ADAPTIVE REGRET FOR BANDITS MADE POSSIBLE: TWO QUERIES SUFFICE

Zhou Lu [1], Richard Zhang [2], Xinyi Chen [1,2], Fred Zhang [3], David Woodruff [2,4], Elad Hazan [1,2]

{ [1] Princeton University, [2] Google, [3] UC Berkeley, [4] Carnegie Mellon University}

Introduction

Background/Related Works

Motivation:

- **Fast changing states pose a significant challenge to online optimization**
- **Want to perform rapid adaptation under limited observation**
- **The classic metric of regret incentivizes static behavior and is not correct in changing environments.**
- **Previous works proposed the notion of (strongly) adaptive regret, defined as the maximum regret over any continuous interval in time.**

Can we efficiently **learn the best learning rate/optimizer that adapts to changing environments** (i.e. learn a schedule)**?**

$$\text{SA-regret}(\mathcal{A}, I) = \max_{s-j=I} \left[\sum_{t=j}^{s} \ell_t^\top e_{x_t} - \min_i \sum_{t=j}^{s} \ell_t^\top e_i \right].$$

Table 1: Adaptive regret bounds and query efficiency in the adversarial multi-armed bandits setting.

Algorithm	Adaptive regret bound	Number of queries
FLH Hazan & Seshadhri (2009)	$\sqrt{nT}$	$O(\log T)$
SAOL Daniely et al. (2015)	$\sqrt{nI}\log T$	$O(\log T)$
EFLH Lu & Hazan (2023)	$I^{\frac{1}{2}+\varepsilon} \cdot \sqrt{n \log T}$	$O\left(\frac{\log \log T}{\varepsilon}\right)$
This paper (Theorem 1)	$\sqrt{nI \log n} \cdot \log^{1.5} T$	2

Results

Two Queries Suffice

- In the bandit setting, there is a **Omega(I) lower bound** for SA-regret!
  - Consider two arms: each interval with sublinear regret guarantee requires trying both arms.
- However, with **two queries per round**, we can achieve an optimal sqrt(I) bandit algorithm.
- Crucially decouples the action and observation distribution, unlike EXP4 or previous algorithms
- For multi-arm bandit, we achieve **optimal dependence on the number of arms.**

Theorem 1 (Adaptive regret minimization for multi-armed bandits). *For the multi-armed bandits problem with n arms and T rounds, Algorithm 1 achieves an expected adaptive regret bound of $O\left(\sqrt{nI \log n} \log^{1.5} T\right)$, using two queries per round.*

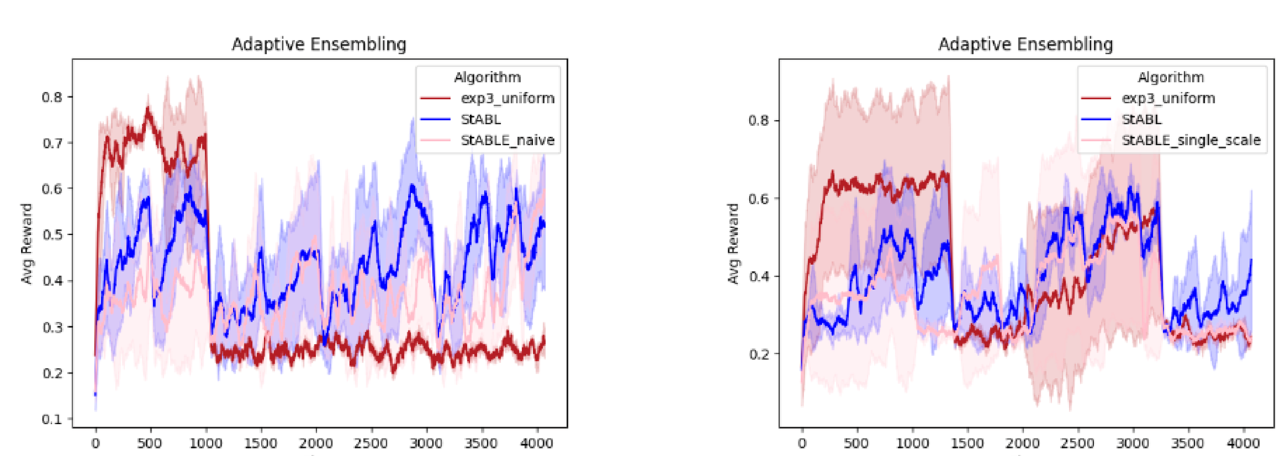

Figure 1: Comparison plots of the algorithm rewards in the learning with expert advice setting.

