

Topic 2: Online allocation

2.1 Basic scenario and performance measures

2.1.1 Online allocation scenario

Online allocation is an abstract online learning scenario proposed by Freund and Schapire (1997). Define the following sets of N -vectors:

$$\mathbb{R}_+^N := \{x = (x_1, \dots, x_N) \in \mathbb{R}^N \mid x_i \geq 0 \text{ for all } i \in [N]\},$$
$$\Delta^{N-1} := \left\{ x = (x_1, \dots, x_N) \in \mathbb{R}_+^N \mid \sum_{i=1}^N x_i = 1 \right\}.$$

The set $\mathbb{R}_+^N$ is the non-negative orthant of $\mathbb{R}^N$, and the set Δ^{N-1} is the probability simplex in $\mathbb{R}^N$. Also define the standard inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$ by

$$\langle x, y \rangle := \sum_{i=1}^N x_i y_i \quad \text{for } x = (x_1, \dots, x_N) \in \mathbb{R}^N \text{ and } y = (y_1, \dots, y_N) \in \mathbb{R}^N.$$

Similar to the expert advice scenario, the online allocation scenario involves a sequence of rounds in which the learner makes choices and observes/incurs consequences. In round t :

- The learner chooses an allocation vector $p_t \in \Delta^{N-1}$.
- A loss vector $\ell_t \in \mathbb{R}_+^N$ is revealed to the learner.
- The learner incurs loss $\langle p_t, \ell_t \rangle$. (Freund and Schapire call this the *mixture loss*.)

Online allocation is an interesting learning scenario in its own right, but there are at least three additional motivations. First, it provides a connection to the expert advice scenario and shows how to seemingly break one of the lower bounds. Second, algorithms for online allocation can be generalized to work in many other learning scenarios (such as multi-armed bandits and online convex optimization). Third, algorithms for online allocation can be used to give simple proofs of important theorems that are used in learning theory.

2.1.2 Performance measures

The cumulative (a.k.a. total, net) loss of the learner after T rounds is

$$L_T := \sum_{t=1}^T \langle p_t, \ell_t \rangle.$$

The number T is called the *horizon*. The cumulative loss of the allocation vector $q \in \Delta^{N-1}$ is

$$L_{T,q} := \sum_{t=1}^T \langle q, \ell_t \rangle.$$

If q is an elementary basis vector¹, then we write $L_{T,i}$ to mean L_{T,e_i} . The regret of the learner to q after T rounds is

$$R_{T,q} := L_T - L_{T,q}$$

(with $R_{T,i} = R_{T,e_i}$), and

$$R_T := \max_{q \in \Delta^{N-1}} R_{T,q} = L_T - \min_{q \in \Delta^{N-1}} L_{T,q}$$

is the regret of the learner (to the best allocation) after T rounds.

Regret is named as such because it measures the extra loss incurred by the learner compared to having used the “best-in-hindsight” allocation in all T rounds. By linearity, we have $\sum_{t=1}^T \langle q, \ell_t \rangle = \langle q, \sum_{t=1}^T \ell_t \rangle$, so the “best-in-hindsight” cumulative loss achievable by a single allocation vector is determined by minimizing a linear function (given by $\sum_{t=1}^T \ell_t$) over the probability simplex. This minimum value will always be achieved at a vertex of the simplex, i.e., one of the N elementary basis vectors. So R_T can also be written as

$$R_T = \max_{i \in [N]} L_T - L_{T,i} = L_T - \min_{i \in [N]} L_{T,i}.$$

2.2 Connection to the expert advice scenario

2.2.1 Randomized Weighted Majority

Consider the following randomized variant of Weighted Majority, unsurprisingly named Randomized Weighted Majority (RWM), also from Littlestone and Warmuth (1994).

Algorithm 2.1: Randomized Weighted Majority for expert advice scenario with N experts.
Hyperparameter: $\beta \in (0, 1)$.

- Initially, set $w_{1,i} := 1$ for all $i \in [N]$.
- In round t :
 - Let $p_t := (p_{t,1}, \dots, p_{t,N}) \in \Delta^{N-1}$ be defined by

$$p_{t,i} := \frac{w_{t,i}}{Z_t} \quad \text{for all } i \in [N],$$

where $Z_t := \sum_{i=1}^N w_{t,i}$ is the total weight (as in the proof of Theorem 1.7).

- Randomly draw $i_t \in [N]$ so that, for each $i \in [N]$, $i_t = i$ with probability $p_{t,i}$.
- Observe experts’ predictions $b_t := (b_{t,1}, \dots, b_{t,N}) \in \{0, 1\}^N$.
- Predict $a_t := b_{t,i_t} \in \{0, 1\}$.
- Observe outcome $y_t \in \{0, 1\}$.
- Update experts’ weights: for each $i \in [N]$,

$$w_{t+1,i} := \begin{cases} w_{t,i} & \text{if } y_t = b_{t,i}, \\ \beta w_{t,i} & \text{if } y_t \neq b_{t,i}. \end{cases}$$

¹The N elementary (or coordinate) basis vectors (in $\mathbb{R}^N$) are $e_1, \dots, e_N$, where e_i has a 1 in the i -th entry and 0 elsewhere.

