Prediction Accuracy of Learning in Games : Follow-the-Regularized-Leader meets Heisenberg

Yi Feng [1] **Georgios Piliouras** [2] **Xiao Wang** [1 3]

Abstract

We investigate the accuracy of prediction in deterministic learning dynamics of zero-sum games with random initializations, specifically focusing on observer uncertainty and its relationship to the evolution of covariances. Zero-sum games are a prominent field of interest in machine learning due to their various applications. Concurrently, the accuracy of prediction in dynamical systems from mechanics has long been a classic subject of investigation since the discovery of the Heisenberg Uncertainty Principle. This principle employs covariance and standard deviation of particle states to measure prediction accuracy. In this study, we bring these two approaches together to analyze the Follow-the-Regularized-Leader (FTRL) algorithm in two-player zero-sum games. We provide growth rates of covariance information for continuous-time FTRL, as well as its two canonical discretization methods (Euler and Symplectic). A Heisenberg-type inequality is established for FTRL. Our analysis and experiments also show that employing Symplectic discretization enhances the accuracy of prediction in learning dynamics.

1. Introduction

In recent years understanding the behavior of learning algorithms in repeated games has attracted increasing interests from the machine learning community (Lanctot et al., 2017; Yang & Wang, 2020). Follow-the-Regularized-Leader (FTRL) algorithm (Abernethy et al., 2009; Shalev-Shwartz et al., 2012), arguably the most well known class of no-regret dynamics, plays a prominent role in analysis of behavior of learning algorithms. The dynamics of such online learning algorithm in zero-sum games has been a particularly intense object of study as zero-sum games related to numerous recent applications and advances in AI such as, achieving super-human performance in Go (Silver et al., 2016), Poker (Brown & Sandholm, 2018) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to name a few.

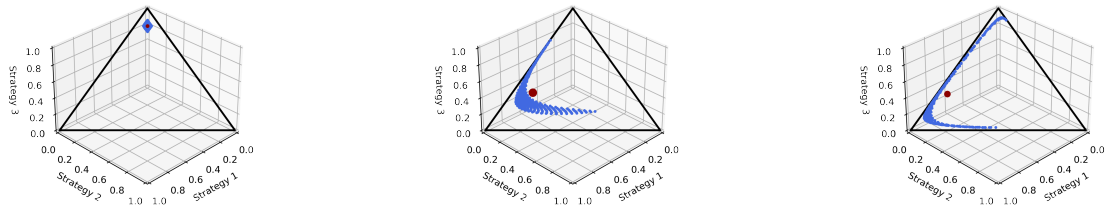
Predicting players' long term behaviors in a repeated game is a fundamental and challenging problem (Nachbar, 1997; Cesa-Bianchi & Lugosi, 2006). The conventional wisdom in this regard is that players' strategies will eventually converge to some equilibrium. However, recent studies have shown that such a belief usually fails, in particular, FTRL dynamics do not converge in zero-sum games and exhibit complex behaviors such as recurrence and divergence (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018). Therefore, one needs to predict the future behaviors of players by tracing their day-to-day (a.k.a. last-iterate) behaviors (Daskalakis et al., 2018; Gidel et al., 2019a;b). By modeling the learning dynamics as a deterministic dynamical system, an observer with the computational ability to trace the learning dynamics can accurately predict future player states by knowing their *exact* current states.

However, in practice, the observer may have some uncertainty about the current states of players. This kind of uncertainty is common both from a game-theoretic perspective (i.e., the unknown external action on agents' preference) as well as from a Machine Learning perspective (i.e., system initialization by sampling from a distribution and noise introduced during training). Thus, the observer may hope that *slightly inaccurate* knowledge of current conditions can lead to *slightly inaccurate* prediction. For example, by tracking the expectation of the evolving distribution along the learning dynamics that model his uncertainty, satisfactory predictions can be obtained.

