Prediction Learning and Games – Exercise Solutions

Ludovic Schwartz

Contents

Chapter 2

Prediction with Expert Advice

Exercise 2.2

Consider a weighted average forecaster based on a potential function

$$\Phi(u) = \psi\left(\sum_{i=1}^{N} \phi(u_i)\right).$$

Assume further that the quantity $C(r_t)$ appearing in the statement of Theorem 2.1 is bounded by a constant for all values of r_t and that the function $\psi(\phi(u))$ is strictly convex. Show that there exists a nonnegative sequnce $\epsilon_n \to 0$ such that the cumulative regret of the forecaster satisfies, for every n and for every outcome sequence y^n,

$$\frac{1}{n}\left(\max_{i=1,\dots N} R_{i,n}\right) \leq \epsilon_n.$$

Solution : We start by assuming that $C(r_t) \leq C$ for all values of r_t. We apply theorem 2.1 and we get that :

$$\Phi(R_n) \leq \Phi(0) + \frac{1}{2}\sum_{t=1}^{n} C(r_t) \leq \Phi(0) + \frac{Cn}{2}.$$

Then, since we know that $\psi \circ \phi$ is non decreasing and strictly convex, it must be increasing and as a result, both $\psi \circ \phi$ and ϕ must be invertible and increasing and we have :

$$\psi\left(\phi\left(\max_{i=1,\dots,N} R_{i,n}\right)\right) \leq \psi\left(\max_{i=1,\dots,N} \phi(R_{i,n})\right) \leq \psi\left(\sum_{i=1}^{N} \phi(R_{i,n})\right) = \Phi(R_n).$$

Hence :

$$\max_{i=1,\dots,N} R_{i,n} \leq \phi^{-1}(\psi^{-1}(\Phi(R_n)))$$
$$\frac{1}{n}\left(\max_{i=1,\dots,N} R_{i,n}\right) \leq \underbrace{\frac{(\psi\circ\phi)^{-1}\left(\Phi(0)+\frac{Cn}{2}\right)}{n}}_{\epsilon_n}.$$

Now we need to show that $\epsilon_n \to 0$. That is the same as saying that for a stricly convex increasing function F, we have $\lim_{x\to\infty}\frac{F^{-1}(x)}{x}=0$ or equivalenty $\lim_{x\to\infty}\frac{F(x)}{x}=+\infty$. This doesn't seem to be true in general. Indeed, the function $F(x)=\sqrt{x^2+1}-1$ is stricly convex but $F(x)\underset{x\to\infty}{\sim} x$. The result would be true if we assume that $\psi\circ\phi$ is strongly convex as it would be lower bounded by a positive quadratic. □

Exercise 2.4

Let $\mathcal{Y}=\{0,1\}, \mathcal{D}=[0,1]$, and $\ell(\bar{p},y)=|\bar{p}-y|$. Prove that the cumulative loss $\hat{L}$ of the exponentially weighted aerage forecaster is always at least as large as the cumulative loss $\min_{i\leq N} L_i$ of the best expert. Show that for other loss functions, such as the square loss $(\bar{p}-y)^2$, this is not necessarily so. *Hint :* Try to reverse the proof of Theorem 2.2.

Solution : As suggested by the hint, we will adapt the proof of Theorem 2.2. Using the same notations, we observe that

$$\begin{aligned}\log\frac{W_n}{W_0} &= \log\left(\sum_{i=1}^N e^{-\eta L_{i,n}}\right) - \log N\\ &\leq \log\left(N\max_{i=1,\dots,N} e^{-\eta L_{i,n}}\right) - \log N\\ &= -\eta\min_{i=1,\dots,N} L_{i,n}.\end{aligned}$$

Here rather than lower bounding the sum by its biggest element, we upper bound it by the number of elements time the maximal elements. Now we will obtain a lower bound of $\log\frac{W_n}{W_0}$. We have for each $t=1,\dots,n$,

$$\begin{aligned}\log\frac{W_t}{W_{t-1}} &= \log\frac{\sum_{i=1}^N e^{-\eta\ell(f_{i,t},y_t)}e^{-\eta L_{i,t-1}}}{\sum_{i=1}^N e^{-\eta L_{i,t-1}}}\\ &= \log\frac{\sum_{i=1}^N w_{i,t-1}e^{-\eta\ell(f_{i,t},y_t)}}{\sum_{i=1}^T w_{i,t-1}}.\end{aligned}$$

