

Lecture 8: Adaptive Regret Bounds for Multi-Armed Bandits

Lecturer: *Mengxiao Zhang*

Week 8

1 Beyond the Worst Case

In the previous lecture, we established that the minimax pseudo-regret in adversarial multi-armed bandits is $\Theta(\sqrt{KT})$. Tsallis-INF achieves this rate by matching the regularizer's local curvature to the importance-weighted estimator. However, this is a worst-case statement: it treats every sequence of T loss vectors as equally difficult. While this gives the learner the most robust guarantee, some environments contain much more structure. In Lecture 5, we have shown that in a stochastic environment with fixed positive gaps, the learner can indeed achieve logarithmic regret. An adversarial sequence may also be easier if the best arm suffers little loss, as you have seen in Homework 1. So our main question is:

Can we design algorithms that have an improved regret guarantee on these easier instances, while simultaneously maintain the worst-case near-optimal guarantees?

For example, when choosing a network route, losses may have stable distributions during normal operation but change when congestion or failures occur. A purely stochastic guarantee gives no protection against such changes, while a worst-case bound does not reflect the easier periods of operation. A guarantee that adapts to the loss sequence lets us exploit favorable structure without first deciding whether a stochastic model is appropriate.

2 Best-of-Both-Worlds with Tsallis-INF

We first discuss the case where the loss sequence can either be stochastic or adversarial. In Lecture 5, we have shown that several algorithms achieve $\mathcal{O}(\sum_{i \neq i^*} \frac{\ln T}{\Delta_i})$ logarithmic instance-dependent regret in stochastic environments, including UCB and arm elimination. However, their guarantees rely on the stochastic environment and do not apply to arbitrary adversarial loss sequences. Our question is can we use one algorithm to achieve logarithmic regret in the stochastic case and the minimax regret bound in the adversarial case, without knowing which environment we face?

Fortunately, the answer is yes, and it is achieved by an algorithm that we have already learned! Recall that Tsallis-INF achieves $\mathcal{O}(\sqrt{KT})$ regret in adversarial environments. We next show that, with a *time-varying learning rate*, Tsallis-INF simultaneously achieves $\mathcal{O}(\sum_{i \neq i^*} \frac{\log T}{\Delta_i})$ pseudo-regret in stochastic environments with a unique best arm.

2.1 Tsallis-INF with an adaptive learning rate

In Lecture 7, we ran Tsallis-INF with the fixed learning rate $\eta = 1/\sqrt{T}$. Here, we use $\eta_t = 1/(2\sqrt{t})$ and keep the same $\psi_{1/2}$ regularizer, importance-weighted estimates, and cumulative estimated losses $\widehat{L}_t$. See Algorithm 1 for the complete update. This schedule requires neither the horizon nor the stochastic gaps. We will account explicitly for the changing regularization coefficient when adapting the potential argument from Lecture 7.

Algorithm 1: Tsallis-INF with a Decreasing Learning Rate

Input: arms $[K]$, $\psi_{1/2}(p) \triangleq 2(1 - \sum_{i=1}^K \sqrt{p_i})$;
Initialize $\widehat{L}_{0,i} = 0$ for every $i \in [K]$;
for $t = 1, 2, \dots$ **do**
 set $\eta_t = 1/(2\sqrt{t})$;
 choose $p_t = \arg \min_{p \in \Delta_K} \left\{ \langle p, \widehat{L}_{t-1} \rangle + \frac{1}{\eta_t} \psi_{1/2}(p) \right\}$;
 draw $I_t \sim p_t$ and observe ℓ_{t,I_t} ;
 set $\widehat{\ell}_{t,i} = \frac{\ell_{t,i}}{p_{t,i}} \mathbf{1}\{I_t = i\}$ for every $i \in [K]$;
 set $\widehat{L}_{t,i} = \widehat{L}_{t-1,i} + \widehat{\ell}_{t,i}$ for every $i \in [K]$;

The same algorithm satisfies the following two guarantees.

