Bandit Algorithms
(part 3)

Tor Lattimore

Menu for the day

- Bandits with experts
- Adversarial linear bandits
- Shortest path problems
- Ranking
- Semibandits

Bandits with experts

- k actions
- Adversary chooses losses $\ell_1, \ldots, \ell_n \in [0,1]^k$
- m experts making recommendations
- Expert i recommends action a_t^i in round t
- Learner chooses an action $A_t \in \{1, \ldots, k\}$
- Regret is

$$\mathfrak{R}_n = \max_{i \in \{1,\ldots,m\}} \mathbb{E}\left[\sum_{t=1}^n \ell_{t,A_t} - \ell_{t,a_t^i}\right]$$

Exp4

- FTRL with negentropy over the experts
- Algorithm samples expert E_t from distribution P_t

$$P_t(i) = \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,a_t^i}\right)}{\sum_{j=1}^{m} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{s,a_t^j}\right)}$$

- Then plays action $A_t = a_t^{E_t}$
- Loss estimate is

$$\hat{\ell}_{t,a} = \frac{\mathbb{1}(A_t = a)\ell_{t,a}}{\sum_{i=1}^{m} \mathbb{1}(a_t^i = a) P_t(i)}$$

Analysis

- Start with the usual bound

$$\mathfrak{R}_n \leq \frac{\log(m)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{m} P_t(i)\hat{\ell}^2_{t,a_t^i}\right]$$

- Variance term

$$\mathbb{E}\left[\sum_{i=1}^{m} P_t(i)\hat{\ell}^2_{t,a_t^i}\right] \leq k\,.$$

- Regret is bounded by

$$\mathfrak{R}_n \leq \frac{\log(m)}{\eta} + \frac{\eta n k}{2} = \sqrt{2nk\log(m)}$$

Application to non-stationary bandits

- Standard bandit setting
- k actions, $\ell_1, \dots, \ell_n \in [0,1]^k$
- Different regret

$$\mathfrak{R}_n = \max_{a_1,\dots,a_n : \sum_{t=1}^{n-1} \mathbb{1}(a_t \neq a_{t+1} \leq c)} \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t} - \ell_{t,a_t}\right]$$

Application to non-stationary bandits

- Standard bandit setting
- k actions, $\ell_1, \ldots, \ell_n \in [0,1]^k$
- Different regret

$$\mathfrak{R}_n = \max_{a_1,\ldots,a_n : \sum_{t=1}^{n-1} \mathbb{1}(a_t \neq a_{t+1} \leq c)} \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t} - \ell_{t,a_t}\right]$$

- Simple algorithm just runs Exp4
- Roughly $m \approx \binom{n}{c} k^c$

Application to non-stationary bandits

- Standard bandit setting
- k actions, $\ell_1, \ldots, \ell_n \in [0,1]^k$
- Different regret

$$\mathfrak{R}_n = \max_{a_1,\ldots,a_n : \sum_{t=1}^{n-1} \mathbb{1}(a_t \neq a_{t+1} \leq c)} \mathbb{E}\left[\sum_{t=1}^{n} \ell_{t,A_t} - \ell_{t,a_t}\right]$$

- Simple algorithm just runs Exp4
- Roughly $m \approx \binom{n}{c} k^c$
- Regret is $O(\sqrt{cnk\log(nk)})$

Adversarial linear bandits

- $\mathcal{A} \subset \mathbb{R}^d$
- Adversary chooses losses $\ell_1, \ldots, \ell_n$
- $\max_{a \in \mathcal{A}} |\langle a, \ell_t \rangle| \leq 1$
- Learner chooses $A_t \in \mathcal{A}$
- Loss for action a is $\ell_t(a) = \langle a, \ell_t \rangle$
- Learner suffers $\ell_t(A_t)$
- Regret is

$$\mathfrak{R}_n = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{n} \langle A_t - a, \ell_t \rangle\right]$$

