

Lecture 6: Adversarial Multi-Armed Bandits

Lecturer: *Mengxiao Zhang*

Week 6

1 Why Adversarial Bandits?

In Lecture 5, we studied stochastic bandits. Each arm had a fixed loss distribution, and observations collected at different times were samples from the same population. However, many sequential decisions do not remain stationary. The performance of an advertisement may change after a competitor launches a campaign. Demand at a posted price may move with seasonality, inventory, or the decisions of other sellers. A routing option may deteriorate during congestion and recover later. In these settings, a single time-invariant distribution may be a poor description of the losses. We therefore ask what can be learned when the loss sequence is allowed to vary arbitrarily over time.

The word *adversarial* describes the absence of a stochastic model. It does not necessarily mean that a malicious person chooses every loss; rather, it means that we impose no distributional assumption on the environment, and a guarantee in this setting must hold for every admissible loss sequence. This returns us to the arbitrary losses of Lecture 2, Section 1, but changes the **feedback**: the learner observes only the loss of the selected arm, rather than the whole vector. Its decision must therefore serve two purposes, earning a small loss and collecting information about the alternatives.

Can the learner compete with the best fixed arm in hindsight while observing only one coordinate of each loss vector?

2 The Adversarial Multi-Armed Bandit Problem

There are $K \geq 2$ arms indexed by $[K] \triangleq \{1, \dots, K\}$ and a horizon of $T \geq 1$ rounds. Before the interaction begins, an *oblivious* environment selects loss vectors $\ell_1, \dots, \ell_T \in [0, 1]^K$.¹ No stationarity, independence across rounds, or fixed arm means are assumed.

At each round $t = 1, \dots, T$,

1. using the past actions and observed losses, the learner selects a probability distribution $p_t \in \Delta_K$, where $\Delta_K \triangleq \{p \in \mathbb{R}^K : \sum_{i=1}^K p_i = 1, p_i \geq 0 \text{ for every } i \in [K]\}$;
2. the learner draws an arm $I_t \sim p_t$;
3. the learner incurs and observes only ℓ_{t, I_t} .

The learner does not observe $\ell_{t, i}$ for $i \neq I_t$. Denote by $p_{t, i}$ the probability assigned to arm i .

¹An oblivious environment fixes the entire loss sequence before the interaction, independently of the learner's random choices. A more general, adaptive environment may choose ℓ_t using the learner's previous actions and observations, but must do so before seeing the current random draw I_t . Our expected hindsight-regret guarantees below concern fixed oblivious sequences. The distinction matters because an adaptive environment can make the hindsight comparator itself depend on the learner's history.

2.1 Regret and pseudo-regret

The comparator is the best single arm for the entire sequence, chosen in hindsight. As in Lecture 2, Section 1, we compare cumulative losses, except that the learner now incurs the loss of a sampled arm:

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

The comparator must use the same arm throughout. Competing with the best arm separately on every round would be a different and generally impossible objective without additional assumptions.

Recall the definition of pseudo-regret in Lecture 5, Section 2.2. The analogous definition here is

$$\overline{\text{Reg}}_T \triangleq \mathbb{E} \left[\sum_{t=1}^T \ell_{t, I_t} \right] - \min_{i \in [K]} \mathbb{E} \left[\sum_{t=1}^T \ell_{t, i} \right]. \quad (2)$$

For a fixed oblivious loss sequence, the comparator's losses are deterministic, so $\overline{\text{Reg}}_T = \mathbb{E}[\text{Reg}_T]$. The expectation is only over the learner's randomization. If the environment also randomizes, the expectation includes that randomness and generally $\overline{\text{Reg}}_T \leq \mathbb{E}[\text{Reg}_T]$, exactly as in the previously studied stochastic setting.

Unlike the stochastic setting, the benchmark does not generally simplify to $T\mu^*$ for fixed, time-independent arm means. Consequently, the fixed-gap counting identity in Lecture 5, Lemma 2.1 does not apply to an arbitrary adversarial sequence. We must compare cumulative losses directly.