Unfortunately, this hope didn't materialize as expected. One remarkable aspect of several learning dynamics, like Multiplicative Weights Updates (FTRL with negative entropy regularizer), is that even slight initial deviation can give rise to significantly divergent strategy trajectories for players over extended periods (Sato et al., 2002; Vilone et al., 2011; Galla & Farmer, 2013). Figure 1 illustrates that in a repeated Rock-Paper-Scissor game, the evolution of many

[1]Shanghai University of Finance and Economics, Shanghai, China [2]Google DeepMind, London, United Kingdom [3]Key Laboratory of Interdisciplinary Research of Computation and Economics, China. Correspondence to: Xiao Wang <wangxiao@sufe.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) t = 0, variance = 0.00020781 (b) t = 50, variance = 0.01038411 (c) t = 80, variance = 0.05732407

(d) t = 130, variance = 0.09969448 (e) t = 150, variance = 0.07895284 (f) t = 200, variance = 0.08126675

Figure 1: Evolution of 400 initial conditions (blue points) for one player using (AltMWU) in R-P-S game, the red point is the expectation of blue points, the variance is calculated on the first pure strategy. At time $t = 0$, these point are sampled in a small square. As time evolves, it appears that these points are randomly located on the simplex and the variance is magnified by a factor of 400. Moreover, the expectation (red point in the figures) cannot accurately predict future outcomes, as sample points in subsequent times may deviate significantly from the expected value. **A demo animation can be found here.**

orbits starting from a small region when players use Alternating Multiplicative Weights Update. Figure 1 shows that even small initial uncertainty can be amplified and make the accurate prediction difficult. Moreover, tracking the expectation fails to accurately prediction as a large portion of points will deviate significantly from the expectation; in terms of statistics, the **(co)variance** of the distribution can be large over time, hindering accurate predictions of players' future behaviors. Motivated by this example, we naturally formulate the following question:

How to track the accuracy of prediction in learning dynamics?

The most relevant paper in this direction is (Cheung et al., 2022), which utilized *differential entropy* as their metric of uncertainty and demonstrated that it grows linearly fast in two-player zero-sum games, quantifying the amount of excess information an observer must gather to keep track of the uncertain system evolution. However, as we will show in following, differential entropy cannot capture the uncertainty evolution of the *alternating* update rule of game dynamics, such as the alternating MWU in Figure 1.

In this work, we will study the evolution of *covariance (standard deviation)* associated with related random variables that govern the dynamics of FTRL, utilizing the framework of the Hamiltonian formulation of FTRL (Bailey & Piliouras, 2019; Wibisono et al., 2022). This perspective studies the evolution of covariance closely related to the well known *Heisenberg Uncertainty Principle* in quantum mechanics, which states that the covariance of the momentum and position of a microscopic particle cannot be small at the same time, i.e., $\Delta p \Delta q \geq$ some positive constant. The Hamiltonian formulation of FTRL endows the *cumulative strategy* and *cumulative payoff* of each agent the roles of position q and momentum p of each particle, and these quantities completely determine the dynamics of the game. Once the initialization is randomized, the deterministic learning dynamics still makes the cumulative strategy and payoff random variables whose mean and covariance matrix is related to that of the initialization. As what is implied from Heisenberg Uncertainty Principle (Busch et al., 2007), information on covariance (standard deviation) measures the *accuracy of prediction* in learning dynamics. We will also demonstrate that this analogy is not vague; similar phenomena to the Heisenberg Uncertainty Principle already exist in FTRL dynamics.

Our contributions. We highlight our main contributions as follows:

- We prove that differential entropy remains constant when two players alternatively update their strategies. Therefore, it is not a strong concept for capturing the evolution

of uncertainty in alternating update, see Proposition 4.1.

- We propose the covariance matrix as an uncertainty measurement which captures both simultaneous and alternating updates, with rate of increasing calculated concretely in Euclidean regularized FTRL, see Theorem 5.1. Our results imply a separation between simultaneous and continuous time/alternating plays. As an immediate application, Corollary 5.3 provides a prediction with quantitative description of risk, i.e., probability of deviating from expectation up to time t;
- For FTRL with general regularizers, a Heisenberg type inequality on variances of cumulative strategy and payoff is obtained, i.e., $\Delta X_{i,\alpha}\Delta y_{i,\alpha} \geq$ positive constant. Similar to the Heisenberg Uncertainty Principle in quantum mechanics, this inequality indicates a trade-off between prediction accuracy in strategy spaces versus payoff spaces for game dynamics, see Theorem 5.4. In Figure 2 we present an example to illustrate this point.

