36 Partial Monitoring

While in a bandit problem the feedback that the learner receives from the environment is the loss (or reward) of the chosen action, in **partial monitoring** the coupling between the loss of the action and the feedback received by the learner is loosened.

To illustrate the ideas we consider the problem of learning to match pennies when feedback is costly. Let $c > 0$ be a known constant. At the start of the game the adversary secretly chooses a sequence $i_1, \dots, i_n \in \{\text{heads}, \text{tails}\}$. In each round the learner chooses action $A_t \in \{$heads, tails, uncertain$\}$ and the loss for choosing action a in round t is

$$y_{ta} = \begin{cases} 0\,, & \text{if } a = i_t\,; \\ c\,, & \text{if } a = \text{uncertain}\,; \\ 1\,, & \text{otherwise}\,. \end{cases}$$

So far this looks like a bandit problem. The difference is that the learner never directly observes y_{tA_t}. Instead, the learner observes nothing unless $A_t = \text{uncertain}$ in which case they observe the value of i_t. As usual, the goal of the regret is to minimize the regret, which is

$$R_n = \mathbb{E}\left[\max_{a \in [K]} \sum_{t=1}^n (y_{tA_t} - y_{ta})\right]\,.$$

How should a learner act in problems like this, where the loss is not directly observed? Can we find a policy with sublinear regret? In this chapter we give nearly complete answers to these questions for a large class of finite adversarial partial monitoring problems.

Matching pennies with costly feedback seems like an esoteric problem. But think about adding contextual information and replace the pennies with emails to be classified as spam or otherwise. The true label is only accessible by asking a human, which replaces the third action.

36.1 Finite adversarial partial monitoring problems

To reduce clutter we slightly abuse notation by using (e_i) to denote the standard basis vectors of Euclidean spaces of potentially different dimensions. A K-action, E-outcome, F-feedback finite adversarial partial monitoring problem is specified by a **loss matrix** $\mathcal{L} \in \mathbb{R}^{K \times E}$ and a **feedback matrix** $\Phi \in [F]^{K \times E}$. At the beginning of the game, the learner gets $\mathcal{L}$ and Φ, while the environment secretly chooses n outcomes $i_1, \dots, i_n$ with $i_t \in [E]$. The loss of action $a \in [K]$ in round t is $y_{ta} = \mathcal{L}_{ai_t}$. In each round t the learner chooses $A_t \in [K]$ and receives feedback $\Phi_t = \Phi_{A_t i_t}$. Given partial monitoring problem $G = (\Phi, \mathcal{L})$ the regret of policy π in environment $i_{1:n} = (i_t)_{t=1}^n$ is

$$R_n(\pi, i_{1:n}, G) = \max_{a \in [K]} \mathbb{E}_{\pi, i_{1:n}, G}\left[\sum_{t=1}^n (y_{tA_t} - y_{ta})\right] .$$

The index of the expectation operator is a reminder that the distribution of $\{y_{tA_t}\}$ is dependent on π, $i_{1:n}$ and G. We will omit these indices and the arguments of R_n when they can be inferred from the context.

36.1.1 Examples

The partial monitoring framework is rich enough to model a wide variety of problems, a few of which are illustrated by the examples that follow. Many of the examples are not very interesting on their own, but are included to highlight the flexibility of the framework and challenges of making the regret small.

EXAMPLE 36.1 (Hopeless problem) Some partial monitoring problems are completely hopeless in the sense one cannot expect to make the regret small. A simple example occurs when $K = E = 2$ and $F = 1$ and

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} . \tag{36.1}$$

Note that rows/columns correspond to choices of the learner/environment respectively. In both rows (corresponding to the actions of the learner), the feedback matrix has identical entries for both columns. As the learner has no way of distinguishing between different sequences of outcomes, there is no way to learn and avoid linear regret.

Two feedback matrices $\Phi, \Phi' \in [F]^{K \times E}$ encode the same information if the pattern of identical entries in each row match. More precisely, if for each row $a \in [K]$ there is an injective function $\sigma : [E] \to [E]$ such that $\Phi'_{ai} = \sigma(\Phi_{ai})$ for all $i \in [E]$.

EXAMPLE 36.2 (Trivial problem) Just as there are hopeless problems, there are also trivial problems. For example, when one action dominates all others as in the following problem:

$$\mathcal{L} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Clearly, in this game the learner can safely ignore the second action and suffer zero regret, regardless of the choices of the adversary.

EXAMPLE 36.3 (Matching pennies) The penny-matching problem mentioned in the introduction has $K = 3$ actions $E = 2$ outcomes and is described by

$$\mathcal{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ c & c \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}. \tag{36.2}$$

Matching pennies is a hard game for $c > 1/2$ in the sense that the adversary can force the regret of any adversary to be at least $\Omega(n^{2/3})$. To see this, consider the randomized adversary that chooses the first outcome with probability p and the second with probability $1 - p$. Let $\varepsilon > 0$ be a small constant to be chosen later and assume p is either $1/2 + \varepsilon$ or $1/2 - \varepsilon$, which determines two environments. The techniques in Chapter 13 show that the learner can only distinguish between these environments by playing the third action about $1/\varepsilon^2$ times. If the learner does not choose to do this, then the regret is expected to be $\Omega(n\varepsilon)$. Taking these together shows the regret is lower bounded by $R_n = \Omega(\min(n\varepsilon,\ (c-1/2+\varepsilon)/\varepsilon^2))$. Choosing $\varepsilon = n^{-1/3}$ leads to a bound of $R_n = \Omega((c - 1/2)n^{2/3})$. Notice the argument fails when $c \le 1/2$. We encourage you to pause for a minute to convince yourself about the correctness of the above argument and to consider what might be the situation when $c \le 1/2$.

EXAMPLE 36.4 (Bandits) Finite-armed adversarial bandits with binary losses can be represented in the partial monitoring framework. When $K = 2$ this is possible with the following matrices:

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 1 & 1 & 2 & 2 \end{pmatrix}.$$

The number of columns for this game is 2^K. For non-binary rewards you would need even more columns. A partial monitoring problem where $\Phi = \mathcal{L}$ can be called a bandit problem because the learner observes the loss of the chosen action. In bandit games we can simply use Exp3 to guarantee a regret of $O(\sqrt{Kn})$.

EXAMPLE 36.5 (Full information problems) One can also represent problems where the learner observes all the losses. With binary losses and two actions we have

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 2 & 3 & 4 \end{pmatrix}.$$

Like for bandits the size of the game grows very quickly as more actions/outcomes are added.

EXAMPLE 36.6 (Dynamic pricing) A charity worker is going door-to-door selling calendars. The marginal cost of a calendar is close to zero, but the wages of the door-knocker represents a fixed cost of $c > 0$ per occupied house. The question is how to price the calendar. Each round corresponds to an attempt to sell a calendar and the action is the seller's asking price from one of E choices. The potential buyer will purchase the calendar if the asking price is low enough. Below we give the corresponding matrices for case where both the candidate asking prices and the possible values for the buyer's private valuations are $\{\$1, \$2, \$3, \$4\}$:

$$\mathcal{L} = \begin{pmatrix} c & c-1 & c-1 & c-1 & c-1 \\ c & c & c-2 & c-2 & c-2 \\ c & c & c & c-3 & c-3 \\ c & c & c & c & c-4 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 2 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 1 & 2 & 2 \\ 1 & 1 & 1 & 1 & 2 \end{pmatrix}.$$

Notice that observing the feedback is sufficient to deduce the loss so the problem could be tackled with a bandit algorithm. But there is additional structure in the losses here because the learner knows that if a calendar did not sell for \$3 then it would not sell for \$4.

36.2 The structure of partial monitoring

The minimax regret of partial monitoring problem $G = (\mathcal{L}, \Phi)$ is

$$R_n^*(G) = \inf_\pi \max_{u_{1:n}} R_n(\pi, u_{1:n}, G)\,.$$

One of the core questions in partial monitoring is to understand the growth of $R_n^*(G)$ as a function of n for different games. We have seen examples where

$$\begin{aligned} R_n^*(G) &= 0 && \text{(Example 36.2)} \\ R_n^*(G) &= \tilde\Theta(\sqrt{n}) && \text{(Example 36.4)} \\ R_n^*(G) &= \Theta(n^{2/3}) && \text{(Example 36.3)} \\ R_n^*(G) &= \Omega(n)\,. && \text{(Example 36.1)} \end{aligned}$$

The main result of this chapter is that there are no other options. A partial monitoring problem is called **trivial** if $R_n^*(G) = 0$, **easy** if $R_n^*(G) = \tilde\Theta(\sqrt{n})$, **hard** if $R_n^*(G) = \Theta(n^{2/3})$ and **hopeless** if $R_n^*(G) = \Omega(n)$. Furthermore, we will show that the category of any G can be deduced from elementary linear algebra.

What makes matching pennies hard and bandits easy? To get a handle on this we need a geometric representation of partial monitoring problems. The next few paragraphs introduce a lot of new terminology that can be hard to grasp all at once. At the end of the section there is an example illustrating the concepts.

The geometry underlying partially monitoring comes from viewing the problem as a linear prediction problem, where both the adversary and the learner play on some simplex. Starting with reworking the adversary's choices, let $u_t = e_{i_t} \in \mathbb{R}^E$, where $e_1, \dots, e_E$ are the standard basis vectors. Then, we can equivalently think of the environment choosing the sequence $\{u_t\}_{t=1}^n$. Letting $\ell_a \in \mathbb{R}^E$ be the ath row of matrix $\mathcal{L}$, where $a \in [K]$, we have that $y_{ta} = \langle \ell_a, u_t \rangle$ is the loss suffered when choosing action a in round t.

Let $\bar{u}_t = \frac{1}{t} \sum_{s=1}^t u_s \in \mathcal{P}_{E-1}$ be the vector of mean frequencies of the adversary's choices over t rounds. An action a is optimal in hindsight if $\max_b \langle \ell_a - \ell_b, \bar{u}_n \rangle = 0$. The **cell** of an action a is subset of $\mathcal{P}_{E-1}$ on which it is optimal:

$$C_a = \left\{ u \in \mathcal{P}_{E-1} : \max_{b \in [K]} \langle \ell_a - \ell_b, u \rangle \leq 0 \right\} ,$$

which is convex polytope. The collection $\{C_a : a \in [K]\}$ is called the **cell decomposition**. Actions with $C_a = \emptyset$ are called **dominated** because they are never optimal, no matter how the adversary plays. For nondominated actions we define the **dimension** of an action to be the dimension of the **affine hull** of C_a. Readers unfamiliar with the affine hull should read Note 3 at the end of the chapter. A nondominated action is called **Pareto optimal** if it has dimension $E - 1$ and **degenerate** otherwise. Actions a and b are **duplicates** if $\ell_a = \ell_b$. Pareto optimal actions a and b are **neighbors** if $C_a \cap C_b$ has dimension $E - 2$. Note that if a and b are Pareto optimal duplicates, then $C_a \cap C_b$ has dimension $E - 1$ and the definition means that a and b are not neighbors. For Pareto optimal action a we let $\mathcal{N}_a$ be the set consisting of a and its neighbors. Given a pair of neighbors (a, b) we let $\mathcal{N}_{ab} = \{c \in [K] : C_a \cap C_b \subseteq C_c\}$, while for Pareto optimal action a we let $\mathcal{N}_{aa} = \emptyset$.

