30 Combinatorial Bandits

A combinatorial bandit is a linear bandit with a special kind of combinatorial action-set $\mathcal{A} \subset \{0,1\}^d$, which for constant $m \in [d]$ satisfies

$$\mathcal{A} \subseteq \left\{a \in \{0,1\}^d : \|a\|_1 \leq m\right\} .$$

The setting is studied in both adversarial and stochastic models. We focus on the former in this chapter and discuss the latter in the notes. As usual the adversary chooses a sequence of loss vectors $y_1, \ldots, y_n$ with $y_t \in \mathbb{R}^d$ and the expected regret is

$$R_n = \mathbb{E}\left[\max_{a \in \mathcal{A}} \sum_{t=1}^{n} \langle A_t - a, y_t \rangle\right] .$$

In Chapters 27 and 28 we assumed that $y_t \in \mathcal{A}^\circ$, which guarantees that $|\langle A_t, y_t \rangle| \leq 1$ for all t. This restriction is not consistent with the applications we have in mind, so instead we assume that $y_t \in [0,1]^d$, which by the definition of $\mathcal{A}$ ensures that $|\langle A_t, y_t \rangle| \leq m$ for all t. In the standard bandit model the learner observes $\langle A_t, y_t \rangle$ in each round. In many applications for which combinatorial bandits are applied there is actually more information available. The simplest is the **full information** setting where the learner observes the whole vector y_t. The full information setup is interesting, but does not have the flavor of a bandit problem and so we do not discuss it further. There is an inbetween setting where the learner receives **semibandit** feedback, which is the vector $(A_{t1} y_{t1}, \ldots, A_{td} y_{td})$. Since $A_{ti} \in \{0,1\}$ this is equivalent to observing y_{ti} for those i when $A_{ti} = 1$.

30.1 Applications

Shortest path problems

The online shortest path problem is a game over n between adversary and learner. Let $G = (V, E)$ be a fixed graph with a finite set of vertices V and edges $E \subseteq V \times V$. At the beginning of the game the adversary chooses the length of each edge in an arbitrary way. In each round the learner chooses a path between fixed vertices $u, v \in V$ with the goal of travelling the shortest distance. The regret of the learner is the difference between the distance they travelled and

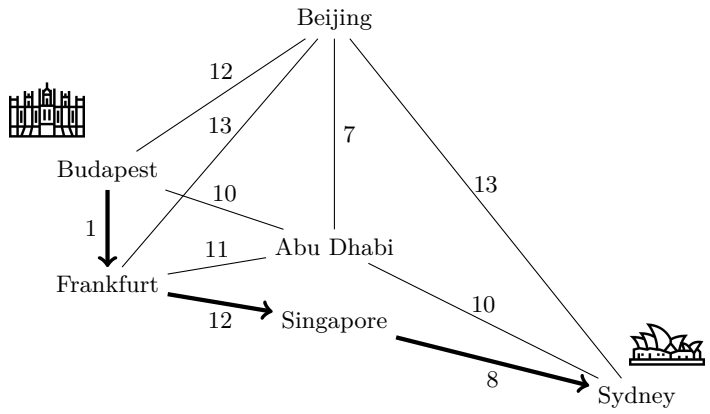

Figure 30.1 Shortest-path problem between Budapest and Sydney. The learner chooses the path Budapest–Frankfurt–Singapore–Sydney. In the bandit setting they observe total travel time (21 hours) while in the semibandit they observe the length of each flight on the route they took (1 hour, 12 hours, 8 hours).

the distance of the optimal path in hindsight. To make things a little formal, let $d = |E|$ and for $t \in [n]$ and $i \in [d]$ let $y_{ti} \in [0,1]$ be the length of the ith path in round t as chosen by the adversary. A path is represented by a vector $a \in \{0,1\}^d$ where $a_i = 1$ if the ith edge is part of the path. Let $\mathcal{A}$ be the set of paths connecting vertices u and v, then the length of path a in round t is $\langle a, y_t \rangle$.

Ranking
Suppose a company has d possible ads they can place and m locations in which to display them. In each round t the learner should choose the m ads to display, which is represented by a vector $A_t \in \{0,1\}^d$ with $\|A_t\|_1 = m$. As before, the adversary chooses $y_t \in [0,1]^d$ that measures the quality of each placement and the learner suffers loss $\langle A_t, y_t \rangle$. This problem could also be called 'selection' because there is no ordering. This is not true in applications like web search where the order of search results is as important as the results themselves. This kind of problem is analyzed in Chapter 32.

