1. HW3 almost graded. soln is up
(D2L → content → HW3-solns)

2. HW4 posted. (course website)
due on May 4, 11:59 pm no late HW will be accepted.

3. Final presentation (preferences on slot Apr 29 | May 4)
send me an email.
discuss presentation in our OHs.

Thm: UCB satisfies:

① $R_T \le \sum_{a:\Delta(a)>0} \frac{16 \ln T}{\Delta(a)} + 3K$. (gap dependent)

② $R_T \le \mathcal{O}(\sqrt{TK})$. (gap independent)

Lemma 2: $\forall a.\ \mathbb{E}[m_T(a)] \le \frac{16 \ln T}{\Delta(a)^2} + 3$.

pf of ①:

given a "cutoff" $\Delta > 0$. group the arms

based on $\Delta(a) \le \Delta$ or $\Delta(a) > \Delta$.

$$R_T = \sum_a \mathbb{E}[n_T(a)] \cdot \Delta(a)$$

$\le \Delta$.

$$= \sum_{a:\, \Delta(a) \in (0,\Delta]} \mathbb{E}[n_T(a)]\, \Delta(a) + \sum_{a:\, \Delta(a) > \Delta} \cdots$$

$$\le T\Delta + \sum_{a:\, \Delta(a) > \Delta} \left(\Delta(a) \frac{16 \ln T}{\Delta(a)^2} + 3\Delta(a) \right)$$

$\le \frac{16 \ln T}{\Delta}$

$$\le T\Delta + \frac{16 K \ln T}{\Delta} + 3K.$$

$$T\Delta = \frac{K \ln T}{\Delta} \Rightarrow \Delta = \sqrt{\frac{K \ln T}{T}}$$

we focus on $T \ge K$. so $\sqrt{TK} \ge K$

$$\Rightarrow R_T \le O\left(\sqrt{TK \ln T}\right).$$

§ adversarial multi-armed bandits

what if the losses are non-stationary. or even does not come from a distribution?

can we still give algorithms w/ regret guarantees?

random.

Recall MAB:

For $t=1,2,\dots T$:

environment gives $l_t \in [0,1]^K$. (l_t's are chosen before round 1. (oblivious adversary))

learner selects action $a_t \in \{1,\dots K\} \sim p_t \in \Delta^{K-1}$ randomly

learner suffer loss $l_t(a_t)$.

Goal: minimize regret.

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} l_t(a_t)\right] - \min_a \sum_{t=1}^{T} l_t(a).$$

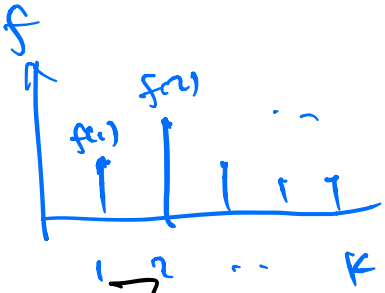
$(l_t(a_t) : a_t = a)$ may not be fid.

How to design algorithms w/ sublinear R_T?

alternative representation of R_T:

$$\mathbb{E}_{a_t \sim p_t} l_t(a_t) = \sum_{a=1}^{K} p_t(a) \cdot l_t(a) = \langle p_t, l_t \rangle$$

$$\min_a \sum_{t=1}^{T} l_t(a) = \min_{p \in \Delta^{K-1}} \langle p, \sum_{t=1}^{T} l_t \rangle$$



$f_t(p) = \langle p, l_t \rangle$ f

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \langle p_t, l_t \rangle - \min_{p \in \Delta^{K-1}} \sum_{t=1}^{T} \langle p, l_t \rangle\right]$$

$\sqrt{T \ln K}$ regret for weather forecasting using OMD (l_t's are known).

Q: can we re-use the OMD algorithm to develop algorithms w/ low R_T ?

$$p_{t+1} = \underset{p \in \Delta^{K-1}}{\operatorname{argmin}} \; \eta \cdot \langle \ell_t, p \rangle + D_\psi(p, p_t)$$

ℓ_t need to be changed.

$a_t \sim p_t$. $\ell_t(a_t)$.

ℓ_t

Thompson sampling

$\ell(a)$.

do OMD on unbiased estimators of ℓ_t's.

$\tilde{\ell}_t$. $\mathbb{E}_{a_t \sim p_t}[\tilde{\ell}_t] = \ell_t$.

Low regret wrt $\tilde{\ell}_t$'s $\Rightarrow$ low regret wrt ℓ_t's. (taking expectation)

Constructing $\tilde{\ell}_t$:

$$\tilde{\ell}_t(a) = \begin{cases} 0 & a_t \neq a. \\ x = \dfrac{\ell_t(a)}{p_t(a)} & a_t = a. \end{cases}$$

($\ell_t(a_t)$ → can be observed.

$$= \frac{\ell_t(a)}{p_t(a_t)}$$

$$\mathbb{E}_{a_t \sim p_t}\left[\tilde{\ell}_t(a)\right]$$

$$= (1 - p_t(a)) \cdot 0 + p_t(a) \cdot \frac{\ell_t(a)}{p_t(a)} = \ell_t(a)$$

$$\tilde{\ell}_t(a) = \frac{\ell_t(a)}{p_t(a)} I(a = a_t).$$

OMD with $\tilde{\ell}_t$'s w/ $\Omega = \Delta^{K-1}$,

$$\psi(p) = \sum_{a=1}^{K} p(a) \ln p(a) \qquad p_1 = \left(\frac{1}{K}, \dots \frac{1}{K}\right).$$

$$p_{t+1}(a) \overset{(*)}{\propto} p_t(a) \exp(-\eta \tilde{\ell}_t(a)) \propto \exp\left(-\eta \sum_{s=1}^{t} \tilde{\ell}_s(a)\right)$$

→ EXP3 algorithm.

achieving exploration / exploitation tradeoff ?

