Multi-armed Bandit

Shuai Li

John Hopcroft Center, Shanghai Jiao Tong University

https://shuaili8.github.io

Adversarial Setting

- similar to the Learning with Expert Advice setting
- In each round t
  - Select one expert A_t
  - Only observe the loss of the selected expert g_{t,A_t} (bandit feedback!)
  - The objective is still to compete with the cumulative loss of the best expert
- Still need randomization!
- Assume the adversary is oblivious (vs. adaptive adversary)
  - He decides the losses of all the rounds before the game starts
- exploration-exploitation trade-off
- Can't directly use OMD or FTRL
  - need the full loss function or their lower bounds

Construct Unbiased Estimator

- Only observe g_{t,A_t}
- Recall that in each round expert i is drawn according to prob. $x_{t,i}$
- $\tilde{g}_{t,i} = \begin{cases} \frac{g_{t,A_t}}{x_{t,A_t}}, & i = A_t \\ 0, & o.w. \end{cases}$
- $\mathbb{E}_{A_t}\left[\tilde{g}_{t,i}\right] = g_{t,i}$
- Run OMD w/ $\psi: \mathbb{R}_+^d \to \mathbb{R}, \psi(x) = \sum_{i=1}^d x_i \ln x_i$, $\|g_t\|_\infty \le L_\infty, x_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right)$ to have

$$\sum_{t=1}^{T} \langle \tilde{g}_t, x_t \rangle - \sum_{t=1}^{T} \langle \tilde{g}_t, u \rangle \le \frac{\ln d}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\tilde{g}_t\|_\infty^2$$

Direct Application of OMD

- Taking expectation
- $\mathbb{E}\left[\sum_{t=1}^{T} g_{t,A_t}\right] - \sum_{t=1}^{T}\langle g_t, u\rangle = \mathbb{E}[\sum_{t=1}^{T}\langle g_t, x_t\rangle] - \sum_{t=1}^{T}\langle g_t, u\rangle$
- $= \mathbb{E}[\sum_{t=1}^{T}\langle \tilde{g}_t, x_t\rangle - \sum_{t=1}^{T}\langle \tilde{g}_t, u\rangle] \leq \frac{\ln d}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T} \mathbb{E}[\|\tilde{g}_t\|_\infty^2]$

$\sum_{i=1}^{d} \frac{g_{t,i}^2}{x_{t,i}}$

- Require: $x_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right), \alpha, \eta > 0$
- For $t = 1:T$ do
  - $\tilde{x}_t = (1-\alpha)x_t + \alpha\left(\frac{1}{d}, \dots, \frac{1}{d}\right)$
  - Draw A_t according to $\mathbb{P}[A_t = i] = \tilde{x}_{t,i}$
  - Select expert A_t
  - Observe only the loss of the selected arm $g_{t,A_t} \in \pm L_\infty$ and pay it
  - Construct estimate $\tilde{g}_{t,i} = \frac{g_{t,i}}{\tilde{x}_{t,i}} \mathbb{I}[A_t = i]$
  - Update $x_{t+1,i} \propto x_{t,i} \exp(-\eta \tilde{g}_{t,i})$

Direct Application of OMD: Result

- $\alpha \propto \sqrt{d^2 L_\infty \eta}, \eta \propto \left(\frac{\ln d}{d L_\infty^{3/2} T}\right)^{2/3}$
- Then

$$\mathbb{E}\left[\sum_{t=1}^{T} g_{t,A_t}\right] - \sum_{t=1}^{T} \langle g_t, u \rangle = O\left(L_\infty (dT)^{2/3} \ln^{1/3} d\right)$$

- Much worse than $O\left(L_\infty \sqrt{T \ln d}\right)$ of the full-information case

OMD Using Local Norm

$$x_{t+1} = \arg\min_{x \in V} \langle g_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$$

- Lemma 6.14. $\tilde{x}_{t+1} \coloneqq \arg\min_{x \in X} \langle g_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$. ψ has positive definite Hessian. Then

$$\ell_t(x_t) - \ell_t(u) \le \frac{B_\psi(u; x_t) - B_\psi(x; x_{t+1})}{\eta_t} + \frac{\eta_t}{2} \min\left(\|g_t\|^2_{(\nabla^2 \psi(z_t))^{-1}}, \|g_t\|^2_{(\nabla^2 \psi(\tilde{z}_t))^{-1}} \right)$$

where $z_t \in [x_t, x_{t+1}]$ and $\tilde{z}_t \in [x_t, \tilde{x}_{t+1}]$

Improved Result

- Require: $x_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right), \alpha, \eta > 0$
- For $t = 1 : T$ do
  - Draw A_t according to $\mathbb{P}[A_t = i] = x_{t,i}$
  - Select expert A_t
  - Observe only the loss of the selected arm $g_{t,A_t} \in [0, L_\infty]$ and pay it
  - Construct estimate $\tilde{g}_{t,i} = \frac{g_{t,i}}{x_{t,i}} \mathbb{I}[A_t = i]$
  - Update $x_{t+1,i} \propto x_{t,i} \exp\left(-\eta \tilde{g}_{t,i}\right)$

- Theorem 10.2. $\eta \propto \sqrt{\frac{\ln d}{L_\infty^2 T}}$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T} g_{t,A_t}\right] - \sum_{t=1}^{T} \langle g_t, u \rangle = O\left(L_\infty \sqrt{dT \ln d}\right)$$

Optimal Regret Using Tsallis Entropy

- Require: $x_1 = \left(\frac{1}{d}, \dots, \frac{1}{d}\right), \alpha, \eta > 0$
- For $t = 1{:}\,T$ do
  - Draw A_t according to $\mathbb{P}[A_t = i] = x_{t,i}$
  - Select expert A_t
  - Observe only the loss of the selected arm $g_{t,A_t} \in [0, L_\infty]$ and pay it
  - Construct estimate $\tilde{g}_{t,i} = \frac{g_{t,i}}{x_{t,i}} \mathbb{I}[A_t = i]$
  - Update $x_{t+1} = \arg\min_{x \in V} \langle \tilde{g}_t, x \rangle + \frac{1}{\eta_t} B_\psi(x; x_t)$

negative Tsallis entropy

- Theorem 10.3. $\psi(x) = \sum_{i=1}^{d} -\sqrt{x_i}$. $\eta \propto \frac{1}{\sqrt{L_\infty^2 T}}$. Then

$$\mathbb{E}\left[\sum_{t=1}^{T} g_{t,A_t}\right] - \sum_{t=1}^{T} \langle g_t, u \rangle = O(L_\infty \sqrt{dT})$$

can be proved to be optimal

Summary

Shuai Li
https://shuaili8.github.io

- Multi-armed bandit setting
  - bandit feedback
  - exploration-exploitation trade-off
- Directly apply OMD $O\left(T^{2/3}\right)$
- OMD w/ local norm $O\left(\sqrt{T}\right)$
- OMD w/ Tsallis entropy, optimal regret bound

Questions?