Hedging in games: Faster convergence of external and swap regrets

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the setting where players run the Hedge algorithm or its optimistic variant [27] to play an n-action game repeatedly for T rounds.

- For two-player games, we show that the regret of optimistic Hedge decays at rate $O(1/T^{5/6})$, improving the previous bound of $O(1/T^{3/4})$ by [27].
- In contrast, we show that the convergence rate of vanilla Hedge is no better than $O(1/\sqrt{T})$, addressing an open question posed in [27].

For general m-player games, we show that the swap regret of each player decays at $O(m^{1/2}(n \log n/T)^{3/4})$ when they combine optimistic Hedge with the classical external-to-internal reduction of Blum and Mansour [6]. Via standard connections, our new (swap) regret bounds imply faster convergence to coarse correlated equilibria in two-player games and to correlated equilibria in multiplayer games.

1 Introduction

Online algorithms for regret minimization play an important role in many applications in machine learning where real-time sequential decision making is crucial [19, 7, 26]. A number of algorithms have been developed, including Hedge / Multiplicative Weights [2], Mirror Decent [19], Follow the Regularized / Perturbed Leader [20], and their power and limits against an adversarial environment have been well understood: The average (external) regret decays at a rate of $O(1/\sqrt{T})$ after T rounds, which is known to be tight for any online algorithm.

What happens if players in a repeated game run one of these algorithms? Given that they are now running against similar algorithms over a fixed game, could the regret of each player decay significantly faster than $1/\sqrt{T}$? This was answered positively in a sequence of works [9, 24, 27]. Among these results, the one that is most relevant to ours is that of Syrgkanis, Agarwal, Luo and Schapire [27]. They showed that if every player in a multiplayer game runs an algorithm that satisfies the RVU (Regret bounded by Variation in Utilities) property, then the regret of each player decays at $O(1/T^{3/4})$. *Can this bound be further improved?*

Besides regret minimization, understanding no-regret dynamics in games is motivated by connections with various equilibrium concepts [15, 13, 12, 18, 6, 17, 22]. For example, if every player runs an algorithm with vanishing regret, then the empirical distribution must converge to a *coarse* correlated equilibrium [7]. Nevertheless, to converge to a more preferred correlated equilibrium [3], a stronger notion of regrets called *swap regrets* (see Section 2) is required [13, 18, 6]. The minimization of swap regrets under the adversarial setting was studied by Blum and Mansour [6]. They gave a generic reduction from regret minimization algorithms which led to a tight $O(\sqrt{n \log n/T})$-bound for the average swap regret. A natural question is *whether a speedup similar to that of* [27] *is possible for swap regrets in the repeated game setting.*

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Our contributions: Faster convergence of swap regrets. We give the first algorithm that achieves an average swap regret that is significantly lower than $O(1/\sqrt{T})$ under the repeated game setting. This algorithm, denoted by `BM-Optimistic-Hedge`, combines the external-to-internal reduction of [6] with the optimistic Hedge algorithm [24, 27] as its regret minimization component. (Optimistic Hedge can be viewed as an instantiation of the optimistic Follow the Regularized Leader algorithm; see Section 2.) We show that if every player in a repeated game of m players and n actions runs `BM-Optimistic-Hedge`, then the average swap regret is at most $O(m^{1/2}(n \log n/T)^{3/4})$; see Theorem 5.1 in Section 5. Via the relationship between correlated equilibria and swap regrets, our result implies faster convergence to a correlated equilibrium. When specialized to two-player games, the empirical distribution of players running `BM-Optimistic-Hedge` converges to an ϵ–correlated equilibrium after $O(n \log n/\epsilon^{4/3})$ rounds, improving the $O(n \log n/\epsilon^2)$ bound of [6].

Our main technical lemma behind Theorem 5.1 shows that strategies produced by the algorithm of [6] with optimistic Hedge moves very slowly in ℓ_1-norm under the adversarial setting (which in turn allows us to apply a stability argument similar to [27]). This came as a surprise because a key component of the algorithm of [6] each round is to compute the stationary distribution of a Markov chain, which is highly sensitive to small changes in the Markov chain. We overcome this difficulty by exploiting the fact that Hedge only incurs small *multiplicative* changes to the Markov chain, which allows us to bound the change in the stationary distribution using the classical Markov chain tree theorem. We further demonstrate the power of this technical ingredient by deriving another fast no-swap regret algorithm, based on a folklore algorithm in [7] and optimistic predictions (see Appendix D). Both of these two algorithms enjoy the benefits of faster convergence when playing with each other, while remain robust against adversaries (see Corollary 5.4 in Appendix C).

Our contributions: Hedge in two-player games. In addition we consider regret minimization in a two-player game with n actions using either vanilla or optimistic Hedge. We show that optimistic Hedge can achieve an average regret of $O(1/T^{5/6})$, improving the bound $O(1/T^{3/4})$ by [27] for two-player games; see Theorem 3.1 in Section 3. In contrast, we show that even under this game-theoretic setting, vanilla Hedge cannot asymptotically outperform the $O(1/\sqrt{T})$ adversarial bound; see Theorem 4.1 in Section 4. This addresses an open question posed by [27] concerning the convergence rate of vanilla Hedge in a repeated game.

The key step in our analysis of optimistic Hedge is to show that, even under the adversarial setting, the trajectory length of strategy movements (in their squared ℓ_1-norm) can be bounded using that of cost vectors (in ℓ_∞-norm); see Lemma 3.2. (Intuitively, it is unlikely for the strategy of optimistic Hedge to change significantly over time while the loss vector stays stable.) This allows us to build a strong relationship between the trajectory length of each player's strategy movements, and then use the RVU property of optimistic Hedge to bound their individual regrets.

Our lower bounds for vanilla Hedge use three very simple 2×2 games to handle different ranges of the learning rate η. For the most intriguing case when η is at least $\Omega(1/\sqrt{n})$ and bounded from above by some constant, we study the zero-sum Matching Pennies game and use it to show that the overall regret of at least one player is $\Omega(\sqrt{T})$. Our analysis is inspired by the result of [5] which shows that the KL divergence of strategies played by Hedge in a two-player zero-sum game is strictly increasing. For Matching Pennies, we start with a quantitative bound on how fast the KL divergence grows in Lemma 4.3. This implies the existence of a window of length $\sqrt{T}$ during which the cost of one of the player grows by $\Omega(1)$ each round; the zero-sum structure of the game allows us to conclude that at least one of the players must have regret at least $\Omega(\sqrt{T})$ at some point in this window.

1.1 Related work

Initiated by Daskalakis, Deckelbaum and Kim [9], there has been a sequence of works that study no-regret learning algorithms in games [24, 27, 14, 29]. Daskalakis et. al. [9] designed an algorithm by adapting Nesterov's accelerated saddle point algorithm to two-player *zero-sum games*, and showed that if both players run this algorithm then their average regrets decay at rate $O(1/T)$, which is optimal. Later Rakhlin and Sridharan [23, 24] developed a simple and intuitive family of algorithms, i.e. *optimistic Mirror Descent* and *optimistic Follow the Regularized Leader*, that incorporate predictions into the strategy. They proved that if both players adopt the algorithm, then their average regrets also decay at rate $O(1/T)$ in *zero sum games*. Syrgkanis et. al. [27] further strengthened this line of works by showing that in a *general* m-player game, if every player runs an algorithm that satisfies the RVU property then the average regret decays at rate $O(1/T^{3/4})$. Syrgkanis et.

al. [27] also considered the convergence of social welfare and proved an even faster rate of $O(1/T)$ in smooth games [25]. Foster et. al. [14] extended [27] and showed that if one only aims for an approximately optimal social welfare, then the class of algorithms allowed can be much broader. Recently, Daskalakis and Panageas [11] proved the last iteration convergence of optimistic Hedge in zero-sum game, i.e., instead of averaging over the trajectory, they showed that optimistic Hedge converges to a Nash equilibrium in a zero-sum game.

There is also a growing body of works [21, 5, 4, 8] on the dynamics of no-regret learning over games in the last few years. Most of these works studied the dynamics of no-regret learning from a dynamical system point of view and provided qualitative intuition on the evolution of no-regret learning. Among them, [4] is most relevant, in which Bailey and Piliouras proved an $\Omega(\sqrt{T})$ lower bound on the convergence rate of online gradient descent [30] for the 2×2 Matching Pennies game. However, we remark that their lower bound only works for online gradient descent and they need to fix the learning rate η to 1. Our lower bound for vanilla Hedge in two-player games holds for arbitrary learning rates.

2 Preliminary

Notation. Given two positive integers $n \le m$, we use $[n]$ to denote $\{1, \ldots, n\}$ and $[n : m]$ to denote $\{n, \ldots, m\}$. We use $D_{\mathrm{KL}}(p\|q)$ to denote the KL divergence with natural logarithm.

Repeated games and regrets. Consider a game G played between m players, where each player $i \in [m]$ has a strategy space S_i with $|S_i| = n$ and a *loss* function $\mathcal{L}_i : S_1 \times \cdots \times S_m \to [0, 1]$ such that $\mathcal{L}_i(\mathbf{s})$ is the loss of player i for each pure strategy profile $\mathbf{s} = (s_1, \ldots, s_n) \in S_1 \times \cdots \times S_m$. A mixed strategy for player i is a probability distribution x_i over S_i, where the jth action is played with probability $x_i(j)$. Given a mixed (or pure) strategy profile $\mathbf{x} = (x_1, \ldots, x_m)$ (or $\mathbf{s} = (s_1, \ldots, s_m)$), we write $\mathbf{x}_{-i}$ (or $\mathbf{s}_{-i}$) to denote the profile after removing x_i (or s_i, respectively).

We consider the scenario where the m players play G repeatedly for T rounds. At the beginning of each round t, $t \in [T]$, each player i picks a mixed strategy x_i^t and let $\mathbf{x}^t = (x_1^t, \ldots, x_m^t)$ be the mixed strategy profile. We consider the *full information* setting where each player observes the *expected* loss of *all* her actions. Formally, player i observes a loss vector ℓ_i^t with $\ell_i^t(j) = \mathbb{E}_{\mathbf{s}_{-i} \sim \mathbf{x}_{-i}^t}[\mathcal{L}_i(j, \mathbf{s}_{-i})]$, and her expected loss is given by $\langle x_i^t, \ell_i^t \rangle$. At the end of round T, the *regret* of player i is

$$\text{regret}_T^i = \sum_{t \in [T]} \langle x_i^t, \ell_i^t \rangle - \min_{j \in [n]} \sum_{t \in [T]} \ell_i^t(j), \tag{1}$$

i.e., the maximum gain one could have obtained by switching to some fixed action. A stronger notion of regret, referred as *swap regret*, is defined as

$$\text{swap-regret}_T^i = \sum_{t \in [T]} \langle x_i^t, \ell_i^t \rangle - \min_{\phi} \sum_{t \in [T]} \sum_{j \in [n]} x_i^t(j) \cdot \ell_i^t(\phi(j)), \tag{2}$$

where the minimum is over all n^n (swap) functions $\phi : [n] \to [n]$ that swap action j with $\phi(j)$. The swap regret equals the maximum gain one could have achieved by using a fixed swap function over its past mixed strategies.

