uai2025

Efficient Algorithms for Logistic Slate Contextual Bandits with Bandit Feedback

Tanmay Goyal[1] Gaurav Sinha [1]

[1]Microsoft Research India

Microsoft

Problem Statement

- Setup: A *slate* consists of N *slots*. At each round $t \in [T]$, the learner receives N sets of items $\mathcal{X}_t^i$ such that $|\mathcal{X}_t^i| = K$.
- Learner's play: The learner chooses an item for each slot i, denoted by $\boldsymbol{x}_t^i$, and plays the slate $\boldsymbol{x}_t = (\boldsymbol{x}_t^1, \dots, \boldsymbol{x}_t^N)$.
- Learner's feedback: The learner receives reward $y_t(\boldsymbol{x}_t)$ such that $\Pr[y_t(\boldsymbol{x}_t) = 1 \mid \boldsymbol{x}_t] = \frac{\exp(\boldsymbol{x}_t^\top \boldsymbol{\theta}^\star)}{1+\exp(\boldsymbol{x}_t^\top \boldsymbol{\theta}^\star)}$ where $\boldsymbol{\theta}^\star \in \mathbb{R}^{Nd}$ is unknown to the learner.
- Learner's goal: Minimize *expected cumulative regret* $R_T = \sum_{t=1}^T \max_{\boldsymbol{x} \in \mathcal{X}_t} \mathbb{E}[y_t(\boldsymbol{x})] - \mathbb{E}[\boldsymbol{y}_t(\boldsymbol{x}_t)]$.
- We wish to develop algorithms with optimal regret guarantees that can achieve (1) Computational Efficiency, and can handle (2) Bandit feedback.
  - Computational Efficiency: Should not enumerate over candidate set with size $2^{\Omega(N)}$.
  - Bandit feedback: A single reward/feedback for the slate played.

Contributions

Algorithms

1. `Slate-GLM-OFU`: based on OFU paradigm. Incurs $\tilde{O}(Nd\sqrt{T})$ regret with high probability.
2. `SLate-GLM-TS`: based on Thompson Sampling principle.
3. `Slate-GLM-TS-Fixed`: Fixed-Arm version of `Slate-GLM-TS`. Incurs $\tilde{O}((Nd)^{3/2}\sqrt{T})$ regret with high probability.

Experiments

1. `Slate-GLM-OFU` incurs lowest regret and `Slate-GLM-TS` is competitive with baselines.
2. Exponential decrease in run time compared to baselines (attributed to *pulling* time).
3. Prompt Tuning: Choose in-context examples for language models for *SST2* and *Yelp Review*. Achieve $\sim 80\%$ accuracy.

Relevant Techniques from Prior Works

1. Online-Proxy Hessian Matrix [Faury et al. 2022]: Replace the optimal Hessian matrix $\boldsymbol{H}_t = \lambda \boldsymbol{I} + \sum_{s \in [t]} \dot{\mu}(\boldsymbol{x}_s^\top \boldsymbol{\theta}^\star) \boldsymbol{x}_s \boldsymbol{x}_s^\top$ with an online proxy matrix, $\boldsymbol{W}_t = \lambda \boldsymbol{I} + \sum_{s \in [t]} \dot{\mu}(\boldsymbol{x}_s^\top \boldsymbol{\theta}_{s+1}) \boldsymbol{x}_s \boldsymbol{x}_s^\top$.
2. Data-driven condition [Faury et al. 2022]: An adaptive condition $\dot{\mu}(\boldsymbol{x}_s^\top \bar{\boldsymbol{\theta}}_s) \leq 2\dot{\mu}(\boldsymbol{x}_s^\top \boldsymbol{\theta}_t^u)$ that helps maintain control over the diameter of the permissible set of parameters.
3. Distribution for Thompson Sampling [Abeille et al. 2017]: A distribution $\mathcal{D}$ which enables the noise $\boldsymbol{\eta}$ to explore enough (*anti-concentration*) but not too much (*concentration*).

