

Lecture 7: Adversarial Bandits, Minimax Bounds, and Tsallis-INF

Lecturer: *Mengxiao Zhang*

Week 7

1 How Good Is EXP3?

In Lecture 6, we obtained an adversarial bandit algorithm by feeding importance-weighted loss estimates to Hedge. The estimates could be arbitrarily large, but the weighted second moment in Hedge's regret bound canceled their inverse sampling probabilities. With $\eta = \sqrt{\ln K / (KT)}$, EXP3 achieved $\mathbb{E}[\text{Reg}_T] \leq 2\sqrt{KT \ln K}$ for every fixed oblivious loss sequence. Compared to the regret bound of $\mathcal{O}(\sqrt{T \ln K})$ in the K -expert problem where the losses of each expert are all available, here we pay for an additional $\sqrt{K}$ factor.

This guarantee leaves the following question: does observing only one arm require the dependence on K to be much larger than in the full-information problem? To answer this question, we first prove a lower bound that applies to every bandit algorithm. We then return to the FTRL framework from Lecture 3, and choose a different regularizer to match that bound.

What is the smallest-in-order worst-case regret achievable with bandit feedback?

1.1 The setting and the minimax value

We use the notation and loss convention from Lecture 6. There are $K \geq 2$ arms and $T \geq K$ rounds. An oblivious environment selects a fixed loss sequence $\ell_1, \dots, \ell_T \in [0, 1]^K$ before the interaction, independently of the learner's randomization. On round t , the learner selects $p_t \in \Delta_K$, draws $I_t \sim p_t$, and observes only ℓ_{t, I_t} . As in Lecture 6, we consider the pseudo-regret

$$\overline{\text{Reg}}_T \triangleq \mathbb{E} \left[\sum_{t=1}^T \ell_{t, I_t} \right] - \min_{i \in [K]} \mathbb{E} \left[\sum_{t=1}^T \ell_{t, i} \right]. \quad (1)$$

For a fixed loss sequence, the expectation is over the learner's randomization, and pseudo-regret equals the expected hindsight regret from Lecture 6. When we use a stochastic loss sequence in the lower-bound construction, the expectation also includes the randomness of the losses. Define

$$V_{T, K}^{\text{bandit}} \triangleq \inf_{\mathcal{A}} \sup_{\ell_1, \dots, \ell_T \in [0, 1]^K} \overline{\text{Reg}}_T(\mathcal{A}, \{\ell_t\}_{t \in [T]}), \quad (2)$$

to be the minimax pseudo-regret. Specifically, the supremum is over the oblivious environments just described. The algorithm may know K and T , but it selects its action using only previous actions, observed losses, and its internal randomization. EXP3 gives an upper bound through its fixed-sequence guarantee. A lower bound must show that every algorithm has some fixed loss sequence on which its pseudo-regret is large.

2 A Minimax Lower Bound

To prove the lower bound, we construct a stochastic environment with one unknown arm having a smaller mean loss. Its entire loss matrix can be sampled before the interaction. We first bound its stochastic pseudo-regret and then extract a fixed loss sequence. This construction should remind you of the full-information construction in Lecture 2. However, that construction cannot give a positive stochastic pseudo-regret lower bound. In that construction, every arm has independent Bernoulli losses with mean $1/2$. The learner and every fixed arm therefore have expected cumulative loss $T/2$, so the pseudo-regret is exactly zero. The hindsight-regret lower bound comes from the random fluctuations of the best realized arm. To capture the additional cost of learning with bandit feedback, we instead give one unknown arm a strictly smaller mean and show that some choice of this arm forces large pseudo-regret.

2.1 A hidden improvement

Similar to the expert case, we consider a stochastic loss sequence. Imagine first that every arm produces independent Bernoulli losses with mean $1/2$. An algorithm has only T observations to distribute among K arms, so the average arm is sampled T/K times. Now improve one unknown arm by changing its mean to $1/2 - \varepsilon$. Each pull of another arm costs ε in expectation, but the learner obtains evidence about the improvement only when it samples the improved arm.

Which value of ε makes this environment hardest? For a small gap, the learner can ignore the improvement and use an arbitrary arm, incurring at most $T\varepsilon$ pseudo-regret. For a large gap, learning the improved arm becomes worthwhile. For $0 < \varepsilon \leq 1/4$, distinguishing Bernoulli means $1/2$ and $1/2 - \varepsilon$ with a constant error probability requires $\Theta(1/\varepsilon^2)$ observations. Since the improved arm is unknown, searching across the other $K - 1$ arms suggests a cost of $(K - 1)/\varepsilon$ in regret: each suboptimal observation costs ε . This scaling heuristic describes the identification cost; it does not specify a complete identification strategy. The learner prefers the smaller of these two costs, so the hardest scale balances them. This gives $T\varepsilon = (K - 1)/\varepsilon$, hence $\varepsilon = \sqrt{(K - 1)/T}$ and a regret order of $\sqrt{(K - 1)T}$. In the proof, we choose a smaller constant $\varepsilon = \frac{1}{8}\sqrt{K/T}$ to keep the history distributions sufficiently close. The following theorem makes this intuition rigorous.