Multitask bandits
Consider playing m multi-armed bandits simultaneously, each with K arms. If the losses for each bandit problem are observed, then it is easy to apply Exp3 or Exp3-IX to each bandit independently. But now suppose the learner only observes the sum of the losses. This problem is represented as a combinatorial bandit by letting $d = mK$ and

$$\mathcal{A} = \left\{ a \in \{0,1\}^d : \sum_{i=1}^{K} a_{i+Kj} = 1 \text{ for all } 0 \leq j < m \right\} .$$

This scenario can arise in practice when a company is making multiple independent interventions, but the quality of the interventions are only observed via a change in revenue.

30.2 Bandits

The easiest approach is to apply the version of Exp3 for linear bandits described in Chapter 27. The only difference is that now $|\langle A_t, y_t \rangle|$ can be as large as m, which increases the regret by a factor of m. We leave the proof of the following theorem to the reader (Exercise 30.1).

THEOREM 30.1 *If Algorithm 14 is run on action-set $\mathcal{A}$ with appropriately chosen learning rate, then*

$$R_n \leq 2m\sqrt{3dn \log |\mathcal{A}|} \leq m^{3/2}\sqrt{12dn \log\left(\frac{ed}{m}\right)}\,.$$

There are two computational issues with this approach. First, the action-set is typically so large that finding the core set of the central minimum volume enclosing ellipsoid that determines the exploration distribution of Algorithm 14 is hopeless. Second, sampling from the resulting exponential weights distribution may be highly nontrivial. There is no golden bullet for these issues. We cannot expect the travelling salesman to get easier when done online and with bandit feedback. There are, however, some special cases where efficient algorithms exist and we give some pointers to the literature at the end of the chapter. One modification that greatly eases computation is to replace the Kiefer–Wolfowitz exploration distribution with something more tractable. Let $\pi : \mathcal{A} \to [0, 1]$ be the exploration distribution used by Algorithm 14 and $Q(\pi) = \sum_{a \in \mathcal{A}} \pi(a) a a^\top$. Then the regret of Algorithm 14 satisfies

$$R_n = O\left(m\sqrt{\max_{a \in \mathcal{A}} \|a\|^2_{Q(\pi)^{-1}} n \log(|\mathcal{A}|)}\right)\,.$$

By Kiefer–Wolfowitz (Theorem 21.1) we know that π can be chosen so that $\|a\|^2_{Q(\pi)^{-1}} = d$ and that if $\text{span}(\mathcal{A}) = \mathbb{R}^d$, then this is optimal. In many cases, however, a similar result can be proven for other exploration distributions with more attractive properties computationally.

30.3 Semibandits

The additional information is easily exploited by noting that y_t can now be estimated in each coordinate. Let

$$\hat{Y}_{ti} = \frac{A_{ti} y_{ti}}{P_{ti}}\,, \tag{30.1}$$

where $P_{ti} = \mathbb{E}[A_{ti} \mid \mathcal{F}_{t-1}]$ with $\mathcal{F}_t = \sigma(A_1, Z_1, \ldots, A_t, Z_t)$. An easy calculation shows that $\mathbb{E}[\hat{Y}_{ti} \mid \mathcal{F}_{t-1}] = y_{ti}$.

Input $\mathcal{A}$, η, F
$\bar{A}_1 = \operatorname{argmin}_{a \in \mathcal{A}} F(a)$
for $t = 1, \ldots, n$ **do**
  Choose P_t on $\mathcal{A}$ such that $\sum_{a \in \mathcal{A}} P_t(a) a = \bar{A}_t$
  Sample $A_t \sim P_t$
  Compute $\hat{Y}_{ti} = \dfrac{A_{ti} y_{ti}}{P_{ti}}$ for all $i \in [d]$
  Update $\bar{A}_{t+1} = \operatorname{argmin}_{a \in \operatorname{co}(\mathcal{A})} \eta \langle a, \hat{Y}_t \rangle + D_F(a, \bar{A}_t)$
end for

Algorithm 16: Online stochastic mirror descent for semibandits

THEOREM 30.2 *Let $F : \mathbb{R}^d \to \mathbb{R}$ be the unnormalized negentropy potential defined by*

$$F(a) = \sum_{i=1}^{d} (a_i \log(a_i) - a_i) \; .$$

If Algorithm 16 is run with $\eta = \sqrt{2m(1 + \log(d/m))/(nd)}$, then

$$R_n \leq \sqrt{2nmd(1 + \log(d/m))} \, .$$

Proof Recall from Chapter 28 that for Legendre potentials the optimization problem for $\bar{A}_{t+1}$ can be written in a two-step process:

$$\begin{aligned} \nabla F(\tilde{A}_{t+1}) &= \nabla F(\bar{A}_t) - \eta \hat{Y}_t \\ \bar{A}_{t+1} &= \operatorname{argmin}_{a \in \operatorname{co}(\mathcal{A})} D_F(a, \tilde{A}_{t+1}) \, . \end{aligned}$$

