28 Follow the Regularized Leader and Mirror Descent

In the last chapter we showed that if $\mathcal{A} \subset \mathbb{R}^d$ has K elements, then the regret of Exp3 with a careful exploration distribution has regret

$$R_n = O(\sqrt{dn\log(K)})\,.$$

When $\mathcal{A}$ is a convex set we also showed the continuous version of this algorithm has regret at most

$$R_n = O(d\sqrt{n\log(n)})\,.$$

Although this algorithm runs in polynomial time, the degree is high and the implementation is complicated and impractical. In many cases this can be improved upon, both in terms of the regret and computation. One of the main results of this chapter is a proof that when $\mathcal{A}$ is the unit ball, then there is an efficient algorithm for which the regret is $R_n = O(\sqrt{dn\log(n)})$. More importantly, however, we introduce a pair of related algorithms called **follow the regularized leader** and **mirror descent**, which have proven to be powerful and flexible tools for the design and analysis of bandit algorithms.

28.1 Online linear optimization

Mirror descent originated in the convex optimization literature. The idea has since been adapted to online learning and specifically to online linear optimization. Online linear optimization is the full information version of the adversarial linear bandit where at the end of each round the learner observes the full vector y_t. Let $\mathcal{A} \subset \mathbb{R}^d$ be a convex set and $\mathcal{L} \subset \mathbb{R}^d$ be an arbitrary set called the **loss space**. Let $y_1, \dots, y_n$ be a sequence of loss vectors with $y_t \in \mathcal{L}$ for all $t \in [n]$. In each round the learner chooses $a_t \in \mathcal{A}$ and subsequently observes y_t. The regret relative to a fixed comparator $a \in \mathcal{A}$ is

$$R_n(a) = \sum_{t=1}^n \langle a_t - a, y_t\rangle$$

and the regret is $R_n = \max_{a\in\mathcal{A}} R_n(a)$. We emphasize that the only difference relative to the adversarial linear bandit is that now y_t is observed rather than $\langle a_t, y_t\rangle$. Actions are not capitalized in this section because the algorithms presented here do not randomize.

Mirror descent

The basic version of mirror descent has two extra parameters beyond n and $\mathcal{A}$. A learning rate $\eta > 0$ and a Legendre function $F : \mathbb{R}^d \to \bar{\mathbb{R}}$ with domain $\mathcal{D} = \operatorname{dom}(F)$. The function F is usually called a **potential function** or **regularizer**. In the first round mirror descent predicts

$$a_1 = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} F(a)\,, \tag{28.1}$$

In subsequent rounds it predicts

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \left(\eta \langle a, y_t \rangle + D_F(a, a_t)\right)\,, \tag{28.2}$$

where $D_F(a, a_t)$ is the F-induced Bregman divergence between a and a_t. The minimization in (28.2) is over $\mathcal{A} \cap \mathcal{D}$ because for $a_t \in \operatorname{int}(\mathcal{D}) \subseteq \operatorname{dom}(\nabla F)$ the domain of $D_F(\cdot, a_t)$ is the same as that of F.

Follow the regularized leader

Like mirror descent, follow the regularized leader depends on a Legendre potential F with domain $\mathcal{D} = \operatorname{dom}(F)$ and predicts $a_1 = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} F(a)$. In subsequent rounds it predicts

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \left(\eta \sum_{s=1}^{t} \langle a, y_s \rangle + F(a)\right)\,. \tag{28.3}$$

The intuition is that the algorithm chooses a_{t+1} to be the action that performed best in hindsight with respect to the regularized loss. As for mirror descent, the regularization serves to stabilize the algorithm, which turns out to be a key property of good algorithms for online linear prediction.

Another algorithm is **follow the leader**, which chooses the action that appears best in hindsight, $a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \sum_{s=1}^{t} \langle a, y_s \rangle$. In general, this algorithm is not well suited for online linear optimization because the absence of regularization makes for an unstable algorithm that can lead to extremely poor performance as you will show in Exercise 28.2.

Equivalence of mirror descent and follow the regularized leader

At first sight these algorithms do not look that similar. To clarify matters let us suppose that F has domain $\mathcal{D} \subseteq \mathcal{A}$. We now show that in this setting mirror descent and follow the regularized leader are identical. Let

$$\Phi_t(a) = \eta \langle a, y_t \rangle + D_F(a, a_t) = \eta \langle a, y_t \rangle + F(a) - F(a_t) - \langle \nabla F(a_t), a - a_t \rangle\,.$$

Now mirror descent chooses a_{t+1} to minimize Φ_t. The reader should check that the assumption that F is Legendre on domain $\mathcal{D} \subseteq \mathcal{A}$ implies that the minimizer occurs in the interior of $\mathcal{D} \subseteq \mathcal{A}$ and that $\nabla \Phi_t(a_{t+1}) = 0$ (see Exercise 28.1). This means that $\eta y_t = \nabla F(a_t) - \nabla F(a_{t+1})$ and so

$$\nabla F(a_{t+1}) = -\eta y_t + \nabla F(a_t) = \nabla F(a_1) - \eta \sum_{s=1}^t y_s = -\eta \sum_{s=1}^t y_s \,,$$

where the last equality is true because a_1 is chosen as the minimizer of F in $\mathcal{A} \cap \mathcal{D} = \mathcal{D}$ and again the fact that F is Legendre ensures this minimum occurs at an interior point where the gradient vanishes. Follow the regularized leader chooses a_{t+1} to minimize $\Phi'_t(a) = \eta \sum_{s=1}^t \langle a, y_s \rangle + F(a)$. The same argument shows that $\nabla \Phi'_t(a_{t+1}) = 0$, which means that

$$\nabla F(a_{t+1}) = -\eta \sum_{s=1}^t y_s \,.$$

The last two displays and the fact that the gradient for Legendre functions is invertible shows that mirror descent and follow the regularized leader are the same in this setting.

The equivalence between these algorithms is far from universal. First of all, it does not hold when F is not Legendre or its domain is larger than $\mathcal{A}$. Second, in many applications of these algorithms the learning rate or potential change with time and in either case the algorithms will typically produce different action sequences. For example, if a learning rate η_t is used rather than η in the definition of Φ_t, then mirror descent chooses $\nabla F(a_{t+1}) = -\sum_{s=1}^t \eta_s y_s$ while follow the regularized leader chooses $\nabla F(a_{t+1}) = -\eta_t \sum_{s=1}^t y_s$. We return to this issue in the notes and exercises.