Theorem 2.1. *For every $K \geq 2$, $T \geq K$, and randomized bandit algorithm $\mathcal{A}$, there exists a stochastic environment with independent Bernoulli losses such that $\overline{\text{Reg}}_T \geq \frac{1}{32}\sqrt{KT}$. Consequently, $V_{T,K}^{\text{bandit}} \geq \frac{1}{32}\sqrt{KT}$.*

To prove the theorem, we need to compare the distributions of the learner's *observed history* under two environments.

2.2 Information in the observed history

Before proving the theorem, we introduce two quantities for comparing history distributions. For distributions $P, Q \in \Delta(\mathcal{H})$ on a finite set $\mathcal{H}$, denote their *relative entropy* and *total variation* distance by $\text{KL}(P\|Q) \triangleq \sum_{h \in \mathcal{H}} P(h) \ln \frac{P(h)}{Q(h)}$ and $\text{TV}(P, Q) \triangleq \frac{1}{2} \sum_{h \in \mathcal{H}} |P(h) - Q(h)|$. Terms with $P(h) = 0$ contribute zero to the relative entropy, and the relative entropy is infinite if $P(h) > 0 = Q(h)$. In fact, relative entropy is the Bregman divergence of the negative-entropy regularizer from Lecture 3, and Pinsker's inequality then gives that $\text{TV}(P, Q) \leq \sqrt{\text{KL}(P\|Q)/2}$.¹ If a statistic Z

¹We in fact show this inequality when we prove that negative entropy is 1-strongly convex in ℓ_1 norm.

takes values in $[0, T]$, we therefore have

$$|\mathbb{E}_P[Z] - \mathbb{E}_Q[Z]| \leq T \cdot \text{TV}(P, Q) \leq T \sqrt{\frac{\text{KL}(P\|Q)}{2}}. \quad (3)$$

Denote by $H_t \triangleq (I_1, \ell_{1,I_1}, \dots, I_t, \ell_{t,I_t})$ the observed history, whose generated sigma-algebras $\mathcal{F}_t \triangleq \sigma(H_t)$ form the observation filtration, and denote by $n_{T,i} \triangleq \sum_{t=1}^T \mathbf{1}\{I_t = i\}$ the total number of pulls of arm i . The following decomposition identifies exactly where the learner acquires information.

Lemma 2.2. *Consider two stochastic bandit environments in which losses are independent across rounds and arms. Arm i has a distribution ν_i in the first environment and ν'_i in the second, each with finite support. Fix the same possibly randomized bandit algorithm in both environments. Denote by P, P' the induced distributions of H_T . Then we have*

$$\text{KL}(P\|P') = \sum_{i=1}^K \mathbb{E}_P[n_{T,i}] \text{KL}(\nu_i\|\nu'_i). \quad (4)$$

Proof. Denote by $\pi_t(i \mid h_{t-1})$ the algorithm's conditional probability of choosing arm i after history h_{t-1} . Denote by $\mathcal{H}_t$ the finite set of possible histories of length t , with h_0 the empty history. For a realized history $h_T = (i_1, \ell_{1,i_1}, \dots, i_T, \ell_{T,i_T})$, denote by h_{t-1} its prefix through round $t-1$. Independence of the fresh losses and the probability chain rule give $P(h_T) = \prod_{t=1}^T \pi_t(i_t \mid h_{t-1}) \nu_{i_t}(\ell_{t,i_t})$, and the same product with ν'_{i_t} gives $P'(h_T)$. Denote by $\mathcal{S}_i$ the union of the supports of ν_i and ν'_i , and denote by $P(h)$ the marginal probability of a prefix $h \in \mathcal{H}_{t-1}$. Given $h_{t-1} = h$, the joint probability of $I_t = i$ and $\ell_{t,I_t} = \ell$ is $\pi_t(i \mid h) \nu_i(\ell)$. Then, we have

