Theoretical Machine Learning
Lecture 9

Instructor: Haipeng Luo

1 Partial Monitoring

In this lecture, we continue to discuss learning with partial information, and focus on a problem called *Partial Monitoring* that allows a very general feedback model and greatly generalizes multi-armed bandits. Indeed, recall that in MAB what the learner observes is exactly the loss she suffers. However, there are many other situations where the learner only gets to observe information indirectly related to the actual loss, and as we will see partial monitoring easily captures these problems.

Specifically, a (finite) partial monitoring problem is defined by a loss matrix $\ell \in [0,1]^{K\times d}$ and a *feedback matrix* $\Phi \in \Sigma^{K\times d}$ where K is the number of actions for the learner, d is the number of outcomes for the environment, and Σ is an arbitrary set of alphabets containing all possible observations for the learner. Both ℓ and Φ are known to the learner. Ahead of time, the environment decides n outcomes $z_1, \ldots, z_n \in [d]$ (hence an oblivious environment). Then the learning protocol proceeds in n rounds. For each round $t = 1, \ldots, n$, the learner selects an action $a_t \in [K]$, suffers loss $\ell(a_t, z_t)$, and importantly only observes $\Phi(a_t, z_t)$. The goal of the learner is as usual to minimize regret against the best fixed action in hindsight:

$$\mathrm{Reg}_n = \sum_{t=1}^{n} \ell(a_t, z_t) - \sum_{t=1}^{n} \ell(a^\star, z_t) \quad \text{where } a^\star \in \operatorname*{argmin}_{a\in[K]} \sum_{t=1}^{n} \ell(a, z_t)$$

The flexibility of having a general feedback matrix Φ allows us to capture many different problems under this framework. Below are a few examples.

Full-information problems. A natural way to encode a learning problem with full information is to set $\Phi(a, z) = z$ for all $a \in [K]$ and $z \in [d]$. That is, no matter what the learner chooses, the actual outcome is observed, which is basically the general online learning protocol we focused on for most of the previous lectures (except that here for simplicity both the action space and the outcome space are finite). Note that, however, this is not the only way to encode a full-information problem. In fact, as long as each row of Φ consists of d distinct elements, then it essentially captures the exact same full-information problem, because clearly the learner can still infer the outcome just based on the observation $\Phi(a, z)$.

Bandit problems. Generalizing the well-known multi-armed bandits problem, the term "bandit feedback/information" usually refers to any problems where the learner observes exactly the loss she suffers. These problems can be modeled as partial monitoring by simply setting $\Phi = \ell$. Take MAB as an example. Assuming losses are all binary (which is in fact without loss of generality), we can set $d = 2^K$ and let the columns of ℓ (which is also Φ) be all the 2^K possible binary vectors in K dimension.

Apple tasting. Apple tasting is one of the simplest problems that are neither full-information nor bandit-information. Imagine a task of classifying a sequence of apples as "good for sale" or "rotten". The loss is 0 if the prediction is correct or 1 otherwise (as in typical binary classification problems). However, only when we predict "rotten" can we actually open the apple and see if it is indeed rotten, otherwise it is sent for sale and we will never know if we predict correctly or not. This can be modeled as partial monitoring by taking

$$\ell = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 2 & 2 \\ 1 & 0 \end{pmatrix},$$

where the first (or second) row corresponds to predicting "good for sale" (or "rotten"), and the first (or second) column corresponds to the apple being actually good (or rotten). An observation from the first row of Φ gives no information at all on the actual outcome, while an observation from the second row reveals everything. Note that there are again many other different ways to represent Φ.

Label efficient learning. Label efficient learning refers to a broad class of problems where querying the true label/outcome is costly and the learner needs to trade-off this cost with information. Consider the following very simple example of classifying emails as spam or not. The spam detector usually does not receive feedback from the users, but in case a very hard instance appears it can choose to ask the user explicitly for answer. Clearly the detector should avoid doing this too often, and to ensure this we can assign a constant loss $c \in (0, 1)$ for each query from the user. Therefore, we can model the problem as partial monitoring by taking

$$\ell = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ c & c \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}, \tag{1}$$

where the last row corresponds to querying the user, which is the only action that reveals the true outcome, but on the other hand incurs loss c no matter what the outcome is.