Method

Main Algorithm (StABL)

Algorithm 1 Strongly Adaptive Bandit Learner (StABL)

Input: general EXP3 algorithm $\mathcal{A}$ and horizon T.

Construct geometric interval set $S = \{[s2^k, (s+1)2^k - 1] \mid 0 \le k \le \log T, s \in \mathbb{N}^+\}$.

Construct $B = \log T$ independent instances of EXP3 algorithm $\mathcal{A}_k$, where $\mathcal{A}_k$ optimizes each $\{I \in S | 2^k = |I|\}$ one after another since they don't overlap.

Denote $w_t(k)$ to be the weight assigned to $\mathcal{A}_k$ at time t by the meta-algorithm.

Denote $v(t,k) \in \mathbb{R}^n$ to be the distribution over arms by $\mathcal{A}_k$ at time t, and $v(t,k)_i$ to be the probability of sampling arm i by $\mathcal{A}_k$ at time t.

Define $\eta_k = \min\left\{1/2\sqrt{n}, 1/\sqrt{n|2^k|}\right\}$, and initialize $w_1(k) = \eta_k$ for all $k \in [B]$.

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

  Let $W_t = \sum_k w_t(k)$ and $p(t) = \frac{1}{W_t}(\ldots, w_t(k), \ldots)$ be the distribution over the base learners.

  For all $i \in [n]$, let

$$\boxed{P(t)_i = \frac{\max_k v(t,k)_i^2}{2\sum_i \max_k v(t,k)_i^2} + \frac{\sum_k v(t,k)_i}{2B}}$$

  `// Observe that this defines a probability distribution over` n `arms.`

  Sample $x_t \sim \sum_k p(t)_k v(t,k)$, and in parallel sample $x_t' \sim P(t)$.
  `// Only the second sample` x_t' `will be used for weight updating.`

  Play x_t, suffer loss $\ell_t^\top e_{x_t}$ and observe loss $\ell_t^\top e_{x_t'}$. Compute loss estimator

$$\boxed{\widehat{\ell}_t = \mathbf{1}_{i=x_t} \frac{1}{P(t)_{x_t'}} \ell_t^\top e_{x_t'}.}$$

  Update each EXP3 instance with the loss estimator ℓ_t, via Algorithm 2.

  Update the meta-algorithm's weights over base learners via the loss estimator $\widehat{\ell}_t$. For each k, update $w_{t+1}(k)$ as follows,

$$w_{t+1}(k) = \begin{cases} \eta_k & 2^k | t+1 \\ w_t(k)\left(1 + \eta_k \widetilde{r}_t(k)\right) & \textbf{else} \end{cases}$$

  where $\widetilde{r}_t(k) = \widehat{\ell}_t^\top \sum_k p(t)_k v(t,k) - \widehat{\ell}_t^\top v(t,k)$.

end for

Algorithm Overview

- EXP3-type algorithms for both the black-box base learners and the meta-algorithm
- Directly using EXP3 in the MAB setting will fail, because the weight distribution might become unbalanced
- Use addition observation to create unbiased estimators of the loss vector with controllable variance
- A naïve (but sub-optimal) choice is to sample uniformly over the arms with O(n^2) variance
- Use importance sampling to reduce variance to O(n)

Conclusion

Conclusion

We study adaptive regret under the limited observation model. Our result not only

- improves the state-of-the-art query efficiency of O(log log T) in Lu & Hazan (2023)
- matches the lower bound of the bandit setting Daniely et al. (2015)
- providing a sharp characterization of the query efficiency of adaptive regret.

This has multiple implications for:

- **Learning rate adaptation**
- **Learning to optimize**
- **Meta-learning**
- **Adaptive gradient methods**
- **Data-driven algorithm design**

Next Steps

- **Explore vs Exploit:** UCB coefficient, random exploration/restarts, algorithm selection between explorative vs exploitative.
- **Multimetric Optimization:** Scalarizations can adaptively explore the Pareto frontier, especially when parts of the Pareto frontier are targeted.
- **Transfer Learning:** Balancing between the algorithms that 1) uses all the prior data and algorithms that 2) ignores all the prior data
- **Batch Setting:** Ensembling acquisitions and strategies in parallel in data-starved environments.
- **Early Stopping:** Early stopping is trading off between 1) never stopping for maximum exploration and 2) stopping early to save resources.