Algorithm and Our Approach

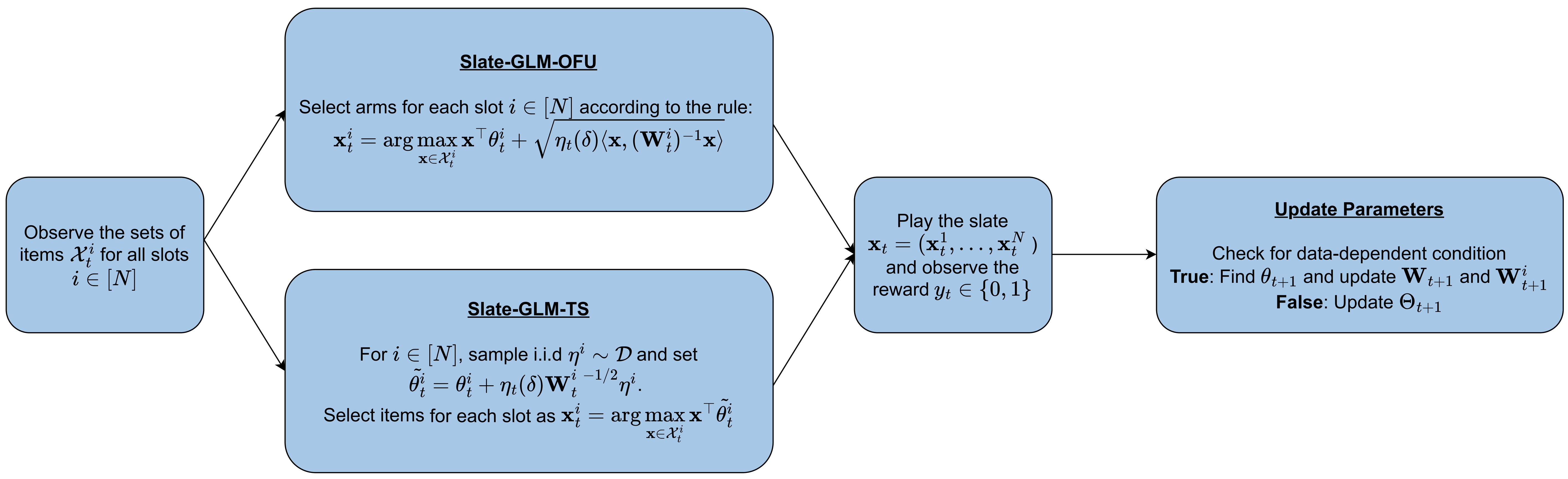

Figure 1. Skeleton of `Slate-GLM-OFU` and `Slate-GLM-TS`

1. To enable independent selection of items for each slot, we require guarantees with respect to the *slot-level online proxy Hessians* $\boldsymbol{W}^i\ \forall i \in [N]$. The inner products are manageable.
2. Note that the proxy matrix $\boldsymbol{W}$ is of the form $\boldsymbol{x}\boldsymbol{x}^\top$:

$$\boldsymbol{x}\boldsymbol{x}^\top = \begin{bmatrix} \boldsymbol{x}^1{\boldsymbol{x}^1}^\top & \boldsymbol{x}^1{\boldsymbol{x}^2}^\top & \dots & \boldsymbol{x}^1{\boldsymbol{x}^N}^\top \\ \boldsymbol{x}^2{\boldsymbol{x}^1}^\top & \boldsymbol{x}^2{\boldsymbol{x}^2}^\top & \dots & \boldsymbol{x}^2{\boldsymbol{x}^N}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}^N{\boldsymbol{x}^1}^\top & \boldsymbol{x}^N{\boldsymbol{x}^2}^\top & \dots & \boldsymbol{x}^N{\boldsymbol{x}^N}^\top \end{bmatrix} = \mathrm{diag}(\boldsymbol{x}^1{\boldsymbol{x}^1}^\top, \dots, \boldsymbol{x}^N{\boldsymbol{x}^N}^\top) + \boldsymbol{A}$$

Achieving an equivalence between $\boldsymbol{W}$ and $\boldsymbol{W}^i$ requires handling the matrix consisting of the off-diagonal terms, denoted by $\boldsymbol{A}$.