Hedge. Consider the adversarial online model where a player has n actions and picks a distribution x^t over them at the beginning of each round t. During round t the player receives a loss vector ℓ^t and pays a loss of $\langle x^t, \ell^t \rangle$. The vanilla Hedge algorithm [16] with learning rate $\eta > 0$ starts by setting x^1 to be the uniform distribution and then keeps applying the following updating rule to obtain x^{t+1} from x^t and the loss vector ℓ^t at the end of round t: for each action $j \in [n]$,

$$x^{t+1}(j) = \frac{x^t(j) \cdot \exp(-\eta \cdot \ell^t(j))}{\sum_{k \in [n]} x^t(k) \cdot \exp(-\eta \cdot \ell^t(k))}.$$

On the other hand, the optimistic Hedge algorithm can be obtained from the *optimistic follow the regularized leader* proposed by [24, 27], and have the following updating rule:

$$x^{t+1}(j) = \frac{x^t(j) \cdot \exp(-\eta(2\ell^t(j) - \ell^{t-1}(j))}{\sum_{k \in [n]} x^t(k) \cdot \exp(-\eta(2\ell^t(k) - \ell^{t-1}(k))}, \tag{3}$$

with $\ell^0 = \mathbf{0}$ being the all-zero vector. We have the following regret bound for optimistic Hedge.

Lemma 2.1 ([24, 27]). *Under the adversarial setting, optimistic Hedge satisfies*

$$\text{regret}_T \leq \frac{2\log n}{\eta} + \eta \sum_{t\in[T]} \|\ell^t - \ell^{t-1}\|_\infty^2 - \frac{1}{4\eta} \sum_{t\in[T]} \|x^{t+1} - x^t\|_1^2. \tag{4}$$

3 Optimistic Hedge in Two-Player Games

In this section we analyze the performance of the optimistic Hedge algorithm when it is used by two players to play a (general, not necessarily zero-sum) $n \times n$ game repeatedly.

Theorem 3.1. *Suppose both players in a two-player game run optimistic Hedge for T rounds with learning rate $\eta = (\log n/T)^{1/6}$. Then the individual regret of each player is $O(T^{1/6}\log^{5/6} n)$.*

We assume without loss of generality that $T \geq \log n$; otherwise, the regret of each player is trivially at most $T \leq T^{1/6}\log^{5/6} n$. The following lemma is essential to our proof of Theorem 3.1. Consider the adversarial online setting where a player runs optimistic Hedge for T rounds. The lemma bounds the trajectory length of the strategy movement using that of cost vectors.

Lemma 3.2. *Suppose that a player runs optimistic Hedge with learning rate η for T rounds. Let $\ell^0, \ell^1, \ldots, \ell^T$ be the cost vectors with $\ell^0 = \mathbf{0}$ and $x^1, \ldots, x^T$ be the strategies played. Then*

$$\sum_{t\in[2:T]} \|x^t - x^{t-1}\|_1^2 \leq O(\log n) + O(\eta + \eta^2) \sum_{t\in[T-1]} \|\ell^t - \ell^{t-1}\|_\infty. \tag{5}$$

We delay the proof of Lemma 3.2 to Appendix A and use it to prove Theorem 3.1.

Proof of Theorem 3.1 assuming Lemma 3.2. Let $G = (A, B)$ be the game, where $A, B \in [0,1]^{n\times n}$ denote the cost matrices of the first and second players, respectively. We use x^t and y^t to denote strategies played by the two players and use ℓ_x^t and ℓ_y^t to denote their cost vectors in the tth round. So we have $\ell_x^t = Ay^t$ and $\ell_y^t = B^T x^t$. Therefore, we have for each $t \geq 2$:

$$\begin{aligned} &\|\ell_y^t - \ell_y^{t-1}\|_\infty = \|B^T(x^t - x^{t-1})\|_\infty \leq \|x^t - x^{t-1}\|_1 \quad \text{and} \\ &\|\ell_x^t - \ell_x^{t-1}\|_\infty = \|A(y^t - y^{t-1})\|_\infty \leq \|y^t - y^{t-1}\|_1. \end{aligned} \tag{6}$$

Without loss of generality it suffices to bound the regret of the second player. Set $\eta = (\log n/T)^{1/6}$ with $T \geq \log n$ so that $\eta \leq 1$. We have

$$\begin{aligned}
\text{regret}_T^y &\leq \frac{2\log n}{\eta} + \eta \sum_{t\in[T]} \|\ell_y^t - \ell_y^{t-1}\|_\infty^2 - \frac{1}{4\eta}\sum_{t\in[T]} \|y^{t+1} - y^t\|_1^2 && \text{Lemma 2.1} \\
&\leq \frac{2\log n}{\eta} + \eta + \eta \sum_{t\in[2:T]} \|x^t - x^{t-1}\|_1^2 - \frac{1}{4\eta} \sum_{t\in[2:T+1]} \|\ell_x^t - \ell_x^{t-1}\|_\infty^2 && \text{using (6)} \\
&\leq \frac{2\log n}{\eta} + \eta + \eta \left(O(\log n) + O(\eta) \sum_{t\in[T-1]} \|\ell_x^t - \ell_x^{t-1}\|_\infty \right) \\
&\qquad - \frac{1}{4\eta} \sum_{t\in[T-1]} \|\ell_x^t - \ell_x^{t-1}\|_\infty^2 + \frac{1}{4\eta} && \text{Lemma 3.2} \\
&= O\left(\frac{\log n}{\eta}\right) + \sum_{t\in[T-1]} \left(O(\eta^2)\cdot \|\ell_x^t - \ell_x^{t-1}\|_\infty - \frac{1}{4\eta}\cdot \|\ell_x^t - \ell_x^{t-1}\|_\infty^2 \right) \\
&\leq O\left(\frac{\log n}{\eta}\right) + T\cdot O(\eta^5) = O\left(T^{1/6}\log^{5/6} n\right).
\end{aligned}$$

This finishes the proof of the theorem. □

4 Lower Bounds for Hedge in Two-Player Games

We prove lower bounds for regrets of players when they both run the vanilla Hedge algorithm. We show that even in games with two actions, vanilla Hedge cannot perform asymptotically better than its guaranteed regret bound of $O(\sqrt{T})$ under the adversarial setting.

Theorem 4.1. *Suppose two players run the vanilla Hedge algorithm to play a two-action game with initial strategy* $(0.4, 0.6)$*. Then for any sufficiently large* T *and any learning rate* $\eta > 0$*, there is a game such that at least one player has regret* $\Omega(\sqrt{T})$ *after* T' *rounds for some* $T' \in [T : T + \sqrt{T}]$*.*

Remark 4.2. *Theorem 4.1 shows that even if players have a good estimation about the number of rounds to play (i.e., between* T *and* $T + \sqrt{T}$*), vanilla Hedge with any learning rate* $\eta(T) > 0$ *picked using* T *cannot promise to achieve a regret bound that is asymptotically lower than* $O(\sqrt{T})$ *for every round* $T' \in [T : T + \sqrt{T}]$*. We would like to point out that the use of* $(0.4, 0.6)$ *as the initial strategy instead of the uniform distribution is not crucial but only to simplify the construction and analysis.*

Let T be a sufficiently large integer. We will use three games $G_i = (A, B_i)$, $i \in \{1, 2, 3\}$, to handle three cases of the learning rate η, where

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}, \quad B_1 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}, \quad B_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad B_3 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

We use G_2 to handle the case when $\eta \le 64/(c_0\sqrt{T})$ (see Appendix B.1) where $c_0 \in (0, 1]$ is a constant introduced below in Lemma 4.3. We use G_3 to handle the case when $\eta \ge 3$ (see Appendix B.2). The most intriguing case is when the learning rate η is between $64/(c_0\sqrt{T})$ and 3. For this case we use the Matching Pennies game $G_1 = (A, B_1)$.

Let x^t and y^t denote strategies played in round t by the first and second players, respectively. Let $x^\star = y^\star = (0.5, 0.5)$. The proof for this case relies on the following lemma, which shows that the KL divergence between $(x^\star, y^\star)$ and (x^T, y^T) after T rounds is at least $\Omega(\sqrt{T}\eta))$.

Lemma 4.3. *Suppose players run vanilla Hedge for* T *rounds with* $\eta : 16/\sqrt{T} \le \eta \le 3$*. Then*

$$D_{KL}(x^\star \| x^T) + D_{KL}(y^\star \| y^T) \ge c_0\sqrt{T}\eta, \quad \textit{for some constant } c_0 \in (0, 1].$$

We are now ready to prove Theorem 4.1 for the main case when $64/(c_0\sqrt{T}) \le \eta \le 3$.

Proof of Theorem 4.1 for the main case. For convenience we let $x_t = x^t(1)$ (or $y_t = y^t(1)$) denote the probability of playing the first action in x^t (or y^t, respectively). We first describe the high level idea behind the proof. Since we know the KL divergence is at least $c_0\sqrt{T}\eta$ at time T by Lemma 4.3, at least one of x_T and y_T is extremely close to either 0 or 1. Assume without loss of generality that this is the case for x_T. As a result, the probability of the first player playing the first action will not change much for the next $\sqrt{T}$ rounds. Consequently, during the next $\sqrt{T}$ rounds, one of the players must keep losing and the other player will keep winning. This can be used to show that one of the two players must have regret at least $\Omega(\sqrt{T})$ at some point T' between T and $T + \sqrt{T}$.

To make this more formal, let ℓ_x^t (or ℓ_y^t) denote the cost vector of the first (or the second) player at round t and define L_x^t and L_y^t to be the total loss up to round t of the two players:

$$L_x^t = \sum_{\tau \in [t]} \langle x^\tau, \ell_x^\tau \rangle \quad \text{and} \quad L_y^t = \sum_{\tau \in [t]} \langle y^\tau, \ell_y^\tau \rangle.$$

Since $G_1 = (A, B_1)$ is zero-sum, we have $\langle x^\tau, \ell_x^\tau \rangle + \langle y^\tau, \ell_y^\tau \rangle = 0$ and thus, $L_x^t + L_y^t = 0$. Moreover, noting that the sum of two rows of A is zero, the first player can always guarantee an overall loss of at most 0 when playing the best fixed action in hindsight. Therefore, $\text{regret}_t^x \ge L_x^t$ and similarly $\text{regret}_t^y \ge L_y^t$. Combining this with $L_x^t + L_y^t = 0$, we have

$$\max\left\{ \text{regret}_t^x, \text{regret}_t^y \right\} \ge |L_x^t| = |L_y^t|.$$

To finish the proof, it suffices to show that

$$\left|L_x^{T'}\right| = \left|L_y^{T'}\right| \ge \Omega(\sqrt{T}), \quad \text{for some } T' \in [T : T + \sqrt{T}]. \tag{7}$$

Let $L = c_0\sqrt{T}/8 \le \sqrt{T}$. We have from Lemma 4.3 that the KL divergence is at least $c_0\sqrt{T}\eta$ (using $\eta \ge 64/(c_0\sqrt{T}) > 16/\sqrt{T}$). We assume without loss of generality that $D_{\mathrm{KL}}(x^\star \| x^T) \ge c_0\sqrt{T}\eta/2$. We further assume without loss of generality that the second term is larger:

$$\frac{1}{2} \cdot \log \frac{1}{2(1-x_T)} \ge \frac{c_0\sqrt{T}\eta}{4}.$$

It follows that x_T is very close to 1: $x_T \ge 1 - \exp(-c_0\sqrt{T}\eta/2)$, and we use this to show that $x_{T+\tau}$ remains close to 1 for all $\tau \in [L]$. To see this is the case, we note that

$$\frac{x_{T+\tau}}{1-x_{T+\tau}} \ge \exp(-2\eta\tau) \cdot \frac{x_T}{1-x_T} \ge \frac{1}{2} \cdot \exp\left(-2\eta L + \frac{c_0\sqrt{T}\eta}{2}\right) = \frac{1}{2} \cdot \exp\left(\frac{c_0\sqrt{T}\eta}{4}\right) \ge 3,$$

where we used $\eta \ge 64/(c_0\sqrt{T})$ in the last inequality. This implies $x_{T+\tau} \ge 3/4$ for all $\tau \in [L]$.