Dominated and degenerate actions can never be uniquely optimal in hindsight, but their presence can make the difference between a hard game and a hopeless one. If $c > 1/2$, then the third action in matching pennies is dominated, but without it the learner would suffer linear regret. Duplicate actions are only duplicate in the sense that they have the same loss. They may have different feedback structures and so cannot be trivially combined.

Let a and b be neighboring actions. The next lemma characterizes actions in $\mathcal{N}_{ab}$ as either a, b, duplicates of a, b or degenerate actions d for which ℓ_d is a convex combination of ℓ_a and ℓ_b. The situation is illustrated when $E = 2$ in Fig. 36.1.

LEMMA 36.1 *Let a, b be neighboring actions and $d \in \mathcal{N}_{ab}$ be an action such that $\ell_d \notin \{\ell_a, \ell_b\}$. Then*

(a) There exists an $\alpha \in (0, 1)$ such that $\ell_d = \alpha \ell_a + (1 - \alpha) \ell_b$.

(b) $C_d = C_a \cap C_b$.

(c) d has dimension $E - 2$.

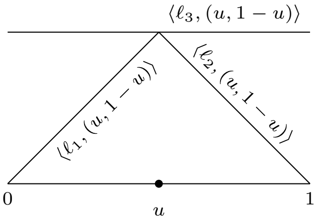

$\langle \ell_3, (u, 1-u) \rangle$
$\langle \ell_1, (u, 1-u) \rangle$
$\langle \ell_2, (u, 1-u) \rangle$
0
u
1

Figure 36.1 The figure shows the situation when $E = 2$ and $\ell_1 = (1, 0)$ and $\ell_2 = (0, 1)$ and $\ell_3 = (1/2, 1/2)$. Then $C_1 = [0, 1/2]$ and $C_2 = [1/2, 1]$, which both have dimension $1 = E - 1$. Then $C_3 = \{1/2\} = C_1 \cap C_2$, which has dimension 0.

Proof We use the fact that if $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathbb{R}^d$ and $\dim(\mathcal{X}) = \dim(\mathcal{Y})$, then $\operatorname{aff}(\mathcal{X}) = \operatorname{aff}(\mathcal{Y})$ (Exercise 36.1). Clearly $C_a \cap C_b \subseteq C_a \cap C_d$ and $\operatorname{aff}(C_a \cap C_b) = \ker(\ell_a - \ell_b)$ and $\operatorname{aff}(C_a \cap C_d) = \ker(\ell_a - \ell_d)$. By assumption $\dim(C_a \cap C_b) = E - 2$. Since $C_a \cap C_b \subseteq C_a \cap C_d$ it holds that $\dim(C_a \cap C_d) \geq E - 2$. Furthermore, $\dim(C_a \cap C_d) \leq E - 2$, since otherwise $\ell_d = \ell_a$. Hence $\ker(\ell_a - \ell_b) = \ker(\ell_a - \ell_d)$, which means that $\ell_a - \ell_b$ is proportional to $\ell_a - \ell_d$ so that $(1-\alpha)(\ell_a - \ell_b) = \ell_a - \ell_d$ for some $\alpha \neq 1$. Rearranging shows that

$$\ell_d = \alpha \ell_a + (1 - \alpha)\ell_b \,.$$

Now we show that $\alpha \in (0, 1)$. First note that $\alpha \notin \{0, 1\}$ since otherwise $\ell_d \in \{\ell_a, \ell_b\}$. Let $u \in C_a$ be such that $\langle \ell_a, u \rangle < \langle \ell_b, u \rangle$, which exists since $\dim(C_a) = E - 1$ and $\dim(C_a \cap C_b) = E - 2$. Then

$$\langle \ell_a, u \rangle \leq \langle \ell_d, u \rangle = \alpha \langle \ell_a, u \rangle + (1 - \alpha)\langle \ell_b, u \rangle = \langle \ell_a, u \rangle + (\alpha - 1)\langle \ell_a - \ell_b, u \rangle \,,$$

which by the positivity of $\langle \ell_a - \ell_b, u \rangle$ implies that $\alpha \leq 1$. A symmetric argument shows that $\alpha > 0$. For (b), it suffices to show that $C_d \subset C_a \cap C_b$. By de Morgan's law for this it suffices to show that $\mathcal{P}_{E-1} \setminus (C_a \cap C_b) \subset \mathcal{P}_{E-1} \setminus C_d$. Thus, pick some $u \in \mathcal{P}_{E-1} \setminus (C_a \cap C_b)$. The goal is to show that $u \notin C_d$. The choice of u implies that there exists an action c such that $\langle \ell_a - \ell_c, u \rangle \geq 0$ and $\langle \ell_b - \ell_c, u \rangle \geq 0$ with a strict inequality for either a or b (or both). Therefore using the fact that $\alpha \in (0, 1)$ we have

$$\langle \ell_d, u \rangle = \alpha \langle \ell_a, u \rangle + (1 - \alpha)\langle \ell_b, u \rangle > \langle \ell_c, u \rangle \,,$$

which by definition means that $u \notin C_d$, completing the proof of (b). Finally, (c) is immediate from (b) and the definition of neighboring actions. □

In order to achieve small regret the learner needs to identify an optimal action. How efficiently this can be done depends on the feedback matrix. First, note that given access to the loss matrix, the learner can restrict the search for the optimal action to the Pareto optimal actions. One way to find the optimal action then could be to estimate $\langle \ell_a, u_t \rangle$ for each Pareto optimal action a and $t \in [n]$ and take differences of the estimates to compare actions. This is asking too much, and a better option is to estimate $\langle \ell_a - \ell_b, u_t \rangle$ directly. This is a better option because on the one hand it is clearly necessary to know the loss differences between Pareto optimal actions, and on the other hand there exist games for which $\langle \ell_a, u_t \rangle$ cannot be estimated, but the differences can. For example, the following game has this property.

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 1 & 2 \\ 1 & 0 & 2 & 1 \end{pmatrix}, \qquad \Phi = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \end{pmatrix}$$

The learner can never tell if the environment is playing in the first two columns or the last two, but the differences between the losses are easily deduced from the feedback. We emphasize once again that only the loss differences between Pareto optimal actions need to be estimated: There are in fact games that are easy yet some loss differences cannot be estimated. For example, there is never any need to estimate the losses of a dominated action.

Having decided we need to estimate the loss differences for Pareto optimal actions, the next question is how can the learner do this? Suppose in round t the learner samples A_t from distribution $P_t \in \operatorname{ri}(\mathcal{P}_{K-1})$. Let a and b be Pareto optimal and suppose we want an estimator $\hat{\Delta}$ of $\Delta = y_{ta} - y_{tb}$. Our estimator $\hat{\Delta}$ should depend on A_t and Φ_t, which suggests defining

$$\hat{\Delta} = \frac{v(A_t, \Phi_t)}{P_{tA_t}},$$

where $v : [K] \times [F] \to \mathbb{R}$ is some suitable chosen function. The division by P_{tA_t} is a convenient normalization and could be pushed into v. The reader can check that $\hat{\Delta}$ is unbiased regardless the choice u_t if and only if

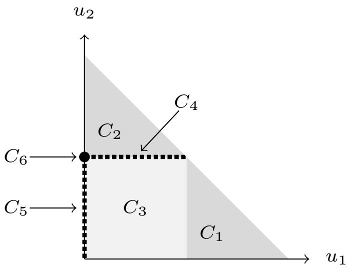
$$\ell_{ai} - \ell_{bi} = \sum_{c=1}^{K} v(c, \Phi_{ci}) \qquad \text{for all } i \in [E]. \tag{36.3}$$

A pair of Pareto optimal actions a and b are called **globally observable** if there exists a function v satisfying Eq. (36.3). They are **locally observable** if the function can be chosen so that $v(c, f) = 0$ whenever $c \notin \mathcal{N}_{ab}$. A partial monitoring problem $G = (\mathcal{L}, \Phi)$ is called globally/locally observable if all pairs of neighboring actions are globally/locally observable. The global/local observability conditions formalize the idea introduced in Example 36.3. Games that are globally observable but not locally observable are hard because the learner cannot identify the optimal action by playing near-optimal actions only. Instead it has to play badly suboptimal actions to gain information and this increases the minimax regret.

EXAMPLE 36.7 The partial monitoring problem below has six actions, three feedbacks and three outcomes. The cell decomposition is shown on the right with the 2-simplex parameterized by its first two coordinates u_1 and u_2 so that

$u_3 = 1 - u_2 - u_1$. Actions 1, 2 and 3 are Pareto optimal. There are no dominated actions while actions 4 and 5 are 1-dimensional and action 6 is 0-dimensional. The neighbors are $(1,3)$ and $(2,3)$, which are both locally observable and so the game is locally observable. Note that $(1,2)$ are *not* neighbors because the intersection of their cells is $(E-3)$-dimensional. Finally, $\mathcal{N}_3 = \{1,2,3\}$ and $\mathcal{N}_1 = \{1,3\}$ and $\mathcal{N}_{23} = \{2,3,4\}$. Think about how we decided on what losses to use to get the cell decomposition shown in the figure!

$$\mathcal{L} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1/2 & 1/2 & 1/2 \\ 3/4 & 1/4 & 3/4 \\ 1 & 1/2 & 1/2 \\ 1 & 1/4 & 3/4 \end{pmatrix} \qquad \Phi = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$



36.3 Classification of finite adversarial partial monitoring

The terminology in the last chapter finally allows us to state the main theorem of this chapter that classifies finite adversarial partial monitoring games.