Examples

- $\mathcal{A} = \{e_1, \dots, e_d\}$
- Just have the usual finite-armed case
- **Fundamental** $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$
- **Practical** $\mathcal{A} =$ finite set
- We can deal with changing action sets as well

Exp3 for linear bandits

- $|\mathcal{A}| = k$
- Algorithm plays FTRL over distribution on $\mathcal{A}$
- Negentropy potential

$$\mathfrak{R}_n \lesssim \mathbb{E}\left[\frac{\log(k)}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\sum_{a \in \mathcal{A}} P_t(a)\hat{\ell}_t(a)^2\right]$$

Estimating ℓ_t

- Last time, $\hat{\ell}_t(a) = \frac{\mathbb{1}(A_t = a)\ell_t(a)}{P_t(a)}$
- Does not use the linear structure

Estimating ℓ_t

- Least squares estimation

$$\hat{\ell}_t = Q_t^{-1} A_t \langle A_t, \ell_t \rangle \qquad Q_t = \sum_{a \in \mathcal{A}} P_t(a) a a^\top$$

- Expectation

$$\mathbb{E}[\hat{\ell}_t \mid P_t] = \sum_{a \in \mathcal{A}} P_t(a) Q_t^{-1} a a^\top \ell_t = Q_t Q_t^{-1} \ell_t = \ell_t$$

Variance

$$M_t = \sum_{a \in \mathcal{A}} P_t(a) \hat{\ell}_t(a)^2$$

Variance

$$
\begin{aligned}
M_t &= \sum_{a \in \mathcal{A}} P_t(a) \hat{\ell}_t(a)^2 \\
&= \sum_{a \in \mathcal{A}} P_t(a) \left(a^\top Q_t^{-1} A_t \langle A_t, \ell_t \rangle \right)^2
\end{aligned}
$$

Variance

$$
\begin{aligned}
M_t &= \sum_{a \in \mathcal{A}} P_t(a) \hat{\ell}_t(a)^2 \\
&= \sum_{a \in \mathcal{A}} P_t(a) \left(a^\top Q_t^{-1} A_t \langle A_t, \ell_t \rangle\right)^2 \\
&\leq \sum_{a \in \mathcal{A}} P_t(a) a^\top Q_t^{-1} A_t A_t^\top Q_t^{-1} a
\end{aligned}
$$

Variance

$$
\begin{aligned}
M_t &= \sum_{a \in \mathcal{A}} P_t(a) \hat{\ell}_t(a)^2 \\
&= \sum_{a \in \mathcal{A}} P_t(a) \left(a^\top Q_t^{-1} A_t \langle A_t, \ell_t \rangle\right)^2 \\
&\leq \sum_{a \in \mathcal{A}} P_t(a) a^\top Q_t^{-1} A_t A_t^\top Q_t^{-1} a \\
&= \sum_{a \in \mathcal{A}} P_t(a) \operatorname{Tr}\left(Q_t^{-1} A_t A_t^\top Q_t^{-1} a a^\top\right)
\end{aligned}
$$

Variance

$$
\begin{aligned}
M_t &= \sum_{a \in \mathcal{A}} P_t(a) \hat{\ell}_t(a)^2 \\
&= \sum_{a \in \mathcal{A}} P_t(a) \left(a^\top Q_t^{-1} A_t \langle A_t, \ell_t \rangle \right)^2 \\
&\leq \sum_{a \in \mathcal{A}} P_t(a) a^\top Q_t^{-1} A_t A_t^\top Q_t^{-1} a \\
&= \sum_{a \in \mathcal{A}} P_t(a) \operatorname{Tr}\left(Q_t^{-1} A_t A_t^\top Q_t^{-1} a a^\top \right) \\
&= \operatorname{Tr}(Q_t^{-1} A_t A_t^\top)
\end{aligned}
$$