Now we turn our attention to the second player. Given that $x_{T+\tau} \ge 3/4$ for all $\tau \in [L]$, $y_{T+\tau}$ keeps growing for all $\tau \in [L]$. As a result there is an interval $I \subseteq [L]$ such that (i) every $y_{T+\tau}$, $\tau \in I$, lies between $1/4$ and $3/4$; (ii) every $y_{T+\tau}$ before I is smaller than $1/4$; and (iii) every $y_{T+\tau}$ after I is larger than $3/4$. Using a similar argument, we show that I cannot be too long. Letting ℓ and r be the left and right endpoints of I, we have

$$3 \ge \frac{y_r}{1-y_r} \ge \exp\left(\frac{\eta(r-\ell)}{2}\right) \cdot \frac{y_\ell}{1-y_\ell} \ge \exp\left(\frac{\eta(r-\ell)}{2}\right) \cdot \frac{1}{3}.$$

As a result, we have $(r-\ell) \le 6/\eta \le (3/32) \cdot c_0\sqrt{T}$ and thus, either (i) or (ii) is of length at least $\Omega(L)$. We focus on the case when (ii) is long; the other case can be handled similarly.

Summarizing what we have so far, there is an interval $J = [\alpha : \beta] \subseteq [L]$ of length $\Omega(L)$ such that for every $\tau \in J$, both $x_{T+\tau}$ and $y_{T+\tau}$ are at least $3/4$. This implies that the total loss of the first player grows by $\Omega(1)$ each round and thus, $L_x^{T+\beta} - L_x^{T+\alpha} \ge \Omega(L)$. Therefore, either $|L_x^{T+\alpha}| \ge \Omega(L)$ or $|L_x^{T+\beta}| \ge \Omega(L)$. This finishes the proof of (7) using $L = \Omega(\sqrt{T})$ and the proof of the theorem. □

5 Faster Convergence of Swap Regrets

Under the adversarial online model, Blum and Mansour [6] gave a black-box reduction showing that any algorithm that achieve good regrets can be converted into an algorithm that achieves good swap regrets. In this section we show that if every player in a repeated game runs their algorithm with optimistic Hedge as its core, then the swap regret of each player can be bounded from above by $O((n\log n)^{3/4}(mT)^{1/4})$, where m is the number of players and n is the number of actions.

We start with an overview on the reduction framework of [6], which we will refer to as the BM algorithm. Let $S = [n]$ be the set of available actions. Given an algorithm $\texttt{ALG}$ that achieves good regrets, the BM algorithm instantiates n copies $\texttt{ALG}_1, \ldots, \texttt{ALG}_n$ of $\texttt{ALG}$ over S. At the beginning of each round $t = 1, \ldots, T$, the BM algorithm receives a distribution q_i^t over S from $\texttt{ALG}_i$ for each $i \in [n]$, and plays x^t, which is the unique distribution over S that satisfies $x^t = x^t Q^t$, where Q^t is the $n \times n$ matrix with row vectors $q_1^t, \ldots, q_n^t$. After receiving the loss vector ℓ^t, the BM algorithm experiences a loss of $\langle x^t, \ell^t \rangle$ and distributes $x^t(i) \cdot \ell^t$ to $\texttt{ALG}_i$ as its loss vector in round t.

We are now ready to state our main theorem of this section:

Theorem 5.1. *Suppose that every player in a repeated game runs the BM algorithm with optimistic Hedge as* `ALG` *and sets the learning rate of the latter to be* $\eta = (n\log n/(m^2T))^{1/4}$. *Then the swap regret of each player is* $O((n\log n)^{3/4} \cdot (m^2T)^{1/4})$.

For convenience we refer to the BM algorithm with optimistic Hedge as `BM-Optimistic-Hedge` in the rest of the section. We first combine the analysis of [6] for the BM algorithm and Lemma 3 to obtain the following bound for the swap regret of `BM-Optimistic-Hedge` under the adversarial setting, in terms of the total path length of cost vectors the player's mixed strategies:

Lemma 5.2. *Suppose that a player runs* `BM-Optimistic-Hedge` *with* $\eta > 0$ *for* T *rounds. Then*

$$\text{swap-regret}_T \le \frac{2n\log n}{\eta} + 2\eta\left(\sum_{t=2}^{T} \|x^t - x^{t-1}\|_1^2 + \sum_{t=1}^{T} \|\ell^t - \ell^{t-1}\|_\infty^2\right), \quad \textit{where } \ell^0 = \mathbf{0}.$$

$$Q = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon' & 1-\epsilon' \end{pmatrix} \;\; x = \begin{pmatrix} \frac{1}{k+1} & \frac{k}{k+1} \end{pmatrix} \;\; \text{vs} \;\; Q = \begin{pmatrix} 1-\epsilon' & \epsilon' \\ \epsilon & 1-\epsilon \end{pmatrix} \;\; x = \begin{pmatrix} \frac{k}{k+1} & \frac{1}{k+1} \end{pmatrix}$$

Figure 1: Let $\epsilon' = \epsilon/k$. Additive perturbations may change the stationary distribution dramatically.

The proof can be found in Appendix C.1. For the repeated game setting, we have for each $t \geq 2$,

$$\|\ell_i^t - \ell_i^{t-1}\|_\infty \leq \|\mathbf{x}_{-i}^t - \mathbf{x}_{-i}^{t-1}\|_1 \leq \sum_{j \neq i} \|\mathbf{x}_j^t - \mathbf{x}_j^{t-1}\|_1$$

where the last inequality used the fact that both $\mathbf{x}_{-i}^t$ and $\mathbf{x}_{-i}^{t-1}$ are product distributions. Combining it with Lemma 5.2, we can bound the swap regret of each player $i \in [m]$ in the game by

$$\text{swap-regret}_T^i \leq \frac{2n \log n}{\eta} + 2\eta + 2\eta m \sum_{j \in [m]} \sum_{t=2}^{T} \|x_j^t - x_j^{t-1}\|_1^2. \tag{8}$$

We prove the following main technical lemma in the rest of the section, which states that the mixed strategy x^t produced by `BM-Optimistic-Hedge` under the adversarial setting moves very slowly (by at most $O(\eta)$ in ℓ_1-distance each round). Theorem 5.1 follows by combining Lemma 5.2 and 5.3.

Lemma 5.3. *Suppose that a player runs* `BM-Optimistic-Hedge` *with rate* $\eta : 0 < \eta \leq 1/6$ *under the adversarial setting. Then we have* $\|x^t - x^{t-1}\|_1 \leq O(\eta)$ *for all* $t \geq 2$.

Proof of Theorem 5.1 Assuming Lemma 5.3. Let $\eta = (n \log n)^{1/4}(m^2 T)^{-1/4}$. For the special case when $\eta > 1/6$, the swap regret of each player is trivially at most $T = O((n \log n)^{3/4} \cdot (m^2 T)^{1/4})$. Assuming $\eta \leq 1/6$, by Lemma 5.2 we have from (8) that

$$\text{swap-regret}_T^i \leq \frac{2n \log n}{\eta} + 2\eta + 2\eta m^2 T \cdot O(\eta^2) = O\left((n \log n)^{3/4} \cdot (m^2 T)^{1/4}\right).$$

This finishes the proof of the theorem. □

The proof of Lemma 5.3 can be found in Appendix C.2. Here we give a high-level description of its proof. Given that `BM-Optimistic-Hedge` runs n copies of optimistic Hedge with rate η, we know that mixed strategies proposed by each $\texttt{ALG}_i$ move very slowly: $\|q_i^t - q_i^{t-1}\|_1 \leq O(\eta)$. However, it is not clear whether this translates into a similar property for x^t since the latter is obtained by solving $x^t = x^t Q^t$. Equivalently, x^t can be viewed as the stationary distribution of the Markov chain Q^t composed by strategies of each individual expert $\texttt{ALG}_i$, and its dependency on Q^t is highly nonlinear. While there is a vast literature on the perturbation analysis of Markov chains, many results require additional assumptions on the underlying Markov chain (e.g. bounded eigenvalue gap) and are not well suited for our setting here. Indeed, it is easy to come up with examples showing that the stationary distrbution is extremely sensitive to small *additive* perturbations (see Figure 1). As a result one cannot hope to prove Lemma 5.3 based on the property $\|q_i^t - q_i^{t-1}\|_1 \leq O(\eta)$ only.

We circumvent this difficulty by noting that optimistic Hedge only incurs small *multiplicative* perturbations on the Markov chain (see Claim C.5), i.e., each entry of Q^t differs from the corresponding entry of Q^{t-1} by no more than a small multiplicative factor of the latter. We present in Lemma C.2 an analysis on stationary distributions of Markov chains under multiplicative perturbations, based on the classical Markov chain tree theorem, and then use it to prove Lemma 5.3.

We further prove that one can design a wrapper for `BM-Optimistic-Hedge` that is robust against adversarial opponents:

Corollary 5.4. *There is an algorithm* `BM-Optimistic-Hedge`* *with the following guarantee. If all players run* `BM-Optimistic-Hedge`**, then the swap regret of each individual is* $\widetilde{O}(n^{3/4}(m^2 T)^{1/4})$*; if the player is facing adversaries, then the swap regret is still at most* $\widetilde{O}((nT)^{1/2} + n^{3/4}(m^2 T)^{1/4})$.

The proof is similar to Corollary 16 in [27]; we present it in Appendix C.3 for completeness.

In the appendix we give two more extensions to our results on swap regrets.

1. In Appendix D, we show that incorporating optimistic Hedge into a folklore algorithm from [7] can also achieve faster convergence of swap regrets, with a slightly worse dependence on n. Interestingly, our analysis of this algorithm also crucially relies on the perturbation analysis of stationary distributions of Markov chains.
2. In Appendix E, we study the convergence to the approximately optimal social welfare (following the definition in [14]) with no-swap regret algorithms, and prove that $O(1/T)$ holds for a wide range of no-swap regret algorithms.

6 Discussion

In this paper, we studied the convergence rate of regrets of the Hedge algorithm and its optimistic variant in two-player games. We obtained a strict separation between vanilla Hedge and optimistic Hedge, i.e., $1/\sqrt{T}$ vs. $1/T^{5/6}$. We also initiated the study on algorithms with faster convergence rates of swap regrets in general multiplayer games and obtained an algorithm with average regret $O(m^{1/2}(n\log n/T)^{3/4})$, improving over the classic result of Blum and Mansour [6].