THEOREM 36.1 *The minimax regret of partial monitoring problem $G = (\mathcal{L}, \Phi)$ falls into one of four categories:*

$$R_n^*(G) = \begin{cases} 0\,, & \textit{if } G \textit{ has no pairs of neighboring actions}\,; \\ \tilde{\Theta}(\sqrt{n})\,, & \textit{if } G \textit{ is locally observable and has neighboring actions}\,; \\ \Theta(n^{2/3})\,, & \textit{if } G \textit{ is globally observable, but not locally observable}\,; \\ \Omega(n)\,, & \textit{otherwise}\,. \end{cases}$$

The Landou notation is used in the traditional mathematical sense and obscures dependence on K, E, F and the finer structure of $G = (\mathcal{L}, \Phi)$.

The proof is split into parts by proving upper and lower bounds for each part. First up is the lower bounds. We then describe a policy for locally observable games and analyze its regret. The upper bound for globally observable games is left as an exercise to the reader (Exercise 36.11).

36.4 Lower bounds

Like for bandits, the lower bounds are most easily proven using a stochastic adversary. In stochastic partial monitoring we assume that $u_1, \dots, u_n$ are chosen at independently at random from the same distribution. To emphasize the randomness we switch to capital letters. Given partial monitoring problem $G = (\mathcal{L}, \Phi)$ and probability vector $u \in \mathcal{P}_{E-1}$ the stochastic partial monitoring environment associated with u samples a sequence of independently and identically distributed random variables $I_1, \dots, I_n$ with $\mathbb{P}(I_t = i) = u_i$ and $U_t = e_{I_t}$. In each round t a policy chooses action A_t and receives feedback $\Phi_t = \Phi_{A_t I_t}$. The regret is

$$R_n(\pi, u, G) = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^n \langle \ell_{A_t} - \ell_a, U_t \rangle\right] = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^n \langle \ell_{A_t} - \ell_a, u \rangle\right] .$$

The reader should check that $R_n^*(G) \geq \min_\pi \max_{u \in \mathcal{P}_{E-1}} R_n(\pi, u, G)$, which allows us to restrict our attention to stochastic partial monitoring problems. Given $u, q \in \mathcal{P}_{E-1}$, let $\mathrm{D}(u, q)$ be the relative entropy between categorical distributions with parameters u and q respectively:

$$\mathrm{D}(u, q) = \sum_{i=1}^K u_i \log\left(\frac{u_i}{q_i}\right) \leq \sum_{i=1}^K \frac{(u_i - q_i)^2}{q_i} , \tag{36.4}$$

where the second inequality follows from the fact that for measures P, Q we have $\mathrm{D}(P, Q) \leq \chi^2(P, Q)$ (see Note 5 in Chapter 13).

THEOREM 36.2 *Let $G = (\mathcal{L}, \Phi)$ be a globally observable partial monitoring problem that is not locally observable. Then there exists a constant $c_G > 0$ such that $R_n^*(G) \geq c_G n^{2/3}$.*

Proof The proof involves several steps. Roughly, we need to define two alternative stochastic partial monitoring problems. We then show these environments are hard to distinguish without playing an action associated with a large loss. Finally we balance the cost of distinguishing the environments against the linear cost of playing randomly.

Step 1: Defining the alternatives

Let a, b be a pair neighboring actions that are not locally observable. Then by definition $C_a \cap C_b$ is a polytope of dimension $E - 2$. Let u be the centroid of $C_a \cap C_b$ and

$$\varepsilon = \min_{c \notin \mathcal{N}_{ab}} \langle \ell_c - \ell_a, u \rangle . \tag{36.5}$$

The value of ε is well defined, since by global observability of G, but nonlocal observability of (a, b) there must exist some action $c \notin \mathcal{N}_{ab}$. Furthermore, since $c \notin \mathcal{N}_{ab}$ it follows that $\varepsilon > 0$. As in the lower bound constructions for stochastic bandits, we now define two stochastic partial monitoring problems. Since (a, b) are not locally observable, there does not exist a function $v : [K] \times [F] \to \mathbb{R}$ such that for all $i \in [E]$,

$$\sum_{c \in N_k} v(c, \Phi_{ci}) = \ell_{ai} - \ell_{bi} . \tag{36.6}$$

In this form it does not seem obvious what the next step should be. To clear things up a little we introduce some linear algebra. Let $S_c \in \{0,1\}^{F \times E}$ be the matrix with $(S_c)_{fi} = \mathbb{I}\{\Phi_{ci} = f\}$, which is chosen so that $S_c e_i = e_{\Phi_{ci}}$. Define the linear map $S : \mathbb{R}^E \to \mathbb{R}^{|\mathcal{N}_{ab}|F}$ by

$$S = \begin{pmatrix} S_a \\ S_b \\ \vdots \\ S_c \end{pmatrix} ,$$

which is the matrix formed by stacking the matrices $\{S_c : c \in \mathcal{N}_{ab}\}$. Then there exists a v satisfying Eq. (36.6) if and only if there exists a $w \in \mathbb{R}^{|\mathcal{N}_{ab}|F}$ such that

$$(\ell_a - \ell_b)^\top = w^\top S .$$

In other words, actions (a, b) are locally observable if and only if $\ell_a - \ell_b \in \operatorname{im}(S^\top)$. Since we have assumed that (a, b) are not locally observable, it means that $\ell_a - \ell_b \notin \operatorname{im}(S^\top)$. Let $z \in \operatorname{im}(S^\top)$ and $w \in \ker(S)$ be such that $\ell_a - \ell_b = z + w$, which is possible since $\operatorname{im}(S^\top) \oplus \ker(S) = \mathbb{R}^E$. Since $\ell_a - \ell_b \notin \operatorname{im}(S^\top)$ it holds that $w \neq 0$ and $\langle \ell_a - \ell_b, w\rangle = \langle z + w, w\rangle = \langle w, w\rangle \neq 0$. Finally let $q = w/\langle \ell_a - \ell_b, w\rangle$. To summarize, we have demonstrated the existence of a vector $q \in \mathbb{R}^E$, $q \neq 0$ such that $Sq = 0$ and $\langle \ell_a - \ell_b, q\rangle = 1$. Let $\Delta > 0$ be some small constant to be tuned subsequently and define $u_a = u - \Delta q$ and $u_b = u + \Delta q$ so that

$$\langle \ell_b - \ell_a, u_a\rangle = \Delta \qquad \text{and} \qquad \langle \ell_a - \ell_b, u_b\rangle = \Delta .$$

We note that if Δ is sufficiently small, then $u_a \in C_a$ and $u_b \in C_b$. This means that action a is optimal if the environment plays u_a on average and b is optimal if the environment plays u_b on average (see Fig. 36.2).

Step 2: Calculating the relative entropy

Given action c and $w \in \mathcal{P}_{E-1}$ let $\mathbb{P}_{cw}$ be the distribution on the feedback observed by the learner when playing action c in stochastic partial monitoring environment determined by w. That is $\mathbb{P}_{cw}(f) = \mathbb{P}_w(\Phi_t = f | A_t = c) = (S_c w)_f$. Further, let $\mathbb{P}_w$ be the distribution on the histories $H_n = (A_1, \Phi_1, \dots, A_n, \Phi_n)$ arising from the interaction of the learner's policy with the stochastic environment determined by w. A modification of Lemma 15.1 shows that

$$\mathrm{D}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) = \sum_{c \in [K]} \mathbb{E}[T_c(n)]\, \mathrm{D}(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) , \tag{36.7}$$

By the definitions of u_a and u_b, we have $S_c u_a = S_c u_b$ for all $c \in \mathcal{N}_{ab}$. Therefore $\mathbb{P}_{cu_a} = \mathbb{P}_{cu_b}$ and so $\mathrm{D}(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) = 0$ for all $c \in \mathcal{N}_{ab}$. On the other hand, if $c \notin \mathcal{N}_{ab}$, then by Eq. (36.4),

$$\mathrm{D}(\mathbb{P}_{cu_a}, \mathbb{P}_{cu_b}) \le \mathrm{D}(u_a, u_b) \le \sum_{i=1}^E \frac{(u_{ai} - u_{bi})^2}{u_{bi}} = 4\Delta^2 \sum_{i=1}^K \frac{q_i^2}{u_i + \Delta q_i} \le C_u \Delta^2 ,$$

where C_u is a suitably large constant. We note that $u \in C_a \cap C_b$ is not on the boundary of $\mathcal{P}_{E-1}$, so $u_i > 0$ for all i. Therefore

$$\mathrm{D}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) \leq c_U \sum_{c \in [K]} \mathbb{E}[T_c(n)]\Delta^2 \,. \tag{36.8}$$

Step 3: Comparing the regret

By Eq. (36.5) and Hölder's inequality, for $c \notin \mathcal{N}_{ab}$ we have $\langle \ell_c - \ell_a, u_a \rangle = \varepsilon + \langle \ell_c - \ell_a, \Delta q \rangle \geq \varepsilon - \Delta \|q\|_1$ and $\langle \ell_c - \ell_b, u_b \rangle \geq \varepsilon - \Delta \|q\|_1$, where, for simplicity, and without the loss of generality, we assumed that the losses lie in $[0, 1]$. Define $\tilde{T}(n)$ to be the number of times an arm not in $\mathcal{N}_{ab}$ is played:

$$\tilde{T}(n) = \sum_{c \notin \mathcal{N}_{ab}} T_c(n) \,.$$

By Lemma 36.1, for each action $c \in \mathcal{N}_{ab}$ there exists an $\alpha \in [0, 1]$ such that $\ell_c = \alpha \ell_a + (1 - \alpha)\ell_b$. Therefore

$$\langle \ell_c - \ell_a, u_a \rangle + \langle \ell_c - \ell_b, u_b \rangle = (1 - \alpha)\langle \ell_b - \ell_a, u_a \rangle + \alpha \langle \ell_a - \ell_b, u_b \rangle = \Delta \,, \tag{36.9}$$

which means that $\max(\langle \ell_c - \ell_a, u_a \rangle, \langle \ell_c - \ell_b, u_b \rangle) \geq \Delta/2$. Define $\bar{T}(n)$ as the number of times an arm in $\mathcal{N}_{ab}$ is played that is at least $\Delta/2$ suboptimal in u_a:

$$\bar{T}(n) = \sum_{c \in \mathcal{N}_{ab}} \mathbb{I}\left\{\langle \ell_c - \ell_a, u_a \rangle \geq \frac{\Delta}{2}\right\} T_c(n) \,.$$

