

Lecture 4: Optimistic Online Learning and Applications to Games

Lecturer: *Mengxiao Zhang*

September 16th, 2026

1 Optimistic Online Learning

1.1 From worst-case losses to predictable losses

In Lecture 2, we studied prediction with expert advice. On each round, the learner chose a distribution over N experts and then observed their losses. Hedge achieved regret of $\mathcal{O}(\sqrt{T \ln N})$, and the matching minimax lower bound showed that this dependence on T and N is unavoidable in the worst case.

The phrase *in the worst case* matters. The lower bound rules out $o(\sqrt{T \ln N})$ regret uniformly over all loss sequences, but it does not say that every sequence is equally difficult. As discussed in Lecture 1, real losses often contain exploitable structure. In **inventory control**, demand may follow seasonal patterns and remain approximately stationary over shorter periods. In **dynamic pricing**, nearby prices may induce similar demand; in **online regression**, consecutive gradients may be correlated; and in **repeated games**, an opponent may adjust her strategy gradually. In each case, information from previous rounds can help predict the next gradient. A guarantee that depends only on T and N ignores this easier structure. This motivates the following question:

Can an online learner retain its worst-case guarantee while performing substantially better in a predictable environment?

1.2 Optimistic FTRL

The answer is affirmative. We augment the OCO protocol from Lecture 3 with a gradient hint: on round t ,

- the learner receives a predictor $m_t \in \mathbb{R}^d$ for the current linear loss ℓ_t ;
- the learner chooses $w_t \in \Omega$;
- the environment reveals ℓ_t , and the learner incurs $\langle w_t, \ell_t \rangle$.

For now, regard m_t as external information. We do not assume that it is accurate; instead, we seek a regret bound that improves with the realized prediction error. A one-line modification of FTRL achieves this goal: *Optimistic FTRL* adds m_t to the observed cumulative loss $L_{t-1} = \sum_{\tau \leq t-1} \ell_\tau$,

$$w_t = \arg \min_{w \in \Omega} \left\{ \langle L_{t-1} + m_t, w \rangle + \frac{1}{\eta} \psi(w) \right\}. \quad (1)$$

As in Lecture 3, ψ is the regularizer and $\eta > 0$ is the learning rate. For comparison, denote the usual FTRL iterate by

$$\hat{w}_t = \arg \min_{w \in \Omega} \left\{ \langle L_{t-1}, w \rangle + \frac{1}{\eta} \psi(w) \right\}, \quad t \geq 1. \quad (2)$$

The update is optimistic because it treats m_t as the current loss. If $m_t = \ell_t$, then $w_t = \hat{w}_{t+1}$: the learner makes exactly the decision that ordinary FTRL would make after seeing the current loss. The Be-the-Leader argument from Lecture 3 then gives non-positive regret for the augmented losses and, consequently, a regret bound for the original losses that is independent of T .

The hint can exploit structure specific to the application:

- **Inventory control:** use the gradient induced by a seasonal or demand forecast.
- **Dynamic pricing:** use a gradient forecast from the current demand-response model.
- **Online regression:** use the previous gradient, or a forecast fitted to recent gradients.
- **Repeated games:** use the loss induced by the opponent's preceding strategy.

None of these hints is assumed accurate: the bound below charges only the realized errors $\ell_t - m_t$. The following decomposition also retains the stability terms that will later control movement in games.

Lemma 1.1 (Optimistic FTRL decomposition). *For every $u \in \Omega$,*

$$\text{Reg}_T(u) \leq \frac{\psi(u) - \psi(\hat{w}_1)}{\eta} + \sum_{t=1}^T \langle \ell_t - m_t, w_t - \hat{w}_{t+1} \rangle - \frac{1}{2\eta} \sum_{t=1}^T \left(\|w_t - \hat{w}_t\|^2 + \|\hat{w}_{t+1} - w_t\|^2 \right). \quad (3)$$