Now we define $\bar{w}_{i,t}=\frac{w_{i,t}}{\sum_{j=1}^N w_{j,t}}$ the normalized exponential weights. We have

:

$$\log \frac{W_t}{W_{t-1}} = \log \sum_{i=1}^{N} \bar{w}_{i,t-1} \cdot e^{-\eta\ell(f_{i,t},y_t)}$$
$$\geq \sum_{i=1}^{N} \bar{w}_{i,t-1} \cdot (-\eta\ell(f_{i,t}, y_t)),$$

where we use the concavity of the logarithm and Jensen's Inequality in the second line. Now, if $y_t = 0$, for any $f \in \mathcal{D}$ we have $\ell(f, y_t) = f$, and if $y_t = 1$, $\ell(f, y_t) = 1 - f$. In particular, the loss function is linear in it's first coordinate. In particular,

$$\log \frac{W_t}{W_{t-1}} \geq -\eta\ell(\sum_{i=1}^{N} \bar{w}_{i,t-1} f_{i.t}, y_t) = -\eta\ell(\hat{p}_t, y_t).$$

Here we used the absolute loss and the particular shape of the decision and outcome spaces. This particular steps is not possible anymore with the square loss. Summing over $t = 1, \ldots, n$ and combining both bounds, we get :

$$\hat{L}_t = \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) \geq \frac{-1}{\eta} \log \frac{W_n}{W_0} \geq \min_{i=1,\ldots,N} L_{i,n}.$$

For a counterexample with the square loss, we consider two experts, the first one always predict 0, and the second one always predicts 1. We then choose y_t to be 0 on even rounds and 1 on odd rounds. We run the algorithm during $2n$ rounds. It is easy to see that the regret of both experts is n(They suffer a regret of 1 during exactly half of the rounds). During odd rounds, our learner puts the same weight on both experts and suffers a regret of $\frac{1}{4}$. During even rounds, our algorithm makes the prediction $\frac{e^\eta}{e^\eta + e^{-\eta}}$. So the exponential weighted forecaster has a regret of $n(\frac{1}{4} + (\frac{e^\eta}{e^\eta + e^{-\eta}})^2)$. This regret is smaller than the one of either experts as long as η is small enough which will be true for the optimal $\eta^* \propto \frac{1}{\sqrt{n}}$ for n big enough.

$\square$

Exercise 2.5 Nonuniform initial weights

By definition, the weighted average forecaster uses uniform initial weights $w_{i,0} = 1$ for all $i = 1, \ldots, N$. However, there is nothing special abut this choice, and the analysis of the regret for this forecaster can be carried out using any set of nonnegative numbers for the inital weights.

Consider the exponentially weighted average forecaster run with arbitrary inital weights $w_{1,0}, \ldots, w_{N,0} > 0$, defined for all $t = 1, 2, \ldots,$ by

$$\hat{p}_t = \frac{\sum_{i=1}^{N} w_{i,t-1} f_{i,t}}{\sum_{j=1}^{N} w_{j,t-1}}, \quad w_{i,t} = w_{i,t-1} \exp -\eta\ell(f_{i,t}, y_t).$$

Under the same conditions as in the statement of Theorem 2.2, show that for every n and for every outcome sequence y^n,

$$\widehat{L_n} \leq \min_{i=1,\ldots,N} \left(L_{i,n} + \frac{1}{\eta} \log \frac{1}{w_{i,0}} \right) + \frac{\log W_0}{\eta} + \frac{\eta}{8} n$$

where $W_0 = w_{1,0} + \cdots + w_{N,0}$.

Solution : As in the proof of theorem 2.2, we define :

$$\begin{aligned} W_t &= \sum_{i=1}^{N} w_{i,t} \\ &= \sum_{i=1}^{N} \exp\left(-\eta L_{i,t}\right) w_{i,0} \\ &= \sum_{i=1}^{N} \exp -\eta \left(L_{i,t} + \frac{1}{\eta} \log \frac{1}{w_{i,0}} \right). \end{aligned}$$

Now as in the proof of theorem 2.2, we analyse $\log \frac{W_t}{W_0}$ and $\log \frac{W_t}{W_{t-1}}$. We have :

$$\begin{aligned} \log \frac{W_n}{W_0} &= \log \left(\sum_{i=1}^{N} \exp -\eta \left(L_{i,n} + \frac{1}{\eta} \log \frac{1}{w_{i,0}} \right) \right) - \log W_0 \\ &\geq \log \left(\max_{i=1,\ldots,N} \exp -\eta \left(L_{i,n} + \frac{1}{\eta} \log \frac{1}{w_{i,0}} \right) \right) \\ &= -\eta \left(\min_{i=1,\ldots,N} L_{i,n} + \frac{1}{\eta} \log \frac{1}{w_{i,0}} \right) - \log W_0. \end{aligned}$$