RWM explicitly uses randomness to make its predictions via the random indices i_t . Specifically, in each round t , the learner's prediction a_t is determined as a function of both i_t and b_t .

Note that use of randomness alone does not allow RWM to subvert the lower bound from Theorem 1.5 (for the expert advice scenario with N experts and the K -mistake assumption). The proof given for Theorem 1.5 was stated for deterministic learning algorithms. However, the argument still holds by allowing the adversary to adaptively pick y_t as a function of a_t in each round t : simply set $y_t := 1 - a_t$ on-the-fly. So the learner makes a mistake in each of the first $2K + 1$ rounds, but one of the “constant” experts makes at most K mistakes.

2.2.2 Non-adaptive adversaries and the connection to online allocation

To subvert the lower bound—and to connect RWM to online allocation—it will suffice to assume that the adversary is non-adaptive: in each round t , the adversary chooses the b_t and y_t in a way that does not depend on the random choices of the learner in round t . The assumption ensures that (b_t, y_t) and i_t are conditionally independent given p_t . In this case,

$$\begin{aligned} \Pr(a_t \neq y_t \mid p_t, b_t, y_t) &= \sum_{i=1}^N \Pr(i_t = i \mid p_t, b_t, y_t) \mathbb{1}\{b_{t,i} \neq y_t\} \\ &= \sum_{i=1}^N \Pr(i_t = i \mid p_t) \mathbb{1}\{b_{t,i} \neq y_t\} \\ &= \sum_{i=1}^N p_{t,i} \mathbb{1}\{b_{t,i} \neq y_t\} \\ &= \langle p_t, \ell_t \rangle \end{aligned}$$

where

$$\ell_{t,i} := \mathbb{1}\{b_{t,i} \neq y_t\} \in \{0, 1\} \quad \text{and} \quad \ell_t := (\ell_{t,1}, \dots, \ell_{t,N}) \in \{0, 1\}^N.$$

RWM chooses p_t based on $\ell_1, \dots, \ell_{t-1}$ from previous rounds (and the initial weights $w_{1,1}, \dots, w_{1,N}$), and hence it can be viewed as a learning algorithm for online allocation where ℓ_t is the loss vector in round t . The total loss of RWM after T rounds is

$$\sum_{t=1}^T \langle p_t, \ell_t \rangle = \sum_{t=1}^T \Pr(a_t \neq y_t \mid p_t, b_t, y_t).$$

The expected value of the above quantity is the expected number of mistakes committed by RWM after T rounds.

2.2.3 Foreshadowing: breaking through the pre-factor of two

For the expert advice scenario with N experts, the K -mistake assumption, and the non-adaptive adversary assumption, the expected number of mistakes of RWM can be made to be at most

$$(1 + \epsilon)K + \frac{2 \ln(N)}{\epsilon}$$

for any $\epsilon > 0$ (with a suitable choice of hyperparameter β depending on ϵ , K , and N). The key point is that the pre-factor on K can be made arbitrarily close to 1. See Corollary 2.2.

2.3 Algorithm for online allocation

2.3.1 Hedge algorithm

We now describe an algorithm called Hedge due to Freund and Schapire (1997) for the online allocation scenario. It is a generalization of RWM (when viewed as an algorithm for a special case of online allocation).

Algorithm 2.2: Hedge for online allocation in N -space.

Hyperparameter: $\eta > 0$.

- Initially, set $w_{1,i} := 1$ for all $i \in [N]$.
- In round t :
 - Choose allocation $p_t := (p_{t,1}, \dots, p_{t,N}) \in \Delta^{N-1}$, defined by

$$p_{t,i} := \frac{w_{t,i}}{Z_t} \quad \text{for all } i \in [N],$$

where $Z_t := \sum_{i=1}^N w_{t,i}$ is the total weight.

- Observe loss vector $\ell_t := (\ell_{t,1}, \dots, \ell_{t,N}) \in \mathbb{R}_+^N$.
- Update weights: for each $i \in [N]$,

$$w_{t+1,i} := w_{t,i} \exp(-\eta \ell_{t,i}).$$

We have changed the hyperparameter from the fraction $\beta \in (0, 1)$ to the positive number $\eta > 0$. But really nothing has changed, since in the special case where $\ell_{t,i} \in \{0, 1\}$, the weight updates for Hedge and RWM are exactly the same when we take $\beta = \exp(-\eta)$. It turns out to be convenient to (hyper-)parameterize the weight updates with η to make certain standard Taylor approximations more recognizable in the analysis.

Note that Hedge is a completely deterministic. Therefore, if the loss vectors in online allocation are chosen by an adversary, the entire sequence of loss vectors may as well be fixed in advance. It is only in Hedge/RWM's interpretation as a randomized learning algorithm for the expert advice scenario where the distinction between adaptive and non-adaptive adversaries makes a difference.