Then by Theorem 28.3 we have

$$R_n \leq \frac{\operatorname{diam}_F(\operatorname{co}(\mathcal{A}))}{\eta} + \frac{1}{\eta} \sum_{t=1}^{n} \mathbb{E}\left[D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t)) \right] \; .$$

The Legendre-Fenchel dual is $F^*(u) = \sum_{i=1}^{d} \exp(u_i)$ and the Bregman divergence with respect to this potential is

$$D_{F^*}(u, v) = \sum_{i=1}^{d} (\exp(u_i) - \exp(v_i)) - \sum_{i=1}^{d} (u_i - v_i) \exp(v_i) \, .$$

Since $\nabla F(a)_i = \log(a_i)$ we have

$$\begin{aligned} D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t)) &= \sum_{i=1}^{d} (\tilde{A}_{t+1,i} - \bar{A}_{ti}) + \sum_{i=1}^{d} \eta \bar{A}_{ti} \hat{Y}_{ti} \\ &= \sum_{i=1}^{d} \bar{A}_{ti} \left(\exp(-\eta \hat{Y}_{ti}) - 1 + \eta \hat{Y}_{ti} \right) \leq \frac{\eta^2}{2} \sum_{i=1}^{d} \bar{A}_{ti} \hat{Y}_{ti}^2 \, . \end{aligned}$$

where the inequality follows from the fact that $\exp(-x) \le 1 - x + x^2/2$ for all $x \ge 0$. Taking the expectation leads to

$$\mathbb{E}\left[\sum_{i=1}^{d} \bar{A}_{ti}\hat{Y}_{ti}^2\right] = \mathbb{E}\left[\sum_{i=1}^{d} \frac{y_{ti}^2 A_{ti}}{\bar{A}_{ti}}\right] \le d\,.$$

The diameter is easily bounded by noting that F is negative in $\mathrm{co}(\mathcal{A})$ and using the Cauchy-Schwartz inequality:

$$\begin{aligned}\mathrm{diam}_F(\mathrm{co}(\mathcal{A})) &= \sup_{a \in \mathrm{co}(\mathcal{A})} \sum_{i=1}^{d} \left(a_i \log(a_i) - a_i + \bar{A}_{1i} + \bar{A}_{1i} \log\left(\frac{1}{\bar{A}_{1i}}\right)\right) \\ &\le m + \sum_{i=1}^{d} \bar{A}_{1i} \log\left(\frac{1}{\bar{A}_{1i}}\right) \le m(1 + \log(d/m))\,.\end{aligned}$$

Putting together the pieces shows that

$$R_n \le \frac{m(1+\log(d/m))}{\eta} + \frac{\eta n d}{2} = \sqrt{2nmd(1+\log(d/m))}\,. \qquad \square$$

Algorithm 16 plays mirror descent on the convex hull of the actions, which has dimension $d-1$. In principle it would be possible to do the same thing on the set of distributions over actions, which has dimension K. Repeating the analysis leads to a suboptimal regret of $O(m\sqrt{dn\log(d/m)})$. We encourage the reader to go through this calculation to see where things go wrong.

Like in Section 30.2, the main problem is computation. There are two challenges: First, in each round the algorithm needs to find a distribution P_t over $\mathcal{A}$ such that $\sum_{a \in \mathcal{A}} P_t(a) = \bar{A}_t$. Feasibility follows from the definition of $\mathrm{co}(\mathcal{A})$ while Carathéodory's theorem proves the support of P_t never needs to be larger than $d+1$. Since $\mathcal{A}$ is finite we can write this problem in terms of linear constraints, but naively the computation complexity is polynomial in K, which is exponential in m. The second difficulty is computing $\bar{A}_{t+1}$ from $\bar{A}_t$ and $\hat{Y}_t$. This is a convex optimization problem, but the computation complexity depends on the representation of $\mathcal{A}$ and may be intractable.

30.4 Follow the perturbed leader

The computational complexity of mirror descent in the previous section can be prohibitively expensive. In this section we describe an efficient algorithm for online combinatorial optimization under the assumption that for all $y \in [0,1]^d$ the optimization problem of finding

$$a = \mathrm{argmin}_{a \in \mathcal{A}} \langle a, y \rangle \tag{30.2}$$

admits an efficient solution. This feels like the minimum one could get away with. If the static problem is too hard it seems unlikely that an online algorithm could be efficient. In fact, an online algorithm with low regret could be used to approximate the solution to the static problem.

