithm for Multinomial Logistic Bandits: `RS-MNL`
  - **Setting**: Adversarial Contextual
  - **Regret**: $\tilde{O}(\sqrt{T})$
  - **Policy Updates**: $\tilde{O}(\log T)$ (improves Sawarni et al. 2024 by logarithmic factors)
  - **Empirical Performance**: Achieves regret better than, or atleast comparable to state-of-the-art logistic and multinomial logistic bandit algorithms.

(Multinomial) Logistic Bandits

- $K+1$ possible outcomes where the probability distribution for the outcomes is given by:

$$\mathbb{P}\{y_t = i \mid \boldsymbol{x}_t, \mathcal{F}_t\} = \begin{cases} z_i(\boldsymbol{x}_t, \boldsymbol{\theta}^\star) = \frac{\exp(\boldsymbol{x}_t^\top \boldsymbol{\theta}_i^\star)}{1+\sum_{j=1}^{K}\exp\left(\boldsymbol{x}_t^\top \boldsymbol{\theta}_j^\star\right)} & 1 \le i \le K, \\ z_0(\boldsymbol{x}_t, \boldsymbol{\theta}^\star) = \frac{1}{1+\sum_{j=1}^{K}\exp\left(\boldsymbol{x}_t^\top \boldsymbol{\theta}_j^\star\right)}, & i = 0, \end{cases}$$

- Hidden Optimal Parameter: $\boldsymbol{\theta}^\star = \left(\boldsymbol{\theta}_1^{\star\top}, \dots, \boldsymbol{\theta}_K^{\star\top}\right)^\top \in \mathbb{R}^{dK}$ such that $\|\boldsymbol{\theta}^\star\| \le S$.
- Known Reward Vector: $\boldsymbol{\rho}$ such that $\|\boldsymbol{\rho}\| \le R$ and $\rho_0 = 0$.
- Link Function $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta}) = (z_1(\boldsymbol{x}, \boldsymbol{\theta}), \dots, z_K(\boldsymbol{x}, \boldsymbol{\theta}))$.
- Gradient of Link Function $\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}) = diag(\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta})) - \boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta})\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta})^\top$.
- Non-linearity parameter κ :

$$\kappa = \sup\left\{\frac{1}{\lambda_{min}(\boldsymbol{A}(\boldsymbol{x}, \boldsymbol{\theta}))} : \boldsymbol{x} \in \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_T, \boldsymbol{\theta} \in \Theta\right\}$$

(Distributional) Optimal Designs

- G-Optimal Design π_G:

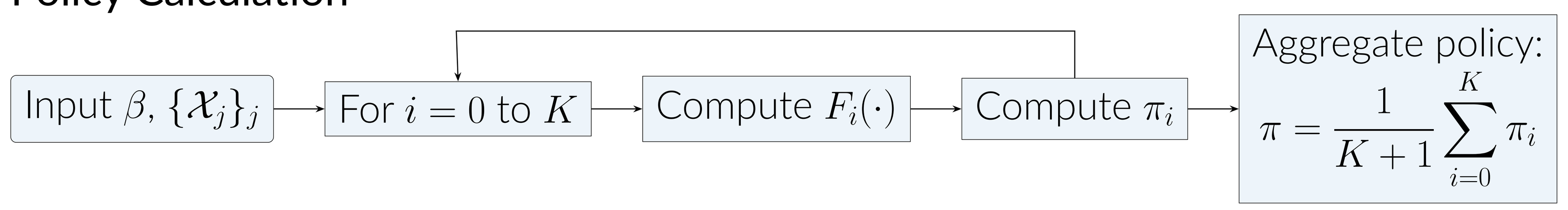
$$\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|^2_{\boldsymbol{V}(\pi_G)^{-1}} \le d, \quad \text{where} \quad \boldsymbol{V}(\pi) = \mathbb{E}_{\boldsymbol{x}\sim\pi}[\boldsymbol{x}\boldsymbol{x}^\top].$$