Theorem 2.1 (Freund and Schapire, 1997). *In the online allocation scenario in N -space, for any T and any sequence of loss vectors $\ell_1, \dots, \ell_T \in \mathbb{R}_+^N$, Hedge with hyperparameter $\eta > 0$ guarantees that*

$$R_T \leq \frac{\ln(N)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \langle p_t, \ell_t^2 \rangle$$

where $\ell_t^2 := (\ell_{t,1}^2, \dots, \ell_{t,N}^2) \in \mathbb{R}_+^N$ for all t .

Corollary 2.1. *In the setup of Theorem 2.1, if $\ell_t \in [0, 1]^N$ for all t , then*

$$R_T \leq \frac{\ln(N)}{\eta} + \frac{\eta T}{2} \quad \text{and} \quad (1 - \eta/2)L_T \leq L^* + \frac{\ln(N)}{\eta}$$

where

$$L^* := \min_{i \in [N]} L_{T,i}.$$

In particular, there is a choice of $\eta > 0$ (depending on N and T) such that

$$R_T \leq \sqrt{2T \ln(N)},$$

and there is a choice of $\eta \in (0, 1)$ (depending on N , T , and L^* , assuming $L^* > 0$) such that

$$R_T \leq \sqrt{2L^* \ln(N) + \ln^2(N)} + \ln(N).$$

Exercise 2.1. Determine the values of η that give the final regret bounds in Corollary 2.1.

We say a learner has *no-regret* if its time-averaged regret to the best allocation after T rounds is non-positive in the limit as $T \rightarrow \infty$:

$$\limsup_{T \rightarrow \infty} \frac{R_T}{T} \leq 0.$$

In other words, the regret is sublinear in T . This is indeed the case for Hedge with an appropriate choice of hyperparameter, as seen in Corollary 2.1. No-regret is an important property of a learning algorithm for online allocation and is taken as the standard for what it means to successfully learn.

The second guarantee in Corollary 2.1 is called an *L^* -bound* or *first-order bound* in that it replaces the dependence on the horizon T with a first-order quantity, namely L^* . (A second-order quantity would be something involving the variability of per-round losses.) In terms of the time-averaged regret, the first-order bound interpolates between two different “rates”:

$$\begin{aligned} L^* = T &\implies \frac{R_T}{T} \leq \sqrt{\frac{2 \ln(N)}{T}}; \\ L^* = 0 &\implies \frac{R_T}{T} \leq \frac{2 \ln(N)}{T}. \end{aligned}$$

The $T^{-1/2}$ rate is referred to as a *slow rate*, while the T^{-1} rate is referred to as a *fast rate*.

Corollary 2.2. *In the expert advice scenario with N experts, the K -mistake assumption, and the non-adaptive adversary assumption, Randomized Weighted Majority with hyperparameter $\beta = \exp(-\eta)$ for $\eta \in (0, 2)$ makes at most*

$$\frac{1}{1 - \eta/2} \left(K + \frac{\ln(N)}{\eta} \right)$$

mistakes in expectation.

The statement of Theorem 2.1 is slightly different from how it is stated by Freund and Schapire (1997), but not in any essential way for our purposes. The proof below is based on that from Freund and Schapire (1999), which uses concepts from information theory, discussed next.

2.3.2 Entropy and relative entropy

The (*Shannon*) *entropy* $H(q)$ of a (discrete) probability distribution $q = (q_1, \dots, q_N) \in \Delta^{N-1}$ over $[N]$ is defined by

$$H(q) := \sum_{i=1}^N q_i \ln\left(\frac{1}{q_i}\right).$$

(Here, we regard $0 \ln(0)$ as 0 rather than $-\infty$.) Entropy is a central concept in information theory and many other fields, with several interpretations. One way to think about entropy is as a measure of average “surprise” one experiences when observing the value of a random variable. The idea is that you experience more surprise when you see rare events than when you see common events. Letting $\ln(1/q_i)$ be the quantitative measure of surprise for the outcome i , we then see $H(q)$ to be exactly the expected surprise. (There are many other interpretations of entropy, such as the average number of bits needed to communicate a random message $X \sim q$, but we will not discuss them here.) The maximum value of $H(q)$ is $\ln(N)$, achieved by the uniform distribution $(1/N, \dots, 1/N)$.

Now consider two probability distributions, $p, q \in \Delta^{N-1}$. The *cross entropy* of q from p is

$$CE(q, p) := \sum_{i=1}^N q_i \ln\left(\frac{1}{p_i}\right).$$

Cross entropy measures the average surprise when outcomes are distributed according to q , but the measure of surprise is (perhaps mistakenly) based on p . The *relative entropy*² of q from p is

$$RE(q, p) := \sum_{i=1}^N q_i \ln\left(\frac{q_i}{p_i}\right) = CE(q, p) - H(q)$$

is the excess average surprise in this scenario. Note that $RE(q, p)$ is well-defined only if q is dominated by p (written $q \ll p$, which means that for any i , we have $p_i = 0$ only if $q_i = 0$ as well). If $q \not\ll p$, then we define $RE(q, p) := +\infty$. Also note that $RE(q, p)$ is potentially (very) different from $RE(p, q)$.

