Achieving Limited Adaptivity for Multinomial Logistic Bandits

Sukruta Prakash Midigeshi, Tanmay Goyal, Gaurav Sinha

Microsoft Research India

Multinomial Logistic Bandits

- Users are offered several (≥ 2) options rather than a single option.
- Several applications, including:
  - E-Commerce
  - Over-the-Top platforms
  - News Platforms
  - Recommendation Systems
- It is unclear if algorithms designed for GLM settings would work for the Multinomial Logistic setting.

Formalism of Notations

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

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

- Hidden Optimal Parameter: $\boldsymbol{\theta}^\star = \left(\boldsymbol{\theta}_1^{\star\top}, \ldots, \boldsymbol{\theta}_K^{\star\top}\right)^\top \in \mathbb{R}^{dK}$ such that $\|\boldsymbol{\theta}^\star\| \leq S$.
- Known Reward Vector: $\boldsymbol{\rho}$ such that $\|\boldsymbol{\rho}\| \leq R$ and $\rho_0 = 0$.
- Link Function $\boldsymbol{z}(\boldsymbol{x}, \boldsymbol{\theta}) = (z_1(\boldsymbol{x}, \boldsymbol{\theta}), \ldots, z_K(\boldsymbol{x}, \boldsymbol{\theta}))$.

Formalism of Notations

- 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 that grows exponentially with the size of parameter sets:

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

- Hessian Matrix: $\boldsymbol{H}_\beta = \lambda \boldsymbol{I} + \sum_{t \in \mathcal{T}_\beta} \boldsymbol{A}(\boldsymbol{x}_t, \boldsymbol{\theta}^\star) \otimes \boldsymbol{x}_t \boldsymbol{x}_t^\top$

Motivation: Limited Adaptivity

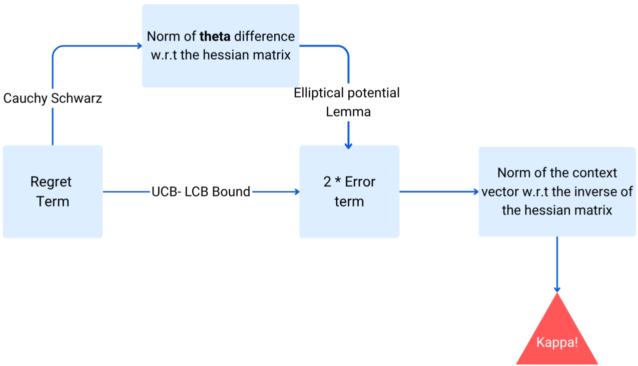
- Practical and computational limitations
- Requirement for parallelism and limited policy updates.
- [Gao et al., 2019] showed that $\Omega(\log \log T)$ policy updates are sufficient to obtain an optimal regret bound of $O(\sqrt{T})$.

Comparision to Prior Works

Type	Work	Update Type	Number of Updates	Regret
Linear	[Ruan et al., 2021]	Batched	$\Omega(\log\log T)$	$\tilde{O}(d\sqrt{T})$
Logistic & GLMs	[Filippi et al., 2010]	×	T	$\tilde{O}(\kappa d\sqrt{T})$
	[Faury et al., 2020]	×	T	$\tilde{O}(d\sqrt{T})$
	[Faury et al., 2022]	×	T	$\tilde{O}(d\sqrt{T})$
	[Sawarni et al., 2024]	Batched	$O(\log\log T)$	$\tilde{O}(d\sqrt{T})$
	[Sawarni et al., 2024]	Rarely-Switching	$\tilde{O}(\log^2 T)$	$\tilde{O}(d\sqrt{T})$
Multinomial Logistic (MNL)	[Amani and Thrampoulidis, 2021]	×	T	$\tilde{O}(Kd\sqrt{\kappa T})$
	[Zhang and Sugiyama, 2023]	×	T	$\tilde{O}(Kd\sqrt{T})$
	Ours	**Batched**	$\boldsymbol{O(\log\log T)}$	$\boldsymbol{\tilde{O}(K^{5/2}d\sqrt{T})}$
	Ours	**Rarely-Switching**	$\boldsymbol{\tilde{O}(\log T)}$	$\boldsymbol{\tilde{O}(K^{3/2}d\sqrt{T})}$

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

The Issue


Norm of **theta** difference w.r.t the hessian matrix
Cauchy Schwarz
Elliptical potential Lemma
Regret Term
UCB- LCB Bound
2 * Error term
Norm of the context vector w.r.t the inverse of the hessian matrix
Kappa!