Our work led to several interesting future directions:

- Our faster convergence result for optimistic Hedge currently only works for two-player games. Can we extend it to multiplayer games? Second, what is the optimal convergence rate for optimistic Hedge and other no-regret algorithms? even for two-player games?
- Regarding swap regrets, it is easy to generalize the result in Section 5 to any algorithm that (1) satisfies the RVU property and (2) makes only multiplicative changes on strategies each iteration. These include optimistic Hedge and optimistic multiplicative weights. However, our current analysis does not apply to general optimistic Mirror Descent or Follow the Regularized Leader. Can we still prove faster convergence of swap regrets via the reduction of [6] without requiring (2) on the regret minimization algorithm? or does there exist some natural gap between these algorithms and optimistic Hedge / multiplicative weights?
- For our result in Appendix E on the convergence to the approximately optimal social welfare, can this fast convergence result be extended to the (exact) optimal social welfare setting (follow the definition in [27])?
- Can we achieve similar convergence rates under partial information models? such as those considered in [24, 14, 29].

References

[1] Venkat Anantharam and Pantelis Tsoucas. A proof of the markov chain tree theorem. *Statistics & Probability Letters*, 8(2):189–192, 1989.

[2] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[3] R.J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1:67–96, 1974.

[4] James Bailey and Georgios Piliouras. Fast and furious learning in zero-sum games: vanishing regret with non-vanishing step sizes. In *Advances in Neural Information Processing Systems*, pages 12977–12987, 2019.

[5] James P Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 321–338, 2018.

[6] Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun):1307–1324, 2007.

[7] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[8] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Conference on Learning Theory*, pages 807–834, 2019.

[9] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254. SIAM, 2011.

[10] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In *International Conference on Learning Representations*, 2018.

[11] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. *arXiv preprint arXiv:1807.04252*, 2018.

[12] Dean P Foster and Rakesh Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29(1-2):7–35, 1999.

[13] Dean P Foster and Rakesh V Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1-2):40, 1997.

[14] Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, pages 4734–4742, 2016.

[15] Yoav Freund and Robert E Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 325–332, 1996.

[16] Yoav Freund and Robert E. Schapire. *J. Comput. System Sci.*, 55(1):119–139, 1997.

[17] Amy Greenwald, Zheng Li, and Warren Schudy. More efficient internal-regret-minimizing algorithms. In *COLT*, pages 239–250, 2008.

[18] Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

[19] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[20] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. *Journal of Computer and System Sciences*, 71(3):291–307, 2005.

[21] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

[22] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[23] Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013.

[24] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[25] Tim Roughgarden. Intrinsic robustness of the price of anarchy. *Journal of the ACM (JACM)*, 62(5):1–42, 2015.

[26] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

[27] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

[28] Jun-Kun Wang and Jacob D Abernethy. Acceleration through optimistic no-regret dynamics. In *Advances in Neural Information Processing Systems*, pages 3824–3834, 2018.

[29] Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory*, pages 1263–1291, 2018.

[30] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Missing proof from Section 3

Proof of Lemma 3.2 For each $t \in [2:T]$, we apply Pinsker's inequality to have

$$
\begin{aligned}
\frac{1}{2} \cdot \|x^t - x^{t-1}\|_1^2 \le D_{\mathrm{KL}}(x^{t-1} \| x^t) &= \sum_{i \in [n]} x^{t-1}(i) \cdot \log\left(\frac{x^{t-1}(i)}{x^t(i)}\right) \\
&= \sum_{i \in [n]} x^{t-1}(i) \cdot \log\left(\sum_{j \in [n]} \exp\left(-\eta\big(2\ell^{t-1}(j) - \ell^{t-2}(j)\big)\right) \cdot x^{t-1}(j)\right) \\
&\qquad + \sum_{i \in [n]} x^{t-1}(i) \cdot \eta\big(2\ell^{t-1}(i) - \ell^{t-2}(i)\big) \\
&= \log\left(\sum_{j \in [n]} \exp\left(-\eta\big(2\ell^{t-1}(j) - \ell^{t-2}(j)\big)\right) \cdot x^{t-1}(j)\right) + \eta\langle x^{t-1}, 2\ell^{t-1} - \ell^{t-2}\rangle \\
&\triangleq \Phi_t + \eta\langle x^{t-1}, 2\ell^{t-1} - \ell^{t-2}\rangle, \qquad (9)
\end{aligned}
$$

where we recall $\ell^0 = \mathbf{0}$. The third step follows from the updating rule of optimistic Hedge. Letting $L^t = \sum_{i \in [t]} \ell^i$, next we use induction to prove the following claim for each $k = 1, \ldots, T$:

$$
\sum_{t \in [k]} \Phi_t = \log\left(\sum_{j \in [n]} x^1(j) \cdot \exp\left(-\eta L^{k-1}(j) - \eta \ell^{k-1}(j)\right)\right). \qquad (10)
$$

The base case holds trivially, as $\Phi_1 = 0$. Suppose the above holds for k. Then for $k+1$ we have

$$
\begin{aligned}
\sum_{t=1}^{k+1} \Phi_t &= \sum_{t=1}^{k} \Phi_t + \Phi_{k+1} \\
&= \log\left(\sum_{j \in [n]} x^1(j) \cdot \exp\left(-\eta L^{k-1}(j) - \eta \ell^{k-1}(j)\right)\right) + \log\left(\sum_{i \in [n]} \exp\left(-\eta\big(2\ell^k(i) - \ell^{k-1}(i)\big)\right) \cdot x^k(i)\right) \\
&= \log\left(\left(\sum_{i \in [n]} \exp\left(-\eta\big(2\ell^k(i) - \ell^{k-1}(i)\big)\right) \cdot x^k(i)\right) \cdot \left(\sum_{j \in [n]} x^1(j) \cdot \exp\left(-\eta L^{k-1}(j) - \eta \ell^{k-1}(j)\right)\right)\right) \\
&= \log\left(\sum_{i \in [n]} \exp\left(-\eta\big(2\ell^k(i) - \ell^{k-1}(i)\big)\right) \cdot x^1(i) \cdot \exp\left(-\eta L^{k-1}(i) - \eta \ell^{k-1}(i)\right)\right) \\
&= \log\left(\sum_{i \in [n]} x^1(i) \cdot \exp\left(-\eta L^k(i) - \eta \ell^k(i)\right)\right),
\end{aligned}
$$

where the third step follows from

$$
x^k(i) = \frac{x^1(i) \cdot \exp\left(-\eta L^{k-1}(i) - \eta \ell^{k-1}(i)\right)}{\sum_{j \in [n]} x^1(j) \cdot \exp\left(-\eta L^{k-1}(j) - \eta \ell^{k-1}(j)\right)}.
$$

Now we have (recall that $\Phi_1 = 0$)

$$\begin{aligned}\frac{1}{2\ln 2}\sum_{t\in[2:T]}\|x^t - x^{t-1}\|_1^2 \leq& \sum_{t\in[2:T]}\Big(\Phi_t + \eta\langle x^{t-1}, 2\ell^{t-1} - \ell^{t-2}\rangle\Big)\\ =& \log\left(\sum_{j\in[n]}\frac{1}{n}\cdot\exp\left(-\eta L^{T-1}(j) - \eta\ell^{T-1}(j)\right)\right) + \sum_{t\in[2:T]}\eta\langle x^{t-1}, 2\ell^{t-1} - \ell^{t-2}\rangle\\ \leq& -\min_{j\in[n]}\Big(\eta L^{T-1}(j) + \eta\ell^{T-1}(j)\Big) + \sum_{t\in[2:T]}\eta\langle x^{t-1}, 2\ell^{t-1} - \ell^{t-2}\rangle\\ \leq& -\eta\min_{j\in[n]}L^{T-1}(j) + \eta\sum_{t\in[T-1]}\langle x^t, \ell^t\rangle + \eta\sum_{t\in[T-1]}\langle x^t, \ell^t - \ell^{t-1}\rangle\\ \leq& \eta\left(\frac{2\log n}{\eta} + \eta\sum_{t\in[T-1]}\|\ell^t - \ell^{t-1}\|_\infty^2\right) + \eta\sum_{t\in[T-1]}\langle x^t, \ell^t - \ell^{t-1}\rangle\\ \leq& 2\log n + \eta^2\sum_{t\in[T-1]}\|\ell^t - \ell^{t-1}\|_\infty^2 + \eta\sum_{t\in[T-1]}\|\ell^t - \ell^{t-1}\|_\infty\\ \leq& 2\log n + (\eta + \eta^2)\sum_{t\in[T-1]}\|\ell^t - \ell^{t-1}\|_\infty.\end{aligned}$$

The first step follows from Eq. (9) and the second step follows from Eq. (10). The fifth step follows from Lemma 2.1. This finishes the proof of the lemma.

B Missing proof from Section 4

B.1 Case when the learning rate is small

We handle the case when $\eta \leq 64/(c_0\sqrt{T}) = O(1/\sqrt{T})$ with the following lemma:

Lemma B.1. *Suppose both players run vanilla Hedge on game $G_2 = (A, B_2)$ with learning rate $\eta = O(1/\sqrt{T})$. Then the regret of the first player is at least $\Omega(\sqrt{T})$ after T rounds.*

Proof. The loss of player 2 is invariant to the strategy of player 1. Thus her strategy stays at $(0.4, 0.6)$. Hence, for any $t \in [T]$, the loss for player 1 is always $\ell = (-0.2, 0.2)$ and we have

$$x^t(1) = \frac{0.4\cdot\exp(0.2\eta t)}{0.4\cdot\exp(0.2\eta t) + 0.6\cdot\exp(-0.2\eta t)} \quad \text{and}$$
$$x^t(2) = \frac{0.6\cdot\exp(-0.2\eta t)}{0.4\cdot\exp(0.2\eta t) + 0.6\cdot\exp(-0.2\eta t)}.$$

One can verify that when $t \leq 1/2\eta$, we have $x^t(1) \leq 0.5 \leq x^t(2)$. Therefore, the regret is

$$\text{regret}_T^x = \sum_{t\in[T]}\langle x^t, \ell\rangle - \sum_{t\in[T]}\ell(1) \geq \sum_{t=1}^{1/2\eta}\langle x^t, \ell\rangle - \sum_{t=1}^{1/2\eta}\ell(1) \geq 0 + \frac{1}{2\eta}\cdot 0.2 = \Omega(\sqrt{T}).$$

Thus we complete the proof. □

B.2 Case when the learning rate is large

We next work on the case when $\eta \geq 3$. Recall that we write $x_t = x^t(1)$ and $y_t = y^t(1)$.