2.2 Why do the stochastic algorithms not suffice?

Arm elimination and UCB were designed to estimate fixed means. With arbitrary losses, an accurate average of previously observed losses says nothing by itself about that arm's future losses. An arm that looks poor early may become the best arm over the full horizon. Permanent elimination can then prevent the learner from ever discovering the change. While UCB keeps every arm available, its old confidence intervals still need not describe any current or future performance.

There is also a precise reason that the versions in Lecture 5 cannot have an adversarial guarantee. With fixed tie-breaking and a fixed sampling order, both algorithms are *deterministic functions* of their observed histories. The construction in Lecture 2, Proposition 2.3 applies here as well. Before play begins, simulate the algorithm recursively and set $\ell_{t, I_t} = 1$ and $\ell_{t, j} = 0$ for $j \neq I_t$ on each simulated round. During actual play the same history occurs, so the learner incurs loss T , while some fixed arm incurs at most T/K . Thus $\text{Reg}_T \geq T(1 - 1/K)$ on a fixed oblivious sequence. Having less feedback does not invalidate that construction. Randomization is therefore necessary for an adversarial bandit guarantee.

2.3 General Approach: Loss Estimation from Bandit Feedback

If we observe only one coordinate and the losses may change arbitrarily between rounds, how can we obtain information about the other coordinates? In fact, exact recovery of the whole current vector is impossible from one observed coordinate. But do we need exact recovery to use a full-information algorithm?

Our goal is an expected-regret bound. This suggests a weaker requirement: construct a vector which recovers the true loss vector *in expectation*. We can then feed this vector to an algorithm such as Hedge and try to transfer its guarantee back to the original losses.

Suppose $p_{t,i} > 0$ for every arm. A first attempt would be to keep the observed loss and put zero in every unobserved coordinate. But $\mathbb{E}_t[\ell_{t,i}\mathbf{1}\{I_t = i\}] = p_{t,i}\ell_{t,i}$, so this systematically underestimates a rarely selected arm's loss. The correction is to divide by the probability that the coordinate is observed. Define the *importance-weighted loss estimator* by

$$\widehat{\ell}_{t,i} \triangleq \frac{\ell_{t,i}}{p_{t,i}}\mathbf{1}\{I_t = i\}, \quad i \in [K]. \quad (3)$$

It is also called an *inverse-propensity-weighted* loss estimator in the literature. Every coordinate is computable: the selected coordinate uses the observed loss, and all unselected coordinates are zero. Denote by $\mathbb{E}_t[\cdot]$ conditional expectation given the history before the draw of I_t , including p_t . For the fixed oblivious sequence, ℓ_t is also fixed under this conditioning. Thus

$$\mathbb{E}_t[\widehat{\ell}_{t,i}] = p_{t,i}\frac{\ell_{t,i}}{p_{t,i}} + (1 - p_{t,i})0 = \ell_{t,i}. \quad (4)$$

Adaptive choices of p_t do not break this identity: after conditioning on the past, the current draw still has the known distribution p_t . This differs from the empirical mean at an adaptively chosen sample count in Lecture 5, where the bias comes from deciding when to stop. Here we correct the current observation probability rather than averaging a randomly selected number of past samples.

The estimator also satisfies the pathwise identity

$$\langle p_t, \widehat{\ell}_t \rangle = \sum_{j=1}^K p_{t,j} \frac{\ell_{t,j}}{p_{t,j}} \mathbf{1}\{I_t = j\} = \ell_{t,I_t}. \quad (5)$$

For any fixed comparator arm i , conditional unbiasedness and the tower property give

$$\mathbb{E} \left[\sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i}) \right] = \mathbb{E} \left[\sum_{t=1}^T \langle p_t - e_i, \widehat{\ell}_t \rangle \right], \quad (6)$$

where e_i denotes the i th coordinate vector. Thus it suffices to control the regret on the estimated losses against each fixed arm. For an oblivious sequence, its best arm is fixed before the learner randomizes, so we may select that arm as the comparator after establishing this identity. We are not interchanging an expectation with a minimum over the random estimates.