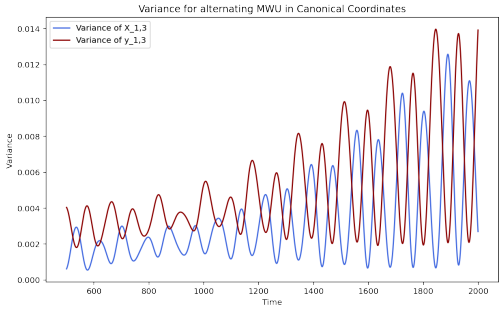


Figure 2: The two curves represent the evolution of $\Delta X_{i,\alpha}$ and $\Delta y_{i,\alpha}$ on 100 random samples of two players when they use (AltMWU). When one curve is decreasing, another curve is increasing. This implies a tradeoff between accuracy in strategy spaces versus payoff spaces.

Technical innovations. The techniques used in drawing the above conclusions come from different areas. The theoretical framework for analyzing the uncertainty evolution is the classic mechanical formulation of games (Bailey & Piliouras, 2019; Wibisono et al., 2022). A rigorous correspondence has been established between symplectic discretization (Haier et al., 2006) and alternating plays to study their properties. To demonstrate that differential entropy is constant in alternating plays, we utilize the volume preservation property of Symplectic discretization. Furthermore, the intuition in deriving the covariance evolution of Symplectic discretization is from (Wang, 1994), and the proof combine tools from application of matrix analysis in dynamical systems (Colonius & Kliemann, 2014). The uncertainty inequality for general FTRL is a consequence of a classic result from symplectic geometry known as non-squeezing theorem (McDuff & Salamon, 2017) and variance analysis methods from multivariate statistics.

2. Preliminaries

Learning in games. A two agent zero-sum game consists of two agent $\mathcal{N} = \{1, 2\}$, where agent i selects a strategy from the strategy space (or primal space) $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ and n_i represents the number of actions available to agent i. Typically, $\mathcal{X}_i$ is chosen to be $\mathbb{R}^{n_i}$, which we called the unconstrained zero-sum game, or it is chosen to be the simplex constrains

$$\Delta_i = \{x \mid \sum_{s=1}^{n_i} x_{i,s} = 1, x_{i,s} \geq 0\}.$$

Utilities of both agents are determined via payoff matrix $A^{(ij)} \in \mathbb{R}^{n_i \times n_j}$, and in a zero-sum game, the pay off matrix satisfy $A^{(ij)} = -A^{(ji)}$. For convenience, we will also use A to refer to $A^{(12)}$, and thus $A^{(21)} = -A^\top$. Given that agent i selects strategy $x_i \in \mathcal{X}_i \subset \mathbb{R}^{n_i}$, agent 1 receives utility $u_1(x_1, x_2) = \langle x_1, Ax_2\rangle$, and agent 2 receives utility $u_2(x_2, x_1) = -\langle x_2, A^\top x_1\rangle$. Naturally agents want to maximize their utility resulting the following max-min problem:

$$\max_{x_1 \in \mathcal{X}_1} \min_{x_2 \in \mathcal{X}_2} x_1^\top A x_2. \qquad \text{(Zero-Sum Game)}$$

Follow-the-Regularized-Leader. Follow-the-Regularized-Leader (FTRL) is a widely used class of no-regret online learning algorithms. In continuous time FTRL, at time t, agent i updates strategies $x_i(t)$ based on the cumulative payoff vector $y_i(t)$,

$$y_i(t) = y_i(0) + \int_0^t A^{(ij)} x_j(s) ds \quad \text{(Continuous FTRL)}$$
$$x_i(t) = \arg\max_{x_i \in \mathcal{X}_i} \{\langle x_i, y_i(t)\rangle - h_i(x_i)\}$$

where h_i is a strongly convex function, which is called the regularizer. It is also well known that

$$x_i(t) = \nabla h_i^* \left(y_i(t)\right), \tag{1}$$

where

$$h_i^*(y_i) = \max_{x_i \in \mathcal{X}_i} \{\langle x_i, y_i\rangle - h_i(x_i)\} \tag{2}$$

is the convex conjugate of h_i (Shalev-Shwartz & Singer, 2006). Therefore, if an observer knows the regularizer used by players, they can convert the information of $y_i(t)$ into the primal space $x_i(t)$.

Gradient descent ascent (GDA) and multiplicative weights updates (MWU) are two of the most well known special cases of FTRL algorithms. For unconstrained GDA, the

regularizers are chosen to be the Euclidean norm, i.e., $h_i(x_i) = \|x_i\|^2$ and $\mathcal{X}_i = \mathbb{R}^{n_i}$. For MWU, the regularizers are chosen to be negative entropy, i.e.,

$$h_i(x_i) = \sum_{j\in[n_i]} x_{i,j} \ln(x_{i,j})$$

and $\mathcal{X}_i$ be the simplex constrains.