$$\begin{aligned} \text{KL}(P\|P') &= \sum_{h_T \in \mathcal{H}_T} P(h_T) \ln \frac{P(h_T)}{P'(h_T)} = \sum_{h_T \in \mathcal{H}_T} P(h_T) \ln \prod_{t=1}^T \frac{\nu_{i_t}(\ell_{t,i_t})}{\nu'_{i_t}(\ell_{t,i_t})} \\ &= \sum_{t=1}^T \sum_{h_T \in \mathcal{H}_T} P(h_T) \ln \frac{\nu_{i_t}(\ell_{t,i_t})}{\nu'_{i_t}(\ell_{t,i_t})} = \sum_{t=1}^T \mathbb{E}_P \left[\ln \frac{\nu_{I_t}(\ell_{t,I_t})}{\nu'_{I_t}(\ell_{t,I_t})} \right] \\ &= \sum_{t=1}^T \mathbb{E}_P \left[\mathbb{E}_P \left[\ln \frac{\nu_{I_t}(\ell_{t,I_t})}{\nu'_{I_t}(\ell_{t,I_t})} \middle| h_{t-1} \right] \right] \quad (\text{tower property}) \\ &= \sum_{t=1}^T \mathbb{E}_P \left[\sum_{i=1}^K \pi_t(i \mid h_{t-1}) \sum_{\ell \in \mathcal{S}_i} \nu_i(\ell) \ln \frac{\nu_i(\ell)}{\nu'_i(\ell)} \right] \quad (\text{conditional law}) \\ &= \sum_{t=1}^T \sum_{h \in \mathcal{H}_{t-1}} P(h) \sum_{i=1}^K \pi_t(i \mid h) \sum_{\ell \in \mathcal{S}_i} \nu_i(\ell) \ln \frac{\nu_i(\ell)}{\nu'_i(\ell)} \quad (\text{history average}) \\ &= \sum_{t=1}^T \sum_{h \in \mathcal{H}_{t-1}} P(h) \sum_{i=1}^K \pi_t(i \mid h) \text{KL}(\nu_i\|\nu'_i) \quad (\text{definition of KL}) \\ &= \sum_{i=1}^K \sum_{t=1}^T \mathbb{P}_P(I_t = i) \text{KL}(\nu_i\|\nu'_i) \quad (\text{total probability}) \\ &= \sum_{i=1}^K \mathbb{E}_P[n_{T,i}] \text{KL}(\nu_i\|\nu'_i). \quad (\text{pull counts}) \end{aligned}$$

The calculation applies when the likelihood ratio is finite on reachable histories. If some reachable observation has positive probability under P but zero probability under P' , both sides of equation (4) are infinite, with $0 \cdot \infty = 0$ for unpulled arms. $\square$

2.3 Proof of the lower bound

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1. Fix an arbitrary algorithm $\mathcal{A}$ and choose $\varepsilon = \frac{1}{8}\sqrt{K/T}$. Since $T \geq K$, we have $0 < \varepsilon \leq 1/8$. Denote by σ_0 the environment in which every loss $\ell_{t,j}$ is an independent $\text{Ber}(1/2)$ random variable. For each $i \in [K]$, denote by σ_i the environment in which arm i has independent $\text{Ber}(1/2 - \varepsilon)$ losses, while all other arms still have independent $\text{Ber}(1/2)$ losses. Denote by P_i the distribution of H_T induced by $\mathcal{A}$ under σ_i , and denote by $\mathbb{E}_i \triangleq \mathbb{E}_{\sigma_i}$ expectation over both this environment and the learner's randomization, for every $i \in \{0, \dots, K\}$. These are randomized oblivious environments: their entire loss sequence $\ell_1, \dots, \ell_T$ is drawn before the learner begins.

The mean-optimal arm in environment i is arm i . Denote the pseudo-regret in this environment by $\overline{\text{Reg}}_T^{(i)}$. As in Lecture 5, we express the pseudo-regret in terms of the gaps:

$$\overline{\text{Reg}}_T^{(i)} = \mathbb{E}_i \left[\sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i}) \right] = \varepsilon (T - \mathbb{E}_i[n_{T,i}]). \quad (5)$$

The equality uses independence of the fresh losses from the learner's past and current action. Each pull of an arm other than i contributes the mean gap ε . It remains to show that the learner misses the improved arm for a constant fraction of the horizon in at least one of these environments.