On the other hand, exactly as in the proof of Theorem 2.2, we get for each $t = 1, \ldots, n$,

$$\begin{aligned} \log \frac{W_t}{W_{t-1}} &= \log \frac{\sum_{i=1}^{N} \exp\left(-\eta \ell(f_{i,t}, y_t)\right) \exp\left(-\eta \left(L_{i,t-1} + \frac{1}{\eta} \log \frac{1}{w_{i,0}}\right)\right)}{\sum_{i=1}^{N} \exp\left(-\eta \left(L_{i,t-1} + \frac{1}{\eta} \log \frac{1}{w_{i,0}}\right)\right)} \\ &= \log \frac{\sum_{i=1}^{N} w_{i,t-1} \exp\left(-\eta \ell(f_{i,t}, y_t)\right)}{\sum_{i=1}^{N} w_{i,t-1}} \\ &\leq -\eta \ell(\hat{p}_t, y_t) + \frac{\eta^2}{8}. \end{aligned}$$

Where the last inequality is obtained by using Hoeffding's Lemma and the convecity of the loss exactly like in the proof of Theorem 2.2. Summing over $t = 1, \ldots, n$, we get

$$\log \frac{W_n}{W_0} \leq -\eta \widehat{L_n} + \frac{\eta^2}{8} \eta.$$

Putting this together with the previous upper bound and solving for $\widehat{L_n}$ we get :

$$\widehat{L_n} \leq \min_{i=1,\dots,N} \left(L_{i,n} + \frac{1}{\eta} \log \frac{1}{w_{i,0}} \right) + \frac{\log W_0}{\eta} + \frac{\eta}{8} n.$$

Exercise 2.6 Many good experts

Sequences of outcomes on which many experts suffer a small loss are intuitively easier to predict. Adapt the proof of Theorem 2.2 to show that the exponentially weighted forecaster satisfies the following property : for every n, for every outcome sequence y^n and for all $L > 0$,

$$\widehat{L_n} \leq L + \frac{1}{\eta} \log \frac{N}{N_L} + \frac{\eta}{8} n,$$

where N_L is the cardinality of the set $\{1 \leq i \leq N : L_{i,n} \leq L\}$

Solution : We adapt the lower bound of $\log \frac{W_n}{W_0}$. We have :

$$\begin{aligned}
\log \frac{W_n}{W_0} &= \log \left(\sum_{i=1}^{N} e^{-\eta L_{i,n}} \right) - \log N \\
&\geq \log \left(\sum_{i=1}^{N} e^{-\eta L_{i,n}} \cdot \mathbb{I}_{L_{i,n} \leq L} \right) - \log N \\
&\geq \log \left(\sum_{i=1}^{N} e^{-\eta L} \cdot \mathbb{I}_{L_{i,n} \leq L} \right) - \log N \\
&= \log \left(N_L e^{-\eta L} \right) - \log N \\
&= -\eta L - \log \frac{N}{N_L}.
\end{aligned}$$

We still have the upper bound of $\log \frac{W_n}{W_0}$ from the proof of Theorem 2.2 :

$$\log \frac{W_n}{W_0} \leq -\eta \widehat{L_n} + \frac{\eta^2}{8} n.$$

Putting everything together, we get :

$$\widehat{L_n} \leq L + \frac{1}{\eta} \log \frac{N}{N_L} + \frac{\eta}{8} n,$$

Exercise 2.8 The doubling trick

Consider the following forecasting strategy("Doubling Trick"): time is divided in periods $(2^m, \ldots, 2^{m+1}-1)$, where $m = 0, 1, 2, \ldots$. In period $(2^m, \ldots, 2^{m+1}-1)$ the strategy uses the exponentially weighted average forecaster initialized at time 2^m with parameter $\eta_m = \sqrt{8(\log N)/2^m}$. Thus, the weighted average forecaster is reset at each time instance that is an integer power of 2 and is restarted with a new value of η. Using Theorem 2.2 prove that, for any sequence $y_1, y_2, \ldots \in \mathcal{Y}$ of outcomes and for any $n \geq 1$, the regret of this forecaster is at most

$$\widehat{L_n} - \min_{i=1,\ldots,N} L_{i,n} \leq \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{\frac{n}{2}\log N}.$$