EXAMPLE 28.1 Let $\mathcal{A} = \mathbb{R}^d$ and $F(a) = \frac{1}{2}\|a\|_2^2$. Then $\nabla F(a) = a$ and $D(a, a_t) = \frac{1}{2}\|a - a_t\|_2^2$. Clearly F is Legendre and $\mathcal{D} = \mathcal{A}$ so mirror descent and follow the regularized leader are the same. By simple calculus we see that

$$a_{t+1} = \operatorname{argmin}_{a \in \mathbb{R}^d} \eta \langle a, y_t \rangle + \frac{1}{2}\|a - a_t\|_2^2 = a_t - \eta y_t \,,$$

which may be familiar as **online gradient descent**.

EXAMPLE 28.2 Let $\mathcal{A}$ be a compact subset of $\mathbb{R}^d$ and $F(a) = \frac{1}{2}\|a\|_2^2$. Then mirror descent chooses

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \eta \langle a, y_t \rangle + \frac{1}{2}\|a - a_t\|_2^2 = \Pi(a_t - \eta y_t) \,, \tag{28.4}$$

where $\Pi(a)$ is the Euclidean projection of a onto $\mathcal{A}$. This algorithm is usually called online projected gradient descent. On the other hand, for follow the regularized leader we have

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A}} \eta \sum_{s=1}^t \langle a, y_s \rangle + \frac{1}{2}\|a - a_t\|_2^2 = \Pi\left(-\eta \sum_{s=1}^t y_s\right) \,,$$

which may be a different choice than mirror descent.

EXAMPLE 28.3 The exponential weights algorithm that appeared on numerous occasions in earlier chapters is a special case of mirror descent corresponding to choosing the constraint set $\mathcal{A}$ as the simplex in $\mathbb{R}^d$ and choosing F to be the unnormalized negentropy function of Example 26.2.

A two-step process for implementation

Solving the optimization problem in Eq. (28.2) is often made easier by using the results in Section 26.6 of Chapter 26. Let $\mathcal{D}^* = \nabla F(\mathcal{D})$ and suppose the following condition holds:

$$\nabla F(x) - \eta y \in \mathcal{D}^* \text{ for all } x \in \mathcal{A} \cap \mathcal{D} \text{ and } y \in \mathcal{L}\,. \tag{28.5}$$

Then solution to Eq. (28.2) can be found using the following two-step procedure.

$$\tilde{a}_{t+1} = \operatorname{argmin}_{a \in \mathcal{D}} \eta \langle a, y_t \rangle + D_F(a, a_t) \qquad \text{and} \tag{28.6}$$

$$a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} D_F(a, \tilde{a}_{t+1})\,. \tag{28.7}$$

Eq. (28.5) means the first optimization problem can be evaluated explicitly as the solution to

$$\eta y_t + \nabla F(\tilde{a}_{t+1}) - \nabla F(a_t) = 0\,. \tag{28.8}$$

Since F is Legendre this means that $\tilde{a}_{t+1} = (\nabla F)^{-1}(\nabla F(a_t) - \eta y_t)$. The optimization problem in Eq. (28.7) is usually harder to calculate analytically, but there are important exceptions as we shall see.

The condition in Eq. (28.5) holds for all choices of potential and losses in this book.

28.2 Regret analysis

We now analyze the regret of mirror descent. The theorem has two parts, the first of which is strictly stronger by a small margin than the second. To minimize clutter we abbreviate D_F by D.

THEOREM 28.1 *Let $\eta > 0$ and F be Legendre with domain $\mathcal{D}$ and $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D})$. Let $a_1, \dots, a_{n+1}$ be the actions chosen by mirror descent. Then for any $a \in \mathcal{A}$ the regret of mirror descent is bounded by:*

$$R_n(a) \le \sum_{t=1}^{n} \langle a_t - a_{t+1}, y_t \rangle + \frac{F(a) - F(a_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^{n} D(a_{t+1}, a_t)\,.$$

Furthermore, suppose that Eq. (28.5) holds and $\tilde{a}_2, \tilde{a}_3, \dots, \tilde{a}_{n+1}$ are given by Eq. (28.6), then

$$R_n(a) \leq \frac{1}{\eta}\left(F(a) - F(a_1) + \sum_{t=1}^n D(a_t, \tilde{a}_{t+1})\right) .$$

Proof of Theorem 28.1 For the first part we split the inner product:

$$\langle a_t - a, y_t \rangle = \langle a_t - a_{t+1}, y_t \rangle + \langle a_{t+1} - a, y_t \rangle .$$

Using the fact that $a_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \eta \langle a, y_t \rangle + D(a, a_t)$ and the first order optimality conditions shows that

$$\langle \eta y_t + \nabla F(a) - \nabla F(a_t), a - a_{t+1} \rangle \geq 0 .$$

By the definition of the Bregman divergence we have

$$\begin{aligned}\langle a_{t+1} - a, y_t \rangle &\leq \frac{1}{\eta} \langle \nabla F(a) - \nabla F(a_t), a - a_{t+1} \rangle \\ &= \frac{1}{\eta} \left(D(a, a_t) - D(a, a_{t+1}) - D(a_{t+1}, a_t)\right) . \qquad (28.9)\end{aligned}$$

Using this, along with the definition of the regret,

$$\begin{aligned}R_n &= \sum_{t=1}^n \langle a_t - a, y_t \rangle \\ &\leq \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{1}{\eta} \sum_{t=1}^n \left(D(a, a_t) - D(a, a_{t+1}) - D(a_{t+1}, a_t)\right) \\ &= \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{1}{\eta}\left(D(a, a_1) - D(a, a_{n+1}) - \sum_{t=1}^n D(a_{t+1}, a_t)\right) \\ &\leq \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle + \frac{F(a) - F(a_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^n D(a_{t+1}, a_t) ,\end{aligned}$$

where the final inequality follows from the fact that $D(a, a_{n+1}) \geq 0$ and $D(a, a_1) \leq F(a) - F(a_1)$, which is true by the first-order optimality conditions for $a_1 = \operatorname{argmin}_{b \in \mathcal{A}} F(b)$. To see the second part note that

$$\begin{aligned}\langle a_t - a_{t+1}, y_t \rangle &= \frac{1}{\eta} \langle a_t - a_{t+1}, \nabla F(a_t) - \nabla F(\tilde{a}_{t+1}) \rangle \\ &= \frac{1}{\eta} \left(D(a_{t+1}, a_t) + D(a_t, \tilde{a}_{t+1}) - D(a_{t+1}, \tilde{a}_{t+1})\right) \\ &\leq \frac{1}{\eta} \left(D(a_{t+1}, a_t) + D(a_t, \tilde{a}_{t+1})\right) .\end{aligned}$$