By Lemma 2.2, only observations of arm i distinguish P_0 from P_i . Moreover, for $0 \leq \varepsilon \leq 1/4$, the Bernoulli probabilities give

$$\begin{aligned} \text{KL}(\text{Ber}(1/2) \parallel \text{Ber}(1/2 - \varepsilon)) &= \frac{1}{2} \ln \frac{1/2}{1/2 - \varepsilon} + \frac{1}{2} \ln \frac{1/2}{1/2 + \varepsilon} \\ &= \frac{1}{2} \ln \frac{1}{(1 - 2\varepsilon)(1 + 2\varepsilon)} = -\frac{1}{2} \ln(1 - 4\varepsilon^2) \leq 4\varepsilon^2. \end{aligned}$$

The inequality uses $-\ln(1 - u) \leq u/(1 - u) \leq 2u$ for $0 \leq u \leq 1/2$, with $u = 4\varepsilon^2$. Consequently,

$$\text{KL}(P_0 \parallel P_i) \leq 4\varepsilon^2 \mathbb{E}_0[n_{T,i}].$$

Apply equation (3) to $n_{T,i} \in [0, T]$. This gives

$$\mathbb{E}_i[n_{T,i}] \leq \mathbb{E}_0[n_{T,i}] + T\varepsilon \sqrt{2\mathbb{E}_0[n_{T,i}]}.$$

Although these counts depend on the observations, their expectations under ν_0 sum to T . Averaging the preceding inequality over i and applying Cauchy–Schwarz therefore yields

$$\frac{1}{K} \sum_{i=1}^K \mathbb{E}_i[n_{T,i}] \leq \frac{T}{K} + T\varepsilon \sqrt{\frac{2T}{K}} \leq \frac{3T}{4}.$$

Here $K \geq 2$ bounds the first term by $T/2$, and the chosen ε bounds the second by $\sqrt{2T}/8$. Substituting into equation (5), we obtain

$$\frac{1}{K} \sum_{i=1}^K \overline{\text{Reg}}_T^{(i)} \geq \frac{\varepsilon T}{4} = \frac{1}{32} \sqrt{KT}.$$

At least one choice of the improved arm therefore gives a stochastic environment with pseudo-regret at least $\frac{1}{32}\sqrt{KT}$. For this choice of i , denote by $\mathbb{E}_{\mathcal{A}}$ expectation over the learner's randomization with the loss sequence fixed. Comparing with the fixed arm i and then averaging the loss sequence sampled from ν_i gives

$$\begin{aligned} \sup_{\ell_1, \dots, \ell_T \in [0,1]^K} \overline{\text{Reg}}_T(\mathcal{A}, \{\ell_t\}_{t \in [T]}) &\geq \sup_{\ell_1, \dots, \ell_T \in [0,1]^K} \mathbb{E}_{\mathcal{A}} \left[\sum_{t=1}^T (\ell_{t, I_t} - \ell_{t, i}) \right] \\ &\geq \mathbb{E}_{\ell_1, \dots, \ell_T \sim \nu_i} \left[\mathbb{E}_{\mathcal{A}} \left[\sum_{t=1}^T (\ell_{t, I_t} - \ell_{t, i}) \right] \right] \\ &= \mathbb{E}_i \left[\sum_{t=1}^T (\ell_{t, I_t} - \ell_{t, i}) \right] \\ &= \overline{\text{Reg}}_T^{(i)} \geq \frac{1}{32}\sqrt{KT}. \end{aligned}$$

Since $\mathcal{A}$ is arbitrary, taking the infimum over $\mathcal{A}$ gives $V_{T,K}^{\text{bandit}} \geq \frac{1}{32}\sqrt{KT}$. $\square$

For $T \geq K$, the lower bound and EXP3's guarantee place the minimax value between a multiple of $\sqrt{KT}$ and $2\sqrt{KT \ln K}$. Thus the dependence on $\sqrt{T}$ is settled, but the factor $\sqrt{\ln K}$ is still unresolved. Is the factor $\sqrt{\ln K}$ fundamental to MAB, or does there exist an algorithm with regret $\mathcal{O}(\sqrt{KT})$? We now return to the FTRL framework and ask how changing the regularizer can close this gap.

3 Tsallis-INF

In Lecture 6, we introduced EXP3, which is FTRL with the negative Shannon entropy regularizer from Lecture 3. Its $\ln K$ regularizer range and K expected quadratic term together give the $\sqrt{K \ln K}$ factor. To remove it, we consider a different regularizer, which is *negative Tsallis entropy*.

3.1 Tsallis entropy

Specifically, for a parameter $\beta \in (0, 1)$, define the negative Tsallis entropy regularizer by

$$\psi_{\beta}(p) \triangleq \frac{1 - \sum_{i=1}^K p_i^{\beta}}{1 - \beta}, \quad (6)$$

for $p \in \Delta_K$, with $0^{\beta} = 0$. A direct calculation shows that each term $-p_i^{\beta}$ is convex, so this is a convex regularizer. Its gradient and Hessian on the positive orthant are

$$\frac{\partial \psi_{\beta}(p)}{\partial p_i} = -\frac{\beta}{1 - \beta} p_i^{\beta-1}, \quad \nabla^2 \psi_{\beta}(p) = \beta \text{diag}(p_i^{\beta-2}),$$

for every $i \in [K]$.