In discrete time, FTRL has two kinds of implementations : simultaneous and alternating. Let $y_1^t = Ax_2^t,\ y_2^t = A^\top x_1^t$. In the case of GDA and MWU, the update rules with step size η are :

$$x_1^t = x_1^{t-1} + \eta y_1^{t-1},\ x_2^t = x_2^{t-1} - \eta y_2^{t-1}. \quad \text{(GDA)}$$
$$x_2^t = x_2^{t-1} - \eta y_2^{t-1},\ x_1^t = x_1^{t-1} + \eta y_1^t. \quad \text{(AltGDA)}$$

and

$$x_1^t = \left(\frac{x_{1,s}^{t-1} e^{\eta y_{1,s}^{t-1}}}{\sum_{j=1}^{n_1} x_{1,j}^{t-1} e^{\eta y_{1,j}^{t-1}}}\right)_s \quad x_2^t = \left(\frac{x_{2,s}^{t-1} e^{-\eta y_{2,s}^{t-1}}}{\sum_{j=1}^{n_2} x_{2,j}^{t-1} e^{-\eta y_{2,j}^{t-1}}}\right)_s \quad \text{(MWU)}$$

$$x_2^t = \left(\frac{x_{2,s}^{t-1} e^{-\eta y_{2,s}^{t-1}}}{\sum_{j=1}^{n_2} x_{2,j}^{t-1} e^{-\eta y_{2,j}^{t-1}}}\right)_s \quad x_1^t = \left(\frac{x_{1,s}^{t-1} e^{\eta y_{1,s}^{t}}}{\sum_{j=1}^{n_1} x_{1,j}^{t-1} e^{\eta y_{1,j}^{t}}}\right)_s \quad \text{(AltMWU)}$$

for $s = \{1, 2, ..., n_i\}$ and $i = 1$ or 2. Note that in simultaneous updates, both two players use the payoff feedback from round $t-1$ to update their strategies in round t, while in alternating updates, Player 2 first update his strategy x_2^t, and Player 1 subsequently updates her strategy x_1^t based on payoff feedback of x_2^t. Comparing to its simultaneous partner, a series of recent works show that alternating update rule has a slower regret growth rate (Bailey et al., 2020; Wibisono et al., 2022; Cevher et al., 2023).

Dynamical system. A system of ordinary differential equations $\dot{x} = f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a differentiable dynamical system. $f(x)$ is called the vector field of the dynamical system. If f is Lipschitz continuous, there exists a continuous map

$$\varphi(t, x_0) : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$$

such that for all $x_0 \in \mathbb{R}^n$, $\varphi(t, x_0)$ is the unique solution of the initial condition problem $\{\dot{x} = f(x), x(0) = x_0\}$. The solution $\varphi(t, x_0)$ is called a *trajectory* or *orbit* of the dynamical system.

Hamiltonian systems. A Hamiltonian system is a class of differential equations describing the evolution of momentums and positions of particles by a scalar function $H(X, Y)$ called Hamiltonian function. The state of the system, the momentum $Y = (y_1, ..., y_n)^\top$ and position $X = (x_1, ..., x_n)^\top$ evolves according to the following Hamilton's equations:

$$\frac{dx_i}{dt} = \frac{\partial H}{\partial y_i}, \quad \frac{dy_i}{dt} = -\frac{\partial H}{\partial x_i}, \quad \text{for} \quad i \in [n]. \tag{3}$$

The solution $\varphi(t, \cdot)$ of a Hamiltonian system is called a *symplectic* map which is a special case of volume-preserving maps, thus the absolute value of determinant of the Jacobian matrix equals to 1. The variables (X, Y) are also referred to the *canonical coordinates* of the system (Arnold, 2013).