The result follows by substituting this into Eq. (28.9) and completing as for the first part. □

The assumption that a_1 minimizes the potential was only used to bound $D(a, a_1) \leq F(a) - F(a_1)$. For a different initialization the following bound still holds:

$$R_n(a) \leq \frac{1}{\eta}\left(D(a, a_1) + \sum_{t=1}^{n} D(a_t, \tilde{a}_{t+1})\right). \tag{28.10}$$

As we shall see in Chapter 31, this is useful when using mirror descent to analyze nonstationary bandits.

The first part of Theorem 28.1 also holds for follow the regularized leader as stated in the next result, the proof of which is left for Exercise 28.6.

THEOREM 28.2 *Let $\eta > 0$ and F be Legendre with domain $\mathcal{D}$ and $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D})$. Then for any $a \in \mathcal{A}$ the regret of follow the regularized leader is bounded by*

$$R_n(a) \leq \sum_{t=1}^{n} \langle a_t - a_{t+1}, y_t \rangle + \frac{F(a) - F(a_1)}{\eta} - \frac{1}{\eta}\sum_{t=1}^{n} D(a_{t+1}, a_t).$$

We now give two examples of how to apply Theorem 28.1. Let $\operatorname{diam}_F(\mathcal{A}) = \max_{a,b \in \mathcal{A}} F(a) - F(b)$ be the diameter of $\mathcal{A}$ with respect to F.

PROPOSITION 28.1 *Let $\mathcal{A} = B_2^d = \{a \in \mathbb{R}^d : \|a\|_2 \leq 1\}$ be the standard unit ball and assume $y_t \in B_2^d$ for all t. Then mirror descent with potential $F(a) = \frac{1}{2}\|a\|_2^2$ and $\eta = \sqrt{1/n}$ satisfies $R_n \leq \sqrt{n}$.*

Proof By Eq. (28.8) we have $\tilde{a}_{t+1} = a_t - \eta y_t$ so

$$D(a_t, \tilde{a}_{t+1}) = \frac{1}{2}\|\tilde{a}_{t+1} - a_t\|_2^2 = \frac{\eta^2}{2}\|y_t\|_2^2.$$

Therefore since $\operatorname{diam}_F(\mathcal{A}) = 1/2$ and $\|y_t\|_2 \leq 1$ for all t,

$$R_n \leq \frac{\operatorname{diam}_F(\mathcal{A})}{\eta} + \frac{\eta}{2}\sum_{t=1}^{n}\|y_t\|_2^2 \leq \frac{1}{2\eta} + \frac{\eta n}{2} = \sqrt{n}. \qquad \square$$

PROPOSITION 28.2 *Let $\mathcal{A} = \{a \in [0,1]^d : \sum_{i=1}^d a_i = 1\}$ be the probability simplex and $y_t \in \mathcal{A}^\circ$ for all t. Then mirror descent with the unnormalized negentropy potential and $\eta = \sqrt{2\log(d)/n}$ satisfies $R_n \leq \sqrt{2n\log(d)}$.*

Proof The Bregman divergence with respect to the unnormalized negentropy potential for $a, b \in \mathcal{A}$ is $D(a,b) = \sum_{i=1}^d a_i \log(a_i/b_i)$. Therefore

$$\begin{aligned} R_n(a) &\le \frac{F(a)-F(a_1)}{\eta} + \sum_{t=1}^n \langle a_t - a_{t+1}, y_t \rangle - \frac{1}{\eta} \sum_{t=1}^n D(a_{t+1}, a_t) \\ &\le \frac{\log(d)}{\eta} + \sum_{t=1}^n \|a_t - a_{t+1}\|_1 \|y_t\|_\infty - \frac{1}{\eta} \sum_{t=1}^n \frac{1}{2} \|a_t - a_{t+1}\|_1^2 \\ &\le \frac{\log(d)}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \|y_t\|_\infty^2 \le \frac{\log(d)}{\eta} + \frac{\eta n}{2} = \sqrt{2n \log(d)}\,. \end{aligned}$$

where the first inequality follows from Theorem 28.1, the second from Pinsker's inequality and the facts that $\operatorname{diam}_F(\mathcal{A}) = \log(d)$. In the third inequality we used that fact that $ax - bx^2/2 \le a^2/(2b)$ for all x. The last inequality follows from the assumption that $\|y_t\|_\infty \le 1$. □

The last few steps in the above proof are so routine that we summarize their use in a corollary, the proof of which we leave to the reader (Exercise 28.3).

COROLLARY 28.1 *Let F be a Legendre potential and $\|\cdot\|_t$ be a norm on $\mathbb{R}^d$ for each $t \in [n]$ such that $D_F(a_{t+1}, a_t) \ge \frac{1}{2}\|a_{t+1} - a_t\|_t^2$. Then the regret of mirror descent or follow the regularized leader satisfies*

$$R_n \le \frac{\operatorname{diam}_F(\mathcal{A})}{\eta} + \frac{\eta}{2} \sum_{t=1}^n \left(\|y_t\|_t^*\right)^2 ,$$

where $\|y\|_t^* = \max_{x : \|x\|_t \le 1} \langle x, y \rangle$ is the ***dual norm*** *of $\|\cdot\|_t$.*

It often happens that the easiest way to bound the regret of mirror descent is to find a norm that satisfies the conditions of Corollary 28.1. To illustrate a suboptimal application of mirror descent and this result, suppose we had chosen $F(a) = \frac{1}{2}\|a\|_2^2$ in the setting of Proposition 28.2. Then $D_F(a_{t+1}, a_t) = \frac{1}{2}\|a_{t+1} - a_t\|^2$ suggests choosing $\|\cdot\|_t$ to be the standard Euclidean norm. Since $\operatorname{diam}_F(\mathcal{A}) = 1/2$ and $\|\cdot\|_2^* = \|\cdot\|_2$, applying Corollary 28.1 shows that

$$R_n \le \frac{1}{2\eta} + \frac{\eta}{2} \sum_{t=1}^n \|y_t\|_2^2\,.$$

But now we see that $\|y_t\|_2^2$ can be as large as d and tuning η would lead to a rate of $O(\sqrt{nd})$ rather than $O(\sqrt{n \log(d)})$.