Theorem 2.1. *Fix $K \geq 2$. For every fixed oblivious loss sequence $\ell_1, \dots, \ell_T \in [0, 1]^K$, Algorithm 1 guarantees that*

$$\overline{\text{Reg}}_T \leq \mathcal{O}\left(\sqrt{KT}\right).$$

If instead the loss vectors are independent and identically distributed in $[0, 1]^K$, with a unique best mean-loss arm i^* and gaps $\Delta_i \triangleq \mathbb{E}[\ell_{t,i}] - \mathbb{E}[\ell_{t,i^*}] > 0$ for $i \neq i^*$, the same algorithm guarantees that

$$\overline{\text{Reg}}_T \leq \mathcal{O}\left(\sum_{i \neq i^*} \frac{\log T}{\Delta_i}\right).$$

Both guarantees follow from the same intermediate bound. We keep the dependence on the sampling probabilities and exclude the comparator arm, so that the bound can be used with both arbitrary losses and stochastic gaps. Denote by e_j the j th standard basis vector. We first state the bound and deduce the two guarantees; we then prove it by adapting the analysis from Lecture 7.

Lemma 2.2. *Fix $K \geq 2$. There exists a universal numerical constant $C > 0$ such that Algorithm 1 guarantees that*

$$\mathbb{E} \left[\sum_{t=1}^T \langle p_t - e_j, \ell_t \rangle \right] \leq C \sum_{t=1}^T \frac{1}{\sqrt{t}} \sum_{\substack{i \in [K] \\ i \neq j}} \mathbb{E}[\sqrt{p_{t,i}}], \quad (1)$$

for every fixed oblivious loss sequence in $[0, 1]^K$ and every fixed arm $j \in [K]$.

Before proving Lemma 2.2, we use it to prove Theorem 2.1.

Proof of Theorem 2.1. For a fixed oblivious loss sequence, choose an arm j minimizing cumulative loss. Cauchy–Schwarz inequality gives $\sum_{i \in [K], i \neq j} \sqrt{p_{t,i}} \leq \sqrt{K-1}$. Applying Lemma 2.2 and using

$\sum_{t=1}^T t^{-1/2} \leq 2\sqrt{T}$, we obtain

$$\overline{\text{Reg}}_T \leq C\sqrt{K-1} \sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2C\sqrt{(K-1)T} = \mathcal{O}(\sqrt{KT}).$$

For the stochastic guarantee, use the same lemma with $j = i^*$. The gap identity from Lecture 5 gives

$$\overline{\text{Reg}}_T = \sum_{t=1}^T \sum_{i \neq i^*} \Delta_i \mathbb{E}[p_{t,i}]. \quad (2)$$

Although we stated Lemma 2.2 for fixed oblivious losses, it also applies after averaging an independent stochastic loss sequence sampled before the interaction. Conditional on that sequence, its bound holds for every fixed comparator; we then take expectation over the losses. Choose $j = i^*$ and combine the lemma with equation (2). The lemma upper bounds regret using $\sqrt{p_{t,i}}$, while the gap identity expresses the same regret using $\Delta_i p_{t,i}$. Applying the lemma to $2\overline{\text{Reg}}_T$ and subtracting the gap identity gives

$$\begin{aligned} \overline{\text{Reg}}_T &= 2\overline{\text{Reg}}_T - \overline{\text{Reg}}_T \\ &\leq \sum_{t=1}^T \sum_{i \neq i^*} \mathbb{E} \left[\frac{2C}{\sqrt{t}} \sqrt{p_{t,i}} - \Delta_i p_{t,i} \right] \\ &\leq \sum_{t=1}^T \sum_{i \neq i^*} \frac{C^2}{t\Delta_i} \leq C^2(1 + \ln T) \sum_{i \neq i^*} \frac{1}{\Delta_i}. \end{aligned}$$

The third line uses $2Cx/\sqrt{t} - \Delta_i x^2 \leq C^2/(t\Delta_i)$ for $x \geq 0$, by completing the square. Thus each gap pays for the corresponding square-root probability term, and summing $1/t$ produces the logarithmic dependence on T . This proves the stochastic guarantee in Theorem 2.1. The algorithm uses neither the gaps nor the knowledge that the losses are stochastic. $\square$

It remains to prove Lemma 2.2. We adapt the FTRL potential argument from Lecture 7.