Lemma B.2. *Suppose both players run vanilla Hedge on game $G_3 = (A, B_3)$ with learning rate $\eta \geq 3$ Then the regret of the first player is at least $\Omega(T)$ after T rounds.*

Proof. Intuitively, (A, B_3) is a cooperation game, and it is beneficial for both players if they choose to cooperate on one single action (by playing either $(1, 2)$ or $(2, 1)$). However, when the learning rate is too large, they actually mismatch in every iterations. Formally, we have

$$\begin{aligned} x_{t+1} &= \frac{x_t \cdot \exp(\eta(1-2y_t))}{x_t \cdot \exp(\eta(1-2y_t)) + (1-x_t) \cdot \exp(\eta(2y_t-1))} \\ &= \frac{x_t \cdot \exp(\eta(1-2x_t))}{x_t \cdot \exp(\eta(1-2x_t)) + (1-x_t) \cdot \exp(\eta(2x_t-1))}. \end{aligned}$$

The second step follows from $x_t = y_t$ for all t because $A = B_3$ in the game. Motivated by this, we define a sequence $a_0, a_1, \ldots$ where $a_0 = x_0 = 0.4$ and

$$a_{t+1} = \frac{(1-a_t) \cdot \exp(\eta(2a_t-1))}{a_t \cdot \exp(\eta(1-2a_t)) + (1-a_t) \cdot \exp(\eta(2a_t-1))}, \quad \text{for each } t \geq 0.$$

Then $a_t = x_t$ if t is even and $a_t = 1 - x_t$ when t is odd. Furthermore, by Claim B.3 below, we have $\eta \exp(-2\eta) \leq a_t \leq 0.4$ for all t when $\eta \geq 3$. Hence, we have

$$\text{regret}_T^x \geq \sum_{t \in [T]} \langle x^t, \ell_x^t \rangle = \sum_{t \in [T]} (2x_t - 1)^2 = \sum_{t \in [T]} (2a_t - 1)^2 \geq \Omega(T).$$

This finishes the proof of the lemma. □

Claim B.3. *When $\eta \geq 3$, we have $\eta \exp(-2\eta) \leq a_t \leq 0.4$ for all $t \geq 0$.*

Proof. We prove by induction on t. The base case holds trivially for $t = 0$. Suppose the inequality holds up to t. Then for $t + 1$, we have

$$\frac{a_{t+1}}{1-a_{t+1}} = \frac{1-a_t}{a_t} \cdot \exp\big(\eta(4a_t - 2)\big) \triangleq f(a_t).$$

By simple calculation, we know that $f(a_t)$ takes maximium at $\eta \exp(-2\eta)$ or 0.4. Thus,

$$\frac{a_{t+1}}{1-a_{t+1}} \leq \max\Big\{f(0.4), f(\eta \exp(-2\eta))\Big\} \leq \frac{2}{3},$$

which implies that $a_{t+1} \leq 0.4$. The second step above follows from

$$f(0.4) = \frac{3}{2} \cdot \exp(-0.4\eta) \leq \frac{2}{3},$$

using $\eta \geq 3$ and

$$f\big(\eta \exp(-2\eta)\big) \leq \frac{1}{\eta} \exp(2\eta) \cdot \exp\big(4\eta^2 \exp(-2\eta) - 2\eta\big) = \frac{1}{\eta} \cdot \exp\big(4\eta^2 \exp(-2\eta)\big) \leq \frac{2}{3}.$$

Moreover, $f(a_t)$ takes minimum at the smaller solution a of $4\eta a(1-a) = 1$. Thus,

$$\frac{a_{t+1}}{1-a_{t+1}} \geq \frac{1-a}{a} \cdot \exp\big(\eta(4a-2)\big) \geq \frac{4}{3} \cdot \eta \exp(-2\eta),$$

where the second step used $\exp(\eta(4a-2)) \geq \exp(-2\eta)$, $a \leq 1/2\eta$ and $a \leq 1/3$. This shows that $a_{t+1} \geq \eta \exp(-2\eta)$ using $\eta \geq 3$, and finishes the induction. □

B.3 Proof of Lemma 4.3

Note that the Matching Pennies game $G_1 = (A, B_1)$ is zero-sum. It is known (see [5]) that the KL divergence of vanilla Hedge in zero-sum games is strictly increasing. We give a careful analysis on its increment each round when playing G_1. (Recall that $x^\star = y^\star = (0.5, 0.5)$.)

Lemma B.4. *Suppose both players run vanilla Hedge with $\eta \leq 3$ on G_1. Then for each $t \geq 0$,*

$$\begin{aligned} &D_{KL}(x^\star \| x^{t+1}) + D_{KL}(y^\star \| y^{t+1}) - \big(D_{KL}(x^\star \| x^t) + D_{KL}(y^\star \| y^t)\big) \\ &\qquad \geq\ e^{-7}\eta^2 x_t(1-x_t)(2y_t-1)^2 + e^{-7}\eta^2 y_t(1-y_t)(2x_t-1)^2. \end{aligned}$$

Proof. Focusing on the first player, we have

$$
\begin{aligned}
&D_{\mathrm{KL}}(x^\star \| x^{t+1}) - D_{\mathrm{KL}}(x^\star \| x^t) \\
&= \sum_{i\in[2]} x^\star(i)\cdot \log\left(\frac{x^\star(i)}{x^{t+1}(i)}\right) - \sum_{i\in[2]} x^\star(i)\cdot \log\left(\frac{x^\star(i)}{x^{t}(i)}\right) \\
&= \sum_{i\in[2]} x^\star(i)\cdot \log\left(\frac{x^t(i)}{x^{t+1}(i)}\right) \\
&= \sum_{i\in[2]} x^\star(i)\cdot \eta\ell^t(i) + \sum_{i\in[2]} x^\star(i)\cdot \log\left(\sum_{j\in[2]} x^t(j)\cdot \exp(-\eta\ell^t(j))\right) \\
&= \log\left(\sum_{j\in[2]} x^t(j)\cdot \exp(-\eta\ell^t(j))\right) \\
&= \log\Big(x_t \cdot \exp(-\eta(2y_t-1)) + (1-x_t)\cdot \exp(-\eta(1-2y_t))\Big) \\
&\geq x_t\cdot(-\eta(2y_t-1)) + (1-x_t)\cdot(-\eta(1-2y_t)) + \frac{1}{2e^6}x_t(1-x_t)\left(e^{-\eta(2y_t-1)} - e^{-\eta(1-2y_t)}\right)^2 \\
&\geq \eta(2y_t-1)(1-2x_t) + e^{-7}\eta^2 x_t(1-x_t)(2y_t-1)^2. \qquad (11)
\end{aligned}
$$

The third step follows from the updating rule of vanilla Hedge. The fourth step uses $x^\star(1) = x^\star(2) = 0.5$ and $\ell^t(1) + \ell^t(2) = (2y_t - 1) + (1 - 2y_t) = 0$. The sixth step uses the fact that $f(x) = -\log x$ is e^{-6}-strongly convex on $(0, e^3)$. Similarly, we can prove

$$
D_{\mathrm{KL}}(y^\star \| y^{t+1}) - D_{\mathrm{KL}}(y^\star \| y^t) \geq \eta(2x_t-1)(2y_t-1) + e^{-7}\eta^2 y_t(1-y_t)(2x_t-1)^2. \qquad (12)
$$

The lemma follows by combining (11) and (12). □

We are now ready to prove Lemma 4.3.

Proof of Lemma 4.3. We first prove that within $O(1/\eta^2)$ steps, the KL divergence $D_{\mathrm{KL}}(x^\star\|x^t) + D_{\mathrm{KL}}(y^\star\|y^t)$ becomes at least 20. The proof follows directly from Lemma B.4, as for any t with $D_{\mathrm{KL}}(x^\star\|x^t) + D_{\mathrm{KL}}(y^\star\|y^t) \leq 20$, we have

$$
\begin{aligned}
&D_{\mathrm{KL}}(x^\star \| x^{t+1}) + D_{\mathrm{KL}}(y^\star \| y^{t+1}) - \big(D_{\mathrm{KL}}(x^\star \| x^t) + D_{\mathrm{KL}}(y^\star \| y^t)\big) \\
&\qquad \geq e^{-7}\eta^2 x_t(1-x_t)(2y_t-1)^2 + e^{-7}\eta^2 y_t(1-y_t)(2x_t-1)^2 \geq \Omega(\eta^2). \qquad (13)
\end{aligned}
$$

The second step follows from the fact that both x_t and y_t are bounded away from 0 and 1 given the divergence at t is at most 20; it also used $\max\{|2x_t-1|, |2y_t-1|\} \geq 0.2$ given that the divergence is strictly increasing.

Let $T_0 = O(1/\eta^2)$ be the first time when the divergence becomes at least 20. If $T/2 \leq T_0$, it follows from (13) that the divergence at T is $\Omega(T\eta^2) = \Omega(\sqrt{T}\eta)$ using the assumption that $\eta \geq 16/\sqrt{T}$. So we focus on the case $T_0 \leq T/2$ and thus, $T = T_0 + L$ with $L \geq T/2$. We prove

Claim B.5. *At round $t = T_0 + \tau^2$, the KL divergence has $D_{KL}(x^\star\|x^t) + D_{KL}(y^\star\|y^t) \geq 10^{-10}\tau\eta$.*

Setting $\tau = \sqrt{T/2}$ so that $T_0 + \tau^2 \leq T$, we have

$$
D_{\mathrm{KL}}(x^\star\|x^T) + D_{\mathrm{KL}}(y^\star\|y^T) \geq \Omega(\sqrt{T}\eta),
$$

and this finishes the proof of the lemma. □

Proof of Claim B.5. We proceed to use induction on τ. The cases with $\tau \leq 16/\eta$ holds trivially as the KL divergence at T_0 is already at least 20. For the induction step, suppose the claim holds up to k for some $k \geq 64/\eta$ at time $t_0 = T_0 + k^2$. We show that at time $T_0 + (k+1)^2$ the KL divergence is at least $10^{-10}(k+1)\eta$. Without loss of generality, we assume that $x_{t_0}, y_{t_0} \geq 0.5$; the other three cases can be handled similarly. In this region, x_t with $t = t_0 + 1, \ldots$ will keep decreasing and y_t will keep increasing, until the moment when x_t drops below 0.5.

Let t_2 denote the first round $t_2 > t_0$ such that $x_t \le 0.5$. We first show that it will take no more than $k/2$ rounds for x_t to drop below 0.5: $t_2 - t_0 \le k/2$. To this end, we use t_1 to denote the first round $t_1 \ge t_0$ such that $y_t \ge 3/4$ and note that $t_1 \le t_2$ (since otherwise at $t = t_2 - 1$, we have $1/2 \le y_t \le 3/4$ and $1/2 \le x_t \le e^6$ in order for x_t to go below $1/2$ with $\eta \le 3$ in the next round; this contradicts with the fact that the KL divergence is at least 20 after T_0).

We break the proof of $t_2 - t_0 \le k/2$ into two phases: $t_1 - t_0 \le k/4$ and $t_2 - t_1 \le k/4$.

Phase 1. First we prove that it takes no more than $k/4$ steps for y_t to get larger than $3/4$. To this end, we notice that for all $t \in [t_0 : t_1 - 1]$, we have $y_t \le 3/4$ and thus, $x_t \ge 3/4$ since the KL divergence is at least 20. During all these rounds the loss vector ℓ_y^t of the second player satisfies $\ell_y^t(1) \le -3/4 + 1/4 \le -0.5$ and $\ell_y^t(2) \ge 0.5$. Thus we have (using $0.5 \le y_{t_0} \le y_{t_1 - 1} \le 3/4$)

$$3 \ge \frac{y_{t_1-1}}{1 - y_{t_1-1}} \ge \exp\big(\eta(t_1 - t_0 - 1)\big) \cdot \frac{y_{t_0}}{1 - y_{t_0}} \ge \exp\big(\eta(t_1 - t_0 - 1)\big).$$

Thus $t_1 - t_0 \le (2/\eta) + 1 \le k/4$ using $k \ge 64/\eta$ and $\eta \le 3$.