3. Using the diversity conditions $\left(\mathbb{E}[\boldsymbol{x}_t^i \mid \mathcal{F}_t] = 0 \text{ and } \mathbb{E}[\boldsymbol{x}_t^i {\boldsymbol{x}_t^i}^\top \mid \mathcal{F}_t] \succeq \rho\kappa \boldsymbol{I}\right)$ enables us to show that $\boldsymbol{A} \preceq C \cdot \mathrm{diag}(\boldsymbol{x}^1{\boldsymbol{x}^1}^\top, \dots, \boldsymbol{x}^N{\boldsymbol{x}^N}^\top)$.
4. Hence, we end up with a *multiplicative equivalence* between $\boldsymbol{W}$ and $\boldsymbol{W}^i\ \forall i \in [N]$, i.e, $(1-C) \cdot \mathrm{diag}(\boldsymbol{W}^1, \dots, \boldsymbol{W}^N) \preceq \boldsymbol{W} \preceq (1+C) \cdot \mathrm{diag}(\boldsymbol{W}^1, \dots, \boldsymbol{W}^N)$.
5. We thus have

$$\|\boldsymbol{x}\|_{\boldsymbol{W}^{-1}} = \left\| \sum_{i=1}^N \boldsymbol{x}^i \otimes \boldsymbol{e}_i \right\|_{\boldsymbol{W}^{-1}} \leq \frac{1}{1-C} \sum_{i=1}^N \|\boldsymbol{x}^i \otimes \boldsymbol{e}_i\|_{(\mathrm{diag}(\boldsymbol{W}^1, \dots, \boldsymbol{W}^N))^{-1}} = \frac{1}{1-C} \sum_{i=1}^N \|\boldsymbol{x}^i\|_{(\boldsymbol{W}^i)^{-1}}$$

allowing us to select items for each slot independent of the others.