- Ruan et al. 2021 introduced **distributional optimal designs** to improve this bound

$$\mathbb{P}\left(\mathop{\mathbb{E}}_{\mathcal{X}\sim\mathcal{D}}\left[\max_{\boldsymbol{x}\in\mathcal{X}} \|\boldsymbol{x}\|_{\boldsymbol{V}^{-1}}\right] \le O(\sqrt{d\log d})\right) \ge 1 - \delta(d)$$

Batched Multinomial Contextual Bandit Algorithm: `B-MNL-CB`

1. Stochastic Contextual Setting:
   - Each arm set $\mathcal{X}$ is sampled from the same (unknown) distribution $\mathcal{D}$.
   - Goal: minimize $R(T) = \mathbb{E}\left[\sum_{t=1}^{T}\left[\max_{\boldsymbol{x}\in\mathcal{X}_t} \boldsymbol{\rho}^\top \boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta}^\star) - \boldsymbol{\rho}^\top \boldsymbol{z}(\boldsymbol{x}_t, \boldsymbol{\theta}^\star)\right]\right]$
2. Algorithmic Skeleton:
   For each batch,
   - Successive Elimination: $\mathcal{X} \leftarrow \{\boldsymbol{x} \in \mathcal{X} : \mathrm{UCB}(\boldsymbol{x}) \ge \max_{\boldsymbol{x}\in\mathcal{X}} \mathrm{LCB}(\boldsymbol{x})\}$
   - Learner's Play: $\boldsymbol{x}_t \sim \pi(\mathcal{X})$
   - Policy Updates: Divide batch indices into two disjoint sets C and D:
     - C -> update estimates.
     - D -> parameters.
3. Policy Calculation



4. **Directionally Scaled Sets**: Form K different sets $F_i(\{\mathcal{X}_t\}_{t\in D}, \beta)$ $\forall i \in [K]$.

$$F_i(\{\mathcal{X}_t\}_{t\in D}, \beta) = \left\{\left\{\frac{\boldsymbol{A}(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_\beta)^{\frac{1}{2}}}{\sqrt{B_\beta(\boldsymbol{x})}}\boldsymbol{e}_i \otimes \boldsymbol{x} : \boldsymbol{x} \in \mathcal{X}_t\right\} : t \in D\right\}.$$

5. **Regret Guarantee**: With high probability, at the end of T rounds, the regret incurred by `B-MNL-CB` is bounded by

$$R(T) = \tilde{O}\left(RS^{5/4}K^{5/2}d\sqrt{T}\right)$$

Proof Sketch

- Using ideas from Zhang et al. 2023, we can decompose the regret as

$$R(T) \le 4\sum \mathbb{E}\left[\max_{\boldsymbol{x}\in\mathcal{X}} \epsilon_1(\boldsymbol{x}) + \max_{\boldsymbol{x}\in\mathcal{X}} \epsilon_2(\boldsymbol{x})\right].$$

- We define $\tilde{\boldsymbol{X}} = \frac{\boldsymbol{A}(\boldsymbol{x},\boldsymbol{\theta})^{1/2}}{B(\boldsymbol{x})} \otimes \boldsymbol{x}$ such that

$$\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top = \sum_{i=1}^{K}\left\{\frac{\boldsymbol{A}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})^{\frac{1}{2}}}{\sqrt{B(\boldsymbol{x})}}\boldsymbol{e}_i \otimes \boldsymbol{x}\right\}\left\{\frac{\boldsymbol{A}(\boldsymbol{x}, \hat{\boldsymbol{\theta}})^{\frac{1}{2}}}{\sqrt{B(\boldsymbol{x})}}\boldsymbol{e}_i \otimes \boldsymbol{x}\right\}^T.$$

We use this to combine the optimal designs learned for the K different sets.