Proof of Lemma 2.2. Denote by $\Phi_t(\eta) \triangleq \min_{p \in \Delta_K} \{ \langle p, \hat{L}_t \rangle + \psi_{1/2}(p)/\eta \}$ the regularized minimum from Lecture 7, and by $p_{t+1}(\eta)$ its minimizer. In Lecture 7, we proved the identity

$$\Phi_t(\eta) - \Phi_{t-1}(\eta) = \langle p_{t+1}(\eta), \hat{\ell}_t \rangle + \frac{1}{\eta} D_{\psi_{1/2}}(p_{t+1}(\eta), p_t(\eta)). \quad (3)$$

We now set $\eta = \eta_t$. Then $p_t(\eta_t) = p_t$, and $p_{t+1}(\eta_t)$ is the next minimizer with the current rate held fixed; the algorithm's next decision uses η_{t+1} . Denote by

$$S_t \triangleq \langle p_t - p_{t+1}(\eta_t), \hat{\ell}_t \rangle - \frac{1}{\eta_t} D_{\psi_{1/2}}(p_{t+1}(\eta_t), p_t)$$

the same stability contribution as in Lecture 7. By equation (3), we have $\langle p_t, \hat{\ell}_t \rangle = \Phi_t(\eta_t) - \Phi_{t-1}(\eta_t) + S_t$.

Summing this identity over all rounds gives

$$\begin{aligned}\sum_{t=1}^T \langle p_t, \hat{\ell}_t \rangle &= \sum_{t=1}^T (\Phi_t(\eta_t) - \Phi_{t-1}(\eta_t)) + \sum_{t=1}^T S_t \\ &= \Phi_T(\eta_T) - \Phi_0(\eta_1) + \sum_{t=2}^T (\Phi_{t-1}(\eta_{t-1}) - \Phi_{t-1}(\eta_t)) + \sum_{t=1}^T S_t.\end{aligned}$$

Since $\hat{L}_0 = 0$, the initial potential is $\Phi_0(\eta_1) = \psi_{1/2}(p_1)/\eta_1$: p_1 is exactly the minimizer in its definition. To bound the differences caused by changing rates, evaluate the old-rate minimum at p_t :

$$\Phi_{t-1}(\eta_{t-1}) \leq \langle p_t, \hat{L}_{t-1} \rangle + \frac{\psi_{1/2}(p_t)}{\eta_{t-1}} = \Phi_{t-1}(\eta_t) + \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t} \right) \psi_{1/2}(p_t).$$

The equality uses that p_t minimizes the objective with rate η_t . Also, $\Phi_T(\eta_T) \leq \hat{L}_{T,j}$ because $\psi_{1/2}(e_j) = 0$. Substituting these bounds into the sum above gives

$$\sum_{t=1}^T \langle p_t - e_j, \hat{\ell}_t \rangle \leq \frac{-\psi_{1/2}(p_1)}{\eta_1} + \sum_{t=1}^T S_t + \sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) (-\psi_{1/2}(p_t)). \quad (4)$$

With a fixed rate, the last sum vanishes, recovering the decomposition from Lecture 7. For our schedule, $1/\eta_t - 1/\eta_{t-1} = 2(\sqrt{t} - \sqrt{t-1}) \leq 2/\sqrt{t}$. Together with $-\psi_{1/2}(p) \leq 2 \sum_{i \neq j} \sqrt{p_i}$ and $\eta_1 = 1/2$, this bounds the regularization terms in equation (4) by $4 \sum_{t=1}^T t^{-1/2} \sum_{i \neq j} \sqrt{p_{t,i}}$.