It also follows from (36.9) that if $c \in \mathcal{N}_{ab}$ and $\langle \ell_c - \ell_a, u_a \rangle < \frac{\Delta}{2}$ then $\langle \ell_c - \ell_b, u_b \rangle \geq \frac{\Delta}{2}$. Hence, under u_b the random pseudo-regret, $\sum_c T_c(n) \langle \ell_c - \ell_b, u_b \rangle$, is at least $(n - \bar{T}(n))\Delta/2$. Assume that Δ is chosen sufficiently small so that $\Delta \|q\|_1 \leq \varepsilon/2$. Then, by the above,

$$\begin{aligned}
&R_n(\pi, u_a, G) + R_n(\pi, u_b, G) \\
&\quad = \mathbb{E}_{u_a}\left[\sum_{c \in [K]} T_c(n) \langle \ell_c - \ell_a, u_a \rangle\right] + \mathbb{E}_{u_b}\left[\sum_{c \in [K]} T_c(n) \langle \ell_c - \ell_b, u_b \rangle\right] \\
&\quad \geq \frac{\varepsilon}{2}\mathbb{E}_{u_a}\left[\tilde{T}(n)\right] + \frac{n\Delta}{4}\left(\mathbb{P}_{u_a}(\bar{T}(n) \geq n/2) + \mathbb{P}_{u_b}(\bar{T}(n) < n/2)\right) \\
&\quad \geq \frac{\varepsilon}{2}\mathbb{E}_{u_a}\left[\tilde{T}(n)\right] + \frac{n\Delta}{8}\exp\left(-\mathrm{D}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b})\right) \\
&\quad \geq \frac{\varepsilon}{2}\mathbb{E}_{u_a}\left[\tilde{T}(n)\right] + \frac{n\Delta}{8}\exp\left(-C_u \Delta^2 \mathbb{E}_{u_a}\left[\tilde{T}(n)\right]\right) \,,
\end{aligned}$$

where the second inequality follows from Theorem 14.2 and the third from Eqs. (36.7) and (36.8). The bound is completed by choosing $\Delta = \varepsilon/(2\|q\|_1 n^{1/3})$ (which is finite since $q \neq 0$) and straightforward optimization (Exercise 36.5). □

We leave the following theorems as exercises for the reader (Exercises 36.6 and 36.7).

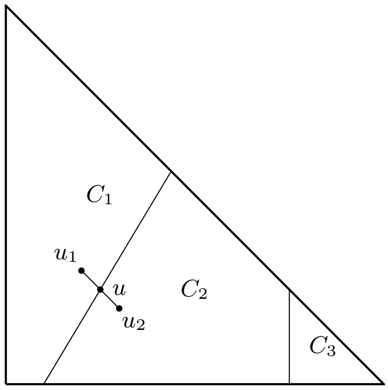


Figure 36.2 Lower bound construction for hard partial monitoring problems

THEOREM 36.3 *If G is not globally observable and has at least two non-dominated actions, then there exists a constant $c_G > 0$ such that $R_n^*(G) \geq c_G n$.*

Proof sketch Since G is not globally observable there exists a pair of neighboring actions (a, b) that are not globally observable. Let u be the centroid of $C_a \cap C_b$. Let $S \in \mathbb{R}^{KF \times E}$ be the stack of matrices from $\{S_c : c \in [K]\}$ (all actions). Then using the same argument as the previous proof we have $\ell_a - \ell_b \notin \operatorname{im}(S^\top)$. Now define $q \in \mathbb{R}^E$ such that $\langle \ell_a - \ell_b, q \rangle = 1$ and $Sq = 0$. Let $\Delta > 0$ be sufficiently small and $u_a = u - \Delta q$ and $u_b = u + \Delta q$. Show that $\mathrm{D}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) = 0$ for all policies and complete the proof in the same fashion as the proof of Theorem 36.2. □

THEOREM 36.4 *Let $G = (\mathcal{L}, \Phi)$ be locally observable and have at least one pair of neighbours. Then there exists a constant $c_G > 0$ such that for all large enough n the minimax regret satisfies $R_n^*(G) \geq c_G \sqrt{n}$.*

Proof sketch By assumption there exists a pair of neighbouring actions (a, b). Define u as the centroid of $C_a \cap C_b$ and let $q = (\ell_a - \ell_b)/\|\ell_a - \ell_b\|^2$. For sufficiently small $\Delta > 0$ let $u_a = u - \Delta q$ and $u_b = u + \Delta q$. Then

$$\mathrm{D}(\mathbb{P}_{u_a}, \mathbb{P}_{u_b}) \leq n \sum_{i=1}^{E} \frac{(u_{ai} - u_{bi})^2}{u_{bi}} \leq C_G n \Delta^2 \,,$$

where $C_G > 0$ is a game-dependent constant. Let $\Delta = 1/\sqrt{n}$ and apply the ideas in the proof of Theorem 36.2. □

36.5 Policy for easy games

Fix a locally observable game $G = (\mathcal{L}, \Phi)$ with at least one pair of neighboring actions. We describe a policy called NeighborhoodWatch2. In every round the policy always chooses $A_t \in \cup_{a,b} \mathcal{N}_{ab}$ where the union is over pairs of neighboring actions. For example, in the partial monitoring game described in Example 36.7 the policy would only play actions 1, 2, 3 and 4. Removing degenerate actions can only increase the minimax regret, so from now on we assume that $[K] = \cup_{a,b \text{ neighbors}} \mathcal{N}_{ab}$. We let $\mathcal{A}$ be an arbitrary largest subset of Pareto optimal actions such that $\mathcal{A}$ does not contain actions that are duplicates of each other and $\mathcal{D} = [K] \setminus \mathcal{A}$ be the remaining actions. In each round t the policy performs four steps as described below.

Step 1 (Local games)

For each $k \in \mathcal{A}$ the policy maintains an exponential weights distribution over $\mathcal{A} \cup \mathcal{D} = [K]$, but concentrated on the intersection of the neighborhood $\mathcal{N}_k$ of k and $\mathcal{A}$ (recall that $\mathcal{N}_k$ contains the neighbors of k, some of which may be duplicates of each other). We denote this distribution by $Q_{tk} \in \mathcal{P}_{K-1}$. For $a \in [K]$, the value of Q_{tka} is given

$$Q_{tka} = \frac{\mathbb{I}_{\mathcal{N}_k \cap \mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)}{\sum_{b \in \mathcal{N}_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{skb}\right)},$$

where $\eta > 0$ is the learning rate and the $\tilde{Z}_{ska}$ are estimators of the loss difference $y_{sa} - y_{sk}$ and will be introduced in step four below. For actions $k \in \mathcal{D}$ we define $Q_{tka} = \mathbb{I}_{\mathcal{A}}(a)/|\mathcal{A}|$ to be the uniform distribution over $\mathcal{A}$.

Step 2 (Global game)

The next step is to merge the local distributions over small neighborhoods into a global distribution over $[K] = \mathcal{A} \cup \mathcal{D}$. A square matrix is **right stochastic** if it has positive entries and its rows sum to one. Such a $d \times d$ matrix describes a homogeneous Markov chain with state-space $[d]$ and row $i \in [d]$ of the matrix defines the distribution over the next-states. We have briefly met homogeneous Markov chains in Section 3.2. The following result is all that we need, the proof of which is to the reader (Exercise 36.8).

LEMMA 36.2 *Let Q_t be the right stochastic matrix with kth row equal to Q_{tk}. Then there exists a unique distribution $\tilde{P}_t$ such that $\tilde{P}_t^\top = \tilde{P}_t^\top Q_t$. Furthermore, this distribution is supported on $\mathcal{A}$.*

The distribution $\tilde{P}_t$ is called the **stationary distribution** of the Markov chain with kernel Q_t. It is supported on $\mathcal{A}$ because following Q_t never transitions to states outside of $\mathcal{A}$. The reader may at this point wonder about why were the actions in $\mathcal{D}$ even included in the first place: The answer is that we want $\tilde{P}_t$ to be defined over $[K]$ merely to simplify some expressions that follow. By rewriting the matrix multiplication we see that

$$\tilde{P}_{tk} = \sum_{a \in \mathcal{A}} \tilde{P}_{ta} Q_{tak}, \tag{36.10}$$

which we use repeatedly in the analysis that follows. In particular, this identity plays a key role in relating the regret to a weighted sum of 'local regrets'.

Step 3 (Redistribution)
Now $\tilde{P}_t$ is rebalanced to a new distribution P_t for which duplicate and degenerate actions in $\mathcal{D}$ are played with sufficient probability. This is done iteratively, starting with $P_t = \tilde{P}_t$. Then for each $d \in \mathcal{D}$ the algorithm finds actions $a, b \in \mathcal{A}$ such that $\ell_d = \alpha \ell_a + (1-\alpha)\ell_b$ for some $\alpha \in [0,1]$, which is possible by Lemma 36.1. Then P_t is updated so that some of the probability assigned to actions a and b is transferred to action d. After mass has been assigned to all degenerate actions the algorithm incorporates a small amount of fixed exploration. The complete procedure is given in Algorithm 22. This is done in such a way that the expected loss of playing according to P_t is approximately the same as $\tilde{P}_t$. The next lemma formalizes the properties of P_t that will be critical in what follows. The proof is left to the reader (Exercise 36.9).

LEMMA 36.3 *Assume $\gamma \in [0, 1/2]$, let $u \in \mathcal{P}_{E-1}$ and let $a, k \in \mathcal{A}$ be arbitrary neighbors. Then $P_t \in \mathcal{P}_{K-1}$ is a probability vector and the following hold:*

(a) $P_{ta} \geq \tilde{P}_{ta}/4$.

(b) $\left|\sum_{a=1}^{K} (P_{ta} - \tilde{P}_{ta}) \langle \ell_a, u \rangle \right| \leq \gamma$.

(c) $P_{tb} \geq \dfrac{\tilde{P}_{tk} Q_{tka}}{4K}$ *for any non-duplicate* $b \in \mathcal{N}_{ka}$.

(d) $P_{ta} \geq \gamma/K$.

(e) $P_{td} \geq \frac{\tilde{P}_{tk}}{4K}$ *for any* $d \in [K]$ *such that* $\ell_d = \ell_k$.

Step 4 (Acting and estimating)
By the definition of local observability, for each pair of neighboring actions a, b there exists a function $v^{ab} : [K] \times [F] \to \mathbb{R}$ satisfying Eq. (36.3) and with $v^{ab}(c, f) = 0$ whenever $c \notin \mathcal{N}_{ab}$. Even though a is not a neighbor of itself, for notational convenience we define $v^{aa}(c, f) = 0$ for all c, f. While the policy will work for any admissible choice of v^{ab}, the analysis suggests minimizing

$$V = \max_{a,b} \|v^{ab}\|_\infty$$

with the maximum over all pairs of neighbors.