Both theorems were presented for the oblivious case where (y_t) are chosen in advance. This assumption was not used, however, and in fact the bounds in this section continue to hold when y_t are chosen strategically as a function of $y_1, x_1, \ldots, y_{t-1}, x_t$. This is analogous to how the basic regret bound for exponential weights continues to hold in the face of strategic losses. But be cautioned, this result does not carry immediately to the application of mirror descent to bandits as discussed at the end in Note 7.

28.3 Online learning for bandits

We now consider the application of mirror descent to bandit problems. Like in the previous chapter the adversary chooses a sequence of vectors $y_1, \dots, y_n$ with $y_t \in \mathcal{L} \subset \mathbb{R}^d$. In each round the learner chooses $A_t \in \mathcal{A} \subset \mathbb{R}^d$ where $\mathcal{A}$ is convex and observes $\langle A_t, y_t \rangle$. The regret relative to action a is

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

The regret is $R_n = \max_{a \in \mathcal{A}} R_n(a)$. The application of mirror descent and follow the regularized leader to linear bandits requires two new ideas. First, because the learner only observes $\langle A_t, y_t \rangle$, the loss vectors need to be estimated from data and it is these estimators that will be used by the mirror descent algorithm. Because estimation of y_t is only possible using randomization, the algorithm cannot play the suggested action of mirror descent, but instead plays a distribution over actions with the same mean as the proposed action. Since the losses are linear, the expected additional regret by playing according to the distribution vanishes. The algorithm is summarized in Algorithm 15. Notice we have switched to capital letters because of the randomization.

```
Input  Legendre potential $F$ with domain $\mathcal{D}$, action set $\mathcal{A}$ and learning rate $\eta > 0$
Choose $\bar{A}_1 = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} F(a)$
for $t = 1, \dots, n$ do
    Choose measure $P_t$ on $\mathcal{A}$ with mean $\bar{A}_t$
    Sample action $A_t$ from $P_t$ and observe $\langle A_t, y_t \rangle$
    Compute estimate $\hat{Y}_t$
    Let $\bar{A}_{t+1} = \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} \eta \langle a, \hat{Y}_t \rangle + D_F(a, \bar{A}_t)$
end for
```

Algorithm 15: Online stochastic mirror descent

THEOREM 28.3 *Provided that $\mathbb{E}[\hat{Y}_t \mid \bar{A}_t] = y_t$ and $a \in \mathcal{A}$ and $\mathcal{A} \subseteq \mathrm{cl}(\mathcal{D})$, then*

$$R_n(a) \leq \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - \bar{A}_{t+1}, \hat{Y}_t \rangle + \frac{F(a) - F(\bar{A}_1)}{\eta} - \frac{1}{\eta} \sum_{t=1}^n D(\bar{A}_{t+1}, \bar{A}_t)\right] .$$

Furthermore, letting $\tilde{A}_{t+1} = \operatorname{argmin}_{a \in \mathcal{D}} \eta \langle a, \hat{Y}_t \rangle + D_F(a, \bar{A}_t)$ and assuming that $\eta \hat{Y}_t + \nabla F(x) \in \mathcal{D}^*$ for all $x \in \mathcal{A}$ almost surely, then

$$R_n \leq \frac{\mathrm{diam}_F(\mathcal{A})}{\eta} + \frac{1}{\eta} \sum_{t=1}^n \mathbb{E}\left[D(\bar{A}_t, \tilde{A}_{t+1})\right] .$$

Proof Using the definition of the algorithm and the assumption that $\hat{Y}_t$ is unbiased given $\bar{A}_t$ and P_t has mean $\bar{A}_t$ leads to

$$\mathbb{E}\left[\langle A_t, y_t\rangle\right] = \mathbb{E}\left[\langle \bar{A}_t, y_t\rangle\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \bar{A}_t, y_t\rangle \mid \bar{A}_t\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \bar{A}_t, \hat{Y}_t\rangle \,\middle|\, \bar{A}_t\right]\right],$$

where the last equality used the linearity of expectations. Hence,

$$R_n(x) = \mathbb{E}\left[\sum_{t=1}^n \langle A_t, y_t\rangle - \langle x, y_t\rangle\right] = \mathbb{E}\left[\sum_{t=1}^n \langle \bar{A}_t - x, \hat{Y}_t\rangle\right],$$

which is the expected random regret of mirror descent on the recursively constructed sequence $\hat{Y}_t$. The result follows from Theorem 28.1 and the note at the end of the last chapter that says that this theorem continues to hold even for recursively constructed sequences. □

The same style of proof also works for follow the regularized leader.

28.4 The unit ball

In the previous chapter we showed that continuous exponential weights on the unit ball has a regret of

$$R_n = O(d\sqrt{n\log(n)})\,.$$

The reader now knows that this is a version of mirror descent with the negentropy potential. Somewhat surprisingly, the dependence on the dimension can be reduced to $\sqrt{d}$ using a more carefully chosen potential. For the remainder of this section let $\mathcal{A} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ be the standard Euclidean ball. In order to instantiate the mirror descent algorithm for bandits we need a potential, sampling rule, an unbiased estimator and a learning rate. We start with the sampling rule and estimator. Recall that in round t we need to choose a distribution on $\mathcal{A}$ with mean $\bar{A}_t$ and sufficient variability that the variance of the estimator is not too large. Let $E_t \in \{0, 1\}$ satisfy $\mathbb{E}_t[E_t] = (1 - \|\bar{A}_t\|)$ and U_t be an independent and uniformly distributed on $\{\pm e_1, \ldots, \pm e_d\}$. Then let

$$A_t = E_t U_t + \frac{(1 - E_t)\bar{A}_t}{\|\bar{A}_t\|}\,.$$

Then the sampling distribution is just the law of A_t, which clearly satisfies $\mathbb{E}_t[A_t] = \bar{A}_t$. For the estimator we use a variant of the importance-weighted estimator from the last chapter:

$$\hat{Y}_t = \frac{dE_t A_t \langle A_t, y_t\rangle}{1 - \|\bar{A}_t\|}\,. \tag{28.11}$$

The reader can check for themselves that this estimator is unbiased. Next we inspect the contents of our magicians hat and select the potential

$$F(a) = -\log\left(1 - \|a\|\right) - \|a\|\,.$$

There is one more modification. Rather than instantiating mirror descent with action-set $\mathcal{A}$, we use $\tilde{\mathcal{A}} = \{x \in \mathbb{R}^d : \|x\|_2 \le r\}$ where $r < 1$ is a radius to be tuned subsequently. The reason for this modification is to control the variance of the estimator in Eq. (28.11), which blows up as $\bar{A}_t$ gets close to the boundary. Note that the exploration means that the algorithm often plays actions that are not in $\tilde{\mathcal{A}}$, but mirror descent always chooses $\bar{A}_t \in \tilde{\mathcal{A}}$.