So we will try to design an algorithm for which the only nontrivial computation is solving Eq. (30.2). The **follow-the-perturbed leader** (FTPL) algorithm operates by estimating the cumulative losses observed so far. In each round the estimates are perturbed by some random amount and the algorithm solves Eq. (30.2) using the perturbed estimates. Let $\hat{L}_t = \sum_{s=1}^{t} \hat{Y}_t$ be the cumulative loss estimates after round t, then FTPL chooses

$$A_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_t - Z_{t+1} \rangle \,, \tag{30.3}$$

where $\eta > 0$ is the learning rate and $Z_{t+1} \in \mathbb{R}^d$ is a random perturbation sampled from distribution Q to be chosen later. The random perturbations introduce the exploration, which if for appropriate perturbation distributions is sufficient to guarantee small regret. Notice that if η is small, then the effect of Z_{t+1} is larger and the algorithm can be expected to explore more, which is consistent with the learning rate used in mirror descent or exponential weights studied in previous chapters.

We still need to define the loss estimates and perturbation distribution. First we make a connection between this algorithm and mirror descent. Given Legendre potential F with $\operatorname{dom}(\nabla F) = \operatorname{int}(\mathcal{A})$ online stochastic mirror descent chooses $\bar{A}_{t+1}$ so that

$$\bar{A}_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{Y}_t \rangle + D_F(a, \bar{A}_t) \,.$$

Taking derivatives and using the fact that $\operatorname{dom}(\nabla F) = \operatorname{int}(\mathcal{A})$ we have

$$\nabla F(\bar{A}_{t+1}) = \nabla F(\bar{A}_t) - \eta \hat{Y}_t = -\eta \hat{L}_t \,.$$

By duality this implies that $\bar{A}_{t+1} = \nabla F^*(-\eta \hat{L}_t)$ where $F^*(x) = \sup_{a \in \mathcal{A}}(\langle x, a \rangle - F(a))$ is the Fenchel conjugate of F. Examining Eq. (30.3) we see that

$$\bar{A}_{t+1} = \mathbb{E}[A_{t+1} \mid \mathcal{F}_t] = \mathbb{E}\left[\operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_t - Z_{t+1} \rangle\right] \,.$$

If we are to view follow-the-perturbed leader as an instance of mirror descent we must find a Legendre potential F with

$$\nabla F^*(-\eta \hat{L}_t) = \mathbb{E}\left[\operatorname{argmin}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_t - Z \rangle\right] = \mathbb{E}\left[\operatorname{argmax}_{a \in \mathcal{A}} \langle a, Z - \eta \hat{L}_t \rangle\right] \,,$$

which is equivalent to $\nabla F^*(x) = \mathbb{E}[\operatorname{argmax}_{a \in \mathcal{A}} \langle a, x + Z \rangle]$. In order to remove clutter in the notation we define

$$a(x) = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, x \rangle \,.$$

Readers with some familiarity with convex analysis will remember that if $\phi(x) = \max_{a \in \mathcal{A}} \langle a, x \rangle$ is the support function and $\mathcal{A}$ has a smooth boundary, then $\nabla \phi(x) = a(x)$. For combinatorial bandits $\mathcal{A}$ is not smooth, but if Q is

absolutely continuous with respect to the Lebesgue measure, then you will show in Exercise 30.3 that nevertheless it is true that

$$\nabla \mathbb{E}\left[\phi(x+Z)\right] = \mathbb{E}\left[a(x+Z)\right] .$$

All this shows that follow-the-perturbed-leader can be interpreted as mirror descent with potential F defined in terms of its Fenchel dual.

$$F^*(x) = \mathbb{E}\left[\phi(x+Z)\right] . \tag{30.4}$$

There are more reasons for making this connection than mere curiosity. The classical analysis of FTPL is highly probabilistic and involves at least one 'leap of faith' in the analysis. In contrast, the analysis via the mirror descent interpretation is more mechanical. Recall that mirror descent depends on choosing a potential, an exploration distribution and an estimator. The exploration distribution is a distribution P_t on $\mathcal{A}$ such that

$$\bar{A}_t = \sum_{a \in \mathcal{A}} P_t(a) a \,,$$

which in our case is simply defined by

$$P_t(a) = \mathbb{P}(a(Z - \eta \hat{L}_{t-1}) = a \mid \mathcal{F}_{t-1}) \,.$$