In Exercise 36.12 you will show that if $|\mathcal{N}_{ab}| = 2$, then v^{ab} can be chosen so that $\|v^{ab}\|_\infty \leq 1 + F$ and that in the worst case this bound is tight. This result no longer holds for larger $\mathcal{N}_{ab}$ as discussed in the exercise.

In round t, the action A_t is chosen at random from P_t. The loss difference estimators are then computed by

$$\tilde{Z}_{tka} = \hat{Z}_{tka} - \beta_{tka} \,,$$

where $\hat{Z}_{tka}$ is an unbiased estimator of $y_{ta} - y_{tk}$ and β_{tka} is a bias term:

$$\hat{Z}_{tka} = \frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}} \quad \text{and} \quad \beta_{tka} = \eta V^2 \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}} . \tag{36.11}$$

The four steps described so far are summarized in Algorithm 22 below.

```
Input  $\mathcal{L}$, $\Phi$, $\eta$, $\gamma$
for $t \in 1, \dots, n$ do
    For $a, k \in [K]$ let
        $Q_{tka} = \mathbb{I}_{\mathcal{A}}(k) \frac{\mathbb{I}_{\mathcal{N}_k \cap \mathcal{A}}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)}{\sum_{b \in \mathcal{N}_k \cap \mathcal{A}} \exp\left(-\eta \sum_{s=1}^{t-1} \tilde{Z}_{ska}\right)} + \mathbb{I}_{\mathcal{D}}(k) \frac{\mathbb{I}_{\mathcal{A}}(a)}{|\mathcal{A}|}$
    Find distribution $\tilde{P}_t$ such that $\tilde{P}_t^\top = \tilde{P}_t^\top Q_t$
    Compute $P_t = (1-\gamma)\textsc{Redistribute}(\tilde{P}_t) + \frac{\gamma}{K}\mathbf{1}$ and sample $A_t \sim P_t$
    Compute loss-difference estimators for each $k \in \mathcal{A}$ and $a \in \mathcal{N}_k \cap \mathcal{A}$.
        $\hat{Z}_{tka} = \frac{\tilde{P}_{tk} v^{ak}(A_t, \Phi_t)}{P_{tA_t}}$
        $\beta_{tka} = \eta V^2 \sum_{b \in \mathcal{N}_{ak}} \frac{\tilde{P}_{tk}^2}{P_{tb}}$          (36.12)
        $\tilde{Z}_{tka} = \hat{Z}_{tka} - \beta_{tka}$
end for
function $\textsc{Redistribute}(p)$
    $q \leftarrow p$
    for $d \in \mathcal{D}$ do
        Find $a, b$ with $d \in \mathcal{N}_{ab}$ and $\alpha \in [0,1]$ such that $\ell_d = \alpha \ell_a + (1-\alpha)\ell_b$
        $c_a \leftarrow \frac{\alpha q_b}{\alpha q_b + (1-\alpha) q_a}$ and $c_b \leftarrow 1 - c_a$ and $\rho \leftarrow \frac{1}{2K} \min\left\{\frac{p_a}{q_a c_a}, \frac{p_b}{q_b c_b}\right\}$
        $q_d \leftarrow \rho c_a q_a + \rho c_b q_b$ and $q_a \leftarrow (1 - \rho c_a) q_a$ and $q_b \leftarrow (1 - \rho c_b) q_b$
    end for
    return $q$
end function
```

Algorithm 22: NeighborhoodWatch2

The next theorem bounds the regret of NeighborhoodWatch2 with high probability for locally observable games.

THEOREM 36.5 *Let $\hat{R}_n$ be the random regret*

$$\hat{R}_n = \max_{b \in [K]} \sum_{t=1}^n \langle \ell_{A_t} - \ell_b, u_t \rangle .$$

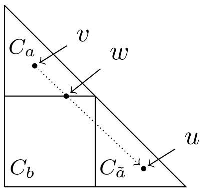


Figure 36.3 The construction used in the proof of Lemma 36.4.

Suppose that Algorithm 22 is run on locally observable $G = (\mathcal{L}, \Phi)$ and

$$\eta = \frac{1}{V}\sqrt{\frac{\log(K/\delta)}{nK}} \qquad \textit{and} \qquad \gamma = VK\eta\,.$$

Let $0 < \delta < 1$. Then with probability at least $1-\delta$ the regret is bounded by $\hat R_n \le C_G\sqrt{n\log(e/\delta)})$, where C_G is a constant depending on G, but not n, or δ.

By choosing $\delta = 1/n$ the following corollary is obtained.

COROLLARY 36.1 *Suppose that Algorithm 22 is run on locally observable $G = (\mathcal{L}, \Phi)$ with the same choices of η and γ as Theorem 36.5 and $\delta = 1/n$, then there exists a constant C'_G depending on G, but not n such that*

$$R_n \le C'_G\sqrt{n\log(n)}\,.$$

36.6 Upper bound for easy games

The first step is a simple lemma showing the regret can be localised to the neighborhood of the played action.

LEMMA 36.4 *There exists a constant $\varepsilon_G > 0$ depending only on G such that for all pairs of actions $a, \tilde a \in \mathcal{A}$ and $u \in C_{\tilde a}$ there exists an action $b \in \mathcal{N}_a \cap \mathcal{A}$ such that $\langle \ell_a - \ell_{\tilde a}, u\rangle \le \langle \ell_a - \ell_b, u\rangle/\varepsilon_G$.*

Proof Since $u \in C_{\tilde a}$, $0 \le \langle \ell_a - \ell_{\tilde a}, u\rangle$. The result is trivial if $a, \tilde a$ are neighbors or $\langle \ell_a - \ell_{\tilde a}, u\rangle = 0$. From now on assume that $\langle \ell_a - \ell_{\tilde a}, u\rangle > 0$ and that $a, \tilde a$ are not neighbors. Let v be the centroid of C_a. The idea is to choose $b \in \mathcal{N}_a \cap \mathcal{A}$ as that neighbor of a whose cell is the one that the line segment that connects v and u enters when leaving C_a. To be precise, if w lies in the intersection of the line segment connecting v and u and the boundary of C_a then b is a neighbor of a in $\mathcal{A}$ so that $w \in C_a \cap C_b$. Note that w is well-defined by the Jordan-Brouwer separation theorem (see the notes at the end of the chapter), and b is well-defined because $\mathcal{A}$ is a maximal duplicate-free subset of the Pareto optimal actions. Using twice that $\langle \ell_a - \ell_b, w\rangle = 0$, we calculate

$$\begin{aligned}\langle \ell_a - \ell_b, u\rangle = \langle \ell_a - \ell_b, u - w\rangle &= \frac{\|u-w\|_2}{\|v-w\|_2}\langle \ell_a - \ell_b, w - v\rangle \\ &= \frac{\|u-w\|_2}{\|v-w\|_2}\langle \ell_b - \ell_a, v\rangle > 0\,, \end{aligned} \tag{36.13}$$

where the second equality used that $w \neq v$ is a point of the line segment connecting v and u, hence $w - v$ and $u - w$ are parallel and share the same direction and $\|v - w\|_2 > 0$ (see Fig. 36.3), and the last inequality follows because v is the centroid of C_a and a, b are distinct Pareto optimal actions.

Let v_c be the centroid of C_c for any $c \in \mathcal{A}$. Then,

$$\begin{aligned}\frac{\langle \ell_a - \ell_{\tilde a}, u\rangle}{\langle \ell_a - \ell_b, u\rangle} = \frac{\langle \ell_a - \ell_{\tilde a}, w + u - w\rangle}{\langle \ell_a - \ell_b, u\rangle} &\overset{(a)}{\leq} \frac{\langle \ell_a - \ell_b, w\rangle + \langle \ell_a - \ell_{\tilde a}, u - w\rangle}{\langle \ell_a - \ell_b, u\rangle} \\ &\overset{(b)}{=} \frac{\langle \ell_a - \ell_{\tilde a}, u - w\rangle}{\langle \ell_a - \ell_b, u\rangle} \overset{(c)}{=} \frac{\|v-w\|_2\langle \ell_a - \ell_{\tilde a}, u - w\rangle}{\|u-w\|_2\langle \ell_b - \ell_a, v\rangle} \\ &\overset{(d)}{\leq} \frac{\|v-w\|_2\|\ell_a - \ell_{\tilde a}\|_2}{\langle \ell_b - \ell_a, v\rangle} \overset{(e)}{\leq} \frac{\sqrt{2E}}{\min_{c\in\mathcal{A}}\min_{d\in\mathcal{N}_c}\langle \ell_d - \ell_c, v_c\rangle} = \frac{1}{\varepsilon_G}\,,\end{aligned}$$

where (a) follows since by (36.13), $\langle \ell_a - \ell_b, u\rangle > 0$ and also because $w \in C_b$ implies that $\langle \ell_a - \ell_{\tilde a}, w\rangle \leq \langle \ell_a - \ell_b, w\rangle$, (b) follows since $\langle \ell_a - \ell_b, w\rangle = 0$, which is used in other steps as well. (c) uses (36.13), (d) is by Cauchy-Schwartz and in (e) we bounded $\|w - v\|_2 \leq \sqrt{2}$ and used that $\|\ell_a - \ell_{\tilde a}\|_2 \leq \sqrt{E}$ and $\langle \ell_b - \ell_a, v\rangle = \langle \ell_b - \ell_a, v_a\rangle \geq \min_{c\in\mathcal{A}}\min_{d\in\mathcal{N}_c}\langle \ell_d - \ell_c, v_c\rangle > 0$. The final equality serves as the definition of $1/\varepsilon_G$. □

LEMMA 36.5 *Let $\mathcal{H}$ be the set of functions $\phi : \mathcal{A} \to \mathcal{A}$ with $\phi(a) \in \mathcal{N}_a$ for all $a \in \mathcal{A}$ and define $a_n^* = \operatorname{argmin}_{a\in[K]} \sum_{t=1}^n \langle \ell_a, u_t\rangle$. Then, for any $(B_t)_{1\leq t\leq n}$ sequence of actions in $\mathcal{A}$,*

$$\sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle \leq \frac{1}{\varepsilon_G}\max_{\phi\in\mathcal{H}}\sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\,.$$

Lemma 36.5 With no loss of generality, we can assume that $a_n^* \in \mathcal{A}$ because $\mathcal{A}$ is a maximal duplicate-free subset of Pareto optimal actions. Apply the previous lemma on subsequences of rounds where $B_t = a$ for each $a \in \mathcal{A}$. □