Denote by $\mathbb{E}_t$ conditional expectation given the past and the current loss vector, before drawing I_t . Since S_t holds η_t fixed, the nonnegative-loss argument from Lecture 7 still gives

$$\begin{aligned}S_t &\leq \eta_t \sum_{i=1}^K p_{t,i}^{3/2} \hat{\ell}_{t,i}^2, \\ \mathbb{E}_t[S_t] &\leq \eta_t \sum_{i=1}^K \sqrt{p_{t,i}} \ell_{t,i}^2 \leq \eta_t \sum_{i=1}^K \sqrt{p_{t,i}}.\end{aligned}$$

This suffices for the adversarial bound, but the term $\sqrt{p_{t,j}}$ need not become small as $p_{t,j}$ approaches one. To prove Lemma 2.2, we need a bound that excludes arm j .

We now extend this local-norm bound to loss vectors whose coordinates are at least -1 . This is the lower bound that will hold after shifting the estimates by a loss in $[0, 1]$. Fix $0 < \eta \leq 1/2$, an interior $p \in \Delta_K$, and $g \in [-1, \infty)^K$. We will prove that

$$\langle p - q, g \rangle - \frac{1}{\eta} D_{\psi_{1/2}}(q, p) \leq 2\eta \sum_{i=1}^K p_i^{3/2} g_i^2,$$

for every $q \in \Delta_K$.

We now repeat the positive-orthant comparison from Lecture 7. Extend $\psi_{1/2}(p) = 2(1 - \sum_{i=1}^K \sqrt{p_i})$ to the nonnegative orthant. Expanding the definition gives

$$D_{\psi_{1/2}}(q, p) = \sum_{i=1}^K \left(2\sqrt{p_i} - 2\sqrt{q_i} + \frac{q_i - p_i}{\sqrt{p_i}} \right) = \sum_{i=1}^K \frac{(\sqrt{q_i} - \sqrt{p_i})^2}{\sqrt{p_i}},$$

for every $p \in \mathbb{R}_{>0}^K \triangleq \{p \in \mathbb{R}^K, p_i > 0, \forall i \in [K]\}$ and $q \in \mathbb{R}_+^K \triangleq \{q \in \mathbb{R}^K, q_i \geq 0, \forall i \in [K]\}$. The FTRL minimizers are interior because the slope of $-2\sqrt{x}$ tends to $-\infty$ at zero. As in Lecture 7, denote by

$$u \triangleq \arg \min_{v \in \mathbb{R}_+^K} \left\{ \langle v, g \rangle + \frac{1}{\eta} D_{\psi_{1/2}}(v, p) \right\}$$

the minimizer over the nonnegative orthant. Its first-order condition is $\nabla \psi_{1/2}(u) = \nabla \psi_{1/2}(p) - \eta g$ by applying the Lagrangian form. Since the i -th gradient coordinate is $-1/\sqrt{p_i}$, this gives $1/\sqrt{u_i} = 1/\sqrt{p_i} + \eta g_i$, and hence

$$u_i = \frac{p_i}{(1 + \eta\sqrt{p_i}g_i)^2}.$$

Here $1 + \eta\sqrt{p_i}g_i \geq 1 - \eta\sqrt{p_i} \geq 1/2$, so this formula defines a positive u . To evaluate the objective at u , add the two Bregman divergences. The function values $\psi_{1/2}(p)$ and $\psi_{1/2}(u)$ cancel, giving

$$D_{\psi_{1/2}}(p, u) + D_{\psi_{1/2}}(u, p) = \langle \nabla \psi_{1/2}(p), p - u \rangle - \langle \nabla \psi_{1/2}(u), p - u \rangle = \eta \langle p - u, g \rangle.$$