Proof. Recall the notation from Lecture 3: $F_t(w) = \langle L_t, w \rangle + \psi(w)/\eta$ and $\Phi_t = \min_{w \in \Omega} F_t(w)$. The proof of Lemma 3.3 there used the following consequence of strong convexity: if x minimizes a $(1/\eta)$ -strongly convex function H over Ω , then

$$H(y) \geq H(x) + \frac{1}{2\eta} \|y - x\|^2 \quad \text{for every } y \in \Omega.$$

We apply this inequality twice. Since $\hat{w}_t$ minimizes F_{t-1} ,

$$F_{t-1}(w_t) \geq F_{t-1}(\hat{w}_t) + \frac{1}{2\eta} \|w_t - \hat{w}_t\|^2 = \Phi_{t-1} + \frac{1}{2\eta} \|w_t - \hat{w}_t\|^2.$$

Moreover, w_t minimizes $F_{t-1}(w) + \langle m_t, w \rangle$, so

$$F_{t-1}(\hat{w}_{t+1}) + \langle m_t, \hat{w}_{t+1} \rangle \geq F_{t-1}(w_t) + \langle m_t, w_t \rangle + \frac{1}{2\eta} \|\hat{w}_{t+1} - w_t\|^2.$$

Adding $\langle \ell_t - m_t, \hat{w}_{t+1} \rangle$ to the last inequality and then using the first one gives

$$\Phi_t \geq \Phi_{t-1} + \langle \ell_t, w_t \rangle - \langle \ell_t - m_t, w_t - \hat{w}_{t+1} \rangle + \frac{1}{2\eta} \|w_t - \hat{w}_t\|^2 + \frac{1}{2\eta} \|\hat{w}_{t+1} - w_t\|^2.$$

Rearranging and summing over t yields

$$\sum_{t=1}^T \langle \ell_t, w_t \rangle \leq \Phi_T - \Phi_0 + \sum_{t=1}^T \langle \ell_t - m_t, w_t - \hat{w}_{t+1} \rangle - \frac{1}{2\eta} \sum_{t=1}^T \left(\|w_t - \hat{w}_t\|^2 + \|\hat{w}_{t+1} - w_t\|^2 \right).$$

Since $\Phi_T \leq \langle L_T, u \rangle + \psi(u)/\eta$ and $\Phi_0 = \psi(\hat{w}_1)/\eta$, subtracting $\langle L_T, u \rangle$ from both sides gives (3). $\square$

The prediction-error term measures how well the hints anticipate the losses, while the negative term rewards stable decisions. The following theorem converts this decomposition into a bound that trades cumulative prediction error against the learner's movement; the negative movement term will be essential in the game application.

Theorem 1.2 (Prediction error minus movement). *Suppose that ψ is 1-strongly convex with respect to $\|\cdot\|$ and has range $B_\psi \triangleq \max_{u,v \in \Omega} (\psi(u) - \psi(v))$. Then Optimistic FTRL guarantees*

$$\text{Reg}_T \leq \frac{B_\psi}{\eta} + \eta \sum_{t=1}^T \|\ell_t - m_t\|_*^2 - \frac{1}{4\eta} \sum_{t=2}^T \|w_t - w_{t-1}\|^2. \quad (4)$$

Proof. Write $\delta_t \triangleq \ell_t - m_t$. The two objectives minimized by w_t and $\hat{w}_{t+1}$ differ only by the linear term $\langle \delta_t, w \rangle$. Adding the corresponding strong-convexity inequalities, exactly as in the proof of Lemma 3.3 from Lecture 3, gives

$$\|w_t - \hat{w}_{t+1}\| \leq \eta \|\delta_t\|_*.$$

The dual-norm inequality therefore implies

$$\langle \delta_t, w_t - \hat{w}_{t+1} \rangle \leq \|\delta_t\|_* \|w_t - \hat{w}_{t+1}\| \leq \eta \|\delta_t\|_*^2.$$

Next, pair the two stability terms that meet at $\hat{w}_t$:

$$\sum_{t=1}^T \left(\|w_t - \hat{w}_t\|^2 + \|\hat{w}_{t+1} - w_t\|^2 \right) \geq \sum_{t=2}^T \left(\|w_t - \hat{w}_t\|^2 + \|\hat{w}_t - w_{t-1}\|^2 \right) \geq \frac{1}{2} \sum_{t=2}^T \|w_t - w_{t-1}\|^2.$$

