Non-stationary multi-armed bandit

Non-stationary bandits

- Background
  - The reward distribution evolves in an un modeled non-stationary way
  - Internet advertisement:
    - Select which ad to display
    - Targeted user's preference and interest might change
  - Stock
    - Changes in market conditions
- Challenge:
  - Previous observation might be useless

Abrupt changing environment: Discounted UCB

- Reward
  - Independent random variables from potentially different distributions
  - Might vary over time
  - Abrupt changes at unknown instance (breakpoints)
  - Remain stationary during intervals
- Discount empirical average
  - Consider older data in a discount fashion
  - Exponential decay version of UCB

Abrupt changing environment: Discounted UCB

- Discount empirical average

$$\bar{X}_t(\gamma, i) = \frac{1}{N_t(\gamma, i)} \sum_{s=1}^{t} \gamma^{t-s} X_s(i) \mathbb{1}_{\{I_s = i\}} \,, \quad N_t(\gamma, i) = \sum_{s=1}^{t} \gamma^{t-s} \mathbb{1}_{\{I_s = i\}},$$

$$c_t(\gamma, i) = 2B \sqrt{\frac{\xi \log n_t(\gamma)}{N_t(\gamma, i)}} \,, \quad n_t(\gamma) = \sum_{i=1}^{K} N_t(\gamma, i)$$

$$I_t = \arg\max_{1 \le i \le K} \bar{X}_t(\tau, i) + c_t(\tau, i),$$

- Strong regret
  - Track the best arm at each step
  - Upper bound: $O\left(T^{(1+\beta)/2} \log T\right)$ $\Upsilon_T = O(T^{\beta})$ *for some* $\beta \in [0, 1)$

Abrupt changing environment: Sliding-window UCB

- Local empirical average
  - Use only the τ last plays

$$\bar{X}_t(\tau,i) = \frac{1}{N_t(\tau,i)} \sum_{s=t-\tau+1}^{t} X_s(i)\mathbb{1}_{\{I_s=i\}} \,, \quad N_t(\tau,i) = \sum_{s=t-\tau+1}^{t} \mathbb{1}_{\{I_s=i\}}$$

$$c_t(\tau,i) = B\sqrt{\frac{\xi \log(t \wedge \tau)}{N_t(\tau,i)}}$$

$$I_t = \arg\max_{1\le i\le K} \bar{X}_t(\tau,i) + c_t(\tau,i)$$

- Strong regret
  - Upper bound: $O\left(T^{(1+\beta)/2}\sqrt{\log T}\right)$ $\Upsilon_T = O(T^\beta)$ *for some* $\beta \in [0,1)$
  - If number of breakpoints is upper-bounded: $O\left(\sqrt{\Upsilon T \log T}\right)$
- Slightly better than pure discount approach

Abrupt changing environment: Piecewise stationary

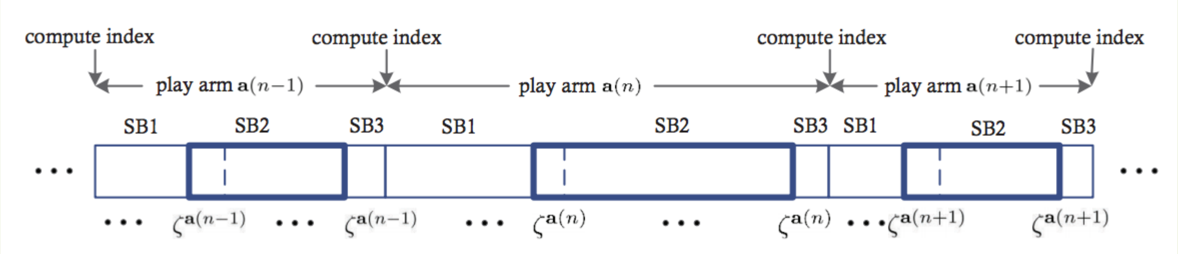
- Abrupt changes at arbitrary intervals
- Number of breakpoint grows linearly with time T
- Partial observation:
  - Query observations on a set of arms not picked before
  - Action is a function of reward observations in the set
- Reset the sub algorithm at the appropriate instants
  - Assume changes are lower bounded $|\beta_{\nu_j}(i) - \beta_{\nu_{j+1}}(i)| > 2\epsilon.$
  - Assume the sub algorithm can guarantee a regret
  - Query the set of arms that received the fewest queries
  - Detect a mean shift w.r.t. a threshold
- Strong regret $O(kn\log(T))$