Proposition 2.1 (Gibbs’ inequality). $RE(q, p) \geq 0$ with equality if and only if $q = p$.

Proof. A Taylor expansion of $\ln$ gives

$$\ln(1+x) \leq x \quad \text{for all } x > -1. \quad (2.1)$$

Without loss of generality, assume $q \ll p$. Then

$$\begin{aligned} RE(q, p) &= - \sum_{i=1}^N q_i \ln\left(1 + \frac{p_i}{q_i} - 1\right) \\ &\geq - \sum_{i=1}^N q_i \left(\frac{p_i}{q_i} - 1\right) && \text{(by (2.1))} \\ &= - \sum_{i=1}^N p_i + q_i = 0. \end{aligned}$$

Equality holds iff $p_i/q_i - 1 = 0$ for all i , which is equivalent to $p_i = q_i$ for all i . □

²Relative entropy is also known as *Kullback-Leibler divergence* and *I-divergence*.

2.3.3 Proof of Theorem 2.1

To prove Theorem 2.1, it suffices to show that, for any $q \in \Delta^{N-1}$, we have

$$\sum_{t=1}^T \langle p_t - q, \ell_t \rangle \leq \frac{\ln(N)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \langle p_t, \ell_t^2 \rangle.$$

The following is the key lemma in the proof.

Lemma 2.1. Fix any $p \in \Delta^{N-1}$ and any $\ell \in \mathbb{R}_+^N$, and define $p' \in \Delta^{N-1}$ by

$$p'_i := \frac{p_i \exp(-\eta \ell_i)}{\sum_{j=1}^N p_j \exp(-\eta \ell_j)} \quad \text{for all } i \in [N].$$

Then for any $q \in \Delta^{N-1}$ such that $q \ll p$,

$$\text{RE}(q, p') - \text{RE}(q, p) \leq \eta \langle q - p, \ell \rangle + \frac{\eta^2}{2} \langle p, \ell^2 \rangle.$$

where $\ell^2 := (\ell_1^2, \dots, \ell_N^2)$.

Proof. Taylor expansion of exp gives

$$\exp(x) \leq 1 + x + \frac{x^2}{2} \quad \text{for all } x \leq 0. \quad (2.2)$$

Therefore

$$\begin{aligned} \text{RE}(q, p') - \text{RE}(q, p) &= \sum_{i=1}^N q_i \ln \left(\frac{p_i}{p'_i} \right) = \sum_{i=1}^N q_i \ln \left(\frac{\sum_{j=1}^N p_j \exp(-\eta \ell_j)}{\exp(-\eta \ell_i)} \right) \\ &= \eta \langle q, \ell \rangle + \ln \left(\sum_{j=1}^N p_j \exp(-\eta \ell_j) \right) \\ &\leq \eta \langle q, \ell \rangle + \ln \left(\sum_{j=1}^N p_j \left(1 - \eta \ell_j + \frac{\eta^2}{2} \ell_j^2 \right) \right) \quad (\text{by (2.2)}) \\ &= \eta \langle q, \ell \rangle + \ln \left(1 - \eta \langle p, \ell \rangle + \frac{\eta^2}{2} \langle p, \ell^2 \rangle \right) \\ &\leq \eta \langle q, \ell \rangle - \eta \langle p, \ell \rangle + \frac{\eta^2}{2} \langle p, \ell^2 \rangle \quad (\text{by (2.1)}). \quad \square \end{aligned}$$

Using Lemma 2.1, we have for any t and any $q \in \Delta^{N-1}$,

$$\eta \langle p_t - q, \ell_t \rangle \leq \text{RE}(q, p_t) - \text{RE}(q, p_{t+1}) + \frac{\eta^2}{2} \langle p_t, \ell_t^2 \rangle.$$

Sum both sides over $t = 1, \dots, T$ to obtain

$$\begin{aligned}
\eta \sum_{t=1}^T \langle p_t - q, \ell_t \rangle &\leq \sum_{t=1}^T \left(\text{RE}(q, p_t) - \text{RE}(q, p_{t+1}) + \frac{\eta^2}{2} \langle p_t, \ell_t^2 \rangle \right) \\
&= \underbrace{\text{RE}(q, p_1)}_{=\ln(N)} - \underbrace{\text{RE}(q, p_{T+1})}_{\geq 0} + \frac{\eta^2}{2} \sum_{t=1}^T \langle p_t, \ell_t^2 \rangle \\
&\leq \ln(N) + \frac{\eta^2}{2} \sum_{t=1}^T \langle p_t, \ell_t^2 \rangle
\end{aligned}$$

where the last inequality uses Proposition 2.1 and the definition of p_1 . $\square$

Exercise 2.2. Suppose we consider a generalization of online allocation in which loss vectors may have components that are negative. Explain how the analysis of (and performance guarantees for) Hedge in Theorem 2.1 and Corollary 2.1 should change. For Corollary 2.1, you may assume that $\max_{i \in [N]} \ell_{t,i} - \min_{i \in [N]} \ell_{t,i} \leq 1$ for all t .