Background - Optimal Designs

- **G-Optimal Design** π_G:

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

- [Ruan et al., 2021] introduced **distributional optimal designs**

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

Simulating a matrix distributional optimal design

Crux of the problem :

$$\mathop{\mathbb{E}}_{\mathcal{X}\sim\mathcal{D}_{\beta+1}}\left[\max_{\boldsymbol{x}\in\mathcal{X}}\left|\left|\tilde{\boldsymbol{X}}_{\beta}^{\top}\boldsymbol{H}_{\beta}^{-1/2}\right|\right|_2\right] \text{ where } \tilde{\boldsymbol{X}}_{\beta} = \frac{\boldsymbol{A}(\boldsymbol{x},\hat{\boldsymbol{\theta}}_{\beta})^{\frac{1}{2}}}{\sqrt{B_{\beta}(\boldsymbol{x})}} \otimes \boldsymbol{x}$$

Simulating a matrix distributional optimal design

Crux of the problem :

$$\mathop{\mathbb{E}}_{\mathcal{X}\sim\mathcal{D}_{\beta+1}}\left[\max_{\boldsymbol{x}\in\mathcal{X}}\left|\left|\tilde{\boldsymbol{X}}_\beta^\top \boldsymbol{H}_\beta^{-1/2}\right|\right|_2\right] \text{ where } \tilde{\boldsymbol{X}}_\beta = \frac{\boldsymbol{A}(\boldsymbol{x},\hat{\boldsymbol{\theta}}_\beta)^{\frac{1}{2}}}{\sqrt{B_\beta(\boldsymbol{x})}} \otimes \boldsymbol{x}$$

$$\tilde{\boldsymbol{X}}_\beta \tilde{\boldsymbol{X}}_\beta^\top = \sum_{i=1}^{K} \tilde{\boldsymbol{x}}_\beta^{(i)} \tilde{\boldsymbol{x}}_\beta^{(i)\,\top}$$

$$\tilde{\boldsymbol{x}}_\beta^{(i)} = \frac{\boldsymbol{A}(\boldsymbol{x},\hat{\boldsymbol{\theta}}_\beta)^{\frac{1}{2}}}{\sqrt{B_\beta(\boldsymbol{x})}} \boldsymbol{e}_i \otimes \boldsymbol{x}$$

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

Input and initialize the parameters
for batches $\beta \in [M]$ **do**
  for each round $t \in \mathcal{T}_\beta$ **do**
    for $j = 1$ to $\beta - 1$ **do**
      Update arm set $\mathcal{X}_t \leftarrow \mathrm{UL}_j(\mathcal{X}_t)$
    end for
    Sample $\boldsymbol{x}_t \sim \pi_{\beta-1}(\mathcal{X}_t)$ and obtain the corresponding reward.
  end for
  Divide $\mathcal{T}_\beta$ into two sets C and D of equal sizes.
  Compute $\hat{\boldsymbol{\theta}}_\beta \leftarrow \arg\min \sum_{s \in C} \ell(\boldsymbol{\theta}, \boldsymbol{x}_s, y_s)$, $\boldsymbol{H}_\beta = \lambda \boldsymbol{I} + \sum_{t \in C} \frac{\boldsymbol{A}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_\beta) \otimes \boldsymbol{x}_t \boldsymbol{x}_t^\top}{B_\beta(\boldsymbol{x}_t)}$,
  and π_β using Algorithm 2 with the inputs $(\beta, \{\mathcal{X}_t\}_{t \in D})$
end for