LEMMA 36.6 *Let $\delta \in (0,1)$. Then with probability at least $1 - 2\delta$ it holds that*

$$\hat R_n \leq \gamma n + \frac{1}{\varepsilon_G}\sum_{k\in\mathcal{A}}\max_{b\in\mathcal{N}_k\cap\mathcal{A}}\sum_{t=1}^n \tilde P_{tk}\sum_{a\in\mathcal{A}} Q_{tka}\left(y_{ta} - y_{tb}\right) + \sqrt{8n\log(|\mathcal{H}|/\delta)}\,.$$

Proof For $t \in [n]$, let $B_t \sim \tilde P_t$. Define the surrogate regret $\hat R_n' = \sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle$. By the definition of A_t and B_t and Lemma 36.3 we have $\mathbb{E}_{t-1}[\langle \ell_{A_t} - \ell_{B_t}, u_t\rangle] \leq \gamma$. Furthermore, $|\langle \ell_a - \ell_b, u_t\rangle| \leq 1$ for all a, b. Therefore, by Hoeffding-Azuma, with probability at least $1 - \delta$,

$$\hat R_n \leq \hat R_n' + \gamma n + \sqrt{2n\log(1/\delta)}\,. \tag{36.14}$$

By Lemma 36.5, the surrogate regret is bounded in terms of the local regret:

$$\hat{R}'_n = \sum_{t=1}^n \langle \ell_{B_t} - \ell_{a_n^*}, u_t\rangle \leq \frac{1}{\varepsilon_G} \max_{\phi \in \mathcal{H}} \sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\,. \tag{36.15}$$

We prepare to use Hoeffding-Azuma again. Fix $\phi \in \mathcal{H}$ arbitrarily. Then,

$$\begin{aligned}\mathbb{E}_{t-1}\big[\langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle\big] &= \sum_{k\in\mathcal{A}} \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}\langle \ell_a - \ell_{\phi(k)}, u_t\rangle \\ &= \sum_{k\in\mathcal{A}} \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)})\,,\end{aligned}$$

where we used the fact that $\tilde{P}_{ta} = \sum_k \tilde{P}_{tk} Q_{tka}$. Hoeffding-Azuma's inequality now shows that with probability at least $1 - \delta/|\mathcal{H}|$,

$$\sum_{t=1}^n \langle \ell_{B_t} - \ell_{\phi(B_t)}, u_t\rangle \leq \sum_{k\in\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)}) + \sqrt{2n\log(|\mathcal{H}|/\delta)}\,.$$

The result is completed via a union bound over all $\phi \in \mathcal{H}$ and chaining with Eqs. (36.14) and (36.15), and noting that

$$\begin{aligned}\max_\phi \sum_{k\in\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)}) &\leq \sum_{k\in\mathcal{A}} \max_\phi \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{t\phi(k)}) \\ = \sum_{k\in\mathcal{A}} \underbrace{\max_{b\in\mathcal{N}_k\cap\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}(y_{ta} - y_{tb})}_{\hat{R}_{nk}}\,. \end{aligned} \tag{36.16}$$

□

Proof of Theorem 36.5 The proof has two steps: Bounding the local regret $\hat{R}_{nk}$ for each $k \in \mathcal{A}$, and then merging the bounds.

Step 1: Bounding the local regret

For the remainder of this step we fix $k \in \mathcal{A}$ and bound the local regret $\hat{R}_{nk}$. First, we need to massage the local regret into a form in which we can apply the result of Exercise 12.2 in Chapter 12. Let $Z_{tka} = \tilde{P}_{tk}(y_{ta} - y_{tk})$ and $\mathcal{G}_t$ be the σ-algebra generated by $(A_1, \ldots, A_t)$. Let $\mathcal{G} = (\mathcal{G}_t)_{t=0}^n$ be the associated filtration. A simple rewriting shows that

$$\hat{R}_{nk} = \max_{b\in\mathcal{N}_k\cap\mathcal{A}} \sum_{t=1}^n \tilde{P}_{tk} \sum_{a\in\mathcal{A}} Q_{tka}\left(y_{ta} - y_{tb}\right) = \max_{b\in\mathcal{N}_k\cap\mathcal{A}} \sum_{t=1}^n \sum_{a\in\mathcal{A}} Q_{tka}\left(Z_{tka} - Z_{tkb}\right)\,.$$

In order to apply the result in Exercise 12.2 we need to check the conditions. Since $(P_t)_t$ and $(\tilde{P}_t)_t$ are $\mathcal{G}$-predictable it follows that $(\beta_t)_t$ and $(Z_t)_t$ are also $\mathcal{G}$-predictable. Similarly, $(\hat{Z}_t)_t$ is $\mathcal{G}$-adapted because $(A_t)_t$ and $(\Phi_t)_t$ are $\mathcal{G}$-adapted. It remains to show that assumptions *(a–d)* are satisfied. For *(a)* let $a \in \mathcal{N}_k \cap \mathcal{A}$. By part (d) of Lemma 36.3 we have $P_{tb} \geq \gamma/K$ for all t and $b \in [K]$. Furthermore, $|v^{ak}(A_t, \Phi_t)| \leq V$ so that $\eta|\hat{Z}_{tka}| = |\eta\tilde{P}_{tk}v^{ak}(A_t,\Phi_t)/P_{tA_t}| \leq \eta VK/\gamma = 1$, where the equality follows from the choice of γ. Assumption *(b)* is satisfied in a similar way with $\eta\beta_{tka} = \eta^2 V^2 \sum_{b\in\mathcal{N}_{ak}} \tilde{P}_{tk}^2/P_{tb} \le \eta^2 K^2 V^2/\gamma = \eta V \le 1$, where in the last inequality we used the definition of η and assumed that $n \ge \log(K/\delta)$. To make sure that the regret bound holds even for smaller values of n, we require $C_G \ge K\sqrt{\log(eK)}$ so that when $n < K^2\log(K/\delta)$, the regret bound is trivial. For assumption *(c)*, we have

$$\mathbb{E}_{t-1}[\hat{Z}_{tka}^2] = \mathbb{E}_{t-1}\left[\left(\frac{\tilde{P}_{tk}v^{ak}(A_t,\Phi_t)}{P_{tA_t}}\right)^2\right] \le V^2\tilde{P}_{tk}^2\mathbb{E}_{t-1}\left[\frac{\mathbb{I}_{\mathcal{N}_{ak}}(A_t)}{P_{tA_t}^2}\right]$$
$$= V^2\sum_{b\in\mathcal{N}_{ak}}\frac{\tilde{P}_{tk}^2}{P_{tb}} = \frac{\beta_{tka}}{\eta}\,.$$

Finally *(d)* is satisfied by the definition of v^{ak} and the fact that $P_t \in \operatorname{ri}(\mathcal{P}_{K-1})$. The result of Exercise 12.2 shows that with probability at least $1-(K+1)\delta$,

$$\hat{R}_{nk} \le \frac{3\log(1/\delta)}{\eta} + 5\sum_{t=1}^n\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\beta_{tka} + \eta\sum_{t=1}^n\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\hat{Z}_{tka}^2\,.$$

Step 2: Aggregating the local regret
Using the result from the previous step in combination with a union bound over $k\in\mathcal{A}$ we have that with probability at least $1-K(K+1)\delta$,

$$\sum_{k\in\mathcal{A}}\hat{R}_{nk} \le \frac{3K\log(1/\delta)}{\eta} + 5\sum_{t=1}^n\sum_{k\in\mathcal{A}}\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\beta_{tka} + \eta\sum_{t=1}^n\sum_{k\in\mathcal{A}}\sum_{a\in\mathcal{N}_a\cap\mathcal{A}} Q_{tka}\hat{Z}_{tka}^2\,. \tag{36.17}$$

For bounding the second term we use the definition of β_{tka} from (36.11) and write

$$\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\beta_{tka} = \eta V^2\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\sum_{b\in\mathcal{N}_{ak}}\frac{\tilde{P}_{tk}^2}{P_{tb}} = \eta V^2\tilde{P}_{tk}\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\sum_{b\in\mathcal{N}_{ak}}\frac{\tilde{P}_{tk}}{P_{tb}}\,.$$

We now split the sum that runs over $b\in\mathcal{N}_{ak}$ into two, separating duplicates of k and the rest:

$$\begin{aligned}\sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\sum_{b\in\mathcal{N}_{ak}}\frac{\tilde{P}_{tk}}{P_{tb}} &= \sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\sum_{b:\ell_b=\ell_k}\frac{\tilde{P}_{tk}}{P_{tb}} + \sum_{a\in\mathcal{N}_k\cap\mathcal{A}} Q_{tka}\sum_{b\in\mathcal{N}_{ak}:\ell_b\neq\ell_k}\frac{\tilde{P}_{tk}}{P_{tb}}\\ &= \sum_{b:\ell_b=\ell_k}\frac{\tilde{P}_{tk}}{P_{tb}} + \sum_{a\in\mathcal{N}_k\cap\mathcal{A}}\sum_{b\in\mathcal{N}_{ak}:\ell_b\neq\ell_k}\frac{Q_{tka}\tilde{P}_{tk}}{P_{tb}}\\ &\le 4K\left(\sum_{b:\ell_b=\ell_k}1 + \sum_{a\in\mathcal{N}_k\cap\mathcal{A}}\sum_{b\in\mathcal{N}_{ak}:\ell_b\neq\ell_k}1\right) \le 4K^2\,,\end{aligned} \tag{36.18}$$

where the first equality used that $\sum_a Q_{tka} = 1$, the second to last inequality follows using parts (c) and (e) of Lemma 36.3, stationarity of $\tilde{P}_t$, and the last inequality uses a simple counting argument. Details of the arguments needed to show the last two inequalities are left to reader in Exercise 36.10. Summing over all rounds and $k \in \mathcal{A}$ yields

$$5 \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \beta_{tka} \leq 20 \eta n K^2 V^2 \,.$$