The first inequality omits only the nonnegative terms $\|w_1 - \hat{w}_1\|^2$ and $\|\hat{w}_{T+1} - w_T\|^2$. The second follows from the triangle inequality and $a^2 + b^2 \geq (a+b)^2/2$. Substituting these bounds into (3) and using $\psi(u) - \psi(\hat{w}_1) \leq B_\psi$ proves the theorem. $\square$

If the total prediction error is positive and known in advance, ignoring the favorable movement term (which we will utilize soon) and choosing $\eta = \sqrt{B_\psi / \sum_{t=1}^T \|\ell_t - m_t\|_*^2}$ gives

$$\text{Reg}_T \leq 2 \sqrt{B_\psi \sum_{t=1}^T \|\ell_t - m_t\|_*^2}.$$

The above choice of η is valid since Theorem 1.2 holds for every $\eta > 0$. The regret therefore depends on *cumulative prediction error* rather than directly on T . If the error is $o(T)$, the bound improves on the worst-case $\mathcal{O}(\sqrt{T})$ rate; if it is $\Theta(T)$, the usual rate is recovered. When the losses and hints are uniformly bounded in the dual norm, the cumulative squared prediction error is $\mathcal{O}(T)$, so suitable learning-rate tuning recovers the usual worst-case rate.

The learner can also construct the hints from past observations. The simplest choice, $m_t \triangleq \ell_{t-1}$ with $\ell_0 \triangleq 0$, yields a bound in terms of the squared path length $\sum_{t=1}^T \|\ell_t - \ell_{t-1}\|_*^2$. Thus slowly changing loss sequences are easier even without an external predictor. Unknown prediction error can be handled by a doubling scheme or a time-varying learning rate.

1.3 Examples: Optimistic Hedge and Optimistic Lazy OGD

Optimistic Hedge. Lecture 3 showed that negative-entropy FTRL on $\Delta(N)$ is exactly Hedge. Applying the same regularizer to Optimistic FTRL gives

$$p_{t,i} = \frac{\exp[-\eta(L_{t-1,i} + m_{t,i})]}{\sum_{j=1}^N \exp[-\eta(L_{t-1,j} + m_{t,j})]}, \quad (5)$$

or, equivalently,

$$p_{t,i} = \frac{\hat{p}_{t,i} e^{-\eta m_{t,i}}}{\sum_{j=1}^N \hat{p}_{t,j} e^{-\eta m_{t,j}}},$$

where $\hat{p}_t$ is the ordinary Hedge distribution based on L_{t-1} . Thus Optimistic Hedge first forms the usual cumulative-loss weights and then tilts them by the predicted current loss. Setting $m_t = 0$ recovers ordinary Hedge.

The following guarantee is obtained directly by specializing Theorem 1.2 to the negative-entropy regularizer.

Theorem 1.3 (Optimistic Hedge bound). *For every $\eta > 0$, Optimistic Hedge guarantees*

$$\text{Reg}_T \leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^T \|\ell_t - m_t\|_\infty^2 - \frac{1}{4\eta} \sum_{t=2}^T \|p_t - p_{t-1}\|_1^2.$$

Proof. For the negative-entropy regularizer $\psi(p) = \sum_{i=1}^N p_i \ln p_i$, the associated Bregman divergence is $D_\psi(p, q) = \sum_{i=1}^N p_i \ln(p_i/q_i)$. Pinsker's inequality gives $D_\psi(p, q) \geq \frac{1}{2} \|p - q\|_1^2$, which is exactly 1-strong convexity with respect to $\|\cdot\|_1$. Moreover, ψ ranges from $-\ln N$ at the uniform distribution to 0 at a vertex of $\Delta(N)$, so $B_\psi = \ln N$, and the dual norm is $\|\cdot\|_\infty$. Substituting these facts into Theorem 1.2 gives the result. $\square$

The negative movement term cannot be optimized independently of the other terms because the sequence $p_1, \dots, p_T$ itself depends on η . Indeed, the same stability argument gives, for $t \geq 2$,

$$\|p_t - p_{t-1}\|_1 \leq \eta \|\ell_{t-1} + m_t - m_{t-1}\|_\infty.$$

Thus, for fixed losses and hints, the squared movement is $\mathcal{O}(\eta^2)$, so the magnitude of its contribution after multiplication by $-1/(4\eta)$ is $\mathcal{O}(\eta)$. One therefore cannot decrease η while treating the movement sum as fixed; as $\eta \rightarrow 0$, the positive term $\ln(N)/\eta$ dominates.

