Single Index Bandits: Generalized Linear Contextual Bandits with Unknown Reward Functions

Yue Kang* Mingshuo Liu Bongsoo Yi Jing Lyu Zhi Zhang Doudou Zhou Yao Li

Microsoft; UC Davis; UNC Chapel Hill; UCLA; NUS

Motivation & Problem Setting

- Contextual bandits are widely used in recommendation, clinical trials, hyperparameter tuning, etc.
- Generalized Linear Bandits (GLBs) assume the reward model:

$$\mathbb{E}[y_t \mid x_t] = f(x_t^\top \theta_*)$$

  where the link function $f(\cdot)$ is *known*.
- In practice, the parametric form of f is often unknown; misspecification can cause severe degradation (linear regret to all existing GLBs).

Setting

At round t, observe an arm set $\mathcal{X}_t = \{x_{t,a} \in \mathbb{R}^d : a \in [K]\}$, pick x_t, and observe

$$y_t = f(x_t^\top \theta_*) + \eta_t,$$

with unknown θ_* and **unknown** (differentiable) $f(\cdot)$.

Key Idea: Stein's Method Estimator

- Score function of context distribution $p(\cdot)$:

$$S(x) = -\nabla_x \log p(x) \in \mathbb{R}^d.$$

Estimator

With exploration samples $\{(x_i, y_i)\}_{i=1}^n$,

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i \, S(x_i).$$

- No iterative optimization; $O(nd)$ time and $O(d)$ memory.
- Only requires finite noise variance.

Algorithms (Monotone Case)

STOR (warm-up)

Explore-then-Commit using Stein estimator for θ_*.
Regret: $\tilde{O}(T^{2/3})$.

ESTOR (main method)

- Epoch-based schedule: alternate **exploration** and **exploitation**.
- In exploration, collect samples from previous epoch and update $\widehat{\theta}$ via Stein estimator.
- In exploitation, pick the arm greedily by

$$x_t = \arg\max_{x \in \mathcal{X}_t} x^\top \widehat{\theta}.$$

- Carefully growing epoch lengths $\Rightarrow$ nearly-optimal regret.

Regret: $\tilde{O}(\sqrt{T})$.

Algorithms (Extensions)

Sparse high-dimensional monotone SIB

- Assume θ_* is s-sparse.
- Stein estimator extends to sparse regime; regret bounds depend on s (not d).

Regret: STOR $\tilde{O}(T^{2/3})$ and ESTOR $\tilde{O}(\sqrt{T})$ (with dimension dependence via sparsity s).

General (possibly non-monotone) f

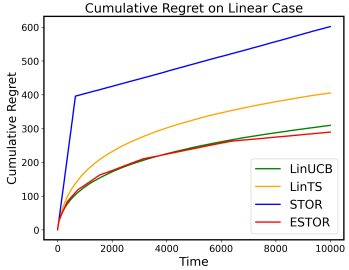
- **GSTOR**: double-explore-then-exploit.
- Learn θ_* via Stein estimator + learn f via kernel regression on the estimated index.

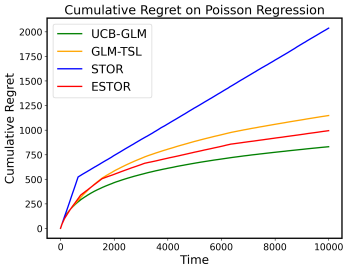
Regret: $\tilde{O}(T^{3/4})$ under Gaussian design.

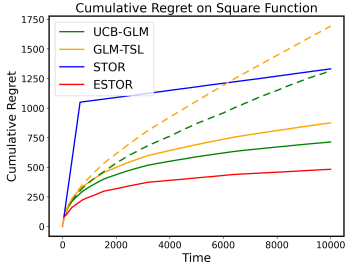
Experiments: Setup & Message

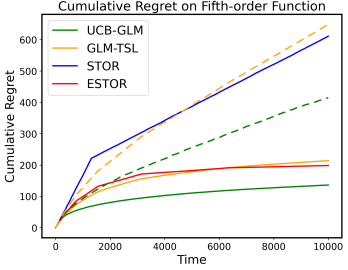
- Horizon: $T = 10{,}000$; average over 20 runs.
- Increasing link functions:
  - (1) Linear: $f(x) = x$ (Gaussian noise)
  - (2) Poisson: $f(x) = \exp(x)$ (Poisson outcomes)
  - (3) Square: $f(x) = \mathrm{sign}(x)\, x^2 + 2x$ (Gaussian noise)
  - (4) Fifth: $f(x) = x^5$ (Gaussian noise)
- Baselines: LinUCB/LinTS (linear); UCB-GLM/GLM-TSL (GLB).
- Conclusions: **Our methods are robust to misspecification**; GLB baselines can break down when fitted with a wrong f.
- More experimental settings with real data are in our paper.

Simulation Results









Effieiency

- In the paper's simulations, STOR/ESTOR are **hundreds to thousands of times faster** than UCB-GLM / GLM-TSL.
- Reason: our updates avoid per-round maximum-likelihood / online Newton steps.
- Practical implication: scalable for large T and higher dimensions.

Conclusion & Future Steps

Conclusion

- Single Index Bandits: GLB with **unknown** reward function $f(\cdot)$.
- Our methods enable efficient learning with strong regret guarantees.
- Robust to link misspecification; competitive in simulations and real data.

Future steps

- Sharper regret for general (non-monotone) f (lower bound).
- Relax distribution assumptions (beyond i.i.d. contexts / Gaussian design).

Thank you!