THEOREM 28.4 *If Algorithm 15 is run using the sampling rule, estimator and potential as described above and the learning rate is $\eta = \sqrt{\log(n)/(3dn)}$ and $r = 1 - 2\eta d$. Then*

$$R_n \le 2\sqrt{3nd\log(n)}\,.$$

Proof The first step is to bound the conditional variance of the estimator.

$$\mathbb{E}_t\left[\|\hat{Y}_t\|^2\right] = \frac{d^2}{(1-\|\bar{A}_t\|)^2}\mathbb{E}_t\left[E_t A_t^\top A_t \langle A_t, y_t\rangle^2\right] = \frac{d\|y_t\|^2}{1-\|\bar{A}_t\|} \le \frac{d}{1-\|\bar{A}_t\|}\,. \tag{28.12}$$

Turning our attention towards the properties of the potential. An easy calculation shows that

$$\nabla F(a) = \frac{a}{1-\|a\|}, \qquad F^*(u) = -\log(1+\|u\|) + \|u\|, \qquad \nabla F^*(u) = \frac{u}{1+\|u\|}\,.$$

Since the domain of F is $\mathcal{D} = \{x : \|x\|_2 < 1\}$ it follows that $\nabla F(\mathcal{D}) = \mathbb{R}^d$, which means that Eq. (28.5) is satisfied. We now use the fact that $\log(x) \ge x - x^2$ for all $x \ge -1/2$, which means that if $(\|u\| - \|v\|)/(1+\|v\|) \ge -1/2$, then

$$\begin{aligned}
&D_{F^*}(u,v) = -\log\left(\frac{1+\|u\|}{1+\|v\|}\right) + \|u\| - \|v\| - \frac{1}{1+\|v\|}\langle v, u-v\rangle \\
&= \frac{1}{1+\|v\|}\left(\|u\| - \|v\| + \|v\|\|u\| - \langle v,u\rangle - (1+\|v\|)\log\left(1 + \frac{\|u\|-\|v\|}{1+\|v\|}\right)\right) \\
&\le \frac{1}{1+\|v\|}\left(\|v\|\|u\| - \langle v,u\rangle + \frac{(\|u\|-\|v\|)^2}{1+\|v\|}\right) \\
&\le \frac{1}{1+\|v\|}\left(\|v\|\|u\| - \langle v,u\rangle + \|u\|^2 + \|v\|^2 - 2\|u\|\|v\|\right) \\
&\le \frac{1}{1+\|v\|}\left(\|u\|^2 + \|v\|^2 - 2\langle v,u\rangle\right) = \frac{\|v-u\|^2}{1+\|v\|}\,.
\end{aligned}$$

Let $\tilde{A}_{t+1}$ be defined as in the statement of Theorem 28.3. By the triangle inequality we have

$$\frac{\|\nabla F(\tilde{A}_{t+1})\| - \|\nabla F(\bar{A}_t)\|}{1+\|\nabla F(\bar{A}_t)\|} \ge -\frac{\|\nabla F(\tilde{A}_{t+1}) - \nabla F(\bar{A}_t)\|}{1+\|\nabla F(\bar{A}_t)\|} \ge -\eta\|\hat{Y}_t\| \ge -\frac{1}{2}\,,$$

where the last inequality follows since $\eta\|\hat{Y}_t\| \le \eta d/(1-r) \le 1/2$. By the remarks at the end of the Section 28.2 and Eq. (28.12) leads to

$$\begin{aligned}
\mathbb{E}\left[D_F(\bar{A}_t, \tilde{A}_{t+1})\right] &= \mathbb{E}\left[D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t))\right] \\
&\leq \mathbb{E}\left[\frac{\|\nabla F(\tilde{A}_{t+1}) - \nabla F(\bar{A}_t)\|^2}{1 + \|\nabla F(\bar{A}_t)\|}\right] \\
&= \mathbb{E}\left[\eta^2(1 - \|\bar{A}_t\|)\|\hat{Y}_t\|^2\right] \leq \eta^2 d\,.
\end{aligned}$$

By Theorem 28.3 for any $a \in \tilde{\mathcal{A}}$ and using the fact that $\bar{A}_1 = 0$ and $F(\bar{A}_1) = 0$,

$$R_n(a) \leq \frac{F(a) - F(\bar{A}_1)}{\eta} + \eta n d \leq \frac{1}{\eta}\log\left(\frac{1}{1 - \|a\|}\right) + \eta n d\,.$$

Let $a^* \in \operatorname{argmin}_{a\in\mathcal{A}} \sum_{t=1}^n \langle a, y_t\rangle$, then

$$\begin{aligned}
R_n(a) = \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a^*, y_t\rangle\right] &= \mathbb{E}\left[\sum_{t=1}^n \langle A_t - ra^*, y_t\rangle\right] + \sum_{t=1}^n \langle ra^* - a^*, y_t\rangle \\
&\leq \frac{1}{\eta}\log\left(\frac{1}{1 - \|a\|}\right) + \eta n d + n(1 - r) \leq \frac{1}{\eta}\log(n) + 3\eta n d\,,
\end{aligned}$$

which completes the proof. □

Surprisingly this is smaller than the regret that we got for stochastic bandits by a factor of at least $\sqrt{d}$. There is no contradiction because the adversarial and stochastic linear bandit models are actually quite different, a topic to which the next chapter is dedicated.

28.5 Notes

1 Finding a_{t+1} for both mirror descent and follow the regularized leader requires solving a convex optimization problem. Provided the dimension is not too large and the action-set and potential are reasonably nice, there exist practical approximation algorithms for this problem. The two-step process described in Eqs. (28.6) and (28.7) is sometimes an easier way to go. Usually (28.6) can be solved analytically while (28.7) can be quite expensive. In some important special cases, however, the projection step can be written in closed form or efficiently approximated.