Quadratic regularization: optimistic lazy OGD. Lecture 3 showed that quadratic FTRL gives *lazy OGD*, also called dual averaging. Optimistic lazy OGD extends that method by using m_t to form the current decision and then using the observed loss ℓ_t to update the cumulative-gradient state. Fix $\hat{w}_1 \in \Omega$. With $\psi(w) = \frac{1}{2} \|w - \hat{w}_1\|_2^2$, the process is:

- initialize $z_1 = \hat{w}_1$;
- after receiving m_t , choose $w_t = \Pi_\Omega(z_t - \eta m_t)$;
- after observing ℓ_t , update $z_{t+1} = z_t - \eta \ell_t$.

Equivalently, $z_t = \hat{w}_1 - \eta L_{t-1}$ and hence

$$w_t = \Pi_\Omega(\hat{w}_1 - \eta(L_{t-1} + m_t)). \quad (6)$$

The generic Optimistic FTRL bound immediately gives the prediction-error guarantee for this update.

Theorem 1.4 (Optimistic lazy-OGD bound). *Suppose that Ω has Euclidean diameter at most D . Then, for every $u \in \Omega$,*

$$\sum_{t=1}^T \langle w_t - u, \ell_t \rangle \leq \frac{D^2}{2\eta} + \eta \sum_{t=1}^T \|\ell_t - m_t\|_2^2 - \frac{1}{4\eta} \sum_{t=2}^T \|w_t - w_{t-1}\|_2^2. \quad (7)$$

Proof. Completing the square in the Optimistic FTRL objective gives

$$\langle L_{t-1} + m_t, w \rangle + \frac{1}{2\eta} \|w - \hat{w}_1\|_2^2 = \frac{1}{2\eta} \|w - (\hat{w}_1 - \eta(L_{t-1} + m_t))\|_2^2 + C_t,$$

where C_t does not depend on w . Its minimizer over Ω is therefore the update in (6).

Moreover, $\psi(w) = \frac{1}{2} \|w - \hat{w}_1\|_2^2$ is 1-strongly convex with respect to $\|\cdot\|_2$. Since $\hat{w}_1$ minimizes ψ on Ω and the diameter of Ω is at most D ,

$$B_\psi = \max_{w \in \Omega} \frac{1}{2} \|w - \hat{w}_1\|_2^2 \leq \frac{D^2}{2}.$$

The dual of the Euclidean norm is again $\|\cdot\|_2$. Substituting these facts into Theorem 1.2 gives

$$\text{Reg}_T(u) \leq \frac{D^2}{2\eta} + \eta \sum_{t=1}^T \|\ell_t - m_t\|_2^2 - \frac{1}{4\eta} \sum_{t=2}^T \|w_t - w_{t-1}\|_2^2.$$

For linear losses, $\text{Reg}_T(u) = \sum_{t=1}^T \langle w_t - u, \ell_t \rangle$, which is exactly the claimed bound. $\square$

When $m_t = 0$, the method reduces to the lazy-OGD update from Lecture 3. On a constrained set, lazy OGD differs from ordinary projected OGD because z_t is not projected after each observed loss; the two updates coincide when $\Omega = \mathbb{R}^d$.

2 Application: Two-Player Zero-Sum Games

We now apply optimistic online learning to repeated two-player zero-sum games. This setting is especially useful because each player's decisions generate the other player's loss sequence, allowing predictability to emerge from the interaction itself.

2.1 Preliminaries

A finite two-player zero-sum game is represented by a matrix $A \in [0, 1]^{N \times M}$. The row player has N actions and minimizes loss, while the column player has M actions and maximizes reward. If they choose actions i and j , respectively, then $A(i, j)$ is the row player's loss and the column player's reward.