Phase 2. Next we prove that, starting from t_1, it takes less than $k/4$ steps for x_t to drop below 0.5. Note that for each $t \in [t_1 : t_2 - 1]$, the loss vector ℓ_x^t of the first player satisfies $\ell_x^t(1) \ge 0.5$ and $\ell_x^t(2) \le -0.5$. Moreover, we assume without loss of generality that $1 - x_{t_1} \ge \exp(-(k+1)\eta/20)$; otherwise the KL divergence at t_1 is already bigger than $10^{-10}(k+1)\eta$ and we are done. Therefore,

$$1 \le \frac{x_{t_2-1}}{1 - x_{t_2-1}} \le \exp\big(-\eta(t_2 - t_1 - 1)\big) \cdot \frac{x_{t_1}}{1 - x_{t_1}} \le \exp\big(\eta(-(t_2 - t_1 - 1) + (k+1)/20)\big)$$

Thus $t_2 - t_1 \ge 1 + (k+1)/20 \le k/4$ using $k \ge 64/\eta \ge 64/3$.

Now we are at time t_2 and we examine the next $R = 3/\eta \le k/2$ rounds $[t_2 : t_2 + R]$; these are the rounds where we will gain a lot in the KL divergence. Given that x_{t_2} just dropped below $1/2$, we have $x_{t_2} \ge 0.5 \cdot \exp(-2\eta)$ and thus, for every $t \in [t_2 : t_2 + R]$,

$$x_t \ge x_{t_2} \cdot \exp(-2\eta \cdot R) \ge 0.5 \cdot e^{-12}.$$

Consequently, we have

$$\begin{aligned}
&\big(D_{\mathrm{KL}}(x^\star \| x^{t_2+R}) + D_{\mathrm{KL}}(y^\star \| y^{t_2+R})\big) - \big(D_{\mathrm{KL}}(x^\star \| x^{t_2}) + D_{\mathrm{KL}}(y^\star \| y^{t_2})\big) \\
&\quad \ge \sum_{t=t_2}^{t_2+R-1} e^{-7}\eta^2 x_t(1 - x_t)(2y_t - 1)^2 + e^{-7}\eta^2 y_t(1 - y_t)(2x_t - 1)^2 \\
&\quad \ge \sum_{t=t_2}^{t_2+R-1} e^{-7}\eta^2 x_t(1 - x_t)(2y_t - 1)^2 \ge \frac{3}{\eta} \cdot e^{-7}\eta^2 \cdot \frac{1}{4}e^{-12} \cdot \frac{1}{4} \ge 10^{-10}\eta.
\end{aligned}$$

So we conclude that after at most $k/4 + k/4 + k/2 = k$ steps, the KL divergence increase at least $10^{-10}\eta$. Thus at time $T_0 + k^2 + k \le T_0 + (k+1)^2$, the KL divergence is at least $10^{-10}k\eta + 10^{-10}\eta = 10^{-10}(k+1)\eta$. This finishes the induction and the proof of the claim. □

C Missing proof from Section 5

C.1 Proof of Lemma 5.2

Fix any swap function $\phi : [n] \to [n]$. By Lemma 2.1, every $\texttt{ALG}_j$ achieves low regret. Thus,

$$\sum_{t\in[T]} \langle q_j^t, x^t(j)\ell^t \rangle \le \sum_{t\in[T]} x^t(j) \cdot \ell^t(\phi(j)) + \frac{2\log n}{\eta} + \eta \sum_{t\in[T]} \|x^t(j)\ell^t - x^{t-1}(j)\ell^{t-1}\|_\infty^2, \quad (14)$$

where we used $x^t = Q^t x^t$, set $\ell^0 = \mathbf{0}$ and $x^0 = \mathbf{1}/n = x^1$. Consequently, we have

$$\begin{aligned}
\sum_{t\in[T]} \langle x^t, \ell^t\rangle &= \sum_{t\in[T]} \langle x^t Q^t, \ell^t\rangle = \sum_{t\in[T]}\sum_{j\in[n]} \langle x^t(j) q_j^t, \ell^t\rangle = \sum_{j\in[n]}\sum_{t\in[T]} \langle q_j^t, x^t(j)\ell^t\rangle \\
&\leq \sum_{j\in[n]} \left(\sum_{t\in[T]} x^t(j)\cdot \ell^t(\phi(j)) + \frac{2\log n}{\eta} + \eta \sum_{t\in[T]} \|x^t(j)\ell^t - x^{t-1}(j)\ell^{t-1}\|_\infty^2 \right) \\
&= \sum_{t\in[T]}\sum_{j\in[n]} x^t(j)\cdot \ell^t(\phi(j)) + \frac{2n\log n}{\eta} + \eta \sum_{t\in[T]}\sum_{j\in[n]} \|x^t(j)\ell^t - x^{t-1}(j)\ell^{t-1}\|_\infty^2
\end{aligned}$$

where the first inequality follows from (14). Furthermore, we have (using $\|\ell^t\|_\infty \leq 1$ and $\|x^t\|_1 = 1$)

$$\begin{aligned}
\sum_{j\in[n]} \|x^t(j)\ell^t - x^{t-1}(j)\ell^{t-1}\|_\infty^2 &\leq \sum_{j\in[n]} \Big(\|x^t(j)\ell^t - x^{t-1}(j)\ell^t\|_\infty + \|x^{t-1}(j)\ell^t - x^{t-1}(j)\ell^{t-1}\|_\infty \Big)^2 \\
&\leq 2\sum_{j\in[n]} \|x^t(j)\ell^t - x^{t-1}(j)\ell^t\|_\infty^2 + 2\sum_{j\in[n]} \|x^{t-1}(j)\ell^t - x^{t-1}(j)\ell^{t-1}\|_\infty^2 \\
&= 2\sum_{j\in[n]} \big(x^t(j) - x^{t-1}(j)\big)^2 \|\ell^t\|_\infty^2 + 2\sum_{j\in[n]} (x^{t-1}(j))^2 \|\ell^t - \ell^{t-1}\|_\infty^2 \\
&= 2\Big(\|x^t - x^{t-1}\|_2^2 \cdot \|\ell^t\|_\infty^2 + \|x^{t-1}\|_2^2 \cdot \|\ell^t - \ell^{t-1}\|_\infty^2 \Big) \\
&\leq 2\Big(\|x^t - x^{t-1}\|_1^2 + \|\ell^t - \ell^{t-1}\|_\infty^2 \Big)
\end{aligned}$$

We can combine all these inequalities (and note that $x^0 = x^1$) to finish the proof of the lemma.

C.2 Proof of Lemma 5.3

We start the proof of Lemma 5.3 with the following definition.

Definition C.1. *Given Markov chains $Q, Q' \in \mathbb{R}^{n\times n}$, we say Q' is $(\eta_1, \ldots, \eta_n)$-approximate to Q if $(1-\eta_i)q'_{i,j} \leq q_{i,j} \leq (1+\eta_i)q'_{i,j}$ for every $i, j \in [n]$, where we write $Q = (q_{i,j})$ and $Q' = (q'_{i,j})$.*

We are ready to state our perturbation analysis on ergodic[1] Markov chains.

Lemma C.2. *Given two ergodic Markov chains Q and Q', where Q' is $(\eta_1, \ldots, \eta_n)$-approximate to Q, the stationary distribution p, p' of Q and Q', respectively, satisfy $\|p - p'\|_1 \leq 8\sum_{i=1}^n \eta_i$.*

The proof of Lemma C.2 relies on the classical Markov chain tree theorem (see [1]). To state it we need the following definition.

Definition C.3. *Suppose Q is an ergodic Markov chain and $G = (V, E)$ with $V = [n]$ is the weighted directed graph associated with Q. We say a subgraph T of G is a* directed tree rooted at $i \in [n]$ *if (1) T does not contain any cycles and (2) Node i has no outgoing edges, while every other node $j \in [n]$ has exactly one outgoing edge. For each node $i \in [n]$, we write $\mathcal{T}_i$ to denote the set of all directed trees rooted at node i. We further define*

$$\Sigma_i = \sum_{T\in\mathcal{T}_i} \prod_{(a,b)\in T} q_{a,b} \quad \textit{and} \quad \Sigma = \sum_{i\in[n]} \Sigma_i,$$

i.e., the weight of T is the product of its edge weights and Σ_i is the sum of weights of trees in $\mathcal{T}_i$.

We can now formally state the Markov chain tree theorem.

Theorem C.4 (Markov chain tree theorem; see [1])**.** *Suppose Q is an erogidc Markov chain and p is its stationary distribution. Then we have $p_i = \Sigma_i/\Sigma$ for every $i \in [n]$.*

We now use the Markov chain tree theorem to prove Lemma C.2.

[1]Note that Q^t used in `BM-Optimistic-Hedge` is always ergodic.

Proof of Lemma C.2. Note that the lemma is trivial when $\sum_{i=1}^n \eta_i > 1/4$ so we assume without loss of generality that $\sum_{i=1}^n \eta_i \le 1/4$. For any $i \in [n]$, we have

$$\Sigma_i = \sum_{T\in\mathcal{T}_i} \prod_{(a,b)\in T} q_{a,b} \le \sum_{T\in\mathcal{T}_i} \prod_{(a,b)\in T} (1+\eta_a)\widetilde{q}_{a,b}$$

$$\le \prod_{j\in[n]} (1+\eta_j) \sum_{T\in\mathcal{T}_i} \prod_{(a,b)\in T} q'_{a,b} = \prod_{j\in[n]} (1+\eta_j)\cdot \Sigma'_i \le \left(1 + 2\sum_{j\in[n]} \eta_j\right)\Sigma'_i. \tag{15}$$

The third step holds because for any tree $T \in \mathcal{T}_i$, each node, other than node i, appears exactly once as a when calculating the weight of T. The last step follows from the fact that when $\sum_{i=1}^n \eta_i \le 1/4$,

$$\prod_{j\in[n]} (1+\eta_j) \le \prod_{j\in[n]} e^{\eta_j} = e^{\sum_{j\in[n]}\eta_j} \le 1 + 2\sum_{j\in[n]}\eta_j.$$

Similarly, we have

$$\Sigma_i \ge \sum_{T\in\mathcal{T}_i} \prod_{(a,b)\in T} (1-\eta_a)\widetilde{q}_{a,b} \ge \prod_{j\in[n]} (1-\eta_j)\cdot \Sigma'_i \ge \left(1 - 2\sum_{j\in[n]} \eta_j\right)\Sigma'_i. \tag{16}$$

The last inequality holds since, for $\sum_{j=1}^n \eta_j \le 1/2$, we have

$$\prod_{j\in[n]} (1-\eta_j) \ge \prod_{j\in[n]} e^{-2\eta_j} = \exp\left(-2\sum_{j\in[n]}\eta_j\right) \ge 1 - 2\sum_{j\in[n]}\eta_j.$$

Since $\Sigma = \sum_i \Sigma_i$, we have $(1 - 2\sum_i \eta_i)\,\widetilde{\Sigma} \le \Sigma \le (1 + 2\sum_i \eta_i)\,\widetilde{\Sigma}$. Applying Theorem C.4,

$$\|p - p'\|_1 = \sum_{i\in[n]} |p_i - p'_i| = \sum_{i\in[n]} \left|\Sigma_i/\Sigma - \Sigma_i{}'/\Sigma'\right| \le \sum_{i\in[n]} \left|\Sigma_i/\Sigma - \Sigma_i/\Sigma'\right| + \sum_{i\in[n]} \left|\Sigma_i/\Sigma' - \Sigma'_i/\Sigma'\right|$$

$$\le \sum_{i\in[n]} \frac{2\sum_{i=1}^n \eta_i}{1 - 2\sum_{i=1}^n \eta_i} \left|\Sigma_i/\Sigma\right| + \sum_{i\in[n]} 2 \sum_{j\in[n]} \eta_j \cdot \left|\Sigma'_i/\Sigma'\right| \le 6\sum_{i\in[n]} \eta_i.$$

This finishes the proof of the lemma. □

Finally we prove Lemma 5.3:

Proof of Lemma 5.3. We start with the following claim, which states that entries of Q^t and Q^{t-1} only differs by a small multiplicative factor.