Therefore the same comparison with u as in Lecture 7 gives

$$\begin{aligned} \langle p - q, g \rangle - \frac{1}{\eta} D_{\psi_{1/2}}(q, p) &\leq \langle p - u, g \rangle - \frac{1}{\eta} D_{\psi_{1/2}}(u, p) = \frac{D_{\psi_{1/2}}(p, u)}{\eta} \\ &= \frac{1}{\eta} \sum_{i=1}^K \frac{(\sqrt{p_i} - \sqrt{u_i})^2}{\sqrt{u_i}} = \eta \sum_{i=1}^K \frac{p_i^{3/2} g_i^2}{1 + \eta\sqrt{p_i}g_i} \\ &\leq 2\eta \sum_{i=1}^K p_i^{3/2} g_i^2. \end{aligned}$$

The first inequality uses the minimizing property of u , and the first equality follows by rearranging the identity just proved. The second equality expands the Bregman divergence using the formula above. For the third equality, substitute $\sqrt{u_i} = \sqrt{p_i}/(1 + \eta\sqrt{p_i}g_i)$ into the divergence formula; the numerator satisfies $\sqrt{p_i} - \sqrt{u_i} = \eta p_i g_i / (1 + \eta\sqrt{p_i}g_i)$. The last inequality uses the denominator bound $1 + \eta\sqrt{p_i}g_i \geq 1/2$. For nonnegative g , the denominators are at least one, which recovers the coefficient η from Lecture 7 for every $\eta > 0$.

The key step is now to apply a constant shift to $\hat{\ell}_t$. Denote by $g_t \triangleq \hat{\ell}_t - \ell_{t,j} \mathbf{1}$ the shifted estimates. Subtracting this common scalar leaves the inner product unchanged because $p_t - p_{t+1}(\eta_t)$ has coordinate sum zero. This shift is only for the analysis; the algorithm does not need to observe $\ell_{t,j}$. Since $g_{t,i} \geq -1$ and $\eta_t = 1/(2\sqrt{t}) \leq 1/2$, the bound applies on every round. Substituting $p = p_t$, $q = p_{t+1}(\eta_t)$, $g = g_t$, and $\eta = \eta_t$ gives

$$S_t = \langle p_t - p_{t+1}(\eta_t), g_t \rangle - \frac{1}{\eta_t} D_{\psi_{1/2}}(p_{t+1}(\eta_t), p_t) \leq 2\eta_t \sum_{i=1}^K p_{t,i}^{3/2} g_{t,i}^2 = 2\eta_t \sum_{i=1}^K p_{t,i}^{3/2} (\hat{\ell}_{t,i} - \ell_{t,j})^2. \quad (5)$$

It remains to see what the shift gains. Unbiasedness and the second moment of the importance-weighted estimates give

$$\begin{aligned} \mathbb{E}_t[g_{t,i}^2] &= \mathbb{E}_t[\hat{\ell}_{t,i}^2] - 2\ell_{t,j} \mathbb{E}_t[\hat{\ell}_{t,i}] + \ell_{t,j}^2 \\ &= \frac{\ell_{t,i}^2}{p_{t,i}} - 2\ell_{t,i}\ell_{t,j} + \ell_{t,j}^2 \\ &= \frac{1 - p_{t,i}}{p_{t,i}} \ell_{t,i}^2 + (\ell_{t,i} - \ell_{t,j})^2. \end{aligned}$$