Rarely-Switching Multinomial Bandit Algorithm: `RS-MNL`

1. Adversarial Contextual Setting:
   - No assumptions on the arm sets.
   - Goal: minimize $R(T) = \sum_{t=1}^{T}\left[\max_{\boldsymbol{x}\in\mathcal{X}_t} \boldsymbol{\rho}^\top \boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta}^\star) - \boldsymbol{\rho}^\top \boldsymbol{z}(\boldsymbol{x}_t, \boldsymbol{\theta}^\star)\right]$.
2. Algorithmic Skeleton:
   At each round t
   - Check if det $\boldsymbol{H}_t \ge 2$ det $\boldsymbol{H}_\tau$ (τ : previous switching round)
   - If true, update estimate of parameters.
   - Choose arm with maximum UCB to play.
3. Improvement over Sawarni et al. 2024:
   - Gets rid of a warm-up switching criterion.
   - Number of switches: $\tilde{O}(\log^2 T) \leftarrow \tilde{O}(\log T)$.
4. **Regret Guarantee**: With high probability, after T rounds, the regret of `RS-MNL` can be bounded by:

$$R_T \le \tilde{O}\left(RK^{3/2}S^{5/4}d\sqrt{T}\right).$$

5. Proof Sketch:
   - Using ideas from Zhang et al. 2023, we can decompose the regret as

$$R(T) \le \sum_{t=1}^{T} 2\epsilon_1(\boldsymbol{x}_t) + 2\epsilon_2(\boldsymbol{x}_t)$$

   where $\epsilon_1(\boldsymbol{x})$ and $\epsilon(\boldsymbol{x})$ are the error terms that appear in the UCB expression.
   - Bound $\epsilon_1(\boldsymbol{x})$ and $\epsilon_2(\boldsymbol{x})$ similar to `B-MNL-CB` (using directionally scaled sets).

Experiments

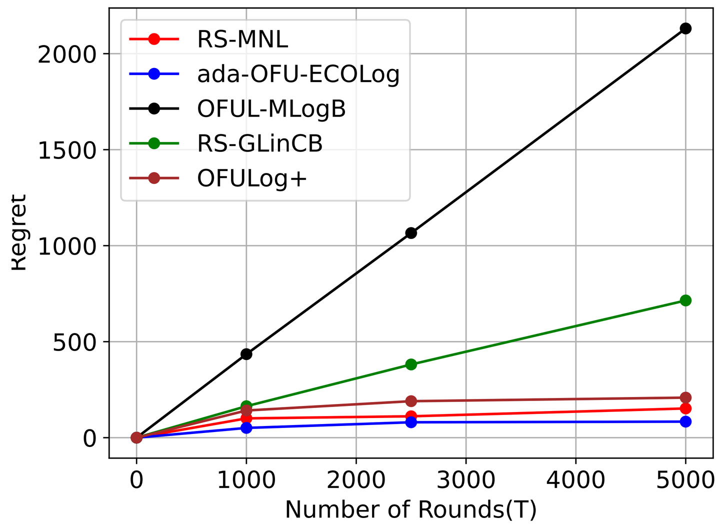
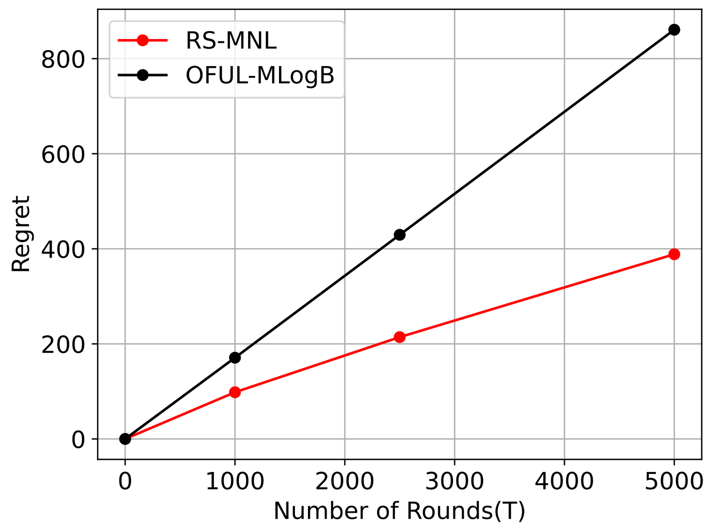


Figure 1. Regret vs. T: Logistic Setting



Regret vs. T: $K = 3$

Correspondence to: t-smidigeshi@microsoft.com, t-tangoyal@microsoft.com

arXiv:2508.03072