2.4 Application to zero-sum games

2.4.1 Zero-sum games

A *zero-sum game* is a two-player game in which the loss of one player is the gain of the other player. The loss/gain is specified by an $m \times n$ payoff matrix A ; for simplicity, assume the entries of A are from $[0, 1]$. The “row” player has m possible actions, and the “column” player has n possible actions. For example, suppose the row player is the goalkeeper, who can dive left, stay put, or dive right; and the column player is the penalty kicker, who can kick left, kick straight, or kick right. The payoff matrix (where the actions are ordered as in the previous sentence) is

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

So, for example, if the goalkeeper stays put, then the penalty kicker gains 1 by kicking left or kicking right (either will do). On the other hand, if the penalty kicker kicks right, then the goalkeeper loses 0 by diving left. Notice that given the action of one player, there is a (possibly non-unique) best action that the other player can choose; this is called the *best response* for that other player.

The players are permitted to use *mixed strategies*, which are probability distributions over their respective action sets. If the row player chooses $p \in \Delta^{m-1}$ and the column player chooses $q \in \Delta^{n-1}$, then the loss of the row player (equivalently, the gain of the column player) is $p^\top A q$. For example, the penalty kicker may choose to kick left and kick right with probability $1/2$ each ($q = (1/2, 0, 1/2)$). This is a best response mixed strategy given that the goalkeeper chooses to stay put ($p = (0, 1, 0)$).

We have been ambiguous about exactly when the choices of the players are made. If the row player chooses p first, and then the column player chooses q in response to seeing p , then the best possible loss of the row player (assuming the column player chooses a best response) is

$$\min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top A q. \quad (2.3)$$

However, if the column player chooses q first, and then the row player chooses p in response to seeing q , then the best possible gain of the column player (assuming the row player chooses a best response) is

$$\max_{q \in \Delta^{n-1}} \min_{p \in \Delta^{m-1}} p^\top Aq. \quad (2.4)$$

Von Neumann's "minimax theorem" states that (2.3) and (2.4) are equal; there is no disadvantage to "going first" in a zero-sum game. This common value is called the value of the game.

Theorem 2.2 (Minimax theorem of von Neumann, 1928). *For any $A \in [0, 1]^{m \times n}$,*

$$\max_{q \in \Delta^{n-1}} \min_{p \in \Delta^{m-1}} p^\top Aq = \min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top Aq.$$

Exercise 2.3. Find the optimal mixed strategies for both players with the payoff matrix given at the start of Section 2.4.1. What is the value of the game?

To prove Theorem 2.2, it suffices to show that the max min is at most the min max, and also that min max is at most the max min. The first part is rather intuitive and easy to prove. The other part is less obvious; Freund and Schapire (1999) showed that this part follows from the existence of no-regret algorithms for online allocation such as Hedge.

2.4.2 Proof of the easy part

Proposition 2.2. *For any $A \in [0, 1]^{m \times n}$,*

$$\max_{q \in \Delta^{n-1}} \min_{p \in \Delta^{m-1}} p^\top Aq \leq \min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top Aq.$$

Proof. Fix any $q' \in \Delta^{n-1}$. Then, for every $p \in \Delta^{m-1}$,

$$p^\top Aq' \leq \max_{q \in \Delta^{n-1}} p^\top Aq.$$

Let p' be the choice of $p \in \Delta^{m-1}$ that minimizes the right-hand side, so

$$\begin{aligned} (p')^\top Aq' &\leq \max_{q \in \Delta^{n-1}} (p')^\top Aq \\ &= \min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top Aq. \end{aligned}$$

We clearly also have

$$(p')^\top Aq' \geq \min_{p \in \Delta^{m-1}} p^\top Aq',$$

so combining the previous two displayed inequalities gives

$$\min_{p \in \Delta^{m-1}} p^\top Aq' \leq \min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top Aq$$

Since this inequality holds for every $q' \in \Delta^{n-1}$, it also holds for the choice of q' that maximizes the left-hand side:

$$\max_{q \in \Delta^{n-1}} \min_{p \in \Delta^{m-1}} p^\top Aq \leq \min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top Aq. \quad \square$$

2.4.3 Proof of the other part via no-regret online allocation

The second part is the “hard” direction, which involves the execution of Hedge in an instance of the online allocation scenario that is constructed on-the-fly.

