Exercise 1 (Hedge and follow the leader)**.** There are $K \geq 2$ experts. At each round $t \geq 1$, the agent chooses a probability vector p_t over $\{1, \dots, K\}$, then the gains $g_t \in [0,1]^K$ of the experts are revealed and the agent receives $\langle p_t, g_t \rangle$. We write $G_0 = 0$ and $G_t = \sum_{s=1}^{t} g_s$ for the vector of the cumulative gains. After T rounds, the regret is

$$R_T = \max_{1 \leq a \leq K} G_T(a) - \sum_{t=1}^{T} \langle p_t, g_t \rangle \ . \tag{1}$$

Follow the leader (FTL) plays the expert of smallest index among those that maximize G_{t-1}. Hedge, with learning rate $\eta > 0$, plays

$$p_t(a) = \frac{e^{\eta G_{t-1}(a)}}{\sum_{b=1}^{K} e^{\eta G_{t-1}(b)}} \ . \tag{2}$$

In questions 1 to 3, $K = 3$, $T = 3$, $\eta = \log 2$, and the gains are

$$g_1 = (1,0,0) \ , \qquad g_2 = (1,1,0) \ , \qquad g_3 = (0,1,1) \ .$$

1. Compute p_1, p_2 and p_3 for Hedge.
2. Compute the total gain of Hedge and its regret R_3.
3. Compute the total gain of FTL and its regret on the same sequence.

In the following questions, $K = 2$ and $T = 2m + 1$ with $m \geq 1$. The gains are $g_1 = (1/2, 0)$, then $g_t = (0,1)$ for even t and $g_t = (1,0)$ for odd $t \geq 3$.

4. Show that $G_T = (1/2 + m, m)$, and that $G_t(1) - G_t(2) = 1/2$ for odd t and $-1/2$ for even t.
5. Show that FTL receives 0 at every round $t \geq 2$, and deduce that its regret is $R_T = m = (T-1)/2$.
6. Show that, for $t \geq 2$, Hedge receives $\langle p_t, g_t \rangle = 1/(1 + e^{\eta/2})$, and deduce that

$$R_T = \frac{1}{4} + m \tanh\left(\frac{\eta}{4}\right) \ .$$

7. For $m = 500$, compute R_T for Hedge with $\eta = \sqrt{8\log(2)/T}$, and compare it with the bound $\sqrt{T \log(K)/2}$, which holds for this learning rate on any sequence. Give the limits of R_T when $\eta \to 0$ and when $\eta \to +\infty$.
   The following approximate values may be used: $\sqrt{8\log(2)/1001} \approx 0.0744$, $\tanh(0.0186) \approx 0.0186$ and $\sqrt{1001 \log(2)/2} \approx 18.6$.

Exercise 2 (The importance sampling estimator)**.** Let $K \geq 2$, $r \in [0,1]^K$ a vector of rewards, and p a probability vector over $\{1, \dots, K\}$ such that $p_a > 0$ for any a. Let A be a random action with distribution p, and, for any a,

$$\hat{r}_a = \frac{r_a \mathbf{1}\{A = a\}}{p_a} \ . \tag{3}$$

Only r_A is observed, but $\hat{r}$ can be computed from A, r_A and p.

1. Show that $\mathbb{E}[\hat{r}_a] = r_a$ for any a.
2. Show that $\mathrm{Var}(\hat{r}_a) = r_a^2 \, \dfrac{1 - p_a}{p_a}$.

3. Let π be another probability vector (the target policy) and $V(\pi) = \sum_a \pi_a r_a$. Show that

$$\hat{V} = \frac{\pi_A}{p_A} r_A = \sum_{a=1}^{K} \pi_a \hat{r}_a \qquad \text{and} \qquad \mathbb{E}[\hat{V}] = V(\pi) \ .$$

4. Assume now that $p_b = 0$ for one action b, and set $\hat{V} = \pi_A r_A / p_A$ (the action b is never drawn). Show that $\mathbb{E}[\hat{V}] = V(\pi) - \pi_b r_b$.
5. The rewards now depend on a context. Let U be a random context, and, conditionally on $U = u$, let A have distribution $\mu(\cdot \mid u)$ and R satisfy $\mathbb{E}[R \mid U = u, A = a] = m(u, a)$. Let $\pi(\cdot \mid u)$ be a target policy, and assume $\mu(a \mid u) > 0$ for any u and a. Show that

$$\mathbb{E}\Big[\frac{\pi(A \mid U)}{\mu(A \mid U)} R\Big] = \mathbb{E}\Big[\sum_{a=1}^{K} \pi(a \mid U)\, m(U, a)\Big] \ ,$$

   the expected reward of a user who receives an action drawn from π.

Exercise 3 (Variance and clipping)**.** There are two actions. The reward of action 1 is a Bernoulli variable with mean $m \in (0, 1]$. The logging policy plays action 1 with probability $q \in (0, 1)$, and the target policy always plays action 1. From n independent logs (A_i, R_i), the importance sampling estimator and its clipped version, for $M > 0$, are

$$\hat{V}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}\{A_i = 1\}}{q} R_i \qquad \text{and} \qquad \hat{V}_n^M = \frac{1}{n} \sum_{i=1}^{n} \min\Big(\frac{1}{q}, M\Big) \mathbf{1}\{A_i = 1\} R_i \ . \tag{4}$$

1. Show that $\mathbb{E}[\hat{V}_n] = m$ and that $\mathrm{Var}(\hat{V}_n) = \frac{1}{n}\Big(\frac{m}{q} - m^2\Big)$.
2. Let $m = 1/2$ and $n = 1000$. Compute the standard deviation of $\hat{V}_n$ for $q = 1/2$ and for $q = 1/100$.
3. With $m = 1/2$, give the number of logs n needed for a standard deviation at most 0.01, for $q = 1/2$ and for $q = 1/100$.
4. Assume $M < 1/q$. Show that the bias and the variance of $\hat{V}_n^M$ are

$$\mathbb{E}[\hat{V}_n^M] - m = -m(1 - Mq) \qquad \text{and} \qquad \mathrm{Var}(\hat{V}_n^M) = \frac{M^2 q m (1 - qm)}{n} \ .$$

5. Deduce that, among $M \in (0, 1/q]$, the mean squared error $\mathbb{E}[(\hat{V}_n^M - m)^2]$ is minimized at

$$M^\star = \frac{nm}{nmq + 1 - qm} \ .$$

   Compute $M^\star$ and the root mean squared errors of $\hat{V}_n^{M^\star}$ and of $\hat{V}_n$ for $m = 1/2$, $q = 1/100$ and $n = 1000$.