For the last term in Eq. (36.17) we use the definition of $\hat{Z}_{tka}$ and Parts (c) and (e) of Lemma 36.3 to show that

$$\begin{aligned}
\eta \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} Q_{tka} \hat{Z}_{tka}^2 &= \eta \sum_{t=1}^{n} \sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \frac{Q_{tka} \tilde{P}_{tk}^2 v^{ak}(A_t, \Phi_t)^2}{P_{tA_t}^2} \\
&\leq \eta V^2 \sum_{t=1}^{n} \frac{1}{P_{tA_t}} \sum_{k \in \mathcal{A}} \tilde{P}_{tk} \sum_{a \in \mathcal{N}_k \cap \mathcal{A}} \frac{Q_{tka} \tilde{P}_{tk} \mathbb{I}_{\mathcal{N}_{ak}}(A_t)}{P_{tA_t}} \\
&\leq 4 \eta K V^2 \sum_{t=1}^{n} \frac{1}{P_{tA_t}} \,,
\end{aligned}$$

where the last step follows by splitting the sum over a into two based on whether A_t is a duplicate of k and following an argument similar to the one used to prove (36.18). Now, from Part (d) of Lemma 36.3, $(\gamma/K)(1/P_{ta}) \leq 1$ for all a, and in particular, holds for $a = A_t$. Furthermore, $\mathbb{E}_{t-1}[1/P_{tA_t}] = K$ and $\mathbb{E}_{t-1}[1/P_{tA_t}^2] = \sum_a 1/P_{ta} \leq K^2/\gamma$. By the result in Exercise 5.17 we get that it holds that with probability at least $1 - \delta$,

$$\sum_{t=1}^{n} \frac{1}{P_{tA_t}} \leq 2nK + \frac{K \log(1/\delta)}{\gamma} \,.$$

Another union bound shows that with probability at least $1 - (1 + K(K+1))\delta$,

$$\sum_{k \in \mathcal{A}} \hat{R}_{nk} \leq \frac{3K \log(1/\delta)}{\eta} + 28 \eta n V^2 K^2 + 4VK \log(1/\delta) \,.$$

The result follows from the definition of η, Lemma 36.6 and the definition of $\hat{R}_{nk}$. □

36.7 Proof of the classification theorem

Almost all the results are now available to prove Theorem 36.1. In Section 36.4 we showed that if G is globally observable and not locally observable, then $R_n^*(G) = \Omega(n^{2/3})$. We also proved that if G is locally observable and has neighbors, then $R_n^*(G) = \Omega(\sqrt{n})$. This last result is complemented by the policy and analysis in Section 36.6 where we showed that for locally observable problems $R_n^*(G) = O(\sqrt{n \log(n)})$. Finally we proved that if G is not globally observable, then $R_n^*(G) = \Omega(n)$. All that remains is to prove that *(a)* if G has no neighboring actions, then $R_n^*(G) = 0$ and *(b)* if G is globally observable, but not locally observable, then $R_n^*(G) = O(n^{2/3})$.

THEOREM 36.6 *If G has no neighboring actions, then $R_n^*(G) = 0$.*

Proof Since G has no neighboring actions, there exists an action a such that $C_a = \mathcal{P}_{E-1}$ and the policy that chooses $A_t = a$ for all rounds suffers no regret. □

THEOREM 36.7 *If G is globally observable, then $R_n^*(G) = O(n^{2/3})$.*

Proof sketch Let $\mathcal{A} \subseteq [K]$ be the set of Pareto optimal actions and $a_\circ \in \mathcal{A}$. Use the definition of global observability to show that each $a \in \mathcal{A}$ there exists a function $h^a : [K] \times [F] \to \mathbb{R}$ such that

$$\sum_{b=1}^K h^a(b, \Phi(b,i)) = \ell_{ai} - \ell_{a_\circ i} \qquad \text{for all } i \in [E]\,.$$

Then define unbiased loss estimator $\hat{\Delta}_{ta} = h^a(A_t, \Phi_t)/P_{tA_t}$, where

$$P_{ta} = (1-\gamma)\frac{\mathbb{I}_{\mathcal{A}}(a)\exp\left(-\eta\sum_{s=1}^{t-1}\hat{\Delta}_{sa}\right)}{\sum_{b\in\mathcal{A}}\exp\left(-\eta\sum_{s=1}^{t-1}\hat{\Delta}_{sb}\right)} + \frac{\gamma}{K}\,.$$

The result is completed by repeating the standard analysis of the exponential weights algorithm (or mirror descent with negentropy potential) and optimizing γ and η. □

36.8 Notes

1 A nonempty set $L \subseteq \mathbb{R}^n$ is a **linear subspace** of $\mathbb{R}^n$ if $\alpha v + \beta w \in L$ for all $\alpha, \beta \in \mathbb{R}$ and $v, w \in L$. If L and M are linear subspaces of $\mathbb{R}^n$, then $L \oplus M = \{v + w : L \in L, w \in M\}$. The **orthogonal complement** of linear subspace L is $L^\perp = \{v \in \mathbb{R}^n : \langle u, v\rangle = 0 \text{ for all } u \in L\}$. The following properties are easily checked: *(i)* $L^\perp$ is a linear subspace, *(ii)* $(L^\perp)^\perp = L$ and *(iii)* $(L \cap M)^\perp = L^\perp \oplus M^\perp$.

2 Let $A \in \mathbb{R}^{m\times n}$ be a matrix and recall that matrices of this form correspond to linear maps from $\mathbb{R}^n \to \mathbb{R}^m$ where the function $A : \mathbb{R}^n \to \mathbb{R}^m$ is given by matrix multiplication, $A(x) = Ax$. The **image** of A is $\operatorname{im}(A) = \{Ax : x \in \mathbb{R}^n\}$ and the **kernel** is $\ker(A) = \{x \in \mathbb{R}^n : Ax = 0\}$. Notice that $\operatorname{im}(A) \subseteq \mathbb{R}^m$ and $\ker(A) \subseteq \mathbb{R}^n$. One can easily check that $\operatorname{im}(A)$ and $\ker(A^\top)$ are linear subspaces and an elementary theorem in linear algebra says that $\operatorname{im}(A) \oplus \ker(A^\top) = \mathbb{R}^m$ for any matrix $A \in \mathbb{R}^{m\times n}$. Finally, if $u \in \operatorname{im}(A)$ and $v \in \ker(A^\top)$, then $\langle u, v\rangle = 0$. There are probably hundreds of introductory texts on linear algebra. A short and intuitive exposition is by Axler [1997].

3 Given a set $A \subseteq \mathbb{R}^d$ the **affine hull** is the set

$$\operatorname{aff}(A) = \left\{ \sum_{i=1}^k \alpha_i x_k : k > 0,\ \alpha \in \mathbb{R}^k,\ x_i \in A \text{ for all } i \in [k] \text{ and } \sum_{i=1}^k \alpha_i = 1 \right\} .$$

Its dimension is the smallest m such that there exist vectors $v_1, \dots, v_m \in \mathbb{R}^d$ such that $\operatorname{aff}(A) = x_\circ + \operatorname{span}(v_1, \dots, v_m)$ for any $x_\circ \in A$.

4 We introduced the stochastic variant of partial monitoring to prove our lower bounds. Of course our upper bounds also apply to this setting, which means the classification theorem also holds in the stochastic case. The interesting question is to understand the problem-dependent regret, which for partial monitoring problem $G = (\mathcal{L}, \Phi)$ is

$$R_n(\pi, u, G) = \max_{a \in [K]} \mathbb{E}\left[\sum_{t=1}^n \langle \ell_{A_t} - \ell_a, U_t \rangle\right] ,$$

where $U, U_1, \dots, U_n$ is a sequence of independent and identically distributed random vectors with $U_t \in \{e_1, \dots, e_E\}$ and $\mathbb{E}[U] = u \in \mathcal{P}_{E-1}$. Provided G is not hopeless one can derive an algorithm for which the regret is logarithmic, and like in bandits there is a sense of asymptotic optimality. The open research question is to understand the in-between regime where the horizon is not yet large enough that the asymptotically optimal logarithmic regret guarantees become meaningful, but not so small that minimax is acceptable.

5 In the proof of Lemma 36.4 we used the overpowered Jordan-Brouwer separation theorem to guarantee that the line segment that connects u with the centroid v of C_a has a nonempty intersection with the boundary of C_a. Here, u was a point that lied outside of C_a. The Jordan-Brouwer separation theorem generalizes the Jordan curve theorem, which states that every simple closed planar curve separates the plane into a bounded interior and an unbounded exterior region so that the boundary of both regions is the said planar curve. The Jordan-Brouwer theorem states that the same holds in higher dimensions where the closed planar curve becomes a topological sphere, which is the image of the unit sphere of $\mathbb{R}^d$ under some continuous injective map from the sphere into $\mathbb{R}^d$. To use the theorem we view the simplex $\mathcal{P}_{E-1}$ as a subset of $\mathbb{R}^{E-1}$ by dropping the last entry in the coordinate representation of the points of $\mathcal{P}_{E-1}$. Then the boundary of C_a can be seen as a topological sphere in $\mathbb{R}^{E-1}$ and v belongs to the interior, while u belongs to the exterior region created by the boundary of C_a. The line segment connecting u and v will pass through the boundary of both regions, which happens to be the boundary of C_a, showing that the intersection of the line segment and the boundary of C_a is nonempty. Note that the argument does not show that the intersection has a single point and we did not need this either. Nevertheless, it is not hard to see that this is also true. The standard proof of the Jordan-Brouwer is an application of algebraic topology [Hatcher, 2002, §2.B].

6 Partial monitoring has many potential applications. We already mentioned

dynamic pricing and spam filtering. In the latter case acquiring the true label comes at a price, which is a typical component of hard partial monitoring problems. In general there are many setups where the learner can pay extra for high quality information. For example, in medical diagnosis the doctor can request additional tests before recommending a treatment plan, but these cost time and money. Yet another potential application is quality testing in factory production where the quality control team can choose which items to test (at great cost).

7 There are many possible extensions to the partial monitoring framework. We have only discussed problems where the number of actions/feedbacks/outcomes are potentially infinite, but nothing prevents studying a more general setting. Suppose the learner chooses a sequence of real-valued outcomes $i_1, \dots, i_n$ with $i_t \in [0,1]$. In each round the learner chooses $A_t \in [K]$ and observes $\Phi_{A_t}(i_t)$ where $\Phi_a : [0,1] \to \Sigma$ is a known feedback function. The loss is determined by a collection of known functions $\mathcal{L}_a : [0,1] \to [0,1]$. We do not know of any systematic study of this setting. The reader can no doubt imagine generalizing this idea to infinite action sets or introducing a linear structure for the loss.