Dynamic pricing. As the last example, consider a vendor trying to dynamically adjust the price of a product with the goal of maximizing revenue. Specifically, each day the vendor first decides a price a_t, say either $1, 2, \ldots$, or K dollars, then a costumer comes with a secrete acceptable price z_t in mind, also in $[K]$ for simplicity (so $d = K$). The costumer purchases the product if and only if $a_t \leq z_t$, in which case the loss of the vendor is $z_t - a_t$, the extra money she would have been able to earn should the product was priced higher at z_t. Otherwise, no transaction happens and the vendor pays for some constant loss c (for storage fee for example). Importantly, the vendor only observes binary feedback: whether the transaction happens or not, and in particular, she does not know the actual loss she suffers if a transaction happens. To model this as partial monitoring, we can take

$$\ell = \begin{pmatrix} 0 & 1 & 2 & \cdots & K-1 \\ c & 0 & 1 & \cdots & K-2 \\ c & c & 0 & \cdots & K-3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c & c & c & \cdots & 0 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} \checkmark & \checkmark & \checkmark & \cdots & \checkmark \\ \text{✗} & \checkmark & \checkmark & \cdots & \checkmark \\ \text{✗} & \text{✗} & \checkmark & \cdots & \checkmark \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{✗} & \text{✗} & \text{✗} & \cdots & \checkmark \end{pmatrix}.$$

2 Classification Theorem

By now you should be convinced that partial monitoring is for sure general enough to capture many problems, but is it too general to be meaningful/solvable? Indeed, clearly there are examples where sublinear regret is impossible, such as

$$\ell = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{2}$$

because essentially the learner receives no feedback at all and cannot figure out the better action. On the other hand, there are also trivial problems such as

$$\ell = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \tag{3}$$

where even though the learner receives no feedback at all, she can ensure exactly 0 regret by always picking the first action (since it is always better than the second one no matter what the outcome is). Furthermore, we have also discussed that for full/bandit-information problems, by applying Hedge/Exp3 we can achieve $\mathcal{O}(\sqrt{n})$ regret and this is optimal in general.

So based on these observations we know that there are different kinds of partial monitoring problems with different minimax regret bounds. A somewhat surprising result is that there are exactly four different kinds of such problems — we have seen three of them, with minimax regret 0, $\Theta(\sqrt{n})$, and $\Theta(n)$, and it turns out there is exactly one more class of problems with minimax regret $\Theta(n^{2/3})$. This is the classification theorem we will discuss next.

To formally introduce this result, we need a couple of concepts (examples are deferred to the end of this section). First, let's consider when an action can be optimal. Note that the benchmark we want to compare with in the regret definition is

$$\min_{a\in[K]} \sum_{t=1}^{n} \ell(a, z_t) = n \min_{a\in[K]} \langle \ell_a, u\rangle$$

where we use ℓ_a to denote the a-th row of ℓ and $u = \frac{1}{n}\sum_{t=1}^{n} e_{z_t} \in \Delta(d)$ is the outcome frequency distribution ($e_1, \ldots, e_d$ are standard basis vectors). The *cell* associated with an action a is then defined as the set of all frequency distributions where a is optimal:

$$C_a = \left\{ u \in \Delta(d) : a \in \operatorname*{argmin}_{a^\star\in[K]} \langle \ell_{a^\star}, u\rangle \right\}$$

All the cells $C_1, \ldots, C_K$ constitute a cell decomposition of $\Delta(d)$. An action a is called

- *dominated*, if $C_a = \emptyset$ (so a is never optimal);
- *degenerate*, if there exists a different action b such that $\emptyset \neq C_a \subsetneq C_b$ (so a is never uniquely optimal);
- *Pareto-optimal*, if it is neither dominated nor degenerate;
- *duplicate*, if there exists a different action a' such that $\ell_a = \ell_b$ (or equivalently $C_a = C_b$).