2.4 Applying Hedge to the estimated losses

Can we now apply the Hedge guarantee? Lecture 2, Theorem 3.2 gives a second-order bound for losses in $[0, 1]^K$, followed by the simpler worst-case bound $\ln K/\eta + \eta T$. The second-order bound is the part we need. Our estimates are nonnegative, but they need not be at most one. We must therefore check the proof rather than invoke the bounded-loss statement without checking its assumptions.

The proof of Lecture 2, Theorem 3.2 uses $e^{-x} \leq 1 - x + x^2$ for $x \geq 0$. This inequality does not require an upper bound on x . The assumption that losses are at most one enters only when the quadratic term is subsequently bounded by T . Keeping that term yields the following extension.

Lemma 2.1 (Hedge with nonnegative loss estimates). *Suppose $w_{1,j} = 1$, $p_{t,j} = w_{t,j} / \sum_{k=1}^K w_{t,k}$, and $w_{t+1,j} = w_{t,j} \exp(-\eta \hat{\ell}_{t,j})$ for finite, nonnegative vectors $\hat{\ell}_t$ and $\eta > 0$. For every comparator arm $i \in [K]$,*

$$\sum_{t=1}^T \langle p_t - e_i, \hat{\ell}_t \rangle \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{j=1}^K p_{t,j} \hat{\ell}_{t,j}^2. \quad (7)$$

The vectors may depend on the previous history and on the current sampled action; this inequality holds for each realized sequence.

Proof. We use exactly the potential from the proof of Lecture 2, Theorem 3.2, with actual losses replaced by estimated losses. Denote $\hat{L}_{t,j} \triangleq \sum_{s=1}^t \hat{\ell}_{s,j}$ and $\Phi_t \triangleq \eta^{-1} \ln(\sum_{j=1}^K \exp(-\eta \hat{L}_{t,j}))$, with $\hat{L}_{0,j} = 0$. The weight update gives

$$\Phi_t - \Phi_{t-1} = \frac{1}{\eta} \ln \left(\sum_{j=1}^K p_{t,j} e^{-\eta \hat{\ell}_{t,j}} \right) \leq -\langle p_t, \hat{\ell}_t \rangle + \eta \sum_{j=1}^K p_{t,j} \hat{\ell}_{t,j}^2.$$

Here we used $e^{-x} \leq 1 - x + x^2$ for nonnegative x and $\ln(1 + u) \leq u$ for $u > -1$. Summing over t telescopes the potential. Since $\Phi_0 = \ln K / \eta$ and $\Phi_T \geq -\hat{L}_{T,i}$ for every arm i , rearranging proves (7). $\square$

The index i in (7) identifies the comparator and the quadratic term sums over *all* arms j . Now we can see the difficulty. A small $p_{t,j}$ can produce a large estimate $\ell_{t,j}/p_{t,j}$. How should we control the scale of the estimates, and which part of their variability does the regret proof actually need?

3 A Naive Reduction with Explicit Exploration

A direct idea is to prevent any sampling probability from becoming too small. Let $q_t \in \Delta_K$ denote the distribution proposed by Hedge, and fix $\gamma \in (0, 1]$. Define the distribution used to sample an arm by

$$p_t \triangleq (1 - \gamma)q_t + \frac{\gamma}{K} \mathbf{1}. \quad (8)$$

With probability $1 - \gamma$, follow Hedge's recommendation and with probability γ , explore uniformly. Since $p_{t,j} \geq \gamma/K$, the estimate satisfies $0 \leq \hat{\ell}_{t,j} \leq K/\gamma$. We can therefore bound every squared estimate by K^2/γ^2 in Lemma 2.1.