Proposition 2.3. *For any $A \in [0, 1]^{m \times n}$,*

$$\min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top A q \leq \max_{q \in \Delta^{n-1}} \min_{p \in \Delta^{m-1}} p^\top A q.$$

Proof. Consider the execution of Hedge for selecting $p_1, p_2, \dots, p_T \in \Delta^{m-1}$ (so $N = m$) with loss vectors $\ell_1, \ell_2, \dots, \ell_T \in [0, 1]^m$ defined by $\ell_t = A q_t$ for all t , where $q_t \in \Delta^{n-1}$ is chosen so that

$$p_t^\top A q_t = \max_{q \in \Delta^{n-1}} p_t^\top A q.$$

By Theorem 2.1, using $\eta = \sqrt{2 \ln(m)/T}$, we have

$$\sum_{t=1}^T \langle p_t, \ell_t \rangle \leq \min_{p \in \Delta^{m-1}} \sum_{t=1}^T \langle p, \ell_t \rangle + \sqrt{2T \ln(m)}.$$

This implies

$$\frac{1}{T} \sum_{t=1}^T p_t^\top A q_t \leq \min_{p \in \Delta^{m-1}} \frac{1}{T} \sum_{t=1}^T p^\top A q_t + \sqrt{\frac{2 \ln(m)}{T}}. \quad (2.5)$$

Define $\bar{p} = \frac{1}{T} \sum_{t=1}^T p_t$ and $\bar{q} = \frac{1}{T} \sum_{t=1}^T q_t$. Then

$$\begin{aligned} \min_{p \in \Delta^{m-1}} \max_{q \in \Delta^{n-1}} p^\top A q &\leq \max_{q \in \Delta^{n-1}} \bar{p}^\top A q \\ &= \max_{q \in \Delta^{n-1}} \frac{1}{T} \sum_{t=1}^T p_t^\top A q \\ &\leq \frac{1}{T} \sum_{t=1}^T \max_{q \in \Delta^{n-1}} p_t^\top A q \\ &= \frac{1}{T} \sum_{t=1}^T p_t^\top A q_t \\ &\leq \min_{p \in \Delta^{m-1}} \frac{1}{T} \sum_{t=1}^T p^\top A q_t + \sqrt{\frac{2 \ln(m)}{T}} \quad (\text{by (2.5)}) \\ &= \min_{p \in \Delta^{m-1}} p^\top A \bar{q} + \sqrt{\frac{2 \ln(m)}{T}} \\ &\leq \max_{q \in \Delta^{n-1}} \min_{p \in \Delta^{m-1}} p^\top A q + \sqrt{\frac{2 \ln(m)}{T}}. \end{aligned}$$

Now let $T \rightarrow \infty$, to conclude the proof. □

The proof of Proposition 2.3 shows how to use Hedge to construct the mixed strategies $\bar{p}$ and $\bar{q}$, essentially by repeatedly simulating the two-player game, where the row player's $p_t \in \Delta^{m-1}$ is controlled by Hedge, and the column player's $q_t \in \Delta^{n-1}$ is the best response to p_t . If $\text{val}(A)$ denotes the value of the game for payoff matrix A , then $\bar{p}$ and $\bar{q}$ satisfy

$$\max_{q \in \Delta^{n-1}} \bar{p}^\top A q \leq \text{val}(A) + \sqrt{\frac{2 \ln(m)}{T}} \quad \text{and} \quad \min_{p \in \Delta^{m-1}} p^\top A \bar{q} \geq \text{val}(A) - \sqrt{\frac{2 \ln(m)}{T}}.$$

The proof (or the above construction) does not rely on Hedge specifically; any no-regret algorithm for online allocation could be used.

2.5 Application to online convex optimization

2.5.1 Convex functions

A function $f: \mathbb{R}^N \rightarrow \mathbb{R}$ is convex³ if, for every $x_o \in \mathbb{R}^N$, there exists $\lambda \in \mathbb{R}^N$ such that

$$f(x) \geq f(x_o) + \langle x - x_o, \lambda \rangle, \quad \forall x \in \mathbb{R}^N. \quad (2.6)$$

Linear and affine functions are convex functions where the inequality can be made to be an equality. That is, if $f(x) = \langle x, m \rangle + b$ for some $m \in \mathbb{R}^N$ and $b \in \mathbb{R}$, then (2.6) holds with equality with $\lambda = m$. Sums of convex functions are also convex functions. The set of vectors $\lambda \in \mathbb{R}^N$ satisfying (2.6), denoted $\partial f(x_o)$, is the set of subgradients of f at x_o . If f is also differentiable at x_o , then $\partial f(x_o)$ contains only the gradient of f evaluated at x_o .

2.5.2 Generalization of online allocation

Consider the following generalization of the online allocation scenario, called online convex optimization (over the probability simplex Δ^{N-1}). In round t :

- The learner chooses an allocation vector $p_t \in \Delta^{N-1}$.
- A convex loss function $f_t \in \mathbb{R}^N \rightarrow \mathbb{R}$ is revealed to the learner. (It is not obvious how such a function is “revealed” to the learner, but we shall make this more precise later.)
- The learner incurs loss $f_t(p_t)$.