2 We saw that mirror descent with a carefully chosen potential function achieves $O(\sqrt{dn\log(n)})$ regret on the ℓ_2-ball. On the ℓ_∞ ball (hypercube) the optimal regret is $O(d\sqrt{n})$. Interestingly, as n tends to infinity the optimal dependence on the dimension is either d or $\sqrt{d}$ with a complete classification given by Bubeck et al. [2018].

3 Adversarial linear bandits on a simplex are equivalent to finite-armed adversarial bandits with d arms. Yet another well-chosen potential function leads to an algorithm with regret $R_n = O(\sqrt{dn})$, which matches the lower bound and

shaves a factor of $\sqrt{\log d}$ from the upper bounds presented in Chapters 11 and 12. For more details see Exercise 28.10.

4 Both mirror descent and follow the regularized leader depend on a carefully chosen potential function. Currently there is no characterization of exactly what this potential should be or how to find it. At least in the full information setting there are quite general universality results showing that if a certain regret is achievable by some algorithm, then that same regret is nearly achievable by mirror descent with some potential [Srebro et al., 2011]. In practice this result is not useful for constructing new potential functions, however. There have been some attempts to develop 'universal' potential functions that exhibit nice behavior for any action sets [Bubeck et al., 2015b, and others]. These can be useful, but as yet we do not know precisely what properties are crucial, especially in the bandit case.

5 When the horizon is unknown the learning rate cannot be tuned ahead of time. One option is to apply the doubling trick. A more elegant solution is to use a decreasing schedule of learning rates. This requires an adaptation of the proofs of Theorems 28.1 and 28.2, which we outline in Exercises 28.7 and 28.8. This is one situation where mirror descent and follow the regularized leader are not the same and where results slightly favour the latter algorithm.

6 In much of the literature the potential is chosen in such a way that mirror descent and follow the regularized leader are the same algorithm. For historical reasons the name mirror descent is more commonly used in the bandit community. We encourage the reader to keep both algorithms in mind, since the analysis of one-or-other can sometimes be slightly easier. Note that **lazy mirror descent** is a variant of mirror descent that is equivalent to follow the regularized leader for all Legendre potentials [Hazan, 2016].

7 We mentioned that for online linear optimization the mirror descent algorithm also works when $y_1, \dots, y_n$ are chosen nonobliviously. This does not translate to the bandit setting for a subtle reason. Let $\hat{R}_n(a) = \sum_{t=1}^n \langle A_t - a, y_t \rangle$ be the random regret so that

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

The second sum is constant when the losses are oblivious, which means the maximum can be brought outside the expectation, which is not true if the loss vectors are nonoblivious. It is still possible to bound the expected loss relative to a fixed comparator a so that

$$R_n(a) = \mathbb{E}\left[\sum_{t=1}^n \langle A_t - a, y_t \rangle\right] \le B \,,$$

where B is whatever bound obtained from the analysis presented above. A little rewriting shows that

$$R_n = \mathbb{E}\left[\max_{a \in \mathcal{A}} \hat{R}_n(a)\right] = B + \mathbb{E}\left[\max_{a \in \mathcal{A}} R_n(a)\right] - \max_{a \in \mathcal{A}} \mathbb{E}\left[R_n(a)\right] .$$

The difference in expectations can be bounded using tools from empirical process theory, but the resulting bound is only $O(\sqrt{n})$ if $\mathbb{V}[\hat{R}_n(a)] = O(n)$. In general, however, the variance can be as large as $n^{3/2}$ so this condition must be checked for each proposed policy. We emphasize again that the nonoblivious regret is a strange measure because it does not capture the reactive nature of the environment. The details of the application of empirical process theory is beyond the scope of this book. For an introduction to that topic we recommend the books by Vaart and Wellner [1996], Dudley [2014] and van de Geer [2000].

8 The price of bandit information on the unit ball is an extra $\sqrt{d \log(n)}$ (compare Proposition 28.1 and Theorem 28.4). Except for log factors this is also true for the simplex (Proposition 28.2 and Note 3). One might wonder if the difference is always about $\sqrt{d}$, but this is not true in general. The price of bandit information can be as high as $\Theta(d)$. Overall the dimension dependence in the regret in terms of the action set is still not well understood except for special cases.

9 The poor behavior of follow the leader in the full information setting depends on (a) the environment being adversarial rather than stochastic and (b) the action set having sharp corners. When either of these factors is missing, follow the leader is a reasonable choice [Huang et al., 2017b]. Note that with bandit feedback the failure is primarily due to a lack of exploration (Exercises 4.5 and 4.6).

10 A simultaneous-action zero sum game is a game between two players. As the name suggests, both players simultaneously choose a distribution $p \in \mathcal{P}_{K-1}$ and $q \in \mathcal{P}_{E-1}$ respectively. The loss to the first player is $\langle p, Gq \rangle$ where $G \in [0,1]^{K \times E}$. The loss for the second player is $-\langle p, Gq \rangle$, which means that the sum of the players losses is zero ('zero sum'). The minimax theorem by von Neumann [1928] implies that

$$\min_p \max_q \langle p, Gq \rangle = \max_q \min_p \langle p, Gq \rangle . \tag{28.13}$$

The result in Eq. (28.13) can be proven using the existence of an algorithm with sublinear regret for online linear optimization as you will show in Exercise 28.12. As an aside, the minimax theorem holds more generally. Given compact convex sets A and B and $f : A \times B \to \mathbb{R}$ be a continuous function with $f(\cdot, y)$ convex for all y and $f(x, \cdot)$ concave for all x. Then

$$\min_{x \in A} \max_{y \in B} f(x,y) = \max_{y \in B} \min_{x \in A} f(x,y) .$$

Convexity/concavity of f can be replaced with quasiconvexity/quasiconcavity and the continuity assumption can be weakened slightly as well. This generalization is called Sion's minimax theorem [Sion, 1958]. An elementary proof is by Komiya [1988].