For Rock–Paper–Scissors, assign row-player loss 1 for losing, 0 for winning, and 1/2 for a tie:

$$A = \begin{array}{c|ccc} & \text{Rock} & \text{Paper} & \text{Scissors} \\ \hline \text{Rock} & \frac{1}{2} & 1 & 0 \\ \text{Paper} & 0 & \frac{1}{2} & 1 \\ \text{Scissors} & 1 & 0 & \frac{1}{2} \end{array}.$$

No fixed action is safe. Randomization is therefore part of the strategic problem, not merely a technical device.

A mixed strategy is a distribution over pure actions. Write $p \in \Delta(N)$ for the row player's strategy and $q \in \Delta(M)$ for the column player's strategy. Their expected payoff is

$$A(p, q) \triangleq p^\top A q = \sum_{i=1}^N \sum_{j=1}^M p_i q_j A(i, j).$$

We also write $A(i, q) \triangleq e_i^\top A q$ and $A(p, j) \triangleq p^\top A e_j$. Linearity implies that a best response can always be chosen to be pure.

Nash equilibrium. A pair (p^*, q^*) is a Nash equilibrium if

$$A(p^*, j) \leq A(p^*, q^*) \leq A(i, q^*) \quad \text{for every } i \in [N], j \in [M].$$

The first inequality says that the maximizing column player cannot gain by deviating; the second says that the minimizing row player cannot reduce his loss by deviating. In Rock–Paper–Scissors, the uniform strategy for each player forms a Nash equilibrium, with value 1/2.

The row player's minimax strategy minimizes the worst-case loss, while the column player's maximin strategy maximizes the worst-case reward. The inequality $\min_p \max_q A(p, q) \geq \max_q \min_p A(p, q)$ is immediate; von Neumann's theorem gives the reverse inequality.

Theorem 2.1 (Minimax theorem). *For every finite two-player zero-sum game,*

$$\min_{p \in \Delta(N)} \max_{q \in \Delta(M)} A(p, q) = \max_{q \in \Delta(M)} \min_{p \in \Delta(N)} A(p, q) \triangleq v(A).$$

Moreover, a pair consists of minimax strategies if and only if it is a Nash equilibrium; in either case its payoff is $v(A)$.

For approximate computation, define the saddle-point gap

$$\text{Gap}(p, q) \triangleq \max_{\tilde{q} \in \Delta(M)} A(p, \tilde{q}) - \min_{\tilde{p} \in \Delta(N)} A(\tilde{p}, q). \quad (8)$$

We call (p, q) an ε -approximate Nash equilibrium if $\text{Gap}(p, q) \leq \varepsilon$. Since the gap is the sum of the two players' possible gains from deviating, this condition bounds each unilateral gain by ε .

If A is known and small, linear programming can compute an equilibrium directly. Learning becomes useful when the matrix is unknown or too large to enumerate, and it can exploit a suboptimal opponent while retaining worst-case protection.

2.2 No-Regret Learning Leads to Approximate Nash Equilibria

A zero-sum game becomes a sequential decision-making problem when the same interaction is repeated. Suppose the game is played for T rounds. On round t , the players use $p_t \in \Delta(N)$ and $q_t \in \Delta(M)$. Afterward, they observe the full loss vectors

$$\ell_t \triangleq Aq_t \in [0, 1]^N, \quad g_t \triangleq -A^\top p_t \in [-1, 0]^M.$$

Each pure row action is now an expert, and the row player's expected loss is $\langle p_t, \ell_t \rangle = A(p_t, q_t)$. The negative sign in g_t converts the column player's reward maximization into loss minimization. This section assumes full-information feedback. Observing only the payoff of a sampled action pair would lead to a bandit game.