This family also connects to the preceding algorithm. Since $\sum_i p_i = 1$, taking β to 1 in equation (6) gives $\sum_i p_i \ln p_i$, with the usual convention $0 \ln 0 = 0$. Hence Shannon entropy appears as a limiting case of Tsallis entropy.

In order to apply the FTRL guarantee to Tsallis entropy, we need to show that it is 1-strongly convex with respect to certain norm $\|\cdot\|$. In Lecture 3, we established that negative Shannon entropy is 1-strongly convex with respect to $\|\cdot\|_1$. This is a fixed norm: it measures a displacement in the same way at every distribution. Tsallis entropy also has global strong convexity. Since $p_i \leq 1$ and $\beta - 2 < 0$, its Hessian satisfies $\nabla^2 \psi_\beta(p) \succeq \beta I$, where I is the identity matrix. Integrating this bound along a segment shows that it is β -strongly convex in $\|\cdot\|_2$, or equivalently 1-strongly convex in $\sqrt{\beta}\|\cdot\|_2$, with extension to the boundary by continuity. However, these fixed-norm bounds discard how the curvature varies with the sampling probabilities.

To account for how the curvature varies with p , we use the Hessian at the current distribution to measure a displacement. For a twice differentiable convex regularizer ψ with positive definite Hessian, denote its Bregman divergence by $D_\psi(q, p) \triangleq \psi(q) - \psi(p) - \langle \nabla \psi(p), q - p \rangle$. Taylor's formula gives $D_\psi(p + \delta x, p) = \delta^2 x^\top \nabla^2 \psi(p) x / 2 + o(\delta^2)$ for directions x along which the perturbation is feasible. Thus the Hessian defines a *local norm* in which the curvature at p is one. Unlike global strong convexity, this pointwise statement alone does not bound a finite displacement; we must also control the Hessian along the segment.

For negative Shannon entropy, the Hessian is $\text{diag}(1/p_i)$, so its squared local norm is $\sum_i x_i^2 / p_i$ and its squared dual norm is $\sum_i p_i x_i^2$. In Lecture 6, we used exactly this weighted quadratic term to analyze EXP3; in Lecture 3, we used the fixed dual norm $\|\cdot\|_\infty$. For Tsallis entropy, at an interior distribution p , denote the squared local norm by $\|x\|_p^2 \triangleq x^\top \nabla^2 \psi_\beta(p) x$ and its squared dual norm by $\|x\|_{*,p}^2 \triangleq x^\top (\nabla^2 \psi_\beta(p))^{-1} x$, for $x \in \mathbb{R}^K$. The Hessian above gives

$$\|x\|_{*,p}^2 = \frac{1}{\beta} \sum_{i=1}^K p_i^{2-\beta} x_i^2. \quad (7)$$

The geometry now depends on p ; a fixed ℓ_1 or ℓ_2 norm does not capture this dependence. This dependence on p enters the stability analysis below.

3.2 The algorithm

Denote the cumulative estimated losses by $\hat{L}_t \triangleq \sum_{s=1}^t \hat{\ell}_s$, with $\hat{L}_0 = 0$. For a fixed learning rate $\eta > 0$, Tsallis FTRL chooses

$$p_t = \arg \min_{p \in \Delta_K} \left\{ \langle p, \hat{L}_{t-1} \rangle + \frac{1}{\eta} \psi_\beta(p) \right\}. \quad (8)$$

As in EXP3, the learner draws an arm from this distribution, observes its loss, and forms the importance-weighted estimate. Only the map from cumulative estimates to the next distribution changes. We keep $\beta \in (0, 1)$ general and call the algorithm β -Tsallis-INF, where INF stands for *Implicitly Normalized Forecaster*. The name Tsallis-INF without a parameter usually refers to the

choice $\beta = 1/2$.²

Algorithm 1: β -Tsallis-INF with Importance-Weighted Loss Estimates

Input: arms $[K]$, positive learning rate $\eta > 0$, $\beta \in (0, 1)$

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

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

choose $p_t = \arg \min_{p \in \Delta_K} \{ \langle p, \widehat{L}_{t-1} \rangle + \psi_\beta(p)/\eta \}$;

