Understanding the Role of Feedback in Online Learning with Switching Costs

Duo Cheng[1], Xingyu Zhou[2], Bo Ji[1]
[1]Virginia Tech & [2]Wayne State University

ICML International Conference On Machine Learning

WAYNE STATE UNIVERSITY

VIRGINIA TECH

Online Learning with Switching Costs

- T-round repeated game between a learner and an adversary
- For round $t = 1, \dots, T$:
  1. The learner chooses (or plays) one of the K actions, denoted by X_t
  2. The learner suffers the loss of the chosen action, which is determined by the (oblivious) adversary; The learner additionally suffers one unit of loss (i.e., switching cost) if $X_t \neq X_{t-1}$
  3. The learner receives some feedback associated with the losses at this round
  4. The learner uses the feedback to update her policy
- The learner's goal is to minimize ***regret*** (with switching costs)

$$R_T \coloneqq \sum_{t=1}^{T}(\ell_t[X_t] + \mathbb{I}\{X_t \neq X_{t-1}\}) - \min_{k\in[K]} \sum_{t=1}^{T}\ell_t[k]$$

Two Typical Types of Feedback: Bandit and Full-information

- Full-information feedback: Observe the losses of all actions
- Bandit feedback: Observe the loss of the chosen action (i.e., $\ell_t[X_t]$) only
- *Without* switching costs, the minimax regret scales as $\Theta(T^{1/2})$ under *both* types of feedback, in contrast to *a strong separation with switching costs*:

Feedback	Bandit		Full-information
Minimax Regret (w/ SC)	$\widetilde{\Theta}(T^{2/3})$	?	$\Theta(T^{1/2})$

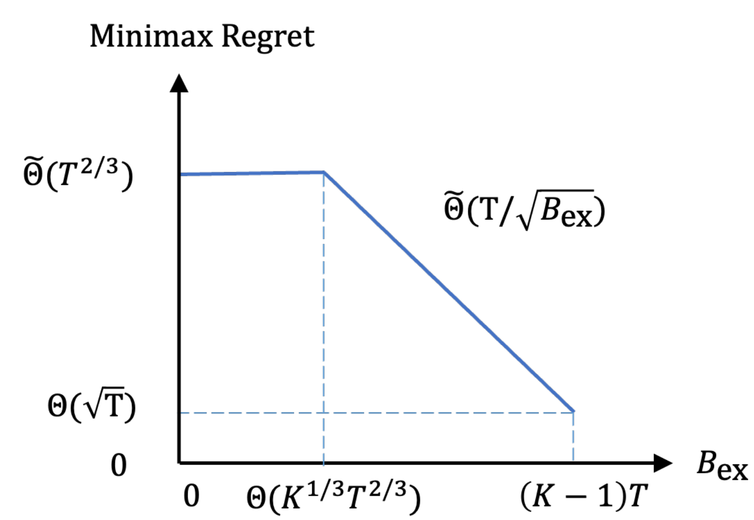
Bandit Learning with Switching Costs under Extra Observation

- Bandit feedback always available
- After receiving bandit feedback, the learner can also observe the loss of any other actions at her choice
- The total number of extra observations should not exceed the given budget B_{ex}
- This incorporates standard bandit and full-information cases as two endpoints:

B_{ex}	$B_{\text{ex}} = 0$ (Bandit)		$B_{\text{ex}} = (K-1)T$ (Full-information)
Minimax Regret (w/ SC)	$\widetilde{\Theta}(T^{2/3})$	?	$\Theta(T^{1/2})$

- Key Question: How do extra observations help improve regret in general?
- We show a ***phase transition*** in terms of how minimax regret scales with B_{ex}:

B_{ex}	0	$O(T^{2/3}K^{1/3})$	$\Omega(T^{2/3}K^{1/3})$	$(K-1)T$
Minimax Regret (w/ SC)	$\widetilde{\Theta}(T^{2/3})$	$\widetilde{\Theta}(T^{2/3})$	$\widetilde{\Theta}(T/\sqrt{B_{\text{ex}}})$	$\Theta(T^{1/2})$



Learning with Switching Costs under Total Observation Budget

- After playing an action, the learner can observe the loss of any actions at her choice (not necessarily including the played action)
- The total number of observation should not exceed the given budget B
- We show that
  1. Adding switching costs does not increase the minimax regret;
  2. How to request feedback (feedback type) matters:

Total Observations	$B \in [K, KT]$ w/o SC	$B \in [K, KT]$ w/ SC
Lower Bound	$\Omega(T/\sqrt{B})$	$\Omega(T/\sqrt{B})$
Upper Bound	$\widetilde{O}(T/\sqrt{B})$	$\widetilde{O}(T/\sqrt{B})$
Minimax Regret	$\widetilde{\Theta}(T/\sqrt{B})$	$\widetilde{\Theta}(T/\sqrt{B})$

Feedback Type	Minimax Regret $B \in [K, KT]$ w/o SC	Minimax Regret $B \in [K, KT]$ w/ SC
Full-information	$\widetilde{\Theta}(T/\sqrt{B})$	
Bandit ($B = O(T^{2/3}K^{1/3})$)		
Bandit ($B = \Omega(T^{2/3}K^{1/3})$)		$\widetilde{\Theta}(T^{2/3})$

Contact: Duo Cheng (duocheng@vt.edu)