Solution : Let $n \geq 1$, and let M be the only integer such that $2^M \leq n \leq 2^{M+1}$. We have :

$$\begin{aligned}
&\widehat{L_n} - \min_{i=1,\ldots,N} L_{i,n} \\
&= \sum_{t=1}^{n} \ell(\hat{p}_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=1}^{n} f_{i,t} \\
&= \left(\sum_{m=0}^{M} \sum_{t=2^m}^{2^{m+1}\wedge n} \ell(\hat{p}_t, y_t) \right) - \min_{i=1,\ldots,N} \left(\sum_{m=0}^{M} \sum_{t=2^m}^{2^{m+1}\wedge n} f_{i,t} \right) \\
&\leq \sum_{m=0}^{M} \left(\sum_{t=2^m}^{2^{m+1}\wedge n} \ell(\hat{p}_t, y_t) \right) - \sum_{m=0}^{M} \left(\min_{i=1,\ldots,N} \sum_{t=2^m}^{2^{m+1}\wedge n} f_{i,t} \right) \\
&= \sum_{m=0}^{M} \left(\underbrace{\sum_{t=2^m}^{2^{m+1}\wedge n} \ell(\hat{p}_t, y_t) - \min_{i=1,\ldots,N} \sum_{t=2^m}^{2^{m+1}\wedge n} f_{i,t}}_{R_{m,n}} \right)
\end{aligned}$$

Where $R_{m,n}$ is the regret suffered by our forecaster during phase m(and up to time n) which has exactly $2^{m+1} \wedge n - 2^m \leq 2^m$ steps. In particular, using the bound of Theorem 2.2, and the choice of η_m we know that

$$\forall m \leq M, R_m \leq \sqrt{(2^m/2)\log N}.$$

Finally :

$$
\begin{aligned}
&\widehat{L_n} - \min_{i=1,\dots,N} L_{i,n} \\
&\le \sum_{m=0}^{M} \sqrt{(2^m/2)\log N} \\
&\le \sqrt{\frac{\log N}{2}} \cdot \frac{\sqrt{(2^{M+1})} - 1}{\sqrt{2}-1} \\
&\le \sqrt{\frac{\log N}{2}} \cdot \frac{\sqrt{(2^{M+1})}}{\sqrt{2}-1} \\
&\le \frac{\sqrt{2}}{\sqrt{2}-1}\sqrt{\frac{n}{2}\log N},
\end{aligned}
$$

where we used $2^{M+1} \le 2n$ on the last line.

Exercise 2.13

Prove Lemma 2.4 ($\forall z \ge \frac{1}{2}, \log(1+z) \ge z - z^2$).

Solution : Let $f(z) = \log(1+z) - z + z^2$. We have that f is differentiable on $[-\frac{1}{2}, +\infty[$. In particular :

$$
f'(z) = \frac{1}{1+z} - 1 + 2z = \frac{1}{1+z} \cdot (1 - (1+z) - 2z(1+z)) = \frac{z(2z+1)}{1+z}.
$$

From this we know that f' is non-positive on $[-\frac{1}{2}, 0]$ and non-negative on $[0, +\infty[$. In particular $\forall z \in [-\frac{1}{2}, +\infty[, f(z) \ge f(0) = 0$. That is to say :

$$
\forall z \ge \frac{1}{2}, \log(1+z) \ge z - z^2.
$$

Exercise 2.16

Prove a regret bound for the multilinear forecaster using the update $w_{i,t} = w_{i,t-1}(1 + \eta r_{i,t})$, where $r_{i,t} = h(f_{i,t}) - h(\hat{p}_t, y_t)$ is the instantaneous regret. What can you say about the evolution of the total weight $W_t = w_{1,t} + \cdots + w_{N,t}$ of the experts ?

Solution : We will prove a bound similar to the bound of Theorem 2.5. We will assume again that the payoff function h is concave in its first argument. We assume that $D = \max_{f_1, f_2 \in \mathcal{D}, y, \in \mathcal{Y}} h(f_1, y) - h(f_2, y) < \infty$ (In particular, that is the case if h is bounded). Then for any n and $0 < \eta < \frac{1}{2D}$, and for all $y_1, \ldots, y_n \in \mathcal{Y}$, the regret of the multilnear forecaster satisfies

$$R_{i,n} \leq \frac{\log N}{\eta} + \eta \sum_{t=1}^{n} r_{i,t}^2 \quad \text{for each } i = 1, \ldots, N.$$