Variance

$$
\begin{aligned}
M_t &= \sum_{a \in \mathcal{A}} P_t(a) \hat{\ell}_t(a)^2 \\
&= \sum_{a \in \mathcal{A}} P_t(a) \left(a^\top Q_t^{-1} A_t \langle A_t, \ell_t \rangle\right)^2 \\
&\leq \sum_{a \in \mathcal{A}} P_t(a) a^\top Q_t^{-1} A_t A_t^\top Q_t^{-1} a \\
&= \sum_{a \in \mathcal{A}} P_t(a) \operatorname{Tr}\left(Q_t^{-1} A_t A_t^\top Q_t^{-1} a a^\top\right) \\
&= \operatorname{Tr}(Q_t^{-1} A_t A_t^\top)
\end{aligned}
$$

Taking the conditional expectation,

$$
\mathbb{E}[M_t \mid P_t] = \sum_{a \in \mathcal{A}} P_t(a) \operatorname{Tr}\left(Q_t^{-1} a a^\top\right) = d
$$

Almost works...

- Plugging in,

$$\begin{aligned}\mathfrak{R}_n &\lesssim \frac{\log(k)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} P_t(a)\hat{\ell}_t(a)^2\right] \\ &\leq \frac{\log(k)}{\eta} + \frac{\eta n d}{2} \\ &\leq \sqrt{2nd\log(k)}\end{aligned}$$

- It's the bound we want, but...

Almost works...

- Plugging in,

$$\begin{aligned}\mathfrak{R}_n &\lesssim \frac{\log(k)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} P_t(a)\hat{\ell}_t(a)^2\right] \\ &\leq \frac{\log(k)}{\eta} + \frac{\eta n d}{2} \\ &\leq \sqrt{2nd\log(k)}\end{aligned}$$

- It's the bound we want, but...
- Taylor's approximation only good when $\eta\hat{\ell}_t(a) \geq -1$

Adding exploration

- FTRL recommends P_t
- Let $\tilde{P}_t = (1-\gamma)P_t + \gamma\pi$
- π is an **exploration distribution**
- $A_t \sim \tilde{P}_t$

Adding exploration

- FTRL recommends P_t
- Let $\tilde{P}_t = (1-\gamma)P_t + \gamma\pi$
- π is an **exploration distribution**
- $A_t \sim \tilde{P}_t$
- $Q_t = \sum_{a \in \mathcal{A}} \tilde{P}_t(a) a a^\top \succ \gamma Q_\pi = \gamma \sum_{a \in \mathcal{A}} \pi(a) a a^\top$

$$\begin{aligned}\hat{\ell}_t(a) &= |a^\top Q_t^{-1} A_t \langle A_t, \ell_t \rangle| \\ &\leq \frac{1}{\gamma} \langle Q_\pi^{-1/2} a, Q_\pi^{-1/2} A_t \rangle \leq \frac{1}{\gamma} \|a\|_{Q_\pi^{-1}} \|A_t\|_{Q_\pi^{-1}} \leq \frac{d}{\gamma}\end{aligned}$$

Kiefer–Wolfowitz theorem

Assume $\mathcal{A}$ spans $\mathbb{R}^d$

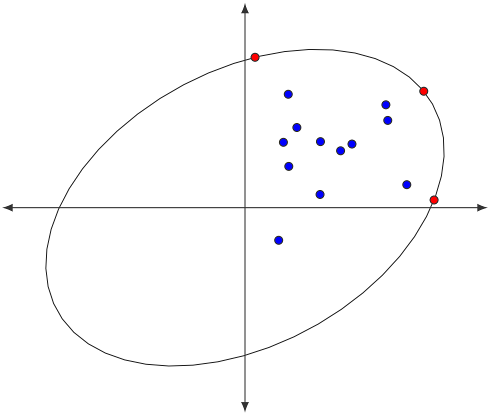
$$f(\pi) = \max_{a \in \mathcal{A}} \log \det Q_\pi \qquad g(\pi) = \max_{a \in \mathcal{A}} \|a\|^2_{Q_\pi^{-1}}$$