8 A pair of Pareto-optimal actions (a,b) are called **weak neighbors** if $C_a \cap C_b \neq \emptyset$ and **pairwise observable** if there exists a function v satisfying Eq. (36.3) and with $v(c,f) = 0$ whenever $c \notin \{a,b\}$. A partial monitoring problem is called a **point-locally observable game** if all weak neighbours are pairwise observable. All point-locally observable games are locally observable, but the converse is not true. Bartók [2013] designed a policy for this type of game for which

$$R_n \leq \frac{1}{\varepsilon_G} \sqrt{K_{\text{loc}} n \log(n)}\,,$$

where $\varepsilon_G > 0$ is a game-dependent constant and K_{loc} is the size of the largest $A \subseteq [K]$ of Pareto optimal actions such that $\cap_{a \in A} C_a \neq \emptyset$. Using a different policy, Lattimore and Szepesvári [2018] have shown that as the horizon grows the game-dependence diminishes so that

$$\lim_{n \to \infty} \frac{R_n}{\sqrt{n}} \leq 8(2+F)\sqrt{2K_{\text{loc}} \log(K)}\,.$$

9 Linear regret is unavoidable in hopeless games, but that does not mean there is nothing to play for. Rustichini considered a version of the regret that captures the performance of policies in this hard setting. Given $p \in \mathcal{P}_{E-1}$ define set $\mathcal{I}(p) \subseteq \mathcal{P}_{E-1}$ by

$$\mathcal{I}(p) = \left\{ q \in \mathcal{P}_{E-1} : \sum_{i=1}^{E} (p_i - q_i) \mathbb{I}\{\Phi_{ai} = f\} \text{ for all } a \in [K] \text{ and } f \in [F] \right\}.$$

This is the set of distributions over the outcomes that are indistinguishable from p by the learner using any actions. Then define

$$f(p) = \max_{q \in \mathcal{I}(p)} \min_{a \in [K]} \sum_{i=1}^{E} q_i \mathcal{L}_{ai} \,.$$

Rustichini [1999] proved there exist policies such that

$$\lim_{n \to \infty} \max_{i_{1:n}} \mathbb{E}\left[\frac{1}{n} \sum_{t=1}^{n} \mathcal{L}_{A_t i_t} - f(\bar{u}_n)\right] = 0 \,,$$

where $\bar{u}_n = \frac{1}{n} \sum_{t=1}^{n} e_{i_t} \in \mathcal{P}_{E-1}$ is the average outcome chosen by the adversary. Intuitively this means the learner does not compete with the best action in hindsight with respect to the actual outcomes. Instead, the learner competes with the best action in hindsight with respect to an outcome sequence that is indistinguishable from the actual outcome sequence. Rustichini did not prove rates on the convergence of the limit. This has been remedied recently and we give some references in the bibliographic remarks.

10 Finally, we want to emphasize that partial monitoring is still quite poorly understood. We do not know how the regret should depend on E, F, K or the structure of G. Lower bounds that depend on these quantities are also missing and the lower bounds proven in Section 36.4 are surely very conservative. We hope this chapter inspires more activity in this area. The setting described in the previous note is even more wide open, with even the dependence on n still not completely nailed down.

36.9 Bibliographical remarks

The first work on partial monitoring is by Rustichini [1999], who focussed on the finding Hannan consistent policies in the adversarial setting. Rustichini shows how to reduce the problem to Blackwell approachability (see Cesa-Bianchi and Lugosi [2006]) and uses this to deduce the existence of a Hannan consistent strategy. Rustichini also used a slightly different notion of regret, which eliminates the hopeless games. The first nonasymptotic result in the setting of this chapter is due to Piccolboni and Schindelhauer [2001] where a policy with regret $O(n^{3/4})$ is given for problems that are not hopeless. Cesa-Bianchi et al. [2006] reduced the dependence to $O(n^{2/3})$ and proved a wide range of other results for specific classes of problems. The classification theorem when $E = 2$ is due to Bartók et al. [2010] (extended version: Antos et al. [2013]). The classification of general partial monitoring games is by Bartók et al. [2014]. The neighborhood watch policy is due to Foster and Rakhlin [2012]. The policy presented here is a simplification of that algorithm [Lattimore and Szepesvári, 2018]. The policies mentioned in Note 8 are due to Bartók [2013] and Lattimore and Szepesvári [2018]. We warn the reader that neighbors are defined differently by Foster and Rakhlin [2012] and Bartók [2013], which can lead to confusion. Additionally, although both papers are largely correct, in both cases the core proofs contain errors that cannot be resolved without changing the policies [Lattimore and Szepesvári, 2018]. There is also a growing literature on the stochastic setting where it is common to study both minimax and asymptotic bounds. In the latter case one can obtain asymptotically optimal logarithmic regret for games that are not hopeless. We refer the reader to papers by Bartók et al. [2012], Vanchinathan et al. [2014], Komiyama et al. [2015b] as a good starting place. As we mentioned, partial monitoring can model problems that lie between bandits and full information. There are now several papers on this topic, but in more restricted settings and consequentially with more practical algorithms and bounds. One such model is when the learner is playing actions corresponding to vertices on a graph and observes the losses associated with the chosen vertex and its neighbours [Mannor and Shamir, 2011, Alon et al., 2013]. A related result is in the finite-armed Gaussian setting where the learner selects an action $A_t \in [K]$ and observes a Gaussian sample from each arm, but with variances depending on the chosen action. Like partial monitoring this problem exhibits many challenges and is not yet well understood [Wu et al., 2015]. We mentioned in Note 9 that for hopeless games the definition of the regret can be refined. A number of authors have studied this setting with sublinear regret guarantees. As usual, the price of generality is that the bounds are correspondingly a bit worse [Perchet, 2011, Mannor and Shimkin, 2003, Mannor et al., 2014].

36.10 Exercises

36.1 Let $\mathcal{X} \subseteq \mathcal{Y} \subseteq \mathbb{R}^d$ and $\dim(\mathcal{X}) = \dim(\mathcal{Y})$. Prove that $\operatorname{aff}(\mathcal{X}) = \operatorname{aff}(\mathcal{Y})$.

36.2 Calculate the neighborhood structure, cell decomposition and action classification for each of the examples in this chapter.

36.3 Apples arrive sequentially from the farm to a processing facility. Most apples are fine, but occasionally there is a rotten one. The only way to figure out whether an apple is good or rotten is to taste it. For some reason customers do not like bite-marks in the apples they buy, which means that tested apples cannot be sold. Good apples yield a unit reward when sold, while the sale of a bad apple costs the company $c > 0$.

(a) Formulate this problem as a partial monitoring problem: Determine $\mathcal{L}$ and Φ.
(b) What is the minimax regret in this problem?
(c) What do you think about this problem? Will actual farmers be excited about your analysis?

36.4 Let $G = (\mathcal{L}, \Phi)$ be a partial monitoring game with $K = 2$ actions. Prove that G is either trivial, hopeless or easy.

36.5 Complete the last step in the proof of Theorem 36.2.

36.6 Prove Theorem 36.4.

36.7 Prove Theorem 36.3.

36.8 In this exercise you will prove the existence of a stationary distribution. Let $P \in [0,1]^{d\times d}$ be right stochastic and $A_n = \frac{1}{n}\sum_{t=0}^{n-1} P^t$. Show that:

(a) A_n is right stochastic.

(b) $A_n + \frac{1}{n}(P^n - I) = A_n P = P A_n$.

(c) $P^* = \lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} P^t$ exists.

(d) $P^*P = PP^* = P^*P^* = P^*$.

(e) There exists a stationary distribution.

(f) Prove Lemma 36.2.

For Parts (c) and (d) you will likely find it useful that the space of right stochastic matrices is compact. Then show that all cluster points of (A_n) are the same.

36.9 Prove Lemma 36.3.

36.10 Prove the last two inequalities shown in Eq. (36.18). In particular, let $k \in \mathcal{A}$ and show that:

(a) For any $b \in [K]$ such that $\ell_b = \ell_k$, $\tilde{P}_{tk}/P_{tb} \le 4K$;

(b) For any $a \in \mathcal{N}_k \cap \mathcal{A}$, $b \in \mathcal{N}_{ak}$ such that $\ell_b \ne \ell_k$, $Q_{tka}\tilde{P}_{tk}/P_{tb} \le 4K$;

(c) The sets $S = \{b \in [K] : \ell_b = \ell_k\}$ and the sets $S_a = \{b \in [K] : b \in \mathcal{N}_{ak}, \ell_b \ne \ell_k\}$ where $a \in \mathcal{N}_k \cap \mathcal{A}$ are all disjoint. Hence,

$$\sum_{b:\ell_b=\ell_k} 1 + \sum_{a\in\mathcal{N}_k\cap\mathcal{A}}\ \sum_{b\in\mathcal{N}_{ak}:\ell_b\ne\ell_k} 1 \le K\,.$$

(d) Put things together and show that the bound of Eq. (36.18) indeed holds.

36.11 Complete the details to prove Theorem 36.7.

36.12 Suppose that a and b are globally observable and let $v : [K]\times[F] \to \mathbb{R}$ be a function satisfying Eq. (36.3).

(a) Show that if a, b are pairwise observable, then v can be chosen so that $\|v\|_\infty \le 1 + F$.

(b) Next let $F = 2$ and construct a game and pair of actions a, b (not pairwise observable) such that for all v satisfying Eq. (36.3), $\|v\|_\infty \ge c^K$ for constant $c > 1$.

36.13 Consider $G = (\mathcal{L}, \Phi)$ given by

$$\mathcal{L} = \begin{pmatrix} 1 & 0 & 1 & 0 & \cdots & 1 & 0 \\ 0 & 1 & 0 & 1 & \cdots & 0 & 1 \end{pmatrix} \qquad \text{and}$$

$$\Phi = \begin{pmatrix} 1 & 2 & 2 & 3 & 3 & 4 & \cdots & F-1 & F-1 & F \\ 1 & 1 & 2 & 2 & 3 & 3 & \cdots & F-2 & F-1 & F-1 \end{pmatrix} .$$

(a) Show this game is locally observable.
(b) Prove there exists a universal constant $c > 0$ such that $R_n^*(G) \geq c(F-1)\sqrt{n}$.

The source for previous exercise is the paper by the authors [Lattimore and Szepesvári, 2018].

36.14 Complete the necessary modification of Lemma 15.1 to show that Eq. (36.7) is true.

36.15 Write a program that accepts as input matrices $\mathcal{L}$ and Φ and outputs the classification of the game.

36.16 In this experiment we test NeighborhoodWatch2 empirically on the spam game with stochastic adversary.

(a) Implement NeighborhoodWatch2.
(b) Apply your algorithm to the spam game for a variety of choices of c and stochastic adversary. Try to stress your algorithm as much as possible (for each c choose the most challenging u).
(c) Plot your results from the previous part. Tell an interesting story.