2.1. Measure of Observer Uncertainty

Differential Entropy. The concept of differential entropy was introduced by Shannon (Shannon, 1948) as a measure of the uncertainty associated with a continuous probability distribution. For a random vector $X \in \mathbb{R}^n$ with probability density function $g(x)$ supported on $\mathcal{X} \subset \mathbb{R}^n$, the differential entropy of X is defined as

$$S(X) = -\int_{\mathcal{X}} g(x) \log g(x) dx. \quad \text{(Differential Entropy)}$$

Covariances of random vectors. Given a random vector $X = (x_1, ..., x_m)^\top$ such that every x_i is a random variable with finite variance and expected value, the covariance matrix $P(X) \in \mathbb{R}^{m\times m}$ of X is a symmetric and positive semi-definite square matrix whose (i, j) entry is the covariance, i.e.,

$$\mathrm{Cov}(x_i, x_j) = \mathbb{E}[(x_i - \mathbb{E}(x_i))(x_j - \mathbb{E}(x_j))].$$

Note that the diagonal elements of $P(x)$ are variances of $\{x_1, ..., x_m\}$.

In general, the differential entropy of a random variable provides a lower bound on the determinant of its covariance matrix. Precisely, if a random vector $X \in \mathbb{R}^m$ has zero mean and covariance matrix $P(X)$, then

$$S(X) \le \frac{1}{2} \log\left((2\pi e)^n \det P(X)\right).$$

3. Setup

In this section we leverage the power of the Hamiltonian formulation of game dynamics (Bailey & Piliouras, 2019; Wibisono et al., 2022). To study the strategy-payoff evolution from the perspective of each agent, it is convenient to apply the Hamiltonian formulation of the continuous time FTRL. We will establish the equivalence between discretization of the Hamiltonian system induced by continuous time FTRL and direct discretization of FTRL, where the latter leads to the GDA or MWU.

Euler discretization. Given an ordinary differential equation $\dot{x} = f(x)$ with initial condition $x(t_0)$ at time t_0, the Euler discretization begin the process by setting $x_0 = x(t_0)$,

next choose a step size η and set $t_n = t_0 + n\eta$, then the Euler discretization $\phi_\eta(\cdot)$ is defined by $x_{n+1} = \phi_\eta(x_n) = x_n + \eta f(x_n)$. The value x_{n+1} is an approximation of the solution of $\dot{x} = f(x)$ at time t_{n+1}.

Symplectic discretization. Given a Hamiltonian system as in (3), a numerical method $\phi_\eta(\cdot)$ is called a Symplectic discretization if when applied to a Hamiltonian system, the discrete flow $\phi_\eta : x \to \phi_\eta(x)$ is a Symplectic map for sufficient small step sizes. In this paper we focus on the following Symplectic discretizations: for $X = (x_1, ..., x_n)^\top, Y = (y_1, ..., y_n)^\top$,

$$\begin{aligned} Y^{t+1} &= Y^t - \eta \nabla_X H(X^t, Y^{t+1}), \\ X^{t+1} &= X^t + \eta \nabla_Y H(X^t, Y^{t+1}), \end{aligned} \qquad \text{(Type I method)}$$

or

$$\begin{aligned} X^{t+1} &= X^t + \eta \nabla_Y H(X^{t+1}, Y^t), \\ Y^{t+1} &= Y^t - \eta \nabla_X H(X^{t+1}, Y^t). \end{aligned} \qquad \text{(Type II method)}$$

Both methods are Symplectic methods, i.e., they make the map $(X^t, Y^t) \to (X^{t+1}, Y^{t+1})$ to be symplectic. More details of Symplectic method can be found in (Haier et al., 2006). Note that although both methods are generally implicit, they become explicit when the Hamiltonian function is separable, i.e., can be expressed as $H(X, Y) = f(X) + g(Y)$ with functions f and g. This property holds for the Hamiltonian formulation of FTRL dynamics.

Canonical coordinates of FTRL dynamics. In this paper we will focus on the dynamics of cumulative strategy and cumulative payoff. The cumulative strategy $X_i(t)$ of agent $i \in [2]$ is defined as follows :

$$X_i(t) = \int_0^t x_i(s) ds. \tag{4}$$

(Bailey & Piliouras, 2019) demonstrates that (Continuous FTRL) can be formulated as a Hamiltonian system through $(X_i(t), y_i(t))$ using the Hamiltonian function

$$H(X_i, y_i) = h_i^*(y_i(t)) + h_j^*(y_j(0) + A^{(ji)} X_i(t)) \tag{5}$$

for $j \neq i$ be the Hamiltonian function [1]. Thus $X_i(t)$ and $y_i(t)$ evolve according to the following Hamiltonian system

$$\frac{dX_i}{dt} = \frac{\partial H(X_i, y_i)}{\partial y_i}, \quad \frac{dy_i}{dt} = -\frac{\partial H(X_i, y_i)}{\partial X_i}. \tag{6}$$

Following the tradition of Hamiltonian mechanics, we will refer to $(X_i(t), y_i(t))$ for $i = 1$ or 2 as the *canonical coordinates* of FTRL dynamics. This can be analogously understood as the position-momentum coordinates used to describe the dynamic of a particle.