Theorem The following are equivalent

- π is a maximizer of f
- π is a minimiser of g
- $g(\pi) = d$

Also, a minimiser of π has support at most $d(d+1)/2$

Geometric intuition



Smallest central ellipsoid containing the $\mathcal{A}$

Linear bandit analysis

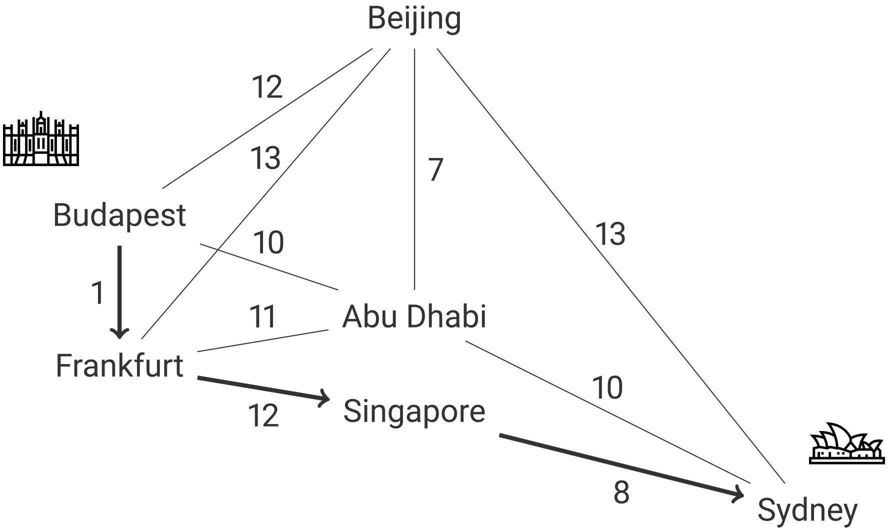
- A little calculation shows that

$$\mathfrak{R}_n \lesssim \frac{\log(k)}{\eta} + n\gamma + \eta n d \qquad \text{with } \gamma \geq \eta d$$

- Optimizing η eventually leads to

$$\mathfrak{R}_n \leq 2\sqrt{3dn\log(k)}$$

Path routing



- d edges in the graph
- A path is a set of edges
- $\mathcal{A} \subset \{0,1\}^d$
- The loss is the length of the whole path
- $\ell_t(a) = \langle a, \ell_t \rangle$
- Assuming $\ell_t \in [0,1]^d$

- d edges in the graph
- A path is a set of edges
- $\mathcal{A} \subset \{0, 1\}^d$
- The loss is the length of the whole path
- $\ell_t(a) = \langle a, \ell_t \rangle$
- Assuming $\ell_t \in [0, 1]^d$
- **Bandit feedback** Observe $\langle A_t, \ell_t \rangle$
- **Semibandit feedback** Observe $A_{t,i}\ell_{t,i}$

A simple ranking problem

- Learner chooses m out of d products to recommend
- $\ell_{t,i} = 0$ if the user would click on product $i \in [d]$
- $\ell_{t,i} = 1$ otherwise
- $\mathcal{A} = \{x \in \{0,1\}^d : \|x\|_1 = m\}$
- Learner observes $A_{t,i}\ell_{t,i}$

Combinatorial semi-bandits

- $\mathcal{A} \subset \{x \in \{0,1\}^d : \|x\|_1 \leq m\}$
- Adversary chooses losses $\ell_t \in [0,1]^d$
- Loss suffered by learner is $\langle \ell_t, A_t \rangle$
- **Bandit feedback** Observe $\langle A_t, \ell_t \rangle$
- **Semibandit feedback** Observe $A_{t,i}\ell_{t,i}$
- Regret as usual

$$\mathfrak{R}_n \leq \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{n} \langle A_t - a, \ell_t \rangle\right]$$