The row and column regrets are

$$\text{Reg}_T^X \triangleq \sum_{t=1}^T A(p_t, q_t) - \min_{p \in \Delta(N)} \sum_{t=1}^T A(p, q_t), \quad (9)$$

$$\text{Reg}_T^Y \triangleq \max_{q \in \Delta(M)} \sum_{t=1}^T A(p_t, q) - \sum_{t=1}^T A(p_t, q_t). \quad (10)$$

What happens when both players use no-regret learning algorithms? We will show that their average strategies form an approximate Nash equilibrium. Define

$$\bar{p}_T \triangleq \frac{1}{T} \sum_{t=1}^T p_t, \quad \bar{q}_T \triangleq \frac{1}{T} \sum_{t=1}^T q_t.$$

Before combining the players, consider the row learner alone. Bilinearity and the definition of regret give

$$\frac{1}{T} \sum_{t=1}^T A(p_t, q_t) = \min_{p \in \Delta(N)} A(p, \bar{q}_T) + \frac{\text{Reg}_T^X}{T} \leq v(A) + \frac{\text{Reg}_T^X}{T}.$$

Thus the row player's average loss is at most Reg_T^X / T above the value of the game against any sequence of opponent strategies. This margin vanishes whenever $\text{Reg}_T^X = o(T)$.

When both players learn, their regrets combine into an exact equilibrium certificate: the saddle-point gap of the average strategies equals the sum of their regrets divided by T .

Theorem 2.2 (Regret-to-equilibrium identity). *For every sequence of mixed strategies,*

$$\text{Gap}(\bar{p}_T, \bar{q}_T) = \frac{\text{Reg}_T^X + \text{Reg}_T^Y}{T}. \quad (11)$$

Consequently, sublinear social regret implies that the average strategies approach the set of Nash equilibria.

Proof. Add (9) and (10). The cumulative payoff cancels, leaving

$$\begin{aligned} \text{Reg}_T^X + \text{Reg}_T^Y &= \max_{q \in \Delta(M)} \sum_{t=1}^T A(p_t, q) - \min_{p \in \Delta(N)} \sum_{t=1}^T A(p, q_t) \\ &= T \left(\max_{q \in \Delta(M)} A(\bar{p}_T, q) - \min_{p \in \Delta(N)} A(p, \bar{q}_T) \right). \end{aligned}$$

The expression in parentheses is exactly $\text{Gap}(\bar{p}_T, \bar{q}_T)$. □

The identity turns decentralized repeated play into a one-shot equilibrium certificate. Each player controls only personal regret, yet their average play approaches the set of Nash equilibria. It also gives an online-learning proof of the minimax theorem: run two no-regret algorithms, let $T \rightarrow \infty$, and combine the vanishing gap with the elementary max–min inequality.

If both players use ordinary Hedge, then $\text{Reg}_T^X = \mathcal{O}(\sqrt{T \ln N})$ and $\text{Reg}_T^Y = \mathcal{O}(\sqrt{T \ln M})$. Theorem 2.2 therefore gives

$$\text{Gap}(\bar{p}_T, \bar{q}_T) = \mathcal{O}\left(\sqrt{\frac{\ln(NM)}{T}}\right).$$

To guarantee gap at most ε , $T = \mathcal{O}(\ln(NM)/\varepsilon^2)$ rounds suffice.

3 Faster Convergence via Optimism

Ordinary Hedge guarantees an ε -approximate Nash equilibrium within $\mathcal{O}(\ln(NM)/\varepsilon^2)$ rounds. Can we converge faster by exploiting the fact that both loss sequences are generated by learning algorithms rather than arbitrary adversaries? In self-play, the row loss is $\ell_t = Aq_t$ and the column loss is $g_t = -A^\top p_t$. If an opponent moves gradually, yesterday’s loss predicts today’s loss well, and Optimistic Hedge can exploit this predictability.

At first the reasoning seems circular:

my stability $\implies$ your losses are predictable $\implies$ your optimistic regret is small.

The negative movement term in Theorem 1.2 closes the loop. Each player’s positive prediction-error term will be paid for by the other player’s movement credit.

Both players therefore use the preceding loss vector as the current hint.

