Online Learning: Partial Information and Bandits

CS6780 – Advanced Machine Learning
Spring 2019
Thorstun Joachims
Cornell University

Reading:

Bandit Learning Model

• Setting
  – N arms named H = \{h_1, ..., h_N\}
  – In each round t, each arm h_i performs an action and incurs loss \Delta_{i,t}.
  – Algorithm can select which arm to pull in each round

• Interaction Model
  – FOR t from 1 to T
    • Algorithm selects arm h_{i,t} according to strategy \pi_{i,t} and follows its action y_t.
    • Arms incur losses \Delta_{i,t} \ldots \Delta_{N,t} (all but \Delta_{i,t} unobserved).
    • Algorithm observes and incurs loss \Delta_{i,t}.
    • Algorithm updates \mathbf{w}_t to \mathbf{w}_{t+1} based on \Delta_{i,t}.

Exponentiated Gradient Algorithm for Bandit Setting (EXP3)

• Initialize \mathbf{w}_1 = \left(\frac{1}{N}, ..., \frac{1}{N}\right), \gamma = \min\left\{1, \frac{\sqrt{N \log N}}{(e-1)\Delta_T}\right\}
• FOR t from 1 to T
  – Algorithm randomly picks \mathbf{i}_t with probability \mathbf{P}_t(\mathbf{i}_t) = (1 - \gamma) \mathbf{w}_{t,i} + \gamma / N
  – Arms incur losses \Delta_{1,t} \ldots \Delta_{N,t}.
  – Algorithm observes and incurs loss \Delta_{i,t}.
  – Algorithm updates \mathbf{w} for bandit \mathbf{i}_t as
    \[ w_{t+1,i_t} = w_{t,i_t} \exp\left(-\Delta_{i,t}/P(i_t)\right) \]
  Then normalize \mathbf{w}_{t+1} so that \sum_j w_{t+1,j} = 1.

Adversarial Bandit Regret

• Idea
  – Compare performance to best arm in hindsight
• Regret
  – Overall loss of best arm \ast in hindsight is
    \[ \Delta_{\ast} = \min_{\mathbf{c} \in \mathbb{R}^N} \sum_{t=1}^T \Delta_{\ast,t} \]
  – Expected loss of algorithm A over sequence of arm selections \mathbf{i}_t is
    \[ E_{\Delta_t} = \mathbf{E}_{\Delta_{\ast,t}} \]
  – Regret is difference between expected loss of algorithm and best fixed arm in hindsight
    \[ E_{\text{ExpectedRegret}}(T) = E_{\Delta_t} - \min_{\mathbf{c} \in \mathbb{R}^N} E_{\Delta_{\ast,t}} \]

EXP3 Regret Bound

• Theorem: For \gamma \in [0,1] and stopping time T
  EXP3 has expected regret of at most
  \[ E_{\text{Regret}}(T) \leq (e-1)\gamma \left(\min_{i} \sum_{t=1}^{T} \Delta_{i,t}\right) + \frac{N \log N}{\gamma} \]
• Corollary: For \Delta_{i,t} \leq \Delta, EXP3 with \gamma as on previous slide has expected regret of at most
  \[ E_{\text{Regret}}(T) \leq 2.63 \sqrt{\Delta TN \log N} \]

Stochastic Bandit Learning Model

• Setting
  – N arms named H = \{h_1, ..., h_N\}
  – In each round t, each arm h_i performs an action and incurs loss \Delta_{i,t} drawn from fixed distribution \mathbf{P}(\Delta|t) with mean \mu_i.
  – Algorithm can select which arm to pull in each round

• Interaction Model
  – FOR t from 1 to T
    • Algorithm selects arm h_{i,t} according to \pi_{i,t} and action y_t.
    • Arms incur losses \Delta_{1,t} \ldots \Delta_{N,t} (all but \Delta_{i,t} unobserved).
    • Algorithm observes and incurs loss \Delta_{i,t}.
    • Algorithm updates \mathbf{w}_t to \mathbf{w}_{t+1} based on \Delta_{i,t}.
Stochastic Bandit Regret

- **Idea**
  - Compare performance to arm with best expected performance

- **Regret**
  - Overall loss of best arm $i^*$ is $\Delta^* = T \min_{i \in [1:N]} \mu_i - T \mu_{i^*}$
  - Expected loss of algorithm $A$ over sequence of arm selections $i_t$ is
    $$E_A \left[ \sum_{t=1}^T \Delta_{i_t} \right]$$
  - Regret is difference between expected loss of algorithm and best fixed arm in hindsight
    $$\text{Expected Regret}(T) = E_A \left[ \sum_{t=1}^T \Delta_{i_t} \right] - T \mu_{i^*}.$$

UCB1 Algorithm

- **Init:**
  - Play each arm $i$ once to get initial values for $w_1 \ldots w_N$.
  - $n = (1, \ldots, 1)$

- **For $t$ from $(N + 1)$ to $T$**
  - Play arm $i_t = \arg\min_i \frac{w_i}{n_i} \sqrt{\frac{2 \log T}{n_i}}$
  - Algorithm observes and incurs loss $\Delta_{i_t, t}$
  - $w_i = w_i + \Delta_{i_t, t}$
  - $n_i = n_i + 1$

UCB1 Regret Bound

- **Theorem:** The expected regret of UCB1 is at most
  $$O \left( \sum_{i \neq i^*} \frac{\log T}{\epsilon_i} \right)$$
  where $i^*$ is the best arm and $\epsilon_i = \mu_{i^*} - \mu_i$.

Other Online Learning Problems

- Contextual Bandits
- Dueling Bandits
- Coactive Learning
- Online Convex Optimization
- Partial Monitoring