Note that none of the dominated, degenerate, or duplicate actions can be simply ignored, because playing them might provide useful feedback. However, for simplicity we assume that there is no degenerate or duplicate actions for this lecture.

Note that C_a is a $(d-1)$-dimensional polytope if a is Pareto-optimal. Two Pareto-optimal actions are *neighbors* if $C_a \cap C_b$ is of $(d-2)$ dimension. We use N_a to denote the set containing a and all its neighbors. The neighborhood action set of two neighboring actions a and b is generally defined as $N_{ab} = \{k \in [K] : C_a \cap C_b \subseteq C_k\}$. However, under our simplifying assumptions, N_{ab} is simply $\{a, b\}$.

The next concepts are related to the feedback matrix Φ (note that all concepts above are only related to the loss matrix ℓ). As we have seen already, there are many equivalent ways to represent the observation set Σ. So to standardize notation, we let the *signal* matrix $S_a \in \{0,1\}^{|\Sigma|\times d}$ associated with an action a be such that $S_a(i, z) = 1$ if and only if $\Phi(a, z)$ is the i-th observation in Σ. Clearly, each column of S_a has exactly one 1. For any subset of actions $N \subseteq [K]$, let $S_{\mathcal{D}} \in \{0,1\}^{(|N||\Sigma|)\times d}$ be the matrix by stacking signal matrices S_a for all $a \in N$.

Now we are ready to define the key concept of observability. A pair of neighboring actions a and b are *globally observable* if $\ell_a - \ell_b \in \text{rowspace}(S_{[K]})$, which is equivalent to either of the following two statements:

- there exists $v_{ab} \in \mathbb{R}^{K|\Sigma|}$ such that $\ell_a - \ell_b = v_{ab}^\top S_{[K]}$;
- there exists a function $v_{ab} : [K] \times \Sigma \to \mathbb{R}$ such that

$$\ell(a, z) - \ell(b, z) = \sum_{k\in[K]} v_{ab}(k, \Phi(k, z)), \quad \forall z \in [d]. \tag{4}$$

Each of these equivalent statements is useful in different ways, but Equation (4) is probably the most intuitive one. Roughly speaking it says that we can always estimate the loss difference between action a and b, no matter what the outcome is. Indeed, suppose at time t the learner selects an action a_t according to a distribution $p_t \in \Delta(K)$, then clearly $\frac{v_{ab}(a_t,\Phi(a_t,z_t))}{p_t(a_t)}$ is an unbiased estimator of $\ell(a, z_t) - \ell(b, z_t)$.

Similarly, a pair of neighboring actions a and b are *locally observable* if $\ell_a - \ell_b \in \text{rowspace}(S_{N_{ab}})$ (which is a stronger condition compared to global observability). In other words, Equation (4) holds with $v_{ab}(k, \cdot) = 0$ for all $k \notin N_{ab}$, and thus we can estimate the loss difference between a and b by *only playing a and b*.

Finally, a partial monitoring problem is called globally (or locally) observable if every pair of neighboring actions is globally (or locally) observable. Note that if a problem is globally observable, then we also have $\ell_a - \ell_b \in \text{rowspace}(S_{[K]})$ for *any* pair of Pareto-optimal actions a and b (not just neighboring pair, since we can find a sequence of neighboring pairs connecting a and b). We are now ready to state the classification theorem.

Theorem 1. *The minimax regret of a partial monitoring problem G is*

$$\inf_{\textit{learner}} \max_{z_{1:n}} \mathbb{E}\left[\text{Reg}_n\right] = \begin{cases} 0, & \textit{if G has only one Pareto-optimal action;} \\ \Theta(\sqrt{n}), & \textit{else if G is locally observable;} \\ \Theta(n^{\frac{2}{3}}), & \textit{else if G is globally observable;} \\ \Theta(n), & \textit{else.} \end{cases}$$

With these concepts and the classification theorem in mind, below we go over each example mentioned in the last section again.