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

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

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

Computation of p_t . Unlike EXP3, the update generally requires solving a scalar equation to normalize the probabilities. Strict convexity gives a unique minimizer, and the derivative of the regularizer tends to $-\infty$ at a zero coordinate, so the minimizer is interior. Specifically, the first-order condition for equation (8) gives

$$\widehat{L}_{t-1,i} + \lambda_t = \frac{\beta}{\eta(1-\beta)} p_{t,i}^{\beta-1}, \quad p_{t,i} = \left(\frac{\beta}{\eta(1-\beta)(\widehat{L}_{t-1,i} + \lambda_t)} \right)^{1/(1-\beta)},$$

for every $i \in [K]$, where λ_t is determined by $\sum_{i=1}^K (\beta/(\eta(1-\beta)(\widehat{L}_{t-1,i} + \lambda_t)))^{1/(1-\beta)} = 1$. On $\lambda_t > -\min_i \widehat{L}_{t-1,i}$, the left-hand side is continuous and strictly decreasing from infinity to zero, so this equation has a unique solution, which we can approximate by bisection. This implicit determination of λ_t is the *implicit normalization* in the name INF.

3.3 Regret analysis

In Lecture 3, we decomposed regret into regularization bias and stability. The bias is the comparator's regularization cost, bounded by B_{ψ_β}/η . A smaller η makes the decisions more stable but increases this bias. We controlled stability using global strong convexity in a fixed norm and combined the two terms to obtain the FTRL regret bound. Here we use the *local norms* introduced above instead: they measure curvature at p_t and account for its dependence on the current distribution. Including the Bregman remainder gives a per-round stability contribution bounded by $\eta \left\| \widehat{\ell}_t \right\|_{*,p_t}^2 / 2$.

The regularizer is maximized at a vertex of the simplex, where it is zero, and minimized at the uniform distribution. Its range is therefore

$$B_{\psi_\beta} \triangleq \max_{p \in \Delta_K} \psi_\beta(p) - \min_{p \in \Delta_K} \psi_\beta(p) = \frac{K^{1-\beta} - 1}{1-\beta}. \quad (9)$$

Compared with EXP3 in Lecture 6, the importance-weighted estimator remains unbiased; the bias here comes from regularization, not loss estimation. Shannon entropy gives regularization bias $\ln K/\eta$ and expected stability at most $\eta K T$ in that analysis. For $\beta = 1/2$, Tsallis entropy has range $2(\sqrt{K} - 1)$, while its dual local norm uses $p_i^{3/2}$ instead of the p_i weights. The expected quadratic

²See [Zimmert and Seldin \(2021\)](#), Section 1, page 3.

term is therefore proportional to $\sum_i \sqrt{p_i}$, which is at most $\sqrt{K}$. We establish the resulting range-stability trade-off after proving the local stability bound. Denote by e_i the i th standard basis vector.

Theorem 3.1. *Fix $\beta \in (0, 1)$ and $\eta > 0$. Algorithm 1 guarantees that*

$$\sum_{t=1}^T \langle p_t - e_i, \hat{\ell}_t \rangle \leq \frac{K^{1-\beta} - 1}{\eta(1-\beta)} + \frac{\eta}{2\beta} \sum_{t=1}^T \sum_{j=1}^K p_{t,j}^{2-\beta} \hat{\ell}_{t,j}^2, \quad (10)$$

for every $i \in [K]$.

Proof. Denote by $\Phi_t \triangleq \min_{p \in \Delta_K} \{ \langle p, \hat{L}_t \rangle + \psi_\beta(p)/\eta \}$ the regularized minimum, and by p_{T+1} the final FTRL minimizer. First, p_t is interior, meaning $p_{t,i} > 0$ for every $i \in [K]$: moving mass $\delta > 0$ from a positive coordinate to a zero coordinate changes the objective by $-\delta^\beta/(\eta(1-\beta)) + \mathcal{O}(\delta) < 0$ for sufficiently small δ , so a boundary point cannot minimize it. Denote the Lagrangian by $\mathcal{L}_t(p, \lambda, \mu) \triangleq \langle p, \hat{L}_{t-1} \rangle + \psi_\beta(p)/\eta + \lambda(\sum_i p_i - 1) - \sum_i \mu_i p_i$, where $\mu_i \geq 0$ are the multipliers for $p_i \geq 0$. Complementary slackness gives $\mu_{t,i} p_{t,i} = 0$, hence $\mu_{t,i} = 0$. Stationarity therefore gives $\hat{L}_{t-1} + \nabla \psi_\beta(p_t)/\eta + \lambda_t \mathbf{1} = 0$. Since both p_t and p_{t+1} have unit mass, taking the inner product with $p_{t+1} - p_t$ gives $\langle \hat{L}_{t-1} + \nabla \psi_\beta(p_t)/\eta, p_{t+1} - p_t \rangle = 0$. Using $\hat{L}_t = \hat{L}_{t-1} + \hat{\ell}_t$ and the definition of Bregman divergence, we obtain