Algorithm 1: Naive Bandit Reduction with Uniform Exploration

Input: arms $[K]$, horizon T , exploration parameter $\gamma \in (0, 1]$, learning rate $\eta > 0$

Initialize Hedge with weights $w_{1,i} = 1$;

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

obtain Hedge's distribution $q_{t,i} = w_{t,i} / \sum_{j=1}^K w_{t,j}$;
 set $p_t = (1 - \gamma)q_t + (\gamma/K)\mathbf{1}$;
 draw $I_t \sim p_t$ and observe ℓ_{t,I_t} ;
 form $\hat{\ell}_{t,i} = (\ell_{t,i}/p_{t,i})\mathbf{1}\{I_t = i\}$ for every $i \in [K]$;
 update $w_{t+1,i} = w_{t,i} \exp(-\eta \hat{\ell}_{t,i})$;

How much does this exploration cost? We must account for the difference between p_t and q_t , as well as the effect of the large estimates on Hedge's quadratic term.

Theorem 3.1. *For every fixed oblivious loss sequence, Algorithm 1 satisfies*

$$\mathbb{E}[\text{Reg}_T] \leq \gamma T + \frac{\ln K}{\eta} + \eta \frac{K^2 T}{\gamma^2}. \quad (9)$$

If $T \geq 4K^2 \ln K$, choose $\gamma = \sqrt{2K}(\ln K/T)^{1/4}$ and $\eta = (\gamma/K)\sqrt{\ln K/T}$. Then $\mathbb{E}[\text{Reg}_T] \leq 2\sqrt{2K} T^{3/4}(\ln K)^{1/4}$.

Proof. Fix an arm i . Lemma 2.1 applies to Hedge's distribution q_t , even though the estimates are generated by sampling from p_t . Since $\sum_j q_{t,j} = 1$ and $\widehat{\ell}_{t,j}^2 \leq K^2/\gamma^2$, it gives the pathwise bound

$$\sum_{t=1}^T \langle q_t - e_i, \widehat{\ell}_t \rangle \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{j=1}^K q_{t,j} \widehat{\ell}_{t,j}^2 \leq \frac{\ln K}{\eta} + \frac{\eta K^2 T}{\gamma^2}.$$

Conditional unbiasedness applies to q_t as well as p_t , because both distributions are determined before I_t is sampled. Therefore,

$$\mathbb{E} \left[\sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i}) \right] = \mathbb{E} \left[\sum_{t=1}^T \langle p_t - q_t, \ell_t \rangle \right] + \mathbb{E} \left[\sum_{t=1}^T \langle q_t - e_i, \widehat{\ell}_t \rangle \right] \leq \gamma T + \frac{\ln K}{\eta} + \frac{\eta K^2 T}{\gamma^2}.$$

Here $\langle p_t - q_t, \ell_t \rangle = \gamma \langle \mathbf{1}/K - q_t, \ell_t \rangle \leq \gamma$, since every coordinate of ℓ_t lies in $[0, 1]$. Taking i to be the best arm of the fixed sequence proves (9). The stated learning rate balances the last two terms, whose sum is $2K\sqrt{T \ln K}/\gamma$. The stated exploration parameter then balances this sum with γT and belongs to $(0, 1]$ under the assumed lower bound on T . $\square$

Is $T^{3/4}$ the best we can achieve, and is this the price of having bandit feedback? We have encountered a similar question before. Explore-then-commit gives a $T^{2/3}$ bound for stochastic MAB (Lecture 5, Theorem 3.1 and equation (6)), while arm elimination gives a $\sqrt{T}$ bound up to logarithms (Lecture 5, Theorem 4.6). Thus a suboptimal exploration argument need not describe the difficulty of the problem itself. The analogous improvement here comes from returning to the weighted quadratic term in Lemma 2.1.

4 EXP3

What happens if we remove the uniform mixture and simply sample from Hedge's own distribution? Then the probability in the quadratic term of Lemma 2.1 is exactly the probability in the estimator's denominator. This matching is what the range-based argument failed to use.