We follow the proof of theorem 2.5, since $\eta \leq \frac{1}{2D}$, we have that $\eta r_{i,t} \geq -\frac{1}{2}$ and we can apply Lemma 2.4 to $\eta r_{i,t}$

$$\begin{aligned}
\log \frac{W_n}{W_0} &= -\log N + \log \prod_{t=1}^{n}(1 + \eta r_{i,t}) \\
&= -\log N + \sum_{t=1}^{n} \log(1 + \eta r_{i,t}) \\
&\leq -\log N + \sum_{t=1}^{n}(\eta r_{i,t} - \eta^2 r_{i,t}^2) \\
&\leq -\log N + \eta R_{i,n} - \eta^2 \sum_{t=1}^{n} r_{i,t}^2.
\end{aligned}$$

On the other hand, defining $\bar{w}_{i,t} = \frac{w_{i,t}}{W_t}$ the normalized weights, we have that for any t,

$$\begin{aligned}
\log \frac{W_t}{W_{t-1}} &= \log \left(\sum_{i=1}^{N} \bar{w}_{i,t}(1 + \eta r_{i,t}) \right) \\
&\leq \eta \sum_{i=1}^{N} \bar{w}_{i,t} r_{i,t} \\
&= \eta (\sum_{i=1}^{N} \bar{w}_{i,t} h(f_{i,t}, y_t) - h(\hat{p}_t, y_t)) \\
&\leq \eta (h(\hat{p}_t, y_t) - h(\hat{p}_t, y_t)) \\
&= 0,
\end{aligned}$$

where the first inequality is a consequence of the elementary inequality $\log(1 + x) \leq x$ for any $x > -1$ and the second inequality comes from the concavity of h in its first argument. In particular we have proven that the total weight W_t is decreasing and that $\log \frac{W_n}{W_0} \leq 0$. Combining that with the lower bound of $\log \frac{W_n}{W_0}$ and dividing by $\eta > 0$, we get the claimed bound. ☐

Exercise 2.17

Prove Lemma 2.5. *Hint: Adapt the proof of Theorem 2.3.*

Lemma 2.5. *Let h be a payoff function concave in its first argument. The exponentially weighted average forecaster, run with any nonincreasing sequence $\eta_1, \eta_2, \dots$ of parameters satisfies, for any $n \geq 1$ and for any sequence $y_1, \dots, y_n$ of outcomes,*

$$H_n^* - \widehat{H}_n \leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1}\right) \log N + \sum_{t=1}^{n} \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t (X_t - \mathbb{E}[X_t])}\right].$$

Where X_t are independent random variable such that $X_t = h(f_{i,t}, y_t)$ with probability $\frac{w_{i,t-1}}{W_{t-1}}$.

Solution : We follow the structure of the proof of Theorem 2.3. We once again define the currently best expert k_t that we will use to lower bound the weight $\log(\frac{w_{k_t,t}}{W_t})$ and $w'_{i,t-1} = \exp \eta_{t-1} H_{i,t-1}$ to denote the weight $w_{i,t-1}$ where the parameter η_t is replaced by η_{t-1}. We write the following :

$$\begin{aligned}
&\frac{1}{\eta_t} \log \frac{w_{k_{t-1},t-1}}{W_{t-1}} - \frac{1}{\eta_{t+1}} \log \frac{w_{k_t,t}}{W_t} \\
&= \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \log \frac{W_t}{w_{k_t,t}}}_{(A)} + \underbrace{\frac{1}{\eta_t} \log \frac{w'_{k_t,t}/W'_t}{w_{k_t,t}/W_t}}_{(B)} + \underbrace{\frac{1}{\eta_t} \log \frac{w_{k_{t-1},t-1}/W_{t-1}}{w'_{k_t,t}/W'_t}}_{(C)}.
\end{aligned}$$

As in the proof of Theorem 2.3 we have that

$$(A) \leq \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \log N.$$

because the sequence η_t is increasing. For (B), we have :

$$\begin{aligned}
(B) &= \frac{1}{\eta_t} \log \frac{w'_{k_t,t}/W'_t}{w_{k_t,t}/W_t} = \frac{1}{\eta_t} \log \frac{\sum_{i=1}^N e^{\eta_{t+1}(H_{i,t} - H_{k_t,t})}}{\sum_{i=1}^N e^{\eta_t (H_{i,t} - H_{k_t,t})}} \\
&\leq \frac{\eta_t - \eta_{t+1}}{\eta_t \eta_{t+1}} \log N = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \log N,
\end{aligned}$$

where we applied Lemma 2.3 with $d_i = H_{k_{t+1},t} - H_t$, $\alpha = \eta_{t+1}$ and $\beta = \eta_t$. As in the proof of theorem 2.3, we have that $d_i \geq 0$ because k_{t+1} is the index of the expert with the highest payoff after the first t rounds and $\sum_{i=1}^N e^{-\eta_{t+1} d_i} \geq 1$ as $d_{k_{t+1}} = 0$.