It remains to choose the loss estimator. A natural choice would be the same as Eq. (30.1), which is

$$\hat{Y}_{ti} = \frac{A_{ti} y_{ti}}{P_{ti}} \,,$$

where $P_{ti} = \mathbb{P}\left(A_{ti} = 1 \mid \mathcal{F}_{t-1}\right)$. The problem is that

$$P_{ti} = \mathbb{P}(a(Z - \eta \hat{L}_{t-1})_i = 1 \mid \mathcal{F}_{t-1}) \,,$$

which does not generally have a closed form solution. If computation were not an issue, then we could simply estimate P_{ti} for each i by sampling. The trick is to notice that we actually only need to estimate the reciprocal $1/P_{ti}$. Let $X \in \{1, 2, \ldots, \}$ be a geometrically distributed random variable with parameter $\theta \in [0, 1]$ so that

$$\mathbb{P}\left(X = k\right) = (1-\theta)^{k-1} \theta^k \,.$$

An easy calculation shows that $\mathbb{E}[X] = 1/\theta$. Let $K_{it} \in \{1, 2, \ldots\}$ be chosen so that $\mathbb{P}\left(K_{it} = k \mid \mathcal{F}_{t-1}\right) = (1 - P_{it})^{k-1} P_{it}$. Then for constant $\beta > 0$ define

$$\hat{Y}_{ti} = \min\left\{\beta, K_{it} A_{it} y_{it}\right\} .$$

The truncation parameter β is needed to ensure that $\hat{Y}_{ti}$ is never too large, but note that without it we would have

$$\mathbb{E}_{t-1}[K_{it} A_{it} y_{it}] = y_{it} \,.$$

We have now provided all the pieces to define mirror descent. The algorithm is summarized in Algorithm 17.

In Chapter 28 we assumed the loss estimator was unbiased, but this is not necessary as we shall see in the analysis.

Input $\mathcal{A}$, n, η, β, Q
$\hat{L}_{0i} = 0$ for each $i \in [d]$
for $t = 1, \ldots, n$ **do**
  Sample $Z_t \sim Q$
  Compute $A_t = \operatorname{argmax}_{a \in \mathcal{A}} \langle a, \eta \hat{L}_{t-1} - Z_t \rangle$
  For each i with $A_{ti} = 1$ sample $K_{ti} \sim \text{Geometric}(P_{ti})$
  $\hat{Y}_{ti} = \min\{\beta, A_{ti} K_{ti} y_{ti}\}$
  $\hat{L}_{ti} = \hat{L}_{t-1,i} + \hat{Y}_{ti}$
end for

Algorithm 17: Follow-the-Perturbed leader for semibandits

THEOREM 30.3 *Let Q have density $q(z) = 2^{-d} \exp(-\|z\|_1)$ and*

$$\eta = \sqrt{\frac{1 + \log(d)}{nd}} \qquad \beta = \frac{1}{\eta m} .$$

Then the regret of Algorithm 17 is bounded by $R_n \leq 2(m \vee e)\sqrt{nd(1 + \log(d))}$.

Proof First we subtract the bias in the loss estimators and apply Theorem 28.1 to show that

$$\begin{aligned}
R_n(a) &= \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a, y_t \rangle\right] = \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t \rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - a, \hat{Y}_t \rangle\right] + \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t - \hat{Y}_t \rangle\right] \\
&\leq \frac{\operatorname{diam}_F(\mathcal{A})}{\eta} + \mathbb{E}\left[\frac{1}{\eta} \sum_{t=1}^n D_F(\bar{A}_t, \bar{A}_{t+1})\right] + \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - a, y_t - \hat{Y}_t \rangle\right] . \qquad (30.5)
\end{aligned}$$

Of the three terms the diameter is most easily bounded.

$$\begin{aligned}
F(a) &= \sup_{x \in \mathbb{R}^d} (\langle x, a \rangle - F^*(x)) = \sup_{x \in \mathbb{R}^d} (\langle x, a \rangle - \mathbb{E}[\max_{b \in \mathcal{A}} \langle x + Z, b \rangle]) \qquad (30.6) \\
&\geq -\mathbb{E}[\max_{b \in \mathcal{A}} \langle Z, b \rangle] \geq -m \mathbb{E}[\|Z\|_\infty] = -m \sum_{i=1}^d \frac{1}{d} \geq -m(1 + \log(d)) ,
\end{aligned}$$

where the first inequality follows by choosing $x = 0$ and the second follows from Holder's inequality. The last equality is nontrivial and is explained in Exercise 30.2. By the convexity of the maximum function and the fact that Z is centered we also have from Eq. (30.6) that $F(a) \leq 0$, which means that

$$\operatorname{diam}_F(\mathcal{A}) = \max_{a,b \in \mathcal{A}} F(a) - F(b) \leq m(1 + \log(d)) . \qquad (30.7)$$