The key is that we only seek expected regret. We do not need a deterministic bound on every squared estimate; we only need to bound the expectation of their weighted sum. Using the conditional expectation $\mathbb{E}_t$ defined above and $\mathbb{E}_t[\mathbf{1}\{I_t = j\}] = p_{t,j}$, we have

$$\mathbb{E}_t \left[\sum_{j=1}^K p_{t,j} \widehat{\ell}_{t,j}^2 \right] = \sum_{j=1}^K p_{t,j} \frac{\ell_{t,j}^2}{p_{t,j}^2} \mathbb{E}_t[\mathbf{1}\{I_t = j\}] = \sum_{j=1}^K \ell_{t,j}^2 \leq K. \quad (10)$$

One factor of $p_{t,j}$ comes from Hedge's quadratic term, and another comes from the probability of observing arm j . Together they cancel the squared denominator. Individual estimates can have large variance, yet the *weighted second moment that appears in the regret bound* is controlled. We do not need to impose a uniform probability floor for this expected-regret analysis.

4.1 The algorithm

This gives EXP3, short for *Exponential-weight algorithm for Exploration and Exploitation*: run Hedge on the importance-weighted loss estimates and sample from its distribution. Initialize $w_{1,i} = 1$ for every arm. Its sampling and update rules are

$$p_{t,i} \triangleq \frac{w_{t,i}}{\sum_{j=1}^K w_{t,j}}, \quad w_{t+1,i} \triangleq w_{t,i} \exp(-\eta \hat{\ell}_{t,i}). \quad (11)$$

Every weight remains strictly positive, so the estimates are well defined at every finite round.

Algorithm 2: EXP3

Input: arms $[K]$, horizon T , learning rate $\eta > 0$

Initialize $w_{1,i} = 1$ for every $i \in [K]$;

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

set $p_{t,i} = w_{t,i} / \sum_{j=1}^K w_{t,j}$ for every $i \in [K]$;
draw $I_t \sim p_t$ and observe ℓ_{t,I_t} ;
set $\hat{\ell}_{t,i} = (\ell_{t,i} / p_{t,i}) \mathbf{1}\{I_t = i\}$ for every $i \in [K]$;
set $w_{t+1,i} = w_{t,i} \exp(-\eta \hat{\ell}_{t,i})$ for every $i \in [K]$;

4.2 Regret guarantee

Theorem 4.1. *For every fixed sequence $\ell_1, \dots, \ell_T \in [0, 1]^K$, EXP3 with any $\eta > 0$ satisfies*

$$\mathbb{E}[\text{Reg}_T] \leq \frac{\ln K}{\eta} + \eta K T. \quad (12)$$

The choice $\eta = \sqrt{\ln K / (KT)}$ gives

$$\mathbb{E}[\text{Reg}_T] \leq 2\sqrt{KT \ln K}. \quad (13)$$

Proof. Fix a comparator arm i . Apply Lemma 2.1 to the estimates generated by EXP3 and take expectations. Equations (6) and (10) give

$$\mathbb{E} \left[\sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i}) \right] \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \mathbb{E} \left[\mathbb{E}_t \left[\sum_{j=1}^K p_{t,j} \hat{\ell}_{t,j}^2 \right] \right] \leq \frac{\ln K}{\eta} + \eta K T.$$

For the fixed oblivious sequence, choose an arm minimizing $\sum_t \ell_{t,i}$. This proves (12). Substituting the stated learning rate balances its two terms and proves (13). $\square$

The proof now has two separate ingredients. The proof of Lecture 2, Theorem 3.2 gives a pathwise inequality for the estimated losses. Importance weighting then transfers the linear term to actual expected regret and controls the weighted quadratic term. We did not need to reconstruct each unobserved loss accurately on each round.