The notions of cumulative loss and regret naturally extend to this scenario, with the loss $\langle q, \ell_t \rangle$ replaced by $f_t(q)$ for all $q \in \Delta^{N-1}$.

In particular, suppose in round t , the learner acquires a vector $\lambda_t \in \partial f_t(p_t)$ after having selected p_t . This is the sense in which the learner needs to “observe” f_t . Then for any $q \in \Delta^{N-1}$, the inequality (2.6) implies

$$f_t(q) \geq f_t(p_t) + \langle q - p_t, \lambda_t \rangle,$$

which rearranges to

$$f_t(p_t) - f_t(q) \leq \langle p_t - q, \lambda_t \rangle.$$

³There are other equivalent definitions of convex functions, but the one based on (2.6) is the most directly relevant for our purposes. For twice-differentiable functions, convexity is equivalent to positive semi-definiteness of the matrix of second derivatives at all points in $\mathbb{R}^N$.

Therefore, a learner that treats λ_t as the loss vector ℓ_t in round t in the original online allocation scenario and guarantees a regret bound will also guarantee a regret bound in online convex optimization.

Corollary 2.3. *In the online convex optimization scenario over the probability simplex Δ^{N-1} , for any T and any sequence of convex loss functions $f_1, \dots, f_T: \mathbb{R}^N \rightarrow \mathbb{R}$, Hedge with hyperparameter $\eta > 0$ —treating $\lambda_t \in \partial f_t(p_t)$ as the loss vector ℓ_t in all rounds t —guarantees that*

$$R_T \leq \frac{\ln(N)}{\eta} + \frac{\eta}{2} \sum_{t=1}^T \langle p_t, \ell_t^2 \rangle$$

where $\ell_t^2 := ((\lambda_{t,1} - m_t)^2, \dots, (\lambda_{t,N} - m_t)^2) \in \mathbb{R}_+^N$ and $m_t := \min_{i \in [N]} \lambda_{t,i}$ for all t .⁴

Exercise 2.4. Suppose in the online convex optimization scenario over the probability simplex Δ^{N-1} , for some positive scalars $\gamma, \theta > 0$, the loss function $f_t: \mathbb{R}^N \rightarrow \mathbb{R}$ in each round t has the form

$$f_t(x) = \max\{0, \gamma - y_t(\langle x, z_t \rangle - \theta)\}$$

for some $z_t \in \{0, 1\}^N$ and $y_t \in \{-1, 1\}$.

- (a) Confirm that each f_t is a convex function.
- (b) Show that Hedge can be used (with some setting of hyperparameter η) to choose $p_1, \dots, p_T \in \Delta^{N-1}$ satisfying

$$\sum_{t=1}^T f_t(p_t) \leq \min_{q \in \Delta^{N-1}} \sum_{t=1}^T f_t(q) + \sqrt{2T \ln(N)}.$$

2.6 Optimality of Hedge

2.6.1 Lower bound for any learning algorithm

While Hedge is a no-regret learning algorithm for online allocation, it is natural to ask if its regret bound can be improved. It turns out that no learning algorithm can guarantee a better regret bound (up to constants) in the worst case that holds for all T and N . This is the content of Theorem 2.3. In this sense, Hedge is an optimal learning algorithm for online allocation.

Theorem 2.3 (Freund and Schapire, 1999; Arora, Hazan, and Kale, 2012). *There exists an absolute constant $C > 0$ such that the following holds. Let A denote a learning algorithm for the online allocation scenario in N -space where loss vectors come from $[0, 1]^N$, and let $R_T(A; \ell_1, \dots, \ell_T)$ denote the regret of A after T rounds with loss vectors $\ell_1, \dots, \ell_T \in [0, 1]^N$. For all sufficiently large N and T , for all learning algorithms A , there exist $\ell_1, \dots, \ell_T \in \{0, 1\}^N$ such that*

$$R_T(A; \ell_1, \dots, \ell_T) \geq C\sqrt{T \ln(N)}.$$

⁴Exercise 2.2 explores how Hedge handles loss vectors that may have negative components.

2.6.2 Proof strategy

The key limitation of learning algorithms is what “information” they have access to when choosing the allocations in each round. This limitation is sufficient to leverage the probabilistic method to establish the lower bound in Theorem 2.3. The main idea is the following:

1. Define a probability distribution for the (random) loss vectors $\ell_1, \dots, \ell_T$, and then show that

$$\Pr\left(R_T(A; \ell_1, \dots, \ell_T) \geq C\sqrt{T \ln(N)}\right) > 0. \quad (2.7)$$

2. Therefore, there must be some element in the sample space, which ultimately determines some $\ell_1, \dots, \ell_T$ and $R_T(A; \ell_1, \dots, \ell_T)$, such that

$$R_T(A; \ell_1, \dots, \ell_T) \geq C\sqrt{T \ln(N)}.$$

So it suffices to show that (2.7) holds for some suitable probability distribution for $\ell_1, \dots, \ell_T$. One distribution that works treats $\ell_1, \dots, \ell_T$ as independent random vectors, with each ℓ_t uniformly distributed over $\{0, 1\}^N$.