It may initially seem strange to trace the dynamics of players in a game based on their cumulative strategies/payoffs rather than their actual strategies $x_i(t)$. However, there is no loss in tracing these variables since $y_i(t)$ can be translated into $x_i(t)$ through the map $\nabla h^*(\cdot)$ introduced in (1), and $X_i(t)$ can be easily translated into $y_j(t)$ through $y_j(t) = y_j(t_0) + A^{(ji)} (X_i(t) - X_i(t_0))$.

Primal-dual correspondence via discretization. Since we use Euler and Symplectic discretization on the Hamiltonian system, which is not obviously equivalent to the conventionally natural update rules in the strategy spaces $\mathcal{X}_i$, we next establish formally that the Euler or Symplectic discretization of continuous time FTRL with Euclidean norm / negative entropy regularizer implies GDA/MWU or AltGDA/AltMWU respectively. This correspondence can be stated in the following proposition.

Proposition 3.1. *For each agent $i \in [2]$, let $\mathcal{X}_i$ denote the strategy spaces. Then following statements holds:*

- *If both players use Euclidean regularizers and $\mathcal{X}_i = \mathbb{R}^{n_i}$, then the Euler discretization of (6) is equivalent to Gradient Descent-Ascent (GDA) on the strategy spaces; and the Symplectic discretization of (6) is equivalent to Alternating Gradient Descent-Ascent (AltGDA) on the strategy spaces.*
- *If both players use entropy regularizers and $\mathcal{X}_i = \Delta_i$ be the simplex, then the Euler discretization of (6) is equivalent to Multiplicative Weights Update (MWU) on the strategy spaces; and the Symplectic discretization of (6) is equivalent to Alternating Multiplicative Weights Update (AltMWU) on the strategy spaces.*

Here equivalent means the variables (X_i^t, y_i^t) getting from the discretizations is the same as the cumulative strategy and payoff of the game dynamics.

It is known that Euler discretization significantly alters the properties of continuous system (Holmes, 2007). This results the differences between continuous FTRL and simultaneous plays. For example, continuous FTRL exhibits cycle behaviors (Mertikopoulos et al., 2018) while simultaneous plays typically diverge (Bailey & Piliouras, 2018). Compared to Euler discretization , Symplectic discretization can preserve the structure of the continuous system (Haier et al., 2006). Therefore, Proposition 3.1 suggests that

The dynamical behaviors of alternating update rules should be consistent with those of continuous learning dynamics.

An example of this phenomenon is (AltGDA) keeps the cycle behaviors of its continuous partner (Bailey et al., 2020).

[1] Intuitive explanation of this Hamiltonian function can be found in Appendix A.1.

To prove Proposition 3.1, we introduce a novel method of discretizing the continuous Hamiltonian system (6) by a combination of two types Symplectic methods, while still keep the symplectic structures on the dynamics of each agents. We believe this method is of independent interests. The detailed proofs of Proposition 3.1 are deferred to Appendix A.

Random initialization. We consider the case when noise is introduced to the canonical coordinates $(X_i(t_0), y_i(t_0))$ at time moment $t_0 > 0$ in continuous FTRL or $(X_i^{t_0}, y_i^{t_0})$ in discrete time settings. The main objective of this paper is to study the evolution of observer uncertainty given the covariance matrix P_0 of the initialization $(X_i(t_0), y_i(t_0))$ or $(X_i^{t_0}, y_i^{t_0})$. Take discrete time FTRL for example, the covariance matrix P_0 consists of variances $\mathrm{Var}(X_{i,\alpha}^{t_0})$, $\mathrm{Var}(y_{i,\alpha}^{t_0})$, and covariances $\mathrm{Cov}(X_{i,\alpha}^{t_0}, y_{i,\beta}^{t_0})$ for all $\alpha, \beta \in [n_i]$. Tracing the evolution of $\mathrm{Var}(X_{i,\alpha}^t)$ and $\mathrm{Var}(y_{i,\alpha}^t)$ in iterations, we are able to quantify how accurate the prediction will be in FTRL dynamics.