4.3 Where does exploration come from?

EXP3 has no separate uniform-exploration parameter in this loss-based formulation. Nevertheless, every arm retains positive sampling probability. When an arm is selected and incurs positive loss, its weight decreases. The weights of the unselected arms do not change because their estimated losses are zero, so their relative probabilities increase. Randomized sampling and the loss update continue to create opportunities to observe other arms.

This does not give a deterministic lower bound on every probability: a probability may become very small. The guarantee comes from the cancellation in (10), rather than from a claim that every estimate has small variance. The exact quantity preserved in the full-information proof is therefore essential. This is also the lesson of Lecture 2, Question 4.4: two bounds with the same worst-case order can behave very differently when their second-order terms are used in a new setting.

Remark 4.2 (Expectation versus high probability). Theorem 4.1 bounds expected regret for each fixed oblivious sequence. The proof does not give a *high-probability* bound of the same order, because rare observations can still create large estimates! We will see later (either in the lecture or in the homework) that EXP3 indeed can have a very large variance and may suffer $\Omega(T)$ regret with a constant probability on suitable loss sequences.

4.4 What If We Observe Rewards Instead of Losses?

Suppose the feedback is a reward $r_{t,I_t} \in [0, 1]$. The simplest reduction is to define $\ell_{t,i} \triangleq 1 - r_{t,i} \in [0, 1]$. This transformation preserves regret exactly, so the loss-based EXP3 analysis above applies directly; reward feedback by itself does not require explicit exploration.

Instead, let us ask what happens if we use the seemingly equivalent signed representation $\ell_{t,i} \triangleq -r_{t,i}$. The objectives are still equivalent: maximizing cumulative reward is the same as minimizing cumulative negative reward. In this subsection, denote by $\text{Reg}_T \triangleq \max_i \sum_{t=1}^T r_{t,i} - \sum_{t=1}^T r_{t,I_t}$ the reward regret, which equals the loss regret for $\ell_t = -r_t$. However, these losses belong to $[-1, 0]^K$, while our analysis above used nonnegative losses.

Intuitively, with $\ell_{t,i} = -r_{t,i}$, the importance-weighted estimate is $\hat{\ell}_{t,i} = -(r_{t,i}/p_{t,i})\mathbf{1}\{I_t = i\}$. If the selected arm receives positive reward, its weight is multiplied by $\exp(\eta r_{t,I_t}/p_{t,I_t}) > 1$, while all unselected weights remain unchanged. Its probability therefore increases, and the probabilities of the unobserved arms decrease. Receiving positive reward encourages the learner to repeat that arm, even though another arm may give a larger reward. Repeated positive feedback can reinforce the same choice without providing information about the alternatives. This is the opposite of the nonnegative-loss update described in the preceding subsection, where a positive observed loss makes the unobserved arms relatively more attractive. This intuition identifies the exploration concern; by itself, it is not a proof of a regret lower bound.

From a technical perspective, the signed-loss extension in full information is doable: for $\ell_{t,i} \in [-1, 1]$ and $\eta \leq 1$, we have $\eta \ell_{t,i} \geq -1$. The inequality $e^{-x} \leq 1 - x + x^2$ holds for $x \geq -1$, so the same potential argument works. Under bandit feedback, however, a negative estimate can be as small as $-1/p_{t,i}$. There is no guarantee that $\eta \hat{\ell}_{t,i} \geq -1$ when the sampling probability is small! Equivalently, the direct reward update contains a positive exponent $\eta r_{t,i}/p_{t,i}$, and $e^x \leq 1 + x + x^2$ does not hold for arbitrarily large positive x . The expected second-moment cancellation still holds, but we cannot use it until we have justified the preceding potential inequality. Thus Theorem 4.1 does not directly apply to negative losses.