Now, we split the term (C) :

$$(C) = \frac{1}{\eta_t} \log \frac{w_{k_{t-1},t-1}/W_{t-1}}{w'_{k_t,t}/W'_t} = \frac{1}{\eta_t} \frac{w_{k_{t-1},t-1}}{w'_{k_t,t}} + \frac{1}{\eta_t} \log \frac{W'_t}{W_{t-1}}.$$

We treat the two subterms of the right-hand side separately. For the first one, we have :

$$\frac{1}{\eta_t} \log \frac{w_{k_{t-1},t-1}}{w'_{k_t,t}} = \frac{1}{\eta_t} \log \frac{e^{\eta_t H_{k_{t-1},t-1}}}{e^{\eta_t H_{k_t,t}}} = H_{k_{t-1},t-1} - H_{k_t,t}.$$

So far, we have not changed anything in the proof of Theorem 2.3, we have just checked that the steps of the proof carry on to the more general signed games setting. Now we will study the second term of the right-hand side of (C) and introduce the variable X_t to replace the Hoeffding's bound that was done in the proof of Theorem 2.3. We have :

$$\frac{1}{\eta_t} \log \frac{W_t'}{W_{t-1}} = \frac{1}{\eta_t} \log \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}} e^{\eta_t h(f_{i,t}, y_t)} = \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t X_t}\right]$$

by definition of X_t. Now we compute $\mathbb{E}[X_t]$:

$$\begin{aligned} \mathbb{E}[X_t] &= \sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}} h(f_{i,t}, y_t) \\ &\leq h\left(\sum_{i=1}^{N} \frac{w_{i,t-1}}{W_{t-1}} f_{i,t}, y_t\right) \\ &= h(\hat{p}_t, y_t). \end{aligned}$$

Where the inequality comes from the concavity of h in its first argument. Coming back to the previous bound, we have :

$$\begin{aligned} \frac{1}{\eta_t} \log \frac{W_t'}{W_{t-1}} &\leq \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t X_t}\right] \\ &\leq \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t X_t}\right] + h(\hat{p}_t, y_t) - \mathbb{E}[X_t] \\ &= \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t X_t}\right] + h(\hat{p}_t, y_t) - \frac{1}{\eta_t} \log e^{\eta_t \mathbb{E}[X_t]} \\ &= \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t (X_t - \mathbb{E}[X_t])}\right] + h(\hat{p}_t, y_t). \end{aligned}$$

Putting everything together, we get :

$$\begin{aligned} &\frac{1}{\eta_t} \log \frac{w_{k_{t-1},t-1}}{W_{t-1}} - \frac{1}{\eta_{t+1}} \log \frac{w_{k_t,t}}{W_t} \\ &\leq 2\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \log N + H_{k_{t-1},t-1} - H_{k_t,t} + \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t (X_t - \mathbb{E}[X_t])}\right] + h(\hat{p}_t, y_t). \end{aligned}$$

Or rearranging :

$$\begin{aligned} &H_{k_{t-1},t-1} - H_{k_t,t} - h(\hat{p}_t, y_t) \\ &\leq 2\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right) \log N + \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t (X_t - \mathbb{E}[X_t])}\right] - \frac{1}{\eta_t} \log \frac{w_{k_{t-1},t-1}}{W_{t-1}} + \frac{1}{\eta_{t+1}} \log \frac{w_{k_t,t}}{W_t}. \end{aligned}$$

Summing all of these inequalities for $t = 1, \dots, n$, we get :

$$\begin{aligned}
H_n^* - \widehat{H}_n &\leq 2\left(\frac{1}{\eta_{n+1}} - \frac{1}{\eta_1}\right)\log N + \sum_{t=1}^{n} \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t(X_t - \mathbb{E}[X_t])}\right] \\
&\quad + \frac{1}{\eta_{n+1}} \log \frac{w_{k_n,n}}{W_n} - \frac{1}{\eta_1} \log \frac{1}{N} \\
&\leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1}\right)\log N + \sum_{t=1}^{n} \frac{1}{\eta_t} \log \mathbb{E}\left[e^{\eta_t(X_t - \mathbb{E}[X_t])}\right],
\end{aligned}$$

where we used that $\frac{w_{k_n,n}}{W_n} \leq 1$.