2.6.3 Most of the proof of Theorem 2.3

Let $\mathbf{p}_t$ denote the allocation vector chosen by A in round t , and recall that $\mathbf{p}_t$ depends only on $\ell_1, \dots, \ell_{t-1}$. By definition of regret and the tower property of expectations,

$$\begin{aligned} \mathbb{E}[R_T(A; \ell_1, \dots, \ell_T)] &= \mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{p}_t, \ell_t \rangle - \min_{i \in [N]} \sum_{t=1}^T \ell_{t,i}\right] \\ &= \sum_{t=1}^T \mathbb{E}\left[\mathbb{E}[\langle \mathbf{p}_t, \ell_t \rangle \mid \ell_1, \dots, \ell_{t-1}]\right] - \mathbb{E}\left[\min_{i \in [N]} \sum_{t=1}^T \ell_{t,i}\right] \\ &= \frac{T}{2} - \mathbb{E}\left[\min_{i \in [N]} \sum_{t=1}^T \ell_{t,i}\right]. \end{aligned}$$

It will be notationally cleaner to introduce the random vectors $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_T$, where $\boldsymbol{\sigma}_t = (\sigma_{t,1}, \dots, \sigma_{t,N})$ is defined by $\sigma_{t,i} := 1 - 2\ell_{t,i}$ for each $i \in [N]$. So the $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_T$ are independent random vectors, and each $\boldsymbol{\sigma}_t$ is uniform distributed over $\{-1, 1\}^N$. Then

$$\min_{i \in [N]} \sum_{t=1}^T \ell_{t,i} = \min_{i \in [N]} \sum_{t=1}^T \frac{1 - \sigma_{t,i}}{2} = \min_{i \in [N]} \frac{T}{2} - \frac{1}{2} \sum_{t=1}^T \sigma_{t,i} = \frac{T}{2} - \frac{1}{2} \max_{i \in [N]} \sum_{t=1}^T \sigma_{t,i}.$$

Therefore, we have

$$\mathbb{E}[R_T(A; \ell_1, \dots, \ell_T)] = \frac{T}{2} - \mathbb{E}\left[\frac{T}{2} - \frac{1}{2} \max_{i \in [N]} \sum_{t=1}^T \sigma_{t,i}\right] = \frac{1}{2} \mathbb{E}\left[\max_{i \in [N]} \sum_{t=1}^T \sigma_{t,i}\right].$$

For each $i \in [N]$, the random variable $\mathbf{s}_{T,i} := \sum_{t=1}^T \sigma_{t,i}$ is the sum of T independent random variables, each uniform over $\{-1, 1\}$. It turns out that, for large enough T and N ,

$$\mathbb{E}\left[\max_{i \in [N]} \mathbf{s}_{T,i}\right] \geq C\sqrt{T \ln(N)}$$

for some absolute constant $C > 0$. We will not prove this, but here is the intuition for why it should be believable. By the Central Limit Theorem, as $T \rightarrow \infty$,

$$\frac{1}{\sqrt{T}}(\mathbf{s}_{T,1}, \dots, \mathbf{s}_{T,N}) \xrightarrow{\text{dist.}} (\mathbf{z}_1, \dots, \mathbf{z}_N) \sim \mathcal{N}(0, I_N).$$

So let us just heuristically pretend that

$$\mathbb{E} \left[\max_{i \in [N]} \mathbf{s}_{T,i} \right] \stackrel{!}{=} \sqrt{T} \mathbb{E} \left[\max_{i \in [N]} \mathbf{z}_i \right].$$

(The smiley is meant to caution you from taking this equality literally.) The distribution of $\max_{i \in [N]} \mathbf{z}_i$ is also well understood in the limit $N \rightarrow \infty$; after a certain centering and scaling, its distribution tends to a Gumbel distribution, which is one of the central distributions that features in the study of extreme value statistics. The results in this literature lead us to

$$\mathbb{E} \left[\max_{i \in [N]} \mathbf{z}_i \right] \stackrel{!}{=} \sqrt{2 \ln(N)} \left(1 - \frac{\ln(\ln(N)) + \ln(4\pi) - 2\gamma}{4 \ln(N)} \right),$$

where $\gamma \approx 0.577215$ is the Euler-Mascheroni constant. So, we find

$$\mathbb{E}[R_T(A; \ell_1, \dots, \ell_T)] \stackrel{!}{=} \sqrt{\frac{T \ln(N)}{2}} \left(1 - \frac{\ln(\ln(N)) + \ln(4\pi) - 2\gamma}{4 \ln(N)} \right).$$

Theorem 2.3 is proved by using the appropriate non-asymptotic versions of these heuristic arguments.