Full-information problems. Since for a full-information problem we can set $\Phi(a, z) = z$ for all $a \in [K]$ and $z \in [d]$, we have that the signal matrix S_a is exactly the identity matrix for any a, and thus any vector in d dimension is in the full-rank row space of S_a. Therefore, regardless the loss matrix, any full-information problem is locally-observable and $\mathcal{O}(\sqrt{n})$ regret is achievable (as we already know since Hedge can be applied).

Bandit problems. Bandit problems are also always locally-observable regardless the loss matrix. To see this, it is more convenient to use Equation (4). Indeed, we can set $v_{ab}(a, \Phi(a, z)) = \Phi(a, z) = \ell(a, z)$, $v_{ab}(b, \Phi(b, z)) = -\Phi(b, z) = -\ell(b, z)$, and $v_{ab}(k, \cdot) = 0$ for all other actions k, so that Equation (4) holds clearly.

Apple tasting. For apple tasting, we have $C_1 = \{u \in \Delta(2) : u_1 \geq u_2\}$ and $C_2 = \{u \in \Delta(2) : u_1 \leq u_2\}$, and both actions are Pareto-optimal and neighbors. Since action 2 reveals all information, S_2 is full-rank and thus $\ell_1 - \ell_2 \in \text{rowspace}(S_2)$. Therefore, apple tasting is locally observable and the minimax regret is $\Theta(\sqrt{n})$.

Label efficient learning. For the simple label efficient learning instance defined in Equation (1), the observability turns out to depend on the value of the query cost c. If $c < 1/2$, then $C_1 = \{u \in \Delta(2) : u_2 \leq c\}$, $C_2 = \{u \in \Delta(2) : u_1 \leq c\}$, and $C_3 = \{u \in \Delta(2) : c \leq \min\{u_1, u_2\}\}$ (try to draw an illustrative figure to help understand this). All actions are Pareto-optimal, and there are two neighboring pairs: actions 1 and 3, and actions 2 and 3. Similarly to apple tasting, since action 3 reveals all information, S_3 is full-rank and thus the two neighboring pairs are both locally observable. The minimax regret is therefore $\Theta(\sqrt{n})$ again.

If c is exactly $1/2$, then C_3 has exactly one point (with $u_1 = u_2 = 1/2$) and is contained in C_1 and C_2. So action 3 is degenerate and this is beyond the scope of this lecture. Finally, if $c > 1/2$, then $C_3 = \emptyset$ and action 3 is dominated. Actions 1 and 2 are still globally observable (because S_3 is full-rank), but they are not locally observable since $\text{rowspace}(S_1)$ and $\text{rowspace}(S_2)$ are both spaces of points with equal coordinates, while $\ell_1 - \ell_2 = (-1, 1)$. This also matches the intuition: playing only actions 1 and 2 is clearly not enough to estimate their loss difference since no information is observed. We conclude that in this case the minimax regret is $\Theta(n^{\frac{2}{3}})$.

Dynamic pricing. For dynamic pricing, it can be verified that regardless of the value of c, every pair of actions is neighbor and is globally observable. However, only the pairs $(1, 2), (2, 3), (3, 4), \ldots$ are locally observable. Therefore the minimax regret is $\Theta(n^{\frac{2}{3}})$. We leave the details as exercise.

Finally, we point out that the hopeless problem defined by Equation (2) is clearly not globally observable, so linear regret is unavoidable, and the trivial problem defined by Equation (3) has only one Pareto-optimal action (action 1), so the minimax regret is 0.

3 Algorithms and Upper Bounds

Note that trivially the minimax regret is in $[0, n]$, and thus to prove Theorem 1, we need to prove three upper bounds and three lower bounds. We focus on upper bounds and provide concrete algorithms for the rest of this lecture (and leave the lower bound proofs for the next lecture). The first case when there is only one Pareto-optimal action is in fact also trivial — the learner simply needs to stick with this Pareto-optimal action for every round to ensure 0 regret since the benchmark in the regret definition is exactly the total loss of this Pareto-optimal action.