Exp3 with Reset

- Variation of mean value is bounded
- Tradeoff: remembering and forgetting old information
- Algorithm
  - Use adversary setting to obtain near-optimal performance
    - Single best arm
  - Exp3 with restart time
- Regret
  - Loss of using single best arm against the dynamic oracle
  - Regret of Exp3 relative to the best static action
- Lower bound $\mathcal{R}^{\pi}(\mathcal{V},T) \geq C\left(KV_T\right)^{1/3}T^{2/3}$.
- Upper bound $\mathcal{R}^{\pi}(\mathcal{V},T) \leq \bar{C}\left(K\log K\cdot V_T\right)^{1/3}T^{2/3}$.

Markovian restless bandits

- Reward evolves as finite Markov chains with unknown transition matrices
  - Discrete-time, finite state, aperiodic, irreducible
- Linear combination of rewards
- Weak regret:
  - Sngle best set of arms
  - Highest expected reward on average $\mu^i = \sum_{z\in S_{i,j}} r_x^i \pi_x^i$ $\gamma^* = \max_{\mathbf{a}\in\mathcal{F}} \sum_{i\in\mathcal{A}_{\mathbf{a}(n)}} a_i\mu^i$
- Only take the observation from a regenerative cycle



- Regret: poly(n)log(T)

Kalman filter

- Optimal estimator
  - Minimize the mean square error of the estimated parameters
  - Linear dynamic system
- State transition equation $\alpha_t = K\alpha_{t-1} + R\eta_t$
- Measuring equation $y_t = Z\alpha_t + \xi_t$
- Nonlinear filters

$$\mathbf{x}_k = f(\mathbf{x}_{k-1}, \mathbf{u}_k) + \mathbf{w}_k$$
$$\mathbf{z}_k = h(\mathbf{x}_k) + \mathbf{v}_k$$

  - Extended Kalman filter
  - Similar to hidden markov model

Kalman filtered bandit

- Normally distributed rewards
- Estimate the mean value of each arm
- Two priori noise
  - Observation noise
  - Transition noise
- Played arm

$$\mu_i[N+1] = \frac{\left(\sigma_i^2[N] + \sigma_{\text{tr}}^2\right) \cdot \tilde{r}_i + \sigma_{\text{ob}}^2 \cdot \mu_i[N]}{\sigma_i^2[N] + \sigma_{\text{tr}}^2 + \sigma_{\text{ob}}^2}$$

$$\sigma_i^2[N+1] = \frac{\left(\sigma_i^2[N] + \sigma_{\text{tr}}^2\right) \sigma_{\text{ob}}^2}{\sigma_i^2[N] + \sigma_{\text{tr}}^2 + \sigma_{\text{ob}}^2}$$

- Non-played arm

$$\mu_j[N+1] = \mu_j[N]$$
$$\sigma_j^2[N+1] = \sigma_j^2[N] + \sigma_{\text{tr}}^2$$

Thompson Sampling/UCB+Kalman Filter

- Combinatorial semi-bandits i.i.d over time
- Linear rewards: the sum of all arms
- TS: prior normal distribution $\bar{w} = \Phi\theta^*$
- UCB: $w_t(e) \in [0,1]$
- Noise: $N(0, \sigma^2)$
- Maintain a mean vector and a covariance matrix
- Use Kalman filtering to update
- Regret
  - independent of number of arms

Thank you!