$\square$

Exercise 2.19

Consider a class $\mathcal{F}$ of simulatable experts. Assume that the set $\mathcal{Y}$ of outcomes is a compact subset of $\mathbb{R}^d$, the decision space $\mathcal{D}$ is convex, and the loss function ℓ is convex and continuous in its first argument. Show that $V_n(\mathcal{F}) = U_n(\mathcal{F})$. *Hint: Check the conditions of Theorem 7.1*

Solution : We define $\mathcal{P}$ the set of all the forecasting strategies on decision space $\mathcal{D}$ and outcome space $\mathcal{Y}$ where a forecasting strategy $P \in \mathcal{P}$ is a sequence $\hat{p}_1, \hat{p}_2 \dots$ of functions

$$\hat{p}_t : \mathcal{Y}^{t-1} \times (\mathcal{D}^N)^t \to \mathcal{D}.$$

We also define $\mathcal{P}(\mathcal{Y})$ the set of all the Borel probability measures on Y. As suggested in the hint, we look at the function :

$$\begin{aligned}
&f : \mathcal{P} \times \mathcal{P}(\mathcal{Y}) \longrightarrow \mathbb{R} \\
&\qquad (P, Q) \longmapsto \int_{\mathcal{Y}^n} \left(\sum_{t=1}^{n} \ell(\hat{p}_t(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t)\right) dQ(y^n).
\end{aligned}$$

And check the conditions of Theorem 7.1. The first step is to check the measurability of the function $g(y) = \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t)$. Since $\mathcal{Y}$ is a compact set, we can find a countable set $\mathcal{F}_d$ such that $g(y) = \inf_{f \in \mathcal{F}_d} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t)$. To see that, we can see that any function $f_t(y^{t-1})$ can be uniformely approximated by a multivariate polynomial with rational coefficients using Stone-Weierstrass Theorem. Then, we only need to approximate n of this functions to get the same infimum(Since the infimum only depends on the predictions up to time n). Now, we also know that ℓ is bounded because $\mathcal{Y}$ is a compact space and ℓ is continuous (Here we make the additional assumption that ℓ is continuous or that ℓ is bounded, we might be able to relax it and only assume that ℓ is continuous in its first argument but I'm not sure of that.). That proves that f is well defined and bounded. Moreover, using dominated convergence, this proves the continuity of f in P(as ℓ is continuous in its first argument). We can easily check that both $\mathcal{P}$ and $\mathcal{P}(\mathcal{Y})$ are convex sets. Now we only need to check the convexity of $f(\cdot, Q)$ and the concavity of $f(P, \cdot)$ for any $P \in \mathcal{P}$ and any $Q \in \mathcal{P}(\mathcal{Y})$. The linearity of $f(P, \cdot)$ directly implies its concavity. Now, let $Q \in \mathcal{P}(\mathcal{Y})$, $P_1, P_2 \in \mathcal{P}$, and $\lambda \in [0, 1]$ We have :

$$
\begin{aligned}
& f((1-\lambda)P_1 + \lambda P_2, Q) \\
&= \int_{\mathcal{Y}^n} \left(\sum_{t=1}^{n} \ell((1-\lambda)\hat{p}_{1,t}(y^{t-1}) + \lambda \hat{p}_{2,t}(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t) \right) dQ(y^n) \\
&\leq \int_{\mathcal{Y}^n} \left(\sum_{t=1}^{n} (1-\lambda)\ell(\hat{p}_{1,t}(y^{t-1}), y_t) + \lambda \ell(\hat{p}_{2,t}(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t) \right) dQ(y^n) \\
&= (1-\lambda) \int_{\mathcal{Y}^n} \left(\sum_{t=1}^{n} \ell(\hat{p}_{1,t}(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t) \right) dQ(y^n) \\
&\quad + \lambda \cdot \int_{\mathcal{Y}^n} \left(\sum_{t=1}^{n} \ell(\hat{p}_{2,t}(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f_t(y^{t-1}), y_t) \right) dQ(y^n) \\
&= (1-\lambda) f(P_1, Q) + \lambda f(P_2, Q).
\end{aligned}
$$

Where the inequality comes from the convexity of ℓ in its first argument and all the integrals manipulations are valid because everything is bounded.