The next step is to bound the Bregman divergence induced by F. We will shortly show that the Hessian $\nabla^2 F^*(x)$ of F^* exists, so by duality and Taylor's theorem there exists an $\alpha \in [0,1]$ and $\xi = -\eta \hat{L}_{t-1} - \alpha\eta\hat{Y}_t$ such that

$$\begin{aligned} D_F(\bar{A}_t, \bar{A}_{t+1}) &= D_{F^*}(\nabla F(\bar{A}_{t+1}), \nabla F(\bar{A}_t)) \\ &= D_{F^*}(-\eta\hat{L}_{t-1} - \eta\hat{Y}_t, \nabla F(-\eta\hat{L}_{t-1})) = \frac{\eta^2}{2}\|\hat{Y}_t\|^2_{\nabla^2 F^*(\xi)}\,, \end{aligned} \tag{30.8}$$

where the last equality follows from Taylor's theorem (see Theorem 26.4). To calculate the Hessian we use a change of variable to avoid applying the gradient to the non-differentiable argmax.

$$\begin{aligned} \nabla^2 F^*(x) &= \nabla(\nabla F(x)) = \nabla \mathbb{E}\left[a(x+Z)\right] = \nabla \int_{\mathbb{R}^d} a(x+z) f(z) dz \\ &= \nabla \int_{\mathbb{R}^d} a(u) f(u-x) du = \int_{\mathbb{R}^d} a(u)(\nabla f(u-x))^\top du \\ &= \int_{\mathbb{R}^d} a(u)\,\mathrm{sign}(u-x)^\top f(u-x) du = \int_{\mathbb{R}^d} a(x+z)\,\mathrm{sign}(z)^\top f(z) dz\,. \end{aligned}$$

Using the definition of ξ and the fact that $a(x)$ is nonnegative,

$$\begin{aligned} \nabla^2 F^*(\xi)_{ij} &= \int_{\mathbb{R}^d} a(\xi+z)_i\,\mathrm{sign}(z)_j f(z) dz \qquad (30.9) \\ &\le \int_{\mathbb{R}^d} a(\xi+z)_i f(z) dz \\ &= \int_{\mathbb{R}^d} a(z - \eta\hat{L}_{t-1} - \alpha\eta\hat{Y}_t)_i f(z) dz \\ &= \int_{\mathbb{R}^d} a(u - \eta\hat{L}_{t-1})_i f(u + \alpha\eta\hat{Y}_t) du \\ &\le \exp\left(\|\alpha\eta\hat{Y}_t\|_1\right) \int_{\mathbb{R}^d} a(u - \eta\hat{L}_{t-1})_i f(u) du \\ &\le e P_{ti}\,, \qquad (30.10) \end{aligned}$$

where the last inequality follows since $\alpha \in [0,1]$ and $\hat{Y}_{ti} \le \beta = 1/(m\eta)$ and $\hat{Y}_t$ has at most m nonzero entries. Continuing on from Eq. (30.8) we have

$$\frac{\eta^2}{2}\|\hat{Y}_t\|^2_{\nabla^2 F^*(\xi)} \le \frac{e\eta^2}{2}\sum_{i=1}^d P_{ti}\hat{Y}_{ti}\sum_{j=1}^d \hat{Y}_{tj} \le \frac{e\eta}{2}\sum_{i=1}^d P_{ti}\hat{Y}_{ti} \le \frac{e\eta}{2}\sum_{i=1}^d P_{ti}K_{ti}\,.$$

Chaining together the parts and taking the expectation shows that

$$\mathbb{E}[D_F(\bar{A}_t, \bar{A}_{t+1})] \le \frac{e\eta}{2}\mathbb{E}\left[\sum_{i=1}^d P_{ti}K_{ti}\right] = \frac{e\eta}{2}\mathbb{E}\left[\sum_{i=1}^d P_{ti}\mathbb{E}[K_{ti} \mid \mathcal{F}_{t-1}]\right] = \frac{ed\eta}{2}\,.$$

The last step is to control the bias term.