3.1 Globally observable problems

Next we discuss the case when the problem is globally observable. Following the idea from last lecture for MAB, our goal is to come up with estimators for unknown information, and then plug them into a full-information algorithm, which can just be Hedge again since the action set is finite. Can we still construct estimators for the loss of each action like MAB though? The answer is no — it is not hard to construct a globally observable problem where it is impossible to estimate the loss itself. For example, if

$$\ell = \begin{pmatrix} 0 & 0.5 & 0.5 & 1 \\ 0.5 & 0 & 1 & 0.5 \end{pmatrix} \qquad \text{and} \qquad \Phi = \begin{pmatrix} 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \end{pmatrix},$$

then the learner can never figure out if the outcome is in $\{1, 2\}$ or $\{3, 4\}$ and thus it is impossible to estimate the loss accurately. However, note the following simple observation: for any reference action b we have

$$\mathrm{Reg}_n = \sum_{t=1}^{n} (\ell(a_t, z_t) - \ell(b, z_t)) - \sum_{t=1}^{n} (\ell(a^\star, z_t) - \ell(b, z_t)).$$

Therefore, it is in fact sufficient to estimate only the loss difference instead of the loss itself. Recall that Equation (4), the condition of global observability, naturally implies an important-weighted estimator for $\ell(a, z_t) - \ell(b, z_t)$ for every pair of Pareto-optimal actions a and b. This suggests that we run Hedge over the set of Pareto-optimal actions (denoted as $\mathcal{A} \subset [K]$), with the natural loss difference estimators. Since we can pick $a^\star$ to be an Pareto-optimal action, applying Hedge's analysis then gives us a regret bound against $a^\star$.

There is a small caveat though. A closer look at the importance-weighted estimators reveals that it requires picking every action with nonzero probability in order to be unbiased, and thus we cannot only play actions in $\mathcal{A}$. This matches with intuition as well. For instance, in the spam detection example with query cost $c > 1/2$, action 3 is not Pareto-optimal, but not playing this action at all is clearly a bad idea since the learner never receives feedback in this case. Of course, we also do not want to play non Pareto-optimal actions too often.

Combining these observations, it suggests that we always pick every action with at least some small probability $\gamma > 0$. One way to ensure this is to mix some uniform exploration with the Hedge distribution. The complete algorithm is shown below.

Algorithm 1: An algorithm for globally observable problems

```
Let $b \in \mathcal{A}$ be any Pareto-optimal action, $\gamma \le 1/K$ and $\eta$ be some parameters,
  $V = \max_{a \in \mathcal{A}} \|v_{ab}\|_\infty$, and $p_1 \in \Delta(K)$ be the uniform distribution.
For $t = 1, \ldots, n$:
    1. Sample $a_t \sim p_t$ and receive feedback $\Phi(a_t, z_t)$;
    2. Construct loss difference estimator $\widehat{\ell}_t(a) = \frac{v_{ab}(a_t, \Phi(a_t, z_t)) + V}{p_t(a_t)}$ for each $a \in \mathcal{A}$;
    3. Update for each $a \in [K]$: $p_t(a) = (1 - \gamma K) p'_t(a) + \gamma$ where
           $p'_t(a) \propto \mathbf{1}\{a \in \mathcal{A}\} \exp\left(-\eta \sum_{\tau=1}^{t} \widehat{\ell}_\tau(a)\right)$.
```

Notice that in this algorithm each action is indeed selected with probability at least γ since $p_t(a) \ge \gamma$. Also note that we include an extra term V in the estimator just to make sure $\widehat{\ell}_t(a)$ is always non-negative. Clearly we have for each $a \in \mathcal{A}$:

$$\mathbb{E}_t\left[\widehat{\ell}_t(a)\right] = \sum_{k=1}^{K} p_t(k)\frac{v_{ab}(k,\Phi(k,z_t))+V}{p_t(k)} = \sum_{k=1}^{K} v_{ab}(k,\Phi(k,z_t))+KV = \ell(a,z_t)-\ell(b,z_t)+KV$$

and thus

$$\mathbb{E}_t\left[\widehat{\ell}_t(a)-\widehat{\ell}_t(a^\star)\right] = \ell(a,z_t)-\ell(a^\star,z_t)$$