Claim C.5. *Suppose that the learning rate $\eta \le 1/6$ and let $x^0 = \mathbf{1}/n = x^1$. Then for any $t \ge 2$, Q^t is a $(\eta_1, \ldots, \eta_n)$-approximate to Q^{t-1}, where $\eta_j = 2\eta x^{t-2}(j) + 4\eta x^{t-1}(j)$ for each $j \in [n]$.*

Combing Claim C.5 and Lemma C.2, we have

$$\|x^t - x^{t-1}\|_1 \le 8\sum_{j\in[n]} \eta_j = 8 \sum_{j\in[n]} \left(2x^{t-2}(j) + 4x^{t-1}(j)\right)\eta = 48\eta.$$

This finishes the proof of Lemma 5.3. □

Proof of Claim C.5. Let $x^0 = \mathbf{1}/n = x^1$. By the updating rule of optimisitic Hedge, we have for any $t \ge 2$, $i, j \in [n]$ that

$$\begin{aligned}
q_j^t(i) &= \frac{\exp(-\eta(2x^{t-1}(j)\ell^{t-1}(i) - x^{t-2}(j)\ell^{t-2}(i)))\cdot q_j^{t-1}(i)}{\sum_{k\in[n]} \exp(-\eta(2x^{t-1}(j)\ell^{t-1}(k) - x^{t-2}(j)\ell^{t-2}(k)))\cdot q_j^{t-1}(k)} \\
&\le \frac{\exp(\eta x^{t-2}(j))\cdot q_j^{t-1}(i)}{\sum_{k\in[n]} \exp(-2\eta x^{t-1}(j))\cdot q_j^{t-1}(k)} \\
&= \exp\left(\eta x^{t-2}(j) + 2\eta x^{t-1}(j)\right)\cdot q_j^{t-1}(i) \\
&\le (1 + 2\eta x^{t-2}(j) + 4\eta x^{t-1}(j))\cdot q_j^{t-1}(i).
\end{aligned}$$

The second step follows from $\ell^t \in [0,1]^n$ and the last step follows from $\exp(a) \leq 1+2a$ for $a \leq 1/2$. The other side holds similarly:

$$\begin{aligned} q_j^t(i) &= \frac{\exp(-\eta(2x^{t-1}(j)\ell^{t-1}(i) - x^{t-2}(j)\ell^{t-2}(i))) \cdot q_j^{t-1}(i)}{\sum_{k\in[n]} \exp(-\eta(2x^{t-1}(j)\ell^{t-1}(k) - x^{t-2}(j)\ell^{t-2}(k))) \cdot q_j^{t-1}(k)} \\ &\geq \frac{\exp(-2\eta x^{t-1}(j)) \cdot q_j^{t-1}(i)}{\sum_{k\in[n]} \exp(\eta x^{t-2}(j)) \cdot q_j^{t-1}(k)} \\ &= \exp\big(-\eta x^{t-2}(j) - 2\eta x^{t-1}(j)\big) \cdot q_j^{t-1}(i) \\ &\geq (1 - \eta x^{t-2}(j) - 2\eta x^{t-1}(j)) \cdot q_j^{t-1}(i). \end{aligned}$$

Thus completing the proof. □

C.3 Proof of Corollary 5.4

The algorithm works as follow. We set

$$\eta = \frac{(n\log n)^{1/4}}{m^{1/2}T^{1/4}}$$

and $B_r = 1$ at initialization, for any player $i \in [m]$ and $\tau = 1, \dots, T$

1. Play x_i^t according to `BM-Optimistic-Hedge`, and receive ℓ_i^t.
2. If $\sum_{t=2}^{\tau} \|\ell_i^t - \ell_i^{t-1}\|_\infty^2 + \sum_{t=2}^{\tau} \|x_i^t - x_i^{t-1}\|_1^2 \geq B_r$.
   (a) Update $B_{r+1} = 2B_r$, $r \leftarrow r+1$, $\eta_r = \min\left\{\sqrt{\frac{n\log n}{B_r}}, \eta\right\}$.
   (b) Start a new run of `BM-Optimistic-Hedge` with learning rate η_r.

For any round r, we use T_r to denote its final iteration and

$$I_r = \sum_{t=T_{r-1}+1}^{T_r} \|x_i^t - x_i^{t-1}\|_1^2 + \sum_{t=T_{r-1}+1}^{T_r} \|\ell_i^t - \ell_i^{t-1}\|_\infty^2.$$

Then we have

$$\begin{aligned} \text{swap-regret}_{T_{r-1}+1:T_r} &\leq \frac{2n\log n}{\eta_r} + 2\eta_r \left(\sum_{t=T_{r-1}+1}^{T_r} \|x_i^t - x_i^{t-1}\|_1^2 + \sum_{t=T_{r-1}+1}^{T_r} \|\ell_i^t - \ell_i^{t-1}\|_\infty^2 \right) \\ &\leq 2(n\log n)^{3/4} \cdot T^{1/4}m^{1/2} + 2\sqrt{n\log n B_r} + 2\eta_r \cdot I_r \\ &\leq 2(n\log n)^{3/4} \cdot T^{1/4}m^{1/2} + 2\sqrt{n\log n B_r} + 2\sqrt{2n\log n I_r} \\ &\leq 2(n\log n)^{3/4} \cdot T^{1/4}m^{1/2} + 4\sqrt{2n\log n I_r} \\ &\leq 2(n\log n)^{3/4} \cdot T^{1/4}m^{1/2} + 4\sqrt{2n\log n} \cdot \sqrt{\left(\sum_{t=2}^{T} \|x_i^t - x_i^{t-1}\|_1^2 + \sum_{t=2}^{T} \|\ell_i^t - \ell_i^{t-1}\|_\infty^2 \right)} \end{aligned}$$

The first step follows from Lemma 5.2, the second step follows from the definition of I_r and the fact

$$\frac{1}{\eta_r} \leq \frac{1}{\eta} + \sqrt{\frac{B_r}{n\log n}} = \frac{m^{1/2}T^{1/4}}{(n\log n)^{1/4}} + \sqrt{\frac{B_r}{n\log n}}$$

The third step follows from $\eta_r \leq \sqrt{\frac{n\log n}{B_r}} \leq \sqrt{\frac{n\log n}{I_r/2}}$, and the last step comes from $\sqrt{B_r} \leq \sqrt{2I_r}$.

Since the number of round is at most $O(\log T)$, we have

$$\text{swap-regret}_T \leq \log T \left(2(n\log n)^{3/4}T^{1/4}m^{1/2} + 4\sqrt{2n\log n} \cdot \sqrt{2\left(\sum_{t=1}^{T} \|x_i^t - x_i^{t-1}\|_1^2 + \sum_{t=1}^{T} \|\ell_i^t - \ell_i^{t-1}\|_\infty^2 \right)} \right)$$

If all players adopt the algorithm, then we know their learning rate is no greater than $\eta = \frac{(n\log n)^{1/4}}{m^{1/2}T^{1/4}}$, thus we know $\|x_i^t - x_i^{t-1}\|_1 \leq O(\eta) = O\left(\frac{(n\log n)^{1/4}}{m^{1/2}T^{1/4}}\right)$ (see Lemma 5.3) and $\|\ell_i^t - \ell_i^{t-1}\|_\infty \leq \sum_{j\neq i}\|x_j^t - x_j^{t-1}\|_1 \leq m\cdot O(\eta) = O\left(\frac{m^{1/2}(n\log n)^{1/4}}{T^{1/4}}\right)$. Thus the swap regret is at most

$$O\left((n\log n)^{3/4}m^{1/2}T^{1/4}\log T\right).$$

If the player is facing an adversary, then $\|x_i^t - x_i^{t-1}\|_1 \leq 2$ and $\|\ell_i^t - \ell_i^{t-1}\|_\infty \leq 1$, thus we conclude its regret is at most

$$O\left(\sqrt{n\log nT}\log T + (n\log n)^{3/4}m^{1/2}T^{1/4}\log T\right).$$

D Another no swap regret algorithm

We prove the optimistic variant of a folklore algorithm, originally appeared in [7], could also achieve fast convergence of swap regret. Our perturbation analysis again plays a key role in the regret analysis.

Define Φ to be all swap functions that map $[n]$ to $[n]$. We have $|\Phi| = n^n$. For any $\phi\in\Phi$, define the swap matrice S^ϕ as: $S^\phi_{i,j} = 1$ if $\phi(i) = j$ and $S^\phi_{i,j} = 0$ otherwise. It is easy to see that S^ϕ contains exactly one 1 each row.

[7] treats each swap matrice S^ϕ as an expert, and run Hedge algorithm on all n^n swap matrices. At time t, the output strategy p^t is determined by these experts via solving a fix point problem[2]. The optimisitic variant of [7] is shown in Algorithm 1. We first analysis the regret,

Algorithm 1

1: **for** $t = 1, 2, \ldots,$ **do**

2: Play p^t and receive the loss vector l^t.

3: Update

$$q^{t+1}(\phi) = \frac{x^t(\phi)\exp(-\eta(2x^tS^\phi\ell^t - x^{t-1}S^\phi\ell^{t-1}))}{\sum_{\phi\in\Phi}x^t(\phi)\exp(-\eta(2x^tS^\phi\ell^t - x^{t-1}S^\phi\ell^{t-1}))} \quad \forall\phi\in\Phi$$

4: Compute $x^{t+1} = x^{t+1}Q^{(t+1)}$, where

$$Q^{(t+1)} = \sum_{\phi\in\Phi} q^{t+1}(\phi)S^\phi.$$

5: **end for**

Lemma D.1. *Algorithm 1 achieves regret*

$$\text{swap-regret}_T \leq \frac{n\log n}{\eta} + 2\eta\sum_{t=2}^T\|x^t - x^{t-1}\|_1^2 + 2\eta\sum_{t=2}^T\|\ell^t - \ell^{t-1}\|_\infty^2.$$

[2]The algorithm is not efficient in general. However, we can turn it into an effiecient one by considering only n^2 swap matrices that are equal to indentical mapping *except* for one coordinate. The regret bound will only blow up by a $\sqrt{n}$ factor.