FTRL for combinatorial semibandits

- Play FTRL with negentropy on $\mathrm{conv}(\mathcal{A})$
- Learner chooses point in $X_t \in \mathrm{conv}(\mathcal{A})$
- Find distribution P_t with $\sum_{a \in \mathcal{A}} P_t(a) a = X_t$
- Estimate losses by

$$\hat{\ell}_{t,i} = \frac{A_{t,i} \ell_{t,i}}{X_{t,i}}$$

FTRL for combinatorial semibandits

- Our standard regret bound

$$\begin{aligned}
\mathfrak{R}_n &\leq \max_{a \in \mathcal{A}} \frac{F(a) - F(X_1)}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^n \sum_{i=1}^d X_{t,i}\hat{\ell}_{t,i}^2\right] \\
&\leq \frac{m(1 + \log(d/m))}{\eta} + \frac{\eta n d}{2} \\
&\leq \sqrt{2nmd(1 + \log(d/m)}
\end{aligned}$$

Drawbacks of FTRL for semibandits

- **Computation seems challenging**
- There are two optimization problems to solve
- Finding the recommendation of FTRL

$$X_t = \operatorname{argmin}_{x \in \operatorname{conv}(\mathcal{A})} \eta \sum_{s=1}^{t-1} \langle x, \hat{\ell}_s \rangle + F(x)$$

- Finding P_t such that $\sum_{a \in \mathcal{A}} P_t(a) a = X_t$

Drawbacks of FTRL for semibandits

- **Computation seems challenging**
- There are two optimization problems to solve
- Finding the recommendation of FTRL

$$X_t = \mathrm{argmin}_{x \in \mathrm{conv}(\mathcal{A})} \, \eta \sum_{s=1}^{t-1} \langle x, \hat{\ell}_s \rangle + F(x)$$

- Finding P_t such that $\sum_{a \in \mathcal{A}} P_t(a) a = X_t$
- The first is convex, the second is linear
- But $\mathcal{A}$ is very large!

Drawbacks of FTRL for semibandits

- A reminder about the regret

$$\mathfrak{R}_n = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{n} \langle A_t - a, \ell_t \rangle\right]$$

- An algorithm with sublinear regret can approximate

$$\min_{a \in \mathcal{A}} \sum_{t=1}^{n} \langle a, \ell_t \rangle$$

- Can we derive an efficient algorithm that solves optimization problems of this kind?

Follow the perturbed leader

- **Follow the perturbed leader**
- Regularize with **randomization**

Follow the perturbed leader

- **Follow the perturbed leader**
- Regularize with **randomization**
- Sample $Z_t \in \mathbb{R}^d$ from carefully chosen distribution

$$A_t = \operatorname{argmin}_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \langle a, \hat{\ell}_s \rangle + \langle a, Z_t \rangle$$

Follow the perturbed leader

- **Follow the perturbed leader**
- Regularize with **randomization**
- Sample $Z_t \in \mathbb{R}^d$ from carefully chosen distribution

$$A_t = \text{argmin}_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \langle a, \hat{\ell}_s \rangle + \langle a, Z_t \rangle$$

- You can prove $\mathfrak{R}_n = O(m\sqrt{nd(1 + \log(d))})$

Follow the perturbed leader

- **Follow the perturbed leader**
- Regularize with **randomization**
- Sample $Z_t \in \mathbb{R}^d$ from carefully chosen distribution

$$A_t = \operatorname{argmin}_{a \in \mathcal{A}} \sum_{s=1}^{t-1} \langle a, \hat{\ell}_s \rangle + \langle a, Z_t \rangle$$

- You can prove $\mathfrak{R}_n = O(m\sqrt{nd(1 + \log(d))})$
- Proof is technical, but very nice
- **Main idea** Write algorithm as FTRL **in expectation**

What else is there?

- A lot!
- How to handle non-stationary environments?
- Delays?
- Other structure (convex bandits, infinite action sets, bandits on graphs, kernelizing linear bandits,...)
- Other settings (pure exploration)
- Partial monitoring
- Bayesian methods