The Algorithm

Algorithm Distributional Optimal Design for MNL bandits

1: **Input** Batch β and collection of arm sets $\{\mathcal{X}_j\}_j$
2: Create the sets $\{F_i(\{\mathcal{X}_j\}_j, \beta)\}_{i=1}^K$.
3: Compute the distributional optimal design policy π_i for each of the sets $F_i(\{\mathcal{X}_j\}_j, \beta)$.
4: Compute the distributional optimal design policy π_0 for the set $\{\mathcal{X}_j\}_j$.
5: **Return** $\pi = \frac{1}{K+1} \sum_{i=0}^{K} \pi_i$

The Algorithm

Algorithm Distributional Optimal Design for MNL bandits

1: **Input** Batch β and collection of arm sets $\{\mathcal{X}_j\}_j$
2: Create the sets $\{F_i(\{\mathcal{X}_j\}_j, \beta)\}_{i=1}^K$.
3: Compute the distributional optimal design policy π_i for each of the sets $F_i(\{\mathcal{X}_j\}_j, \beta)$.
4: Compute the distributional optimal design policy π_0 for the set $\{\mathcal{X}_j\}_j$.
5: **Return** $\pi = \frac{1}{K+1} \sum_{i=0}^{K} \pi_i$

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

Rarely Switching Contextual Bandit Algorithm: `RS-MNL`

Algorithm

Algorithm RS-MNL

Input ρ, S, T and initialize the parameters λ, γ, τ.

for $t = 1, \ldots, T$ **do**

  Observe arm set $\mathcal{X}_t$

  if $\det(\boldsymbol{H}_t) > 2\det(\boldsymbol{H}_\tau)$ **then**

    Set $\tau = t$

    Update $\hat{\boldsymbol{\theta}}_\tau \leftarrow \arg\min_{\boldsymbol{\theta}} \sum_{s \in [t-1]} \ell(\boldsymbol{\theta}, \boldsymbol{x}_s, y_s)$.

    Update $\boldsymbol{H}_t = \sum_{s \in [t-1]} \frac{\boldsymbol{A}(\boldsymbol{x}_s, \hat{\boldsymbol{\theta}}\tau)}{B_\tau(\boldsymbol{x}_s)} \otimes \boldsymbol{x}_s \boldsymbol{x}_s^\top + \lambda \boldsymbol{I}_{Kd}$

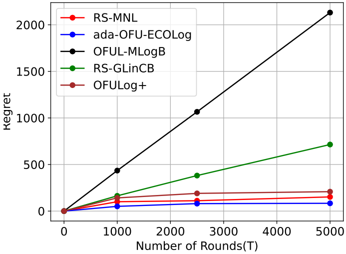
  end if

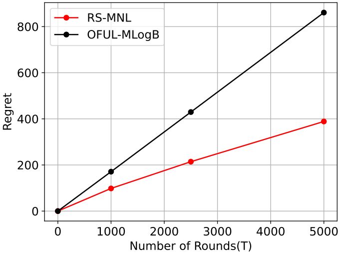
  Select $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x} \in \mathcal{X}_t} \mathrm{UCB}(t, \tau, \boldsymbol{x})$ and observe y_t.

  Update $\boldsymbol{H}_{t+1} \leftarrow \boldsymbol{H}_t + \frac{\boldsymbol{A}(\boldsymbol{x}_t, \hat{\boldsymbol{\theta}}_\tau)}{B_\tau(\boldsymbol{x}_t)} \otimes \boldsymbol{x}_t \boldsymbol{x}_t^\top$

end for

Experimental Results





Conclusion

- Batched Algorithm for MNL Bandits: `B-MNL-CB`
  - Achieves a $\kappa-$independent regret bound of $\tilde{O}(K^{5/2}d\sqrt{T})$.
  - $O(\log \log T)$ policy updates.
- Rarely-Switching algorithm for MNL Bandits: `RS-MNL`
  - Achieves a $\kappa-$independent regret bound of $\tilde{O}(K^{3/2}d\sqrt{T})$.
  - $\tilde{O}(\log T)$ policy updates, which is an improvement over [Sawarni et al., 2024]
  - Improves the per-round time complexity compared to [Sawarni et al., 2024] by re-introducing the UCB selection rule from [Abbasi-Yadkori et al., 2011].