$\square$

Chapter 3

Chapter 03

Exercise 3.2

Let $\mathcal{Y} = \mathcal{D} = [0,1]$, and consider the *Hellinger loss*

$$\ell(z,y) = \frac{1}{2}\left((\sqrt{z}-\sqrt{y})^2 + (\sqrt{1-z}-\sqrt{1-y})^2\right)$$

(see Figure 3.6). Determine the values of η for which the function $F(z)$ defined in Section 3.3 is concave.

Solution : We have

$$F(z) = e^{-\eta\ell(z,y)}.$$

As a result, we have

$$\frac{\partial F}{\partial z} = -\eta\frac{\partial \ell}{\partial z}(z,y)e^{-\eta\ell(z,y)}.$$

and

$$\frac{\partial^2 F}{\partial^2 z} = \left(-\frac{\partial^2 \ell}{\partial z}(z,y) + \eta\left(\frac{\partial \ell}{\partial z}(z,y)\right)^2\right)\eta e^{-\eta\ell(z,y)}.$$

As a result, we know that F is going to be concave as long as $z \to \left(-\frac{\partial^2 \ell}{\partial z}(z,y) + \eta\left(\frac{\partial \ell}{\partial z}(z,y)\right)^2\right)$ is non-positive for all values of $y \in \mathcal{Y}$. We have

$$\frac{\partial \ell}{\partial z}(z,y) = \frac{1}{2}\left(\frac{\sqrt{1-y}}{\sqrt{1-z}} - \frac{\sqrt{y}}{\sqrt{z}}\right),$$

and

$$\frac{\partial^2 \ell}{\partial^2 z}\ell(z,y) = \frac{1}{4}\left(\frac{\sqrt{1-y}}{(1-z)^{\frac{3}{2}}} + \frac{\sqrt{y}}{z^{\frac{3}{2}}}\right).$$

The maximal value of η for which the function F is concave is then the infimum of $\frac{\frac{\partial^2 \ell}{\partial z}(z,y)}{\left(\frac{\partial \ell}{\partial z}(z,y)\right)^2}$. I have not been able to explicitly compute this infimum but I have found a nonnegative value of η that works. Indeed

$$\begin{aligned}
\left(\frac{\partial \ell}{\partial z}(z,y)\right)^2 &\leq \frac{1}{4}\left(\frac{1-y}{1-z} + \frac{y}{z} - 2\sqrt{\frac{y(1-y)}{z(1-z)}}\right) \\
&\leq \frac{1}{4}\left(\frac{1-y}{1-z} + \frac{y}{z}\right) \\
&\leq \frac{1}{4}\left(\frac{\sqrt{1-y}}{(1-z)^{\frac{3}{2}}} + \frac{\sqrt{y}}{z^{\frac{3}{2}}}\right) \\
&= \frac{\partial^2 \ell}{\partial^2 z}(z,y),
\end{aligned}$$

where we used $y \leq \sqrt{y}, 1-y \leq (1-y)$ because $y \in [0,1]$ and $\frac{1}{1-z} \leq \left(\frac{1}{1-z}\right)^{\frac{3}{2}}, \frac{1}{z} \leq \left(\frac{1}{z}\right)^{\frac{3}{2}}$ because $\frac{1}{z} \geq 1, \frac{1}{1-z} \geq 1$. As a result, we have shown that F is concave for $\eta \leq 1$. $\square$

Chapter 9

Chapter 09

Exercise 9.6

Show that

$$\int_0^1 \frac{1}{\sqrt{x(1-x)}}\,dx = \pi$$

Hint : Substitue x by $\sin^2\alpha$

Solution : As suggested by the hint, we consider the change of variable $x = \sin^2\alpha$. This change of variable is well defined and the squared sinus is a one to one differentiable mapping with non vanishing derivative from $]0, \frac{\pi}{2}[$ to $]0, 1[$. Moreover, we have $\frac{dx}{d\alpha} = 2\cos\alpha\sin\alpha$. Hence by the change of variable theorem :

$$\begin{aligned}
\int_0^1 \frac{1}{\sqrt{x(1-x)}}\,dx &= \int_0^{\frac{\pi}{2}} \frac{1}{\sqrt{\sin^2\alpha\cos^2\alpha}} \cdot 2\sin\alpha\cos\alpha\,d\alpha \\
&= 2\int_0^{\frac{\pi}{2}} 1\,d\alpha \\
&= \pi
\end{aligned}$$

$\square$