(recall $\mathbb{E}_t$ denotes the conditional expectation given the history up to the beginning of round t). Using Hedge's analysis it is now not hard to prove the following

Theorem 2. *Algorithm 1 ensures*

$$\mathbb{E}\left[\mathrm{Reg}_n\right] \le \frac{\ln K}{\eta} + \frac{4\eta n K V^2}{\gamma} + \gamma n K.$$

Picking the optimal γ and η gives $\mathbb{E}\left[\mathrm{Reg}_n\right] = \mathcal{O}\left((nKV)^{\frac{2}{3}}(\ln K)^{\frac{1}{3}}\right)$.

Proof. Using Lemma 1 of Lecture 6, we have

$$\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p'_t(a)\widehat{\ell}_t(a) - \sum_{t=1}^{n}\widehat{\ell}_t(a^\star) \le \frac{\ln K}{\eta} + \eta\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p'_t(a)\widehat{\ell}_t^2(a).$$

We bound the second moment as

$$\mathbb{E}_t\left[\widehat{\ell}_t^2(a)\right] = \sum_{k=1}^{K} p_t(k)\frac{(v_{ab}(k,\Phi(k,z_t))+V)^2}{p_t^2(k)} \le \sum_{k=1}^{K}\frac{4V^2}{p_t(k)} \le \frac{4KV^2}{\gamma}.$$

Taking expectation we thus have

$$\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} p'_t(a)\ell(a,z_t) - \sum_{t=1}^{n}\ell(a^\star,z_t)\right] \le \frac{\ln K}{\eta} + \frac{4\eta n K V^2}{\gamma}.$$

Note that the actual expected loss of the learner is $\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in[K]} p_t(a)\ell(a,z_t)\right]$. We thus bound the difference as

$$\mathbb{E}\left[\sum_{t=1}^{n}\sum_{a\in[K]}(p_t(a)-p'_t(a))\ell(a,z_t)\right] \le \gamma\sum_{t=1}^{n}\sum_{a\in[K]}\ell(a,z_t) \le \gamma n K.$$

Combining the last two displayed proves the theorem. □

We point out that importantly, the variance cancellation effect in the Exp3 analysis does not happen here because the second moment $\mathbb{E}_t\left[\widehat{\ell}_t^2(a)\right]$ is in terms of $\sum_k \frac{1}{p_t(k)}$ instead of just $\frac{1}{p_t(a)}$, which is the key reason why we end up getting $\mathcal{O}(n^{\frac{2}{3}})$ regret but not $\mathcal{O}(\sqrt{n})$.

3.2 Locally observable problems

To improve the regret to $\mathcal{O}(\sqrt{n})$ for locally observable problems, we need to make use of the critical condition that we can estimate $\ell_a - \ell_b$ by just playing the neighboring pair a and b. First we note that because of this fact there is no need to play dominated actions any more and we can constrain the learner to pick $a_t \in \mathcal{A}$. However, we still need to deal with a key issue that we do not know what the neighboring actions of $a^\star$ are. This is addressed by the following key lemma that allows us to replace $a^\star$ by a neighboring action of a_t in the benchmark.

Lemma 1. *There exists a constant $B > 0$ (that only depends on the loss matrix ℓ) such that for any pair of Pareto-optimal actions a and b and any frequency distribution $u \in \Delta(d)$, one can find $b' \in N_a$ such that*

$$\langle \ell_a - \ell_b, u\rangle \le B\langle \ell_a - \ell_{b'}, u\rangle.$$

For instance, in the spam detection example with query cost $c < 1/2$, one can verify that $B = 2$ (left as an exercise). We omit the proof of this lemma, but remark that this is true for any partial monitoring problems (not just locally observable ones).