Thank you

References I

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011).
Improved algorithms for linear stochastic bandits.
In *Advances in Neural Information Processing Systems 24 (NeurIPS)*, pages 2312–2320.

Amani, S. and Thrampoulidis, C. (2021).
UCB-based algorithms for multinomial logistic regression bandits.
In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Faury, L., Abeille, M., Calauzenes, C., and Fercoq, O. (2020).
Improved optimistic algorithms for logistic bandits.
In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060. PMLR.

References II

Faury, L., Abeille, M., Jun, K.-S., and Calauzenes, C. (2022).
Jointly efficient and optimal algorithms for logistic bandits.
In Camps-Valls, G., Ruiz, F. J. R., and Valera, I., editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 546–580. PMLR.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010).
Parametric bandits: The generalized linear case.
In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Gao, Z., Han, Y., Ren, Z., and Zhou, Z. (2019).
Batched multi-armed bandits problem.
In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 503–513. Curran Associates, Inc.

References III

Ruan, Y., Yang, J., and Zhou, Y. (2021).
Linear bandits with limited adaptivity and learning distributional optimal design.
In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, page 74–87, New York, NY, USA. Association for Computing Machinery.

Sawarni, A., Das, N., Barman, S., and Sinha, G. (2024).
Generalized linear bandits with limited adaptivity.
In Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., and Zhang, C., editors, *Advances in Neural Information Processing Systems*, volume 37, pages 8329–8369. Curran Associates, Inc.

Zhang, Y.-J. and Sugiyama, M. (2023).
Online (multinomial) logistic bandit: Improved regret and constant computation cost.
In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29741–29782. Curran Associates, Inc.

Appendix

Number of Switches in RS-MNL

$$\frac{\det\ \boldsymbol{H}_t(\hat{\boldsymbol{\theta}}_{\tau_{m-1}})}{\det\ \boldsymbol{H}_{\tau_0}(\boldsymbol{\theta})} = \frac{\det\ \boldsymbol{H}_{\tau_m}(\hat{\boldsymbol{\theta}}_{\tau_{m-1}})}{\det\ \boldsymbol{H}_{\tau_{m-1}}(\hat{\boldsymbol{\theta}}_{\tau_{m-1}})} \times \dots \times \frac{\det\ \boldsymbol{H}_{\tau_1}(\hat{\boldsymbol{\theta}}_{\tau_0})}{\det\ \boldsymbol{H}_{\tau_0}(\boldsymbol{\theta})} \geq 2^m$$

$$\det\ \boldsymbol{H}_t(\hat{\boldsymbol{\theta}}_{\tau_{m-1}}) \geq 2^m \det\ (\lambda \boldsymbol{I}_{Kd}) = 2^m \lambda^{Kd}$$

$$\det\ \boldsymbol{H}_t(\hat{\boldsymbol{\theta}}_{\tau_{m-1}}) \leq \left(\frac{\mathrm{trace}\ \boldsymbol{H}_t(\hat{\boldsymbol{\theta}}_{\tau_{m-1}})}{Kd}\right)^{Kd} \leq \left(\lambda + \frac{t}{d}\right)^{Kd}$$

Combining these facts shows that $m \approx Kd \log(T)$.

Elliptical Potential Lemma

Let $\{\boldsymbol{x}_s\}_{s=1}^{t}$ represent a set of vectors in $\mathbb{R}^d$ and let $||\boldsymbol{x}_s||_2 \leq L$. Let $\boldsymbol{V}_s = \lambda \boldsymbol{I}_{d\times d} + \sum_{m=1}^{s-1} x_m x_m^\top$. Then, for $\lambda \geq 1$

$$\sum_{s=1}^{t} ||\boldsymbol{x}_s||^2_{\boldsymbol{V}_s^{-1}} \leq 2d \log\left(1 + \frac{tL^2}{\lambda d}\right) \leq 4d \log(tL^2)$$