Understanding the Role of Feedback in Online Learning with Switching Costs

Duo Cheng¹, Xingyu Zhou², Bo Ji¹
¹Virginia Tech & ²Wayne State University

Online Learning with Switching Costs

- T-round repeated game between a learner and an adversary
- For round \( t = 1, \ldots, T \):
  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)

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

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:

<table>
<thead>
<tr>
<th>( B_{ex} )</th>
<th>( B_{ex} = 0 ) (Bandit)</th>
<th>( B_{ex} = (K - 1)T ) (Full-information)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimax Regret (w/ SC)</td>
<td>( \tilde{\Theta}(T^{2/3}) )</td>
<td>( \Theta(T^{1/2}) )</td>
</tr>
</tbody>
</table>

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.

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 observations 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:

<table>
<thead>
<tr>
<th>Total Observations</th>
<th>( B \in [K, KT] )</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o SC</td>
<td>w/ SC</td>
</tr>
<tr>
<td>Lower Bound</td>
<td>( \Omega(T/\sqrt{B}) )</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>( \tilde{\Theta}(T/\sqrt{B}) )</td>
</tr>
<tr>
<td>Minimax Regret</td>
<td>( \tilde{\Theta}(T/\sqrt{B_{ex}}) )</td>
</tr>
</tbody>
</table>

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

\[
\begin{array}{c|c|c|c|c}
\hline
\text{Feedback Type} & \text{Minimax Regret} & \text{w/o SC} & \text{w/ SC} \\
\hline
\text{Bandit} (B = \Theta(T^{2/3}K^{1/3})) & \tilde{\Theta}(T/\sqrt{B_{ex}}) & & \\
\text{Bandit} (B = \Omega(T^{2/3}K^{1/3})) & \Theta(T^{2/3}) & & \\
\hline
\end{array}
\]

Contact: Duo Cheng (duocheng@vt.edu)