The power of this lemma is that we can bound the regret as

$$\begin{aligned}
\mathrm{Reg}_n &= \sum_{t=1}^n \ell(a_t, z_t) - \ell(a^\star, z_t) = \sum_{a \in \mathcal{A}} \sum_{t : a_t = a} \ell(a, z_t) - \ell(a^\star, z_t) \\
&= \sum_{a \in \mathcal{A}} m_a \left\langle \ell_a - \ell_{a^\star}, u_a \right\rangle \qquad (m_a = |\{t : a_t = a\}| \text{ and } u_a = \tfrac{1}{m_a} \textstyle\sum_{t:a_t=a} e_{z_t}) \\
&\le B \sum_{a \in \mathcal{A}} m_a \max_{b \in N_a} \left\langle \ell_a - \ell_b, u_a \right\rangle \qquad (\text{Lemma 1}) \\
&= B \max_{\pi \in \Pi} \sum_{a \in \mathcal{A}} m_a \left\langle \ell_a - \ell_{\pi(a)}, u_a \right\rangle \qquad (\Pi = \{\pi : \mathcal{A} \to \mathcal{A} \mid \pi(a) \in N_a\}) \\
&= B \max_{\pi \in \Pi} \sum_{t=1}^n \ell(a_t, z_t) - \ell(\pi(a_t), z_t),
\end{aligned}$$

where we have essentially replaced $a^\star$ by $\pi(a_t)$, some neighboring action of a_t. This new regret measure is in fact known as the *swap regret* if π can be any mapping from $\mathcal{A}$ to itself (which makes the problem even harder). However, the special constraint $\pi(a) \in N_a$ plays a key role here as we will see. By a concentration argument (details omitted), with high probability the last upper bound on the regret is bounded by

$$B \max_{\pi \in \Pi} \sum_{t=1}^n \sum_{k \in \mathcal{A}} p_t(k) \left(\ell(k, z_t) - \ell(\pi(k), z_t)\right) + \mathcal{O}(\sqrt{n}), \tag{5}$$

where as usual $p_t(k)$ is the probability of picking k at time t. Next, we apply a common technique to ensure low swap regret (which, frankly, is not very intuitive). Suppose $p_t(k)$ is such that $p_t(k) = \sum_{a \in \mathcal{A}} Q_{ta}(k) p_t(a)$ for some distributions $Q_{ta} \in \Delta(\mathcal{A})$, or in other words, p_t is the stationary distribution of a transition matrix Q_t defined by stacking $Q_{ta}^\top, \forall a \in \mathcal{A}$ so that $p_t^\top = p_t^\top Q_t$. Then we can further rewrite the first term of Equation (5) as (dropping the constant B for conciseness)

$$\begin{aligned}
&\max_{\pi \in \Pi} \sum_{t=1}^n \left(\sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{A}} Q_{ta}(k) p_t(a) \ell(k, z_t) - \sum_{k \in \mathcal{A}} p_t(k) \ell(\pi(k), z_t) \right) \\
&= \max_{\pi \in \Pi} \sum_{t=1}^n \left(\sum_{k \in \mathcal{A}} \sum_{a \in \mathcal{A}} Q_{ta}(k) p_t(a) \ell(k, z_t) - \sum_{a \in \mathcal{A}} p_t(a) \ell(\pi(a), z_t) \right) \\
&= \max_{\pi \in \Pi} \sum_{a \in \mathcal{A}} \sum_{t=1}^n \left(\left(\sum_{k \in \mathcal{A}} Q_{ta}(k) p_t(a) \ell(k, z_t) \right) - p_t(a) \ell(\pi(a), z_t) \right) \\
&= \sum_{a \in \mathcal{A}} \max_{b \in N_a} \sum_{t=1}^n \left(\left(\sum_{k \in \mathcal{A}} Q_{ta}(k) p_t(a) \ell(k, z_t) \right) - p_t(a) \ell(b, z_t) \right).
\end{aligned}$$

Now if we fix a, the term $\max_{b \in N_a} \sum_{t=1}^n \left(\left(\sum_{k \in \mathcal{A}} Q_{ta}(k) p_t(a) \ell(k, z_t) \right) - p_t(a) \ell(b, z_t) \right)$ is exactly the regret of a sub-problem where the loss matrix becomes $p_t(a)\ell$ at time t, $Q_{ta}(k)$ is the probability of selecting k in this sub-problem, and we only care about regret against some action in N_a.