$$\mathbb{E}\left[\sum_{t=1}^{n}\langle \bar{A}_t - a, y_t - \hat{Y}_t\rangle\right] \le \mathbb{E}\left[\sum_{t=1}^{n}\langle \bar{A}_t, y_t - \hat{Y}_t\rangle\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{i=1}^{d} P_{ti}\left(y_{ti} - \min\left\{P_{ti}\beta, y_{ti}\right\}\right)\right] \le \frac{dn}{2\beta} = \frac{dnm\eta}{2}\,.$$

Putting together all the pieces into Eq. (30.5) leads to

$$R_n \le \frac{m(1+\log(d))}{\eta} + \frac{endm\eta}{2} + \frac{dnm\eta}{2} \le 2(m \vee e)\sqrt{nd(1+\log(d))}\,. \qquad \square$$

30.5 Notes

1 For a long time it was speculated that the dependence of the regret on $m^{3/2}$ in Theorem 30.1 might be improvable to m. Very recently, however, the lower bound was increased to show the upper bound is tight [Cohen et al., 2017]. For semibandits the worst case lower bound is $\Omega(\sqrt{dnm})$ (Exercise 30.5), which is matched up to constant factors by online stochastic mirror descent with a different potential (Exercise 30.4).

2 Algorithm 17 needs to sample K_{ti} for each i with $A_{ti} = 1$. The conditional expected running time for this is A_{ti}/P_{ti}, which has expectation 1. It follows that the expected running time over the whole n rounds is $O(nd)$ calls to the oracle linear optimization algorithm. It can happen that the algorithm is unlucky and chooses $A_{ti} = 1$ for some i with P_{ti} quite small. To avoid catastrophic slowdowns it is possible to truncate the sampling procedure by defining $\tilde{K}_{ti} = \min\{K_{ti}, M\}$ for M suitably large. This introduces a small controllable bias [Neu, 2015a].

3 While FTPL is excellent in the face of semibandit information, we do not know of a general result for the bandit model. The main challenge is controlling the variance of the least squares estimator without using a sophisticated exploration distribution like what is provided by Kiefer–Wolfowitz.

4 Combinatorial bandits can also be studied in a stochastic setting. There are several ways to do this. The first mirrors our assumptions for stochastic linear bandits in Chapter 19 where the loss (more commonly reward) is defined by

$$X_t = \langle A_t, \theta\rangle + \eta_t\,, \tag{30.11}$$

where $\theta \in \mathbb{R}^d$ is fixed and unknown and η_t is the noise on which statistical assumptions are made (for example, conditionally 1-subgaussian). There are at least two alternatives. Suppose that $\theta_1, \ldots, \theta_n$ are sampled independently from some multivariate distribution and define the reward by

$$X_t = \langle A_t, \theta_t\rangle\,. \tag{30.12}$$

This latter version has 'parameter noise' and is more closely related to the

adversarial setup studied in this chapter. Finally, one can assume additionally that the distribution of θ_t is a product distribution so that $(\theta_{1i})_{i=1}^d$ are also independent.

5 For some action-sets the off-diagonal elements of the Hessian in Eq. (30.9) are negative, which improves the dependence on m to just $\sqrt{m}$. An example where this occurs is when $\mathcal{A} = \{a \in \{0,1\}^d : \|a\|_1 = m\}$. Let $i \neq j$ and suppose that $z, \xi \in \mathbb{R}^d$ and $z_j \geq 0$. Then you can check that $a(z+\xi)_i \leq a(z - 2z_j e_j + \xi)_i$ and so

$$\begin{aligned}
\nabla^2 F^*(\xi)_{ij} &= \int_{\mathbb{R}^d} a(z+\xi)_i \operatorname{sign}(z)_j f(z) dz \\
&= \int_{\mathbb{R}^{d-1}} \int_0^\infty (a(z+\xi)_i - a(z - 2z_j e_j + \xi)_i) f(z) dz_j dz_{-j} \\
&\leq 0\,,
\end{aligned}$$

where dz_{-j} is shorthand for $dz_1 dz_2, \ldots dz_{j-1} dz_{j+1}, \ldots, dz_d$. You are asked to complete all the details in Exercise 30.6. This result unfortunately does not hold for every action-set (Exercise 30.7).