For $i \neq j$, this is at most $1/p_{t,i}$ since losses lie in $[0, 1]$. For $i = j$, the squared difference vanishes and the bound improves to $(1 - p_{t,j})/p_{t,j}$. Consequently, equation (5) gives

$$\mathbb{E}_t[S_t] \leq \mathcal{O} \left(\eta_t \left[\sqrt{p_{t,j}}(1 - p_{t,j}) + \sum_{i \neq j} \sqrt{p_{t,i}} \right] \right) \leq \mathcal{O} \left(\eta_t \sum_{i \neq j} \sqrt{p_{t,i}} \right),$$

where the last inequality uses $\sqrt{p_{t,j}}(1 - p_{t,j}) \leq 1 - p_{t,j} = \sum_{i \neq j} p_{t,i} \leq \sum_{i \neq j} \sqrt{p_{t,i}}$. Combining this stability bound with the regularization bound in equation (4), taking expectations, and using unbiasedness proves equation (1). $\square$

3 First-Order Regret with Probability Clipping

The previous section used stochastic gaps to improve the worst-case guarantee. We now return to fixed oblivious losses $\ell_1, \dots, \ell_T \in [0, 1]^K$ and ask for an improvement when some arm has small cumulative loss. Denote by $L_T^* \triangleq \min_{j \in [K]} \sum_{t=1}^T \ell_{t,j}$ the best arm's cumulative loss. A *first-order*, or *small-loss*, bound has a leading term proportional to $\sqrt{KL_T^*}$, up to logarithmic factors. We have already seen this type of adaptive bounds in Homework 1 for full-information expert problem and we will now obtain such a bound for MAB by modifying the sampling rule of EXP3 and reusing its entropy analysis from Lecture 6.

Specifically, in Lecture 6, we established that the second-order analysis of EXP3 gives

$$\overline{\text{Reg}}_T \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^K \ell_{t,i}^2.$$

The inverse sampling probabilities cancel in expectation, but the quadratic term sums losses over all arms. Even if the best arm has zero loss, the other arms can contribute on every round. To obtain a small-loss bound from this inequality, we need to relate this term to the loss of a single comparator.

The GREEN idea (Allenberg et al., 2006) is to freeze arms whose exponential-weights probabilities fall below a threshold and renormalize on the remaining arms. The name GREEN refers to the saying *the grass is always greener on the other side*, because unchosen arms receive zero estimated loss. We use a variant with threshold η , so that the same parameter controls the weight updates and freezing. This rule bounds each estimated-loss increment and, as we will show, keeps the cumulative estimates of different arms close.

3.1 The clipped distribution and the estimation bias

Fix $0 < \eta \leq 1/(4K)$. As in EXP3, form the exponential-weights distribution q_t from the cumulative estimates. Denote by $A_t \triangleq \{i \in [K] : q_{t,i} \geq \eta\}$ the active set and by $F_t \triangleq \sum_{i \in [K] \setminus A_t} q_{t,i}$ the frozen mass. We sample from

$$\begin{aligned} q_{t,i} &\triangleq \frac{e^{-\eta \hat{L}_{t-1,i}}}{\sum_{j=1}^K e^{-\eta \hat{L}_{t-1,j}}}, \\ p_{t,i} &\triangleq \frac{q_{t,i} \mathbf{1}\{i \in A_t\}}{1 - F_t}. \end{aligned} \tag{6}$$

Since $F_t < K\eta \leq 1/4$, the active set is nonempty and p_t is a distribution. Every positive sampling probability is at least η . A frozen arm can become active again later, since its cumulative estimate stops increasing while other arms may accumulate losses.

Algorithm 2: GREEN-Style Clipped EXP3 with a Fixed Learning Rate

Input: arms $[K]$, learning rate $0 < \eta \leq 1/(4K)$

Initialize $\widehat{L}_{0,i} = 0$ for every $i \in [K]$;

for $t = 1, 2, \dots$ **do**

 compute q_t , A_t , F_t , and p_t using equation (6);

 draw $I_t \sim p_t$ and observe ℓ_{t,I_t} ;

 set $\widehat{\ell}_{t,i} = (\ell_{t,I_t}/p_{t,I_t})\mathbf{1}\{I_t = i\}$ for every $i \in [K]$;

 set $\widehat{L}_{t,i} = \widehat{L}_{t-1,i} + \widehat{\ell}_{t,i}$ for every $i \in [K]$;