Algorithm 1: Optimistic Hedge self-play in a zero-sum game

Input: Matrix $A \in [0, 1]^{N \times M}$, learning rate $\eta > 0$, and horizon T

Initialize $\ell_0 \triangleq 0 \in \mathbb{R}^N$ and $g_0 \triangleq 0 \in \mathbb{R}^M$;

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

The row player chooses $p_{t,i} \propto \exp[-\eta(\ell_{t-1,i} + \sum_{s < t} \ell_{s,i})]$;

The column player chooses $q_{t,j} \propto \exp[-\eta(g_{t-1,j} + \sum_{s < t} g_{s,j})]$;

After both choices, reveal $\ell_t = Aq_t$ and $g_t = -A^\top p_t$;

Output: $\bar{p}_T = T^{-1} \sum_{t=1}^T p_t$ and $\bar{q}_T = T^{-1} \sum_{t=1}^T q_t$

The stability–predictability cancellation yields constant social regret and therefore a $1/T$ convergence rate, as stated next.

Theorem 3.1 (Constant social regret and a $1/T$ Nash rate). *If both players run Optimistic Hedge self-play with $\eta = 1/4$, then*

$$\text{Reg}_T^X + \text{Reg}_T^Y \leq 4 \ln(NM) + \frac{1}{2}, \quad \text{Gap}(\bar{p}_T, \bar{q}_T) \leq \frac{4 \ln(NM) + 1/2}{T}.$$

Thus the average strategies form an $\mathcal{O}(\ln(NM)/T)$ -approximate Nash equilibrium.

Proof. Apply Theorem 1.2 to the row player's entropy-FTRL update with hint $m_t = \ell_{t-1}$. Negative entropy has range $\ln N$ and is strongly convex in $\|\cdot\|_1$. Hence

$$\text{Reg}_T^X \leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^T \|\ell_t - \ell_{t-1}\|_\infty^2 - \frac{1}{4\eta} \sum_{t=2}^T \|p_t - p_{t-1}\|_1^2.$$

For $t \geq 2$, the payoff matrix converts the column player's movement into the row player's prediction error:

$$\|\ell_t - \ell_{t-1}\|_\infty = \max_i |\langle A(i, \cdot), q_t - q_{t-1} \rangle| \leq \|q_t - q_{t-1}\|_1.$$

The symmetric bound $\|g_t - g_{t-1}\|_\infty \leq \|p_t - p_{t-1}\|_1$ holds for the column player. Since the first prediction error is at most 1, we obtain

$$\text{Reg}_T^X \leq \frac{\ln N}{\eta} + \eta + \eta \sum_{t=2}^T \|q_t - q_{t-1}\|_1^2 - \frac{1}{4\eta} \sum_{t=2}^T \|p_t - p_{t-1}\|_1^2. \quad (12)$$

The same reasoning for the column player gives

$$\text{Reg}_T^Y \leq \frac{\ln M}{\eta} + \eta + \eta \sum_{t=2}^T \|p_t - p_{t-1}\|_1^2 - \frac{1}{4\eta} \sum_{t=2}^T \|q_t - q_{t-1}\|_1^2. \quad (13)$$

Before fixing η , the coefficient of the combined movement term is $\eta - 1/(4\eta)$, which is nonpositive only when $\eta \leq 1/2$. We choose $\eta = 1/4$ rather than the boundary value $1/2$ so that a strictly negative movement term remains; this residual stability credit is what allows the individual regrets to be controlled as well. Taking $\eta > 1/2$ makes the combined movement coefficient positive, so dropping the movement term no longer gives a time-independent social-regret bound. Adding (12) and (13) with $\eta = 1/4$ now gives

$$\text{Reg}_T^X + \text{Reg}_T^Y \leq 4 \ln(NM) + \frac{1}{2} - \frac{3}{4} \sum_{t=2}^T \left(\|p_t - p_{t-1}\|_1^2 + \|q_t - q_{t-1}\|_1^2 \right) \leq 4 \ln(NM) + \frac{1}{2}.$$

Theorem 2.2 converts this constant social regret into the claimed saddle-point bound. $\square$

The guarantee concerns the average pair $(\bar{p}_T, \bar{q}_T)$, not necessarily the final pair (p_T, q_T) .

Question. Can you show that Reg_T^X and Reg_T^Y are both $\mathcal{O}(\ln(NM))$? Why does the constant bound on their sum not imply these individual bounds directly?