To fix this, we return to Algorithm 1, now using the observed negative reward $\ell_{t,I_t} = -r_{t,I_t}$. Keep

the same mixture $p_t = (1 - \gamma)q_t + (\gamma/K)\mathbf{1}$ and update $w_{t+1,i} = w_{t,i} \exp(-\eta \widehat{\ell}_{t,i})$. Since $p_{t,i} \geq \gamma/K$, choosing $\eta \leq \gamma/K$ ensures $\eta \widehat{\ell}_{t,i} \geq -1$. Explicit exploration now has a precise role: it prevents the positive reward exponent from becoming too large. But do we have to accept the earlier $T^{3/4}$ guarantee? No: that bound used a deterministic range bound on every squared estimate, rather than the expected weighted second moment.

Theorem 4.3 (EXP3 with explicit exploration for rewards). *For every fixed oblivious reward sequence in $[0, 1]^K$, Algorithm 1 applied to losses $\ell_t = -r_t$, with $0 < \gamma \leq 1/2$ and $0 < \eta \leq \gamma/K$, satisfies*

$$\mathbb{E}[\text{Reg}_T] \leq \gamma T + \frac{\ln K}{\eta} + \frac{\eta K T}{1 - \gamma}. \quad (14)$$

Choosing $\gamma = \min\{1/2, \sqrt{K \ln K / T}\}$ and $\eta = \gamma/K$ gives $\mathbb{E}[\text{Reg}_T] = O(\sqrt{KT \ln K})$.

Proof. The condition $\eta \widehat{\ell}_{t,j} \geq -1$ allows the potential proof of Lemma 2.1 to go through with Hedge's distribution q_t , using $e^{-x} \leq 1 - x + x^2$ for $x \geq -1$. Hence, for every fixed arm i ,

$$\sum_{t=1}^T \langle q_t - e_i, \widehat{\ell}_t \rangle \leq \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \sum_{j=1}^K q_{t,j} \widehat{\ell}_{t,j}^2.$$

Unlike the proof of Theorem 3.1, we now take conditional expectations before bounding the quadratic term. The estimator still uses the actual sampling probability $p_{t,j}$, so

$$\mathbb{E}_t \left[\sum_{j=1}^K q_{t,j} \widehat{\ell}_{t,j}^2 \right] = \sum_{j=1}^K \frac{q_{t,j} \ell_{t,j}^2}{p_{t,j}} \leq \frac{1}{1 - \gamma} \sum_{j=1}^K \ell_{t,j}^2 \leq \frac{K}{1 - \gamma}. \quad (15)$$

Here we used $p_{t,j} \geq (1 - \gamma)q_{t,j}$ and $|\ell_{t,j}| \leq 1$. Conditional unbiasedness remains valid for negative losses. Moreover, $\langle p_t - q_t, \ell_t \rangle = \gamma \langle \mathbf{1}/K - q_t, \ell_t \rangle \leq \gamma$, because both averages belong to $[-1, 0]$. The same regret decomposition as in Theorem 3.1 therefore gives

$$\mathbb{E} \left[\sum_{t=1}^T (\ell_{t,I_t} - \ell_{t,i}) \right] \leq \gamma T + \frac{\ln K}{\eta} + \eta \sum_{t=1}^T \mathbb{E} \left[\mathbb{E}_t \left[\sum_{j=1}^K q_{t,j} \widehat{\ell}_{t,j}^2 \right] \right] \leq \gamma T + \frac{\ln K}{\eta} + \frac{\eta K T}{1 - \gamma}.$$

Take i to be the best fixed reward arm. When $T \geq 4K \ln K$, the stated choices give $\gamma T = \ln K / \eta = \sqrt{KT \ln K}$ and $\eta K T / (1 - \gamma) \leq 2\sqrt{KT \ln K}$. For $T < 4K \ln K$, the trivial bound $T < 2\sqrt{KT \ln K}$ suffices. $\square$

The improvement comes from the analysis, not from reconstructing more feedback. The naive proof bounded the quadratic term by $K^2 T / \gamma^2$; equation (15) bounds its expectation by $KT / (1 - \gamma)$. We can therefore choose a smaller exploration probability while keeping the second-order term under control.