28.6 Bibliographic remarks

The results in this chapter come from a wide variety of sources. The online convex optimization framework is due to Zinkevich [2003], where online gradient descent is introduced and analyzed. Mirror descent was first developed by Nemirovski [1979] and Nemirovski and Yudin [1983] for classical optimization while 'follow the regularized leader' appears first in the work by Shalev-Shwartz [2007], Shalev-Shwartz and Singer [2007]. An implicit form of regularization is to add a perturbation of the losses, leading to the 'follow the perturbed leader' algorithm [Hannan, 1957, Kalai and Vempala, 2005a], which is further explored in the context of combinatorial bandit problems in Chapter 30 (and see also Exercise 11.7). Readers interested in an overview of online learning will like the short book by Hazan [2016] while the book by Cesa-Bianchi and Lugosi [2006] has a little more depth (but is also ten years older). As far as we know, the first application of mirror descent to bandits was by Abernethy et al. [2008]. Since then the idea has been used extensively with some examples by Audibert et al. [2013], Abernethy et al. [2015], Bubeck et al. [2018]. Mirror descent has been adapted in a generic way to prove high probability bounds by Abernethy and Rakhlin [2009]. The reader can find (slightly) different proofs of some mirror descent results in the book by Bubeck and Cesa-Bianchi [2012]. The result for the unit ball are from a paper by Bubeck et al. [2012]. Mirror descent can be generalized to Banach spaces. For details see the article by Sridharan and Tewari [2010].

28.7 Exercises

28.1 Show that if F is Legendre with domain $\mathcal{D} \subseteq \mathcal{A} \subset \mathbb{R}^d$ then the minimizer of $\Phi_t(a) = \eta\langle a, y_t\rangle + D_F(a, a_t) = \eta\langle a, y_t\rangle + F(a) - F(a_t) - \langle \nabla F(a_t), a - a_t\rangle$ over $\mathcal{A}$ belongs to the interior of $\mathcal{D}$ and at the minimizer a_{t+1}, $\nabla \Phi_t(a_{t+1}) = 0$ holds.

28.2 Let $\mathcal{A} = [-1, 1]$ and let $y_1 = 1/2$ and $y_s = 1$ for odd $s > 1$ and $y_s = -1$ for even $s > 1$.

(a) Recall that follow the leader (without regularization) chooses $a_t = \operatorname{argmin}_a \sum_{s=1}^{t-1} \langle a, y_s\rangle$. Show that this algorithm suffers linear regret.
(b) Implement follow the regularized leader or mirror descent on this problem with quadratic potential $F(a) = a^2$ and plot a_t as a function of time.

28.3 Prove Corollary 28.1.

28.4 Let $\mathcal{A} = \mathcal{P}_{K-1}$ be the simplex, F the unnormalized negentropy potential and $\eta > 0$. Let $P_1 = \operatorname{argmin}_{p \in \mathcal{A}} F(p)$ and for $t \geq 1$ let

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

where $\hat{Y}_{ti} = \mathbb{I}\{A_t = i\}\, y_{ti}/P_{ti}$ and A_t is sampled from P_t.

(a) Show that the resulting algorithm is exactly Exp3 from Chapter 11.

(b) What happens if you replace mirror descent by follow the regularized leader? That is, if $P_t = \operatorname{argmin}_{p \in \mathcal{A}} \sum_{s=1}^{t-1} \langle p, \hat{Y}_s \rangle + F(p)$.

28.5 Continuing on from the last exercise, in this exercise you will show that the tools in this chapter not only lead to the same algorithm, but also the same bounds.

(a) Let $\tilde{P}_{t+1} = \operatorname{argmin}_{p \in [0,\infty)^K} \eta \langle p, \hat{Y}_t \rangle + D_F(p, P_t)$. Show both relations in the following display:

$$D_F(P_t, \tilde{P}_{t+1}) = \sum_{i=1}^K P_{ti} \left(\exp(-\eta \hat{Y}_{ti}) - 1 + \eta \hat{Y}_{ti} \right) \leq \frac{\eta^2}{2} \sum_{i=1}^K P_{ti} \hat{Y}_{ti}^2 .$$

(b) Show that $\displaystyle \frac{1}{\eta} \mathbb{E}\left[\sum_{t=1}^n D_F(P_t, \tilde{P}_{t+1}) \right] \leq \frac{\eta n K}{2}$.

(c) Show that $\operatorname{diam}_F(\mathcal{P}_{K-1}) = \log(K)$.

(d) Conclude that for appropriately tuned $\eta > 0$ the regret of Exp3 satisfies,

$$R_n \leq \sqrt{2nK \log(K)} .$$

28.6 Prove Theorem 28.2.

28.7 Let $\mathcal{A}$ be closed and convex and $y_1, \ldots, y_n \in \mathcal{L} \subseteq \mathbb{R}^d$. Let F be Legendre with domain $\mathcal{D}$ and assume that $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D})$ and that Eq. (28.5) holds. Let $\eta_0, \eta_1, \ldots, \eta_n$ be a sequence of learning rates where we assume that $\eta_0 = \infty$ and $a_1 = \operatorname{argmin}_{a \in \mathcal{A}} F(a)$ and

$$\begin{aligned} \tilde{a}_{t+1} &= \operatorname{argmin}_{a \in \mathcal{D}} \eta_t \langle a, y_t \rangle + D_F(a, a_t) , \\ a_{t+1} &= \operatorname{argmin}_{a \in \mathcal{A} \cap \mathcal{D}} D_F(a, \tilde{a}_{t+1}) . \end{aligned}$$

Show that for all $a \in \mathcal{A}$:

(a) $\displaystyle R_n(a) = \sum_{t=1}^n \langle a_t - a, y_t \rangle \leq \sum_{t=1}^n \frac{D_F(a_t, \tilde{a}_{t+1})}{\eta_t} + \sum_{t=1}^n \frac{D_F(a, a_t) - D_F(a, \tilde{a}_{t+1})}{\eta_t}$.

(b) $\displaystyle R_n(a) \leq \sum_{t=1}^n \frac{D_F(a_t, \tilde{a}_{t+1})}{\eta_t} + \sum_{t=1}^n D_F(a, a_t) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right)$.