① $\sum_{s=1}^{t} \tilde{\ell}_s(a) \approx \sum_{s=1}^{t} \ell_s(a)$ ⇒ exploit

② $\eta < \infty$.

$a_t = a$.

$$\tilde{\ell}_t = \left(0, \dots 0, \frac{\ell_t(a)}{p_t(a)}, \dots 0\right)$$

$\ell_t(a) \in [-1, 0]$. the algorithm will behave greedy

↓

by (*)

p_{t+1} skew toward action other than a_t that p_t is.

it encourages taking other actions.

Analysis of EXP3:

applying OMD guarantees naively: RHS

$$\sum_{t=1}^{T} \langle p_t, \tilde{\ell}_t \rangle - \sum_{t=1}^{T} \langle p, \tilde{\ell}_t \rangle \le \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \|\tilde{\ell}_t\|_\infty^2$$

$\mathbb{E}_{a_t \sim p_t}\left[\tilde{\ell}_t(a)\right] = \ell_t(a)$

$\mathbb{E}_{a_t \sim p_t}\left(\frac{\ell_t(a_t)}{p_t(a_t)}\right)^2$

$\mathbb{E}\Big[\forall p \in \Delta^{K-1}$

$\mathbb{E}\Big[\sum_{t=1}^{T} \langle p_t, \ell_t \rangle\Big]$

$\mathbb{E}\Big[$

$\sum_{t=1}^{T} \langle p, \ell_t \rangle$

$\le \sum_a \frac{1}{p_t(a)}$

cannot be well controlled.

* get a better regret bound for OMD w/ neg-entropy regularizer.

(local norm bound)

Lemma: OMD w/ $\Omega = \Delta^{K-1}$, $\psi(w) =$ neg. entropy

learning rate η. on $\{f_t(w) = \langle w, g_t \rangle\}_{t=1}^{T}$ where

$g_t \in \mathbb{R}^K$

$g_t \in [0, \infty]$.

$$\sum_{t=1}^{T} \langle w_t, g_t \rangle - \sum_{t=1}^{T} \langle w, g_t \rangle \le \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \|g_t\|^2$$

$\text{diag}(w_t)$

$\sum w_t(a) g_t(a)^2$

$\le \|g_t\|_\infty^2$

Applying Lemma to EXP3:

$$RHS = \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \cdot \tilde{l}_t(a)^2$$

$$= \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} p_t(a_t) \left(\frac{l_t(a_t)}{p_t(a_t)} \right)^2$$

$$\le \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \frac{1}{p_t(a_t)}$$

$\le K$

→ better than $\frac{1}{p_t(a_t)^2}$

$$\mathbb{E}[RHS]$$

$$= \frac{\ln K}{\eta} + \eta \sum_{t=1}^{T} \mathbb{E}\left[\frac{1}{p_t(a_t)} \right]$$

$$\mathbb{E}_{a_t \sim p_t}\left[\frac{1}{p_t(a_t)}\right] = \sum_{a=1}^{K} p_t(a) \cdot \frac{1}{p_t(a)} = K.$$

$$= \frac{\ln K}{\eta} + \eta T \cdot K$$

$$\eta = \sqrt{\frac{\ln K}{TK}}$$

$$= 2\sqrt{TK \ln K}.$$

In conclusion:

Thm: EXP3 w/ $\eta = \sqrt{\frac{\ln K}{TK}}$ has regret $R_T \le O\left(\sqrt{TK \ln K}\right)$.

Pf of Lemma:

Recall in OMD:

$$\langle w_{t+1} - w, \eta g_t \rangle \le D_\varphi(w, w_t) - D_\varphi(w, w_{t+1}) - D_\varphi(w_{t+1}, w_t)$$

w_t

$$\langle w_t - w, \eta g_t \rangle$$

$$\leq \langle w_t - w_{t+1}, \eta g_t \rangle + D_\varphi(w, w_t) - D_\varphi(w, w_{t+1})$$

①

$$w_{t+1}(a) = \frac{w_t(a) \cdot e^{-\eta g_t(a)}}{Z_t} \qquad Z_t \leq 1$$

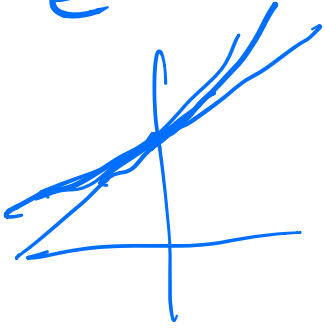
$$\sum_{a'} w_t(a') e^{-\eta g_t(a')}$$

$$\geq w_t(a) \cdot e^{-\eta g_t(a)}.$$

$$① \leq \sum_{a=1}^{K} w_t(a) \left(1 - e^{-\eta g_t(a)}\right) \eta g_t(a).$$

$$\geq 1 - \eta g_t(a)$$

$$e^x \geq 1 + x \qquad \leq \eta^2 \sum_{a=1}^{K} w_t(a) \cdot g_t(a)^2.$$



summing over t, divide by η

$\square$

$\Rightarrow$ learner

Next class:

stochastic linear bandits.

(contextual)

$a_1 \rightarrow \ell_1 = (0 \cdots 0, \underset{\uparrow}{1}\, 0 \cdots 0)$, $\uparrow a_1$

$a_2 \rightarrow \ell_2 = (0 \cdots 0\, \underset{\uparrow}{1}\, 0 \cdots 0)$, $\uparrow a_2$

$\forall$ Alg. $\exists$ adversary.

$R_T \geq \Omega(T)$.

Cover's impossibility result. $\Omega(T)$.