**Conservative Contextual Linear Bandits**

Abbas Kazerouni\(^1\); Mohammad Ghavamzadeh\(^2\); Benjamin Van Roy\(^1\)

\(^{1}\)Stanford University, \(^{2}\)Adobe Research

---

**Motivation**

The manager has a base strategy, but is willing to give us (our policy) a portion of the traffic as long as the loss of us handling that portion instead of him is not more than a threshold.

---

**Set Up**

- A set of available actions at time \(t\): \(\mathcal{A}_t\)
- Each action \(a \in \mathcal{A}_t\) at time \(t\) has a feature vector denoted by \(\phi_a^t \in \mathbb{R}^d\)
- There is an unknown parameter \(\theta^* \in \mathbb{R}^d\) that governs the expected reward of playing action \(a\) at time \(t\):
  \[r^t_a = \langle \theta^*, \phi_a^t \rangle\]
- When action \(a\) is played at time \(t\), we observe
  \[Y_t = r^t_a + \eta_t\]
  where \(\eta_t\) is independent noise
- Manager has a base policy \(\pi_B\) to determine the actions
- Policy \(\pi_B\) is being played for a long time and its actions and their expected rewards are known ahead of time

**Conservative Bandit**

- Select the sequence of actions \(a_1, a_2, \cdots\) to minimize the regret
  \[R_T = \sum_{t=1}^T \langle \theta^*, \phi_{a^*_t} \rangle - \sum_{t=1}^T \langle \theta^*, \phi_{a_t} \rangle\]
  such that the following holds for all \(t = 1, 2, \cdots, T\):
  \[\sum_{i=1}^T r^t_{a^*_i} \geq (1 - \alpha) \sum_{i=1}^T r^t_{b_i}\]
  \(a^*_t = \arg\max_{a \in \mathcal{A}_t} \langle \theta^*, \phi_a \rangle\) is the optimal action at time \(t\)
  \(b_i\) is the action played by \(\pi_B\) at time \(i\)
  \(\alpha \in (0, 1)\) indicates the risk-taking level of the manager
  At any time, the agent has a dilemma among exploration, exploitation and conservativeness

---

**Special Case: Contextual Linear Bandit**

- In Contextual bandits, there is a fixed set of actions \(\mathcal{A}\) and each action \(a \in \mathcal{A}\) has a feature vector \(\psi_a\)
- At time \(t\), a context \(X_t\) arrives
- Time dependent feature of action \(a \in \mathcal{A}\) can be built as
  \[\phi_a^t = f(X_t, \psi_a)\]

---

**Algorithm**

**Algorithm 1 CLUCB**

- **Input**: \(n, A, B\)
- **Initialize**: \(S_0 = 0, S^*_0 = 0, z_0 = 0 \in \mathbb{R}^d\), and \(C_t = B\)
- for \(t = 1, 2, 3, \cdots\) do
  - Find \((\phi^*_t, \phi_t) = \arg\max_{(\phi, \phi^*)} \langle \theta, \phi^* \rangle\)
  - Find \(L_t = \min_{\phi \in \Phi_B} \langle \theta, \phi^* + \phi \rangle\)
  - if \(L_t + \sum_{i=1}^t r^t_{b_i} \geq (1 - \alpha) \sum_{i=1}^t r^t_{a_i}\)
    - Play \(a_t = \phi^*_t\) and observe reward \(Y_t\)
    - Set \(z_t = z_{t-1} + \phi^*_t, S_t = S_{t-1} + t, S^*_t = S^*_{t-1}\)
    - Given \((a_t, Y_t), (a_{t-1}, Y_{t-1})\), update the confidence set \(C_{t+1}\)
  - else
    - Play \(a_t = b_t\) and observe reward \(Y_t\)
    - Set \(z_t = z_{t-1} + \phi^*_t, S_t = S_{t-1} + t, S^*_t = S^*_{t-1} + t\) and \(C_{t+1} = C_t\)
    - end if
  - end for

**Confidence sets are built as follows:**

- Let \(m_t = |S_t|\) be the number of times CLUC has not been conservative
- Let \(\delta = \frac{1}{2} \min_{\phi \in \Phi} \langle \theta, \phi^* \rangle^2\), and \(\gamma(m_t) = \log^{\gamma_m} \left( \frac{|S_t| + 1}{m_t} \right)\)
- Then,
  \[
  \phi_{a_t} = \theta \in C_{t+1} : \left\{ \sum_{\phi \in \Phi} \left( \langle \phi, \phi_{a_t} \rangle - \langle \phi, \theta \rangle \right)^2 \right\} \leq \beta(m_t, \delta)
  \]

---

**Results**

**Assumptions**

- There exist \(0 \leq \Delta_l \leq \Delta_h\) and \(0 < r_l < r_h\) such that for any \(t\)
  \[\Delta_l \leq r^t_{a^*_t} - r^t_{b_t} \leq \Delta_h\]
  and \(r_l \leq r^t_{b_t} \leq r_h\)
- The noise components are \(\sigma^2\) - subgaussian
- \(\theta^* \in B, \phi_{a_t} \in \Phi\) for all \(t\)

**Theorem**

With probability at least \(1 - 2\delta\), CLUCB satisfies all the safety constraints and its regret is bounded by

\[
R_T(CLUCB) \leq 30Dd\sigma \sqrt{T \log \left( \frac{2BT\log \left( \frac{1}{\alpha r_1 + \Delta_1} \right)}{\delta} \right) + K \frac{\Delta_h}{\alpha r_1 + \Delta_1}}
\]

where \(K = O\left( \left[ Dd\sigma \log \left( \frac{1}{\alpha r_1 + \Delta_1} \right) \right]^2 \right)\)

**Simulation**

- CLUCB satisfies all the constraints while LUCB violates 26561 of them
- The agent is conservative only at a finite number of time steps
- CLUCB quickly achieves the same performance as LUCB

---

Contact Information:
abbask@stanford.edu
ghavamza@adobe.com
bvr@stanford.edu

The complete version of this work is accessible through https://arxiv.org/abs/1611.06426