$$\begin{aligned} \Phi_t - \Phi_{t-1} &= \langle p_{t+1}, \hat{\ell}_t \rangle + \langle p_{t+1} - p_t, \hat{L}_{t-1} \rangle + \frac{\psi_\beta(p_{t+1}) - \psi_\beta(p_t)}{\eta} \\ &= \langle p_{t+1}, \hat{\ell}_t \rangle + \frac{\psi_\beta(p_{t+1}) - \psi_\beta(p_t) - \langle \nabla \psi_\beta(p_t), p_{t+1} - p_t \rangle}{\eta} \\ &= \langle p_{t+1}, \hat{\ell}_t \rangle + \frac{1}{\eta} D_{\psi_\beta}(p_{t+1}, p_t). \end{aligned}$$

Summing this identity and using $\Phi_T \leq \langle e_i, \hat{L}_T \rangle + \psi_\beta(e_i)/\eta$ and $\Phi_0 = \psi_\beta(p_1)/\eta$ gives a curvature refinement of the decomposition we used in Lecture 3:

$$\sum_{t=1}^T \langle p_t - e_i, \hat{\ell}_t \rangle \leq \frac{B_{\psi_\beta}}{\eta} + \sum_{t=1}^T \left(\langle p_t - p_{t+1}, \hat{\ell}_t \rangle - \frac{1}{\eta} D_{\psi_\beta}(p_{t+1}, p_t) \right).$$

Extend ψ_β to the nonnegative orthant by the same formula and define

$$q_t \triangleq \arg \min_{q \in \mathbb{R}_+^K} \left\{ \langle q, \hat{\ell}_t \rangle + \frac{1}{\eta} D_{\psi_\beta}(q, p_t) \right\}.$$

The first-order condition gives

$$\nabla \psi_\beta(q_t) = \nabla \psi_\beta(p_t) - \eta \hat{\ell}_t, \quad q_{t,j}^{\beta-1} = p_{t,j}^{\beta-1} + \frac{1-\beta}{\beta} \eta \hat{\ell}_{t,j},$$

for every $j \in [K]$. This equation defines a positive q_t , and strict convexity verifies that it is the minimizer. Since $\hat{\ell}_{t,j} \geq 0$ and $\beta - 1 < 0$, we have $q_{t,j} \leq p_{t,j}$. Thus each coordinate along the segment from p_t to q_t is at most the corresponding coordinate of p_t . Since $\beta - 2 < 0$, we obtain

$$\nabla^2 \psi_\beta((1-s)p_t + sq_t) = \beta \operatorname{diag} \left(((1-s)p_{t,j} + sq_{t,j})^{\beta-2} \right) \succeq \nabla^2 \psi_\beta(p_t),$$

for every $s \in [0, 1]$. Taylor's integral formula therefore gives

$$D_{\psi_\beta}(q_t, p_t) \geq \frac{1}{2} \|p_t - q_t\|_{p_t}^2.$$

Applying the dual-norm inequality we used in Lecture 3 and completing the square yield

$$\begin{aligned} \left\langle p_t - p_{t+1}, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_{\psi_\beta}(p_{t+1}, p_t) &\leq \left\langle p_t - q_t, \widehat{\ell}_t \right\rangle - \frac{1}{\eta} D_{\psi_\beta}(q_t, p_t) \\ &\leq \|p_t - q_t\|_{p_t} \left\| \widehat{\ell}_t \right\|_{*, p_t} - \frac{1}{2\eta} \|p_t - q_t\|_{p_t}^2 \leq \frac{\eta}{2} \left\| \widehat{\ell}_t \right\|_{*, p_t}^2. \end{aligned}$$

Substituting this bound, equation (9), and equation (7) into the decomposition proves equation (10). $\square$

As in Lecture 6, we take expectations to transfer this fixed-comparator bound to pseudo-regret. Denote by $\mathbb{E}_t$ conditional expectation given the past and the fixed oblivious loss sequence, before drawing I_t .

Theorem 3.2. Fix $\beta \in (0, 1)$ and $\eta > 0$. For every fixed oblivious loss sequence $\ell_1, \dots, \ell_T \in [0, 1]^K$, Algorithm 1 guarantees that

$$\overline{\text{Reg}}_T \leq \frac{K^{1-\beta} - 1}{\eta(1-\beta)} + \frac{\eta T K^\beta}{2\beta}. \quad (11)$$

In particular, Algorithm 1 with $\beta = 1/2$ and $\eta = 1/\sqrt{T}$ guarantees that $\overline{\text{Reg}}_T \leq 3\sqrt{KT}$.