As before, $\mathbb{E}_t$ conditions on the past and the current loss vector, before drawing I_t . The estimator now satisfies

$$\begin{aligned}\mathbb{E}_t[\widehat{\ell}_{t,i}] &= \begin{cases} p_{t,i} \frac{\ell_{t,i}}{p_{t,i}} = \ell_{t,i}, & i \in A_t, \\ 0, & i \notin A_t, \end{cases} \\ &= \ell_{t,i} \mathbf{1}\{i \in A_t\} \leq \ell_{t,i}.\end{aligned}$$

Thus freezing introduces a downward bias, which helps when comparing against a fixed arm. Ordinary EXP3 also assigns zero estimates to unchosen arms, but its rare importance-weighted updates make each coordinate unbiased. A frozen arm receives a zero estimate with certainty on that round. Intuitively, clipping stops increasing an arm's cumulative estimate once the arm falls sufficiently far behind a comparator. While the arm is active, its exponential weight bounds that gap; its last positive update can increase the gap by at most $1/\eta$. After that update, its estimate stays fixed while the comparator's estimate cannot decrease. The following lemma makes this argument precise.

Lemma 3.1. Fix $K \geq 2$ and $0 < \eta \leq 1/(4K)$. For losses in $[0, 1]^K$, the estimates generated by Algorithm 2 satisfy the following bound for every pair of arms $i, j \in [K]$ and every $t \geq 0$:

$$\widehat{L}_{t,i} \leq \widehat{L}_{t,j} + \frac{\ln(1/\eta) + 1}{\eta}. \quad (7)$$

Proof. Fix $i, j \in [K]$ and $t \geq 0$. If arm i has received no positive estimate by round t , then $\widehat{L}_{t,i} = 0$ and the claim follows from $\widehat{L}_{t,j} \geq 0$. Otherwise, denote by $s \leq t$ the last round with $\widehat{\ell}_{s,i} > 0$. Its cumulative estimate stays fixed after that round, so $\widehat{L}_{t,i} = \widehat{L}_{s,i}$. Arm i was active on round s , and hence

$$\begin{aligned}\eta &\leq q_{s,i} = \frac{e^{-\eta \widehat{L}_{s-1,i}}}{\sum_{a=1}^K e^{-\eta \widehat{L}_{s-1,a}}} \leq e^{-\eta(\widehat{L}_{s-1,i} - \widehat{L}_{s-1,j})}, \\ \widehat{L}_{t,i} = \widehat{L}_{s,i} &\leq \widehat{L}_{s-1,j} + \frac{\ln(1/\eta)}{\eta} + \frac{1}{\eta} \leq \widehat{L}_{t,j} + \frac{\ln(1/\eta) + 1}{\eta}.\end{aligned}$$

Taking logarithms in the first line gives $\widehat{L}_{s-1,i} \leq \widehat{L}_{s-1,j} + \ln(1/\eta)/\eta$, which bounds the gap just before the last update. The second line adds at most $1/\eta$ for that update, since $\widehat{\ell}_{s,i} \leq 1/p_{s,i} \leq 1/\eta$, and uses $\widehat{L}_{s-1,j} \leq \widehat{L}_{t,j}$ because estimates are nonnegative. $\square$

Summing Lemma 3.1 over arms controls $\sum_{i=1}^K \widehat{L}_{T,i}$ by $K\widehat{L}_{T,j}$ plus an additive term depending only on K and η . The proof below bounds the quadratic sum $\sum_{t=1}^T \sum_{i=1}^K q_{t,i} \widehat{\ell}_{t,i}^2$ by this sum of cumulative estimates. For a best arm j , the downward bias gives $\mathbb{E}[\widehat{L}_{T,j}] \leq L_T^*$. We now combine this relation with the entropy bound from Lecture 6 to obtain the regret guarantee.