28.8 Like in the previous exercise let $\mathcal{A}$ be closed and convex and $y_1, \ldots, y_n \in \mathcal{L} \subseteq \mathbb{R}^d$. Let $F_1, \ldots, F_n$ be a sequence of Legendre functions where $\operatorname{dom}(F_t) = \mathcal{D}_t$ and $\mathcal{A} \subseteq \operatorname{cl}(\mathcal{D}_t)$ for all t. Let $\Phi_t(a) = F_t(a) + \sum_{s=1}^t \langle a, y_s \rangle$ and $a_t = \operatorname{argmin}_{a \in \mathcal{A}} \Phi_{t-1}(a)$. Show that

$$\begin{aligned} R_n(a) \leq \sum_{t=1}^n &\left(\langle a_t - a_{t+1}, y_t \rangle - D_{F_{t-1}}(a_{t+1}, a_t) \right) \\ &+ F_n(a) - F_0(a_1) + \sum_{t=1}^n \left(F_{t-1}(a_{t+1}) - F_t(a_{t+1}) \right) . \end{aligned}$$

28.9 Consider the finite-armed adversarial bandit problem described in Chapter 11 where the adversary chooses $y_1, \dots, y_n$ with $y_t \in [0,1]^K$. Let $P_t \in \mathcal{P}_{K-1}$ be defined by

$$P_{ti} = \frac{\exp\left(-\eta_{t-1} \sum_{s=1}^{t-1} \hat{Y}_{ti}\right)}{\sum_{j=1}^{K} \exp\left(-\eta_{t-1} \sum_{s=1}^{t-1} \hat{Y}_{tj}\right)}\,,$$

where $\eta_0, \eta_1, \dots$ is an infinite sequence of learning rates and $\hat{Y}_{ti} = \mathbb{I}\{A_t = i\}\, y_{ti}/P_{ti}$ and A_t is sampled from P_t.

(a) Let $\mathcal{A} = \mathcal{P}_{K-1}$ be the simplex, F be the negentropy potential, $F_t(p) = F(p)/\eta_t$ and $\Phi_t(p) = F(p)/\eta_t + \sum_{s=1}^{t} \langle p, \hat{Y}_s \rangle$. Show that P_t is the choice of follow the regularized leader with potentials (F_t) and losses $(\hat{Y}_t)$.

(b) Let $P \in \mathcal{P}_{K-1}$ be the standard basis vector with $P_i = 1$ for $i = \operatorname{argmin}_j \sum_{t=1}^{n} y_{tj}$. Use the fact that $\hat{Y}_t$ is an unbiased estimate of y_t and Exercise 28.8 to show that

$$\begin{aligned} R_n \leq \mathbb{E}\left[\sum_{t=1}^{n} \langle P_t - P_{t+1}, \hat{Y}_t \rangle - D_{F_t}(P_t, P_{t+1})\right] \\ + F_n(P) - F_0(P_1) + \sum_{t=1}^{n} (F_{t-1}(P_{t+1}) - F_t(P_{t+1}))\,. \end{aligned}$$

(c) Assume that (η_t) is decreasing and show that

$$F_n(P) - F_0(P_1) + \sum_{t=1}^{n} (F_{t-1}(a_{t+1}) - F_t(a_{t+1})) \leq \frac{\log(K)}{\eta_n}\,.$$

(d) Use Theorem 26.5 in combination with the facts that $\hat{Y}_{ti} \geq 0$ for all i and $\hat{Y}_{ti} = 0$ unless $A_t = i$ to show that

$$\langle P_t - P_{t+1}, \hat{Y}_t \rangle - D_{F_t}(P_t, P_{t+1}) \leq \frac{\eta_t}{2 P_{tA_t}}\,.$$

(e) Prove that $R_n \leq \dfrac{\log(K)}{\eta_n} + \dfrac{K}{2} \displaystyle\sum_{t=1}^{n} \eta_t$.

(f) Choose $\eta_0, \eta_1, \eta_2, \dots$ so that $R_n \leq 2\sqrt{nK \log(K)}$.

28.10 Let $\mathcal{A} = \mathcal{P}_{K-1}$ be the simplex and assume $y_t \in \mathcal{A}^\circ$ for all t and let $F(a) = -2 \sum_{i=1}^{d} \sqrt{a_i}$.

(a) Show that $F^*(u) = -\sum_{i=1}^{K} u_i^{-1}$ whenever $u \in (-\infty, 0]^K$.

(b) Show that for $u, v \in (-\infty, 0]^K$,

$$D_{F^*}(u, v) = -\sum_{i=1}^{K} \frac{(u_i - v_i)^2}{u_i v_i^2}\,.$$

(c) Show that $\operatorname{diam}_F(\mathcal{A}) \leq 2\sqrt{d}$.

(d) Let A_t be chosen so that $\mathbb{P}\left(A_t = i|\bar{A}_t\right) = \bar{A}_{ti}$ and $\hat{Y}_t$ be the importance-weighted estimator

$$\hat{Y}_{ti} = \frac{\mathbb{I}\left\{A_t = i\right\} y_{ti}}{\bar{A}_{ti}}\,.$$

(e) Show that $\tilde{A}_{t+1,A_t} \leq \bar{A}_{t+1,A_t}$ and $\tilde{A}_{t+1,i} = \bar{A}_{t+1,i}$ for $i \neq A_t$.

(f) Show that $\mathbb{E}\left[D_{F^*}(\nabla F(\tilde{A}_{t+1}), \nabla F(\bar{A}_t))\right] \leq \eta^2\sqrt{d}$.

(g) Conclude that the regret of mirror descent with this potential is bounded by

$$R_n \leq \sqrt{8dn}\,.$$

(h) Devise an efficient implementation of this algorithm.

The algorithm in the above exercise is called the **implicitly normalized forecaster** and was introduced by Audibert and Bubeck [2009]. At first it went unnoticed that this algorithm was an instance of mirror descent and the proof was consequentially much more complicated. More details are in the book by Bubeck and Cesa-Bianchi [2012].

28.11 Let F be the unnormalized negentropy potential and consider online mirror descent with $\mathcal{A} = \mathcal{P}_{d-1}$ and loss vectors $y_1, \dots, y_n$ chosen from the hypercube: $y_t \in [0,1]^d$. Prove that $R_n \leq \sqrt{2n\log(K)}$.

28.12 Prove the minimax theorem described in Note 10.

Let $p_1, \dots, p_n \in \mathcal{P}_{K-1}$ be the choices of mirror descent with unnormalized negentropy potential when the losses $y_1, \dots, y_n$ are given by $y_t = Gq_t$ and $q_t = \operatorname{argmax}_q \langle p_t, Gq\rangle$. Then use the result from Exercise 28.11. The minimax theorem is due to von Neumann [1928].