Proof. Fix a comparator arm i and apply equation (10). As we showed in Lecture 6, the linear terms are unbiased, while

$$\mathbb{E}_t \left[\sum_{j=1}^K p_{t,j}^{2-\beta} \widehat{\ell}_{t,j}^2 \right] = \sum_{j=1}^K p_{t,j}^{1-\beta} \ell_{t,j}^2 \leq \sum_{j=1}^K p_{t,j}^{1-\beta} \leq K^\beta.$$

The last inequality follows from concavity of $x^{1-\beta}$. Taking expectations and choosing an arm minimizing expected cumulative loss gives equation (11). Substituting $\beta = 1/2$ and $\eta = 1/\sqrt{T}$ gives $\overline{\text{Reg}}_T \leq 2(\sqrt{K} - 1)\sqrt{T} + \sqrt{KT} \leq 3\sqrt{KT}$. $\square$

For $T \geq K$, this matches the lower bound in Theorem 2.1, giving

$$\frac{1}{32} \sqrt{KT} \leq V_{T,K}^{\text{bandit}} \leq 3\sqrt{KT}. \quad (12)$$

Thus the minimax pseudo-regret is $\Theta(\sqrt{KT})$ in this regime. The factor $\sqrt{\ln K}$ in EXP3 is not essential with bandit feedback.

In addition, for a fixed $\beta \in (0, 1)$, choosing

$$\eta = \sqrt{\frac{2\beta(K^{1-\beta} - 1)}{(1-\beta)TK^\beta}}$$

in equation (11) also gives $\overline{\text{Reg}}_T \leq \sqrt{2KT/(\beta(1-\beta))}$. Thus any fixed β away from the endpoints gives the same rate; $\beta = 1/2$ yields the simple choice $\eta = 1/\sqrt{T}$.

4 Full Information and Bandit Feedback

The lower and upper bounds now give two complementary explanations of the feedback cost. The lower bound shows that finding a hidden improvement requires observations of the improved arm. The upper bound shows how an online learner can use one observed coordinate to reconstruct an unbiased estimator and achieve the minimax optimal regret. We also show that it is essential to choose the correct regularizer to achieve the tightest bound. The following table compares regularizer range and expected squared local norm for the same importance-weighted estimator.

Regularizer	Range B_ψ	$\mathbb{E}_t[\ \widehat{\ell}_t\ _{*,p_t}^2]$
Shannon entropy	$\ln K$	at most K
Tsallis entropy, $\beta = 1/2$	$2(\sqrt{K} - 1)$	at most $2\sqrt{K}$

Balancing the range against the expected squared local norm gives $\sqrt{KT \ln K}$ for Shannon entropy and $\sqrt{KT}$ for Tsallis entropy. The feedback and estimator are the same and the regularizer changes how the estimate affects the next distribution. In the homework, we will study another useful regularizer, the *log-barrier regularizer*. It arises as $\beta \downarrow 0$ after rescaling Tsallis entropy and removing a constant independent of p :

$$\psi_{\log}(p) \triangleq - \sum_{i=1}^K \ln p_i = \lim_{\beta \downarrow 0} \frac{\psi_\beta(p) + (K-1)/(1-\beta)}{\beta},$$

for every interior $p \in \Delta_K$.

To close this topic, we compare the guarantee of Tsallis-INF in the full-information case. With full information, we feed the true bounded vector ℓ_t to FTRL directly, without importance weighting. The stochastic cancellation calculation is no longer needed. For $\psi_{1/2}$, the quadratic term satisfies

$$\sum_{i=1}^K p_{t,i}^{3/2} \ell_{t,i}^2 \leq \sum_{i=1}^K p_{t,i}^{3/2} \leq 1.$$

Thus Theorem 3.1 gives the full-information bound $\text{Reg}_T \leq 2(\sqrt{K} - 1)/\eta + \eta T$ where Reg_T denotes the usual full-information regret. Choosing $\eta = \sqrt{2(\sqrt{K} - 1)/T}$ yields

$$\text{Reg}_T \leq 2\sqrt{2(\sqrt{K} - 1)T} = \mathcal{O}(K^{1/4}\sqrt{T}).$$

By comparison, in Lecture 2, we obtained $\mathcal{O}(\sqrt{T \ln K})$ for Shannon entropy. This shows that the bandit improvement of Tsallis-INF does not transfer into a better full-information guarantee through the same analysis. A regularizer's usefulness depends on the loss vectors or estimates entering its stability term.