Theorem 3.2. *Fix $K \geq 2$ and $0 < \eta \leq 1/(4K)$. For every fixed oblivious loss sequence in $[0, 1]^K$, Algorithm 2 guarantees that*

$$\overline{\text{Reg}}_T \leq \frac{2K\eta L_T^* + \frac{\ln K}{\eta} + K \left(\ln \frac{1}{\eta} + 1 \right)}{1 - K\eta}.$$

Proof. In Lecture 6, we proved the following pathwise entropy bound for arbitrary finite nonnegative estimates. The exponential-weights distribution here is q_t , so that bound gives

$$\sum_{t=1}^T \left\langle q_t, \widehat{\ell}_t \right\rangle - \widehat{L}_{T,j} \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{i=1}^K q_{t,i} \widehat{\ell}_{t,i}^2, \quad (8)$$

for every fixed arm $j \in [K]$. This is a deterministic inequality for the realized estimates. We will use the downward bias when we take expectations.

We first account for the change in sampling distribution. Since the selected arm is active, $q_{t,I_t}/p_{t,I_t} = 1 - F_t$. Therefore,

$$\begin{aligned} \left\langle q_t, \widehat{\ell}_t \right\rangle &= \frac{q_{t,I_t}}{p_{t,I_t}} \ell_{t,I_t} = (1 - F_t) \ell_{t,I_t} \geq (1 - K\eta) \ell_{t,I_t}, \\ \sum_{i=1}^K q_{t,i} \widehat{\ell}_{t,i}^2 &= \frac{q_{t,I_t}}{p_{t,I_t}} \ell_{t,I_t} \widehat{\ell}_{t,I_t} \leq \widehat{\ell}_{t,I_t} = \sum_{i=1}^K \widehat{\ell}_{t,i}. \end{aligned}$$

The second line uses $\ell_{t,I_t} \leq 1$ and $1 - F_t \leq 1$. Denote by $C_T \triangleq \sum_{t=1}^T \ell_{t,I_t}$ the learner's observed cumulative loss. Substituting these two bounds into equation (8) gives

$$(1 - K\eta)C_T - \widehat{L}_{T,j} \leq \frac{\ln K}{\eta} + \eta \sum_{i=1}^K \widehat{L}_{T,i}. \quad (9)$$

Summing Lemma 3.1 over $i \in [K]$ and substituting into equation (9) gives

$$(1 - K\eta)C_T \leq (1 + K\eta)\widehat{L}_{T,j} + \frac{\ln K}{\eta} + K \left(\ln \frac{1}{\eta} + 1 \right).$$

Take expectations and use the downward bias $\mathbb{E}[\widehat{L}_{T,j}] \leq \sum_{t=1}^T \ell_{t,j}$. Choose an arm minimizing cumulative loss and substitute $\mathbb{E}[C_T] = L_T^* + \overline{\text{Reg}}_T$. Rearranging proves the theorem. $\square$

If L_T^* were known, the choice $\eta = \min\{1/(4K), \sqrt{\ln K/(K(L_T^* + 1))}\}$ would give

$$\overline{\text{Reg}}_T \leq \mathcal{O} \left(\sqrt{K(L_T^* + 1) \ln K} + K \ln(K(L_T^* + 1)) \right).$$

This displays the desired small-loss dependence, but the learning-rate choice depends on the unknown L_T^* . An additional doubling trick is needed to tune the algorithm using only observed feedback. We leave the design and analysis of this tuning rule as a homework question.

References

Chamy Allenberg, Peter Auer, László Györfi, and György Ottucsák. Hannan consistency in on-line learning in case of unbounded losses under partial monitoring. In *Proceedings of the 17th International Conference on Algorithmic Learning Theory (ALT)*, 2006.