Experiments

Finite Contexts

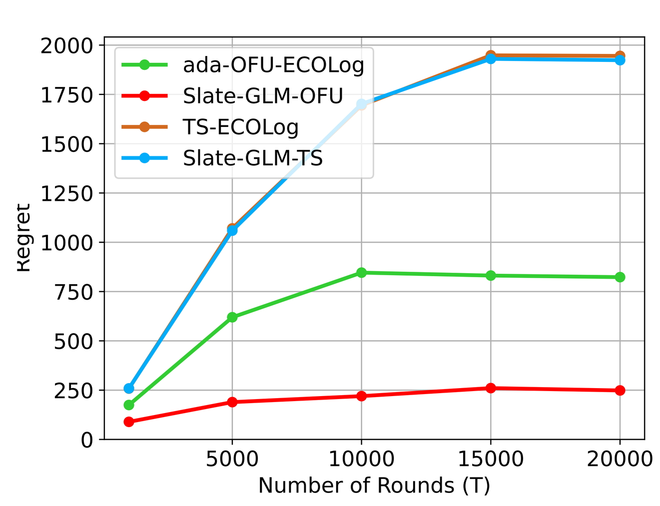

Figure 2. $d = 5, K = 5, N = 3$

Infinite Contexts

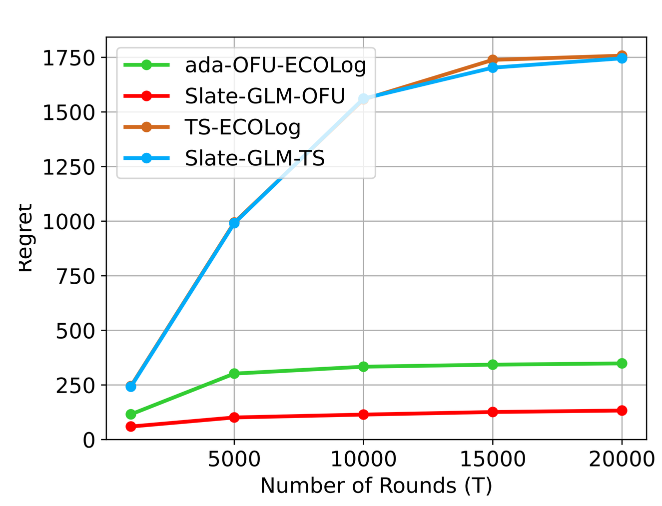

Figure 3. $d = 5, K = 5, N = 3$

Non-Contextual

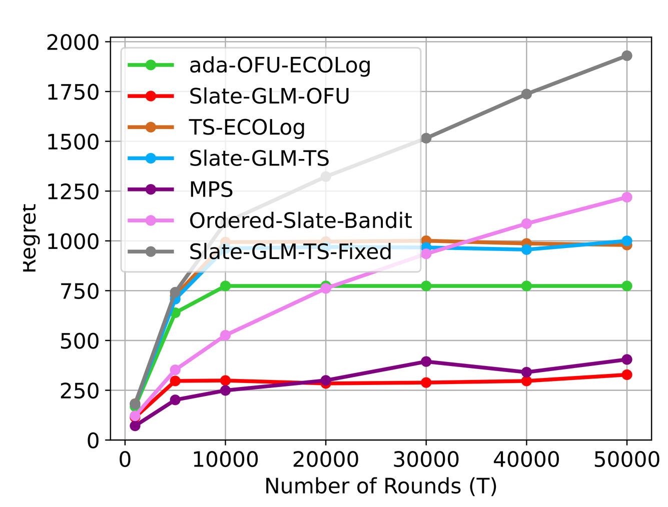

Figure 4. $d = 5, K = 5, N = 3$

(Average) Running Times

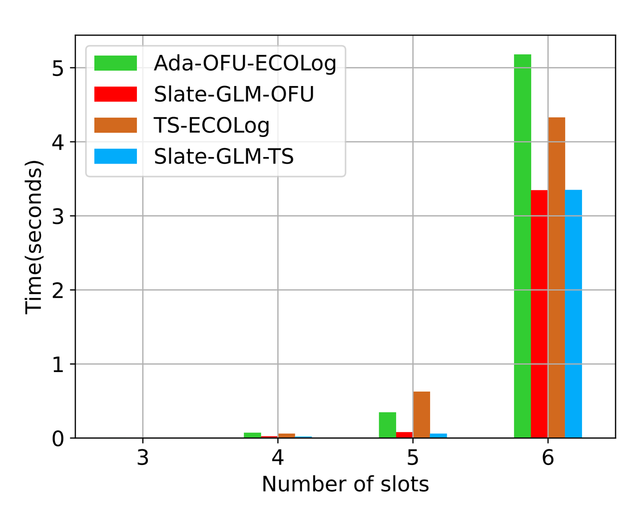

Figure 5. $d = 5, K = 7$

Prompt Tuning

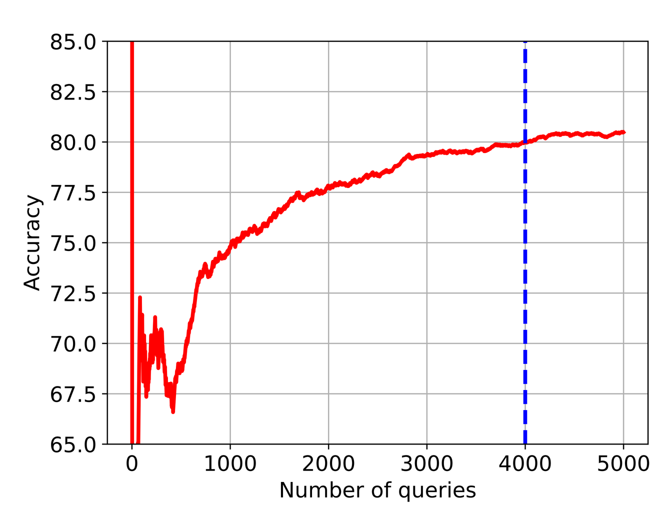

Figure 6. Accuracy v/s. T

arXiv:2506.13163