Proof. According to the updating rule, for any $\phi \in \Phi$, we have

$$\begin{aligned}
\text{swap-regret}_T &= \sum_{t=2}^{T} \langle x^t, \ell^t \rangle - \max_{\phi \in \Phi} \sum_{t=2}^{T} x^t S^{\phi} \ell^t \\
&= \sum_{t=2}^{T} \langle x^t Q^{(t)}, \ell^t \rangle - \max_{\phi \in \Phi} \sum_{t=2}^{T} x^t S^{\phi} \ell^t \\
&= \sum_{t=2}^{T} \sum_{\phi \in \Phi} x^t (q^t(\phi) S^{\phi}) \ell^t - \max_{\phi \in \Phi} \sum_{t=2}^{T} x^t S^{\phi} \ell^t \\
&= \sum_{t=2}^{T} \sum_{\phi \in \Phi} q^t(\phi) \cdot x^t S^{\phi} \ell^t - \max_{\phi \in \Phi} \sum_{t=2}^{T} x^t S^{\phi} \ell^t \\
&\leq \frac{n \log n}{\eta} + \eta \sum_{t=2}^{T} \max_{\phi \in \Phi} \left| x^t S^{\phi} \ell^{t-1} - x^{t-1} S^{\phi} \ell^{t-1} \right\|^2 \\
&\leq \frac{\log n}{\eta} + 2\eta \sum_{t=2}^{T} \|x^t - x^{t-1}\|_1^2 + 2\eta \sum_{t=2}^{T} \|\ell^t - \ell^{t-1}\|_\infty^2.
\end{aligned}$$

The fifth step follows the regret bound of optimistic Hedge and the last step follows from the fact that for any $\phi \in \Phi$,

$$\begin{aligned}
\left| x^t S^{\phi} \ell^t - x^t S^{\phi} \ell^t \right|^2 &= \left| x^t S^{\phi} \ell^t - x^{t-1} S^{\phi} \ell^t + x^{t-1} A_{\phi} \ell^t - x^{t-1} S^{\phi} \ell^{t-1} \right|^2 \\
&\leq 2 \left| x^t S^{\phi} \ell^t - x^{t-1} S^{\phi} \ell^t \right|^2 + 2 |x^{t-1} S^{\phi} \ell^t - x^{t-1} S^{\phi} \ell^{t-1} \right|^2 \\
&= 2 \langle x^t - x^{t-1}, S^{\phi} \ell^t \rangle + 2 \langle x^{t-1} S^{\phi}, \ell^t - l^{t-1} \rangle \\
&\leq 2 \|x^t - x^{t-1}\|_1^2 \|S^{\phi} \ell^t\|_\infty^2 + 2 \|x^{t-1} S^{\phi}\|_1 \|\ell^t - \ell^{t-1}\|_\infty^2 \\
&\leq 2 \|x^t - x^{t-1}\|_1^2 + 2 \|\ell^t - \ell^{t-1}\|_\infty^2.
\end{aligned}$$

Thus completing the proof. □

It remains to show that the environment is stable. Again, since x^t is the stationary distribution of $Q^{(t)}$, we only need some perturbation analysis on $Q^{(t)}$. In particular, we have

Lemma D.2. *For any t, $Q^{(t)}$ is $(6\eta, \ldots, 6\eta)$ approximate to $Q^{(t+1)}$.*

Proof. For any ϕ, we have

$$\begin{aligned}
q^{t+1}(\phi) &= \frac{q^t(\phi) \exp(-\eta(2x^t A_{\phi} \ell^t - x^{t-1} A_{\phi} \ell^{t-1}))}{\sum_{\phi \in \Phi} q^t(\phi) \exp(-\eta(2x^t A_{\phi} \ell^t - x^{t-1} A_{\phi} \ell^{t-1}))} \\
&\leq \frac{q^t(\phi) \exp(\eta)}{\sum_{\phi \in \Phi} q^t(\phi) \exp(-2\eta)} \\
&\leq (1 + 6\eta) q^t(\phi)
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
q^{t+1}(\phi) &= \frac{q^t(\phi) \exp(-\eta(2x^t A_{\phi} \ell^t - x^{t-1} A_{\phi} \ell^{t-1}))}{\sum_{\phi \in \Phi} q^t(\phi) \exp(-\eta(2x^t A_{\phi} \ell^t - x^{t-1} A_{\phi} \ell^{t-1}))} \\
&\geq \frac{q^t(\phi) \exp(-2\eta)}{\sum_{\phi \in \Phi} q^t(\phi) \exp(\eta)} \\
&\geq (1 - 6\eta) q^t(\phi)
\end{aligned}$$

Thus, for any $i, j \in [n]$, we have

$$Q_{i,j}^{(t+1)} = \sum_{\phi \in \Phi} q^{t+1}(\phi) S_{i,j}^{\phi} \leq (1 + 6\eta) \sum_{\phi \in \Phi} q^t(\phi) S_{i,j}^{\phi} = (1 + 6\eta) Q_{i,j}^{(t)}$$

and

$$Q_{i,j}^{(t+1)} = \sum_{\phi\in\Phi} q^{t+1}(\phi) S_{i,j}^{\phi} \geq (1-6\eta) \sum_{\phi\in\Phi} q^{t}(\phi) S_{i,j}^{\phi} \geq (1-6\eta) Q_{i,j}^{(t)}$$

Thus we conclude $Q^{(t)}$ is $(6\eta, \dots, 6\eta)$ approximate to $Q^{(t+1)}$. □

Combining the above results, we have

Theorem D.3. *Suppose every player uses Algorithm 1 and choose $\eta = O\left((\frac{\log n}{nm^2T})^{1/4}\right)$, then each individual's swap regret is at most $O\left(m^{1/2}n^{5/4}(\log n)^{3/4}T^{1/4}\right)$.*

Proof. By Lemma D.1, for any palyer $i \in [m]$, we have

$$\begin{aligned}
\text{swap-regret}_T &\leq \frac{n\log n}{\eta} + 2\eta \sum_{t=2}^{T} \|x_i^t - x_i^{t-1}\|_1^2 + 2\eta \sum_{t=2}^{T} \|\ell_i^t - \ell_i^{t-1}\|_\infty^2 \\
&\leq \frac{n\log n}{\eta} + 2\eta \sum_{t=2}^{T} \|x^t - x^{t-1}\|_1^2 + 2m\eta \sum_{t=2}^{T} \sum_{j\neq i} \|x_j^t - x_j^{t-1}\|_1^2
\end{aligned}$$

where w^t denotes the other player's strategy. Moreover, since $Q^{(t-1)}$ is $(6\eta, \dots, 6\eta)$ approximates to $Q^{(t)}$, we know

$$\|x_i^t - x_i^{t-1}\|_1 \leq 8 \cdot \sum_{i=1}^{n} 6\eta = O(n\eta)$$

holds for any i. Thus we have

$$\begin{aligned}
\text{swap-regret}_T &\leq \frac{n\log n}{\eta} + 2\eta \sum_{t=2}^{T} \|x^t - x^{t-1}\|_1^2 + 2m\eta \sum_{t=2}^{T} \sum_{j\neq i} \|x_j^t - x_j^{t-1}\|_1^2 \\
&\leq \frac{n\log n}{\eta} + O(\eta^3 n^2 m^2 T).
\end{aligned}$$

Choosing $\eta = O\left((\frac{\log n}{nm^2T})^{1/4}\right)$, the regret is

$$\text{swap-regret}_T = O\left(n^{5/4}(\log n)^{3/4}T^{1/4}m^{1/2}\right).$$

□

E Price of anarchy

In this section, we show that a large class of no swap regret algorithm satisfies the *low approximate regret* property (see Definition E.2). Thus when all players adopt such algorithm, they experience fast convergence to an approximately optimal social welfare in *smooth games* (see Definition E.1). In particular, we show that the average social welfare converges to an approximately optimal welfare at rate $O(1/T)$. The proof in this section is straightforward, our aim is to point out that such fast convergence rate generally holds for no-swap regret algorithms. We first introduce the smooth game. Recall $\mathcal{L}(\mathbf{x}) = \sum_{i\in[m]} \mathcal{L}_i(\mathbf{x})$ is the summation of each individual's loss under strategy profile $\mathbf{x}$.

Definition E.1 (Smooth game)**.** *A cost minimization game is (λ, μ)-smooth if for all strategy profiles $\mathbf{x}$ and $\mathbf{x}^\star$, $\sum_i \mathcal{L}_i(x_i^\star, x_{-i}) \leq \lambda \cdot \mathcal{L}(\mathbf{x}^\star) + \mu \cdot \mathcal{L}(\mathbf{x})$.*

A wide range of games belongs to smooth game, including routing games, auctions, etc. We refer interested reader to [25] for detailed coverage.

We next introduce the definition of low approximate regret.

Definition E.2 (Low approximate regret [14])**.** *A learning algorithm satisfies the low approximate regret property for given parameters* $(\epsilon, A(n))$ *, if*

$$(1-\epsilon)\sum_{t=1}^{T}\langle x^t, \ell^t\rangle \leq \min_i L(i) + \frac{A(n)}{\epsilon}.$$

Lemma E.3. *The BM reduction transfers the low approximate regret property. In particular, if we reduce from a no external regret algorithm satisfying low approximate regret with*$(\epsilon, A(n))$*, then the no swap regret algorithm satisfies low approximate regret with* $(\epsilon, nA(n))$*.*

Proof. For any fixed i, using the low approximate regret property, we know

$$(1-\epsilon)\sum_{t=1}^{T}\langle q_j^t, x^t(j)\ell_t\rangle \leq \min_{i'}\sum_{t=1}^{T} x^t(j)\ell^t(i') + \frac{A(n)}{\epsilon} \leq \sum_{t=1}^{T} x^t(j)\ell_t(i) + \frac{A(n)}{\epsilon}.$$

Consequently, we have

$$\begin{aligned}(1-\epsilon)\sum_{t=1}^{T}\langle x^t, \ell^t\rangle &= (1-\epsilon)\sum_{t=1}^{T}\langle x^t Q^{(t)}, \ell^t\rangle \\ &= (1-\epsilon)\sum_{t=1}^{T}\sum_{j=1}^{n}\langle x^t(j) q_j^t, \ell^t\rangle \\ &= (1-\epsilon)\sum_{j=1}^{n}\sum_{t=1}^{T}\langle q_j^t, x^t(j)\ell^t\rangle \\ &\leq \sum_{j=1}^{n}\left(\sum_{t=1}^{T} x^t(j)\ell^t(i) + \frac{A(n)}{\epsilon}\right) \\ &= \sum_{t=1}^{T}\sum_{j=1}^{n} x^t(j)\ell^t(i) + \frac{nA(n)}{\epsilon} \\ &= \sum_{t=1}^{T}\ell^t(i) + \frac{nA(n)}{\epsilon}.\end{aligned}$$

Thus concluding the proof. □

A direct corollary of Lemma E.3 and Theorem 3 in [14] is

Theorem E.4. *In a* (λ, μ)*-smooth game, if all players use no swap regret algorithm generated from BM reduction and a no external regret algorithm satisfying low approximate regret property with parameter* ϵ *and* $A(n) = \log n$*, then we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathcal{L}(\mathbf{x}_t) \leq \frac{\lambda}{1-\mu-\epsilon}\cdot \mathrm{OPT} + \frac{m}{T}\cdot\frac{1}{1-\mu-\epsilon}\cdot\frac{n\log n}{\epsilon}.$$

where OPT *denotes the optimal social welfare, i.e.,* $\min_{\mathbf{x}}\mathcal{L}(\mathbf{x})$*.*