30.6 Bibliographic remarks

The online combinatorial bandit was introduced by Cesa-Bianchi and Lugosi [2012] where the most comprehensive list of known applications for which computation is efficient is given. The analysis presented in Section 30.2 is due to Bubeck and Cesa-Bianchi [2012]. While computational issues remain in the bandit problem, there has been some progress in certain settings [Combesd et al., 2015]. The full information setting has been studied quite extensively [Koolen et al., 2010, and references from/to]. The follow-the-perturbed-leader algorithm was first proposed by Hannan [1957], rediscovered by Kalai and Vempala [2005a,b] and generalized by Poland [2005], Hutter and Poland [2005]. Poland [2005] showed a near-optimal regret for finite-armed adversarial bandits while for combinatorial settings suboptimal rates have been shown by Awerbuch and Kleinberg [2004], McMahan and Blum [2004]. Semibandits seem to have been introduced in the context of shortest-path problems by György et al. [2007]. The general setup and algorithmic analysis of FTPL presented follows the work by Neu [2015a] who also had the idea to estimate the inverse probabilities via a geometric random variable. Our analysis based on mirror descent improves the regret by a factor of $\sqrt{m}$. As far as we know this has not been seen in the literature on combinatorial bandits before, but the approach is heavily inspired by Abernethy et al. [2014] who present the core ideas in the prediction with expert advice setting, Cohen and Hazan [2015] in the combinatorial full information case and Abernethy et al. [2015] for finite-armed bandits. The literature on stochastic combinatorial semibandits is also quite large with algorithms and analysis in the frequentist [Gai et al., 2012, Combesd et al., 2015, Kveton et al., 2015b] and Bayesian settings [Wen et al.,

2015, Russo and Van Roy, 2016]. These works focus on the case where the reward is given by Eq. (30.12) and the components of θ_t are independent. When the reward is given by Eq. (30.11) one can use the tools for stochastic linear bandits developed in Part V.

30.7 Exercises

30.1 Prove Theorem 30.1.

30.2 Let Z be sampled from measure on $\mathbb{R}^d$ with density $f(z) = 2^{-d}\exp(-\|z\|_1)$. The purpose of this exercise is to show that

$$\mathbb{E}[\|Z\|_\infty] = \sum_{i=1}^{d} \frac{1}{i}\,. \tag{30.13}$$

Recall that an exponential with rate λ has density $f(x) = \lambda\exp(-\lambda x)\mathbb{1}_{x\geq 0}$.

(a) Let X be exponential with rate λ. Show that $\mathbb{E}[X] = 1/\lambda$.
(b) Let $X_1, \ldots, X_i$ be independent and exponentially distributed with rate 1. Show that $M = \min_{j\in[i]} X_j$ is exponentially distributed with rate i.
(c) Show that $\|Z\|_\infty$ has the same law as the maximum of d independent standard exponentials.
(d) Let $M_1, \ldots, M_d$ be independent exponentially distributed random variables where M_i has rate i. Show that Z has the same law as $\sum_{i=1}^d M_i$.
(e) Show that Eq. (30.13) holds.

30.3 Let $\mathcal{A} \subset \mathbb{R}^d$ be a compact convex set and $\phi(x) = \max_{a\in\mathcal{A}}\langle a, x\rangle$ its support function. Let Q be a measure on $\mathbb{R}^d$ that is absolutely continuous with respect to the Lebesgue measure and let $Z \sim Q$. Show that

$$\nabla\mathbb{E}[\phi(x+Z)] = \mathbb{E}\left[\operatorname{argmax}_{a\in\mathcal{A}}\langle a, x+Z\rangle\right]\,.$$

30.4 Adapt the analysis in Exercise 28.10 to derive an algorithm for combinatorial bandits with semibandit feedback for which the regret is $R_n \leq C\sqrt{mdn}$ for universal constant $C > 0$.

30.5 Let $m \geq 1$ and that $d = km$ for some $k > 1$. Prove that for any algorithm there exists a combinatorial semibandit such that $R_n \geq c\min\{nm, \sqrt{mdn}\}$ where $c > 0$ is a universal constant.

The most obvious choice is to choose $\mathcal{A} = \{a \in \{0,1\}^d : \|a\|_1 = m\}$, which are sometimes called m-sets. A lower bound does hold for this action-set [Lattimore et al., 2018]. However an easier path is to impose a little additional structure and analyze the multitask bandit setting.

30.6 Use the ideas in Note 5 to prove that FTPL has $R_n = \tilde{O}(\sqrt{mnd})$ regret when $\mathcal{A} = \{a \in \{0,1\}^d : \|a\|_1 = m\}$.

After proving the off-diagonal elements of the Hessian are negative you will also need to tune the learning rate. We do not know of a source for this result, but the full information case was studied by Cohen and Hazan [2015].

30.7 Construct an action-set and $i \neq j$ and $z, \xi \in \mathbb{R}^d$ with $z_j > 0$ such that $a(z + \xi)_i \geq a(z - 2z_j e_j + \xi)_i$.