This suggests the following algorithm: for each action $a \in \mathcal{A}$, run a Hedge algorithm locally among all the actions in N_a to produce Q_{ta} such that

$$Q_{ta}(k) \propto \mathbf{1}\{k \in N_a\} \exp\left(-\eta \sum_{\tau=1}^{t-1} p_\tau(a) \widehat{\ell}_{\tau a}(k) \right), \quad \forall k \in \mathcal{A}$$

where

$$\widehat{\ell}_{ta}(k) = \frac{v_{ka}(a_t, \Phi(a_t, z_t)) + V}{p_t(a_t)} \mathbf{1}\{a_t \in N_{ka}\}$$

is the loss difference estimator with mean

$$\mathbb{E}_t\left[\widehat{\ell}_{ta}(k)\right] = v_{ka}(k, \Phi(k, z_t)) + v_{ka}(a, \Phi(a, z_t)) + 2V = \ell(k, z_t) - \ell(a, z_t) + 2V$$

Here, the last step is by the local observability of neighboring actions k and a. So ignoring the constant shift $(-\ell(a, z_t) + 2V)$, the loss we feed to Hedge is indeed $p_t(a)\ell(\cdot, z_t)$ in expectation as desired. The learner finally aggregates these distributions Q_{ta}'s as a transition matrix Q_t, finds its stationary distribution p_t, and samples $a_t \sim p_t$.

It remains to argue that each Hedge instance ensures the local regret

$$\begin{aligned}&\max_{b \in N_a} \sum_{t=1}^{n} \left(\left(\sum_{k \in \mathcal{A}} Q_{ta}(k) p_t(a) \ell(k, z_t) \right) - p_t(a)\ell(b, z_t) \right) \\ &= \max_{b \in N_a} \sum_{t=1}^{n} \left(\left(\sum_{k \in \mathcal{A}} Q_{ta}(k) p_t(a) \left(\ell(k, z_t) - \ell(a, z_t)\right) \right) - p_t(a)\left(\ell(b, z_t) - \ell(a, z_t)\right) \right).\end{aligned}$$

to be of order $\mathcal{O}(\sqrt{n})$. There is some technicality here since the loss $p_t(a)\ell(\cdot, z_t)$ is now adaptive (as opposed to being oblivious), which requires modifying the algorithm slightly. We skip the details, but point out that in the end it still boils down to bounding the term from the Hedge's lemma:

$$\frac{\ln K}{\eta} + \eta \sum_{t=1}^{n} \sum_{k \in N_a} Q_{ta}(k) p_t^2(a) \widehat{\ell}_{ta}^2(k),$$

and we note that a similar variance cancellation as in Exp3 happens again since

$$\mathbb{E}_t\left[\widehat{\ell}_{ta}^2(k)\right] \le \frac{4V^2}{p_t(a)} + \frac{4V^2}{p_t(k)},$$

and thus by the definition of $p_t(k)$:

$$\begin{aligned}\mathbb{E}_t\left[\sum_{k \in N_a} Q_{ta}(k) p_t^2(a) \widehat{\ell}_{ta}^2(k)\right] &\le 4V^2 \left(p_t(a) \sum_{k \in N_a} Q_{ta}(k) + p_t(a) \sum_{k \in N_a} \frac{Q_{ta}(k) p_t(a)}{\sum_{a' \in \mathcal{A}} Q_{ta'}(k) p_t(a')} \right) \\ &\le 4(K+1)V^2 p_t(a).\end{aligned}$$

Summing up the local regret of all Hedge instances, we arrive at

$$\text{Reg}_n \approx B\left(\frac{K \ln K}{\eta} + 4\eta(K+1)V^2 \sum_{t=1}^{n} \sum_{a \in \mathcal{A}} p_t(a) \right) = B\left(\frac{K \ln K}{\eta} + 4\eta n(K+1)V^2 \right).$$

With the optimal learning rate η, this finally leads to $\mathcal{O}(BKV\sqrt{n \ln K})$ regret for the original problem.