4. Deficiency of Differential Entropy

Differential entropy, as a measure of observer uncertainty, was used in studying the predictability of MWU in zero-sum games (Cheung et al., 2022). In this section we investigate the evolution of differential entropy in FTRL with differential discretization methods. In Proposition 4.1 we show that differential entropy is insufficient in capturing the uncertainty evolution for alternating plays.

Proposition 4.1. *When two players use (AltMWU) with arbitrary step size, the differential entropy of their cumulative strategy and payoff keeps constant, i.e., if $t > t_0$,*

$$S(X_i^t, y_i^t) = S(X_i^{t_0}, y_i^{t_0}). \tag{7}$$

Proposition 4.2. *When two players use (MWU) with step sizes $\eta < \min\{1, 1/\|A\|_2^2\}$, the differential entropy of their cumulative strategy and payoff has linear growth rate, i.e.,*

$$S(X_i^t, y_i^t) \geq S(X_i^{t_0}, y_i^{t_0}) + ct \tag{8}$$

where $c > 0$ is a constant determined by payoff matrix A.

Different from the proof of (Cheung et al., 2022) for the differential entropy evolution of (MWU), which relies a detailed calculation of the Jacobin map of the dynamics, our proof of Proposition 4.1 is established based on the relationship between Symplectic discretization and (AltMWU), as state in Proposition 3.1. In fact, the evolution of differential entropy is determined by the determinant of the Jacobin matrix of the update rule from (X_i^t, y_i^t) to (X_i^{t+1}, y_i^{t+1}), and in each update, the differential entropy is invariant if and only if the absolute value of this determinant equals to 1. As shown in Proposition 3.1, the update rule of (X_i^t, y_i^t) in (AltMWU) constitutes a symplectic map, and the absolute value of the determinant of Jacobin matrix for every symplectic map must equal to 1, therefore we can conclude that the differential entropy in (AltMWU) keeps a constant. The detailed proofs of Propositions 4.2 and 4.1 are deferred to Appendix B, where we also provide numerical examples for these two propositions.

5. Covariance in FTRL

In this section, we are presenting formally the evolution of covariances of cumulative strategies and payoffs. We start with continuous time FTRL with Euclidean regularizers, and proceed in considering Euler and Symplectic discretization of continuous time FTRL. In the end, for general FTRL, covariance evolution follows an inequality derived by using techniques of symplectic geometry.

5.1. Covariance evolution in Euclidean regularizer.

The evolution of covariance matrix with continuous time FTRL can be deduced from the Hamiltonian formulation of learning dynamics. The Euler discretization of each agent's continuous time FTRL exponentially amplifies the covariance in the learning process. In contrast, symplectic discretization, which has been proven equivalent to alternating update in strategies, amplify covariance of cumulative strategies polynomially and keep that of cumulative payoffs bounded. In this section we will focus on the view point of agent 1, and the same results also hold for agent 2 as they are symmetry.

Theorem 5.1. *In two-player zero-sum games, suppose both players use FTRL with Euclidean norm regularizers and unconstrained strategy sets $\mathcal{X}_i = \mathbb{R}^{n_i}$. Suppose at time $t_0 > 0$ the random cumulative strategy and payoff form a random vector $\left(X_i^{t_0}, y_i^{t_0}\right)$ with covariance matrix $P(t_0) \neq 0$. Then for all $t > t_0$ and $\alpha, \beta \in \{1, ..., n_1\}$, the covariance of (X_i^t, y_i^t) evolves in continuous and discrete FTRL according to the following :*

1. *In Euler discretization, for all $\alpha, \beta \in [n_1]$, it holds that $\mathrm{Cov}(X_{1,\alpha}^t, X_{1,\beta}^t)$, and $\mathrm{Cov}(y_{1,\alpha}^t, y_{1,\beta}^t)$ are of $\Theta\left((1+\gamma\eta^2)^{2t}\right)$, where η is the step size and γ is the maximal eigenvalue of $AA^\top$.*
2. *In continuous time and symplectic discretization, for all $\alpha, \beta \in [n_1]$, it holds that*
   - *if $AA^\top$ is non-singular, then $\mathrm{Cov}(X_{1,\alpha}^t, X_{1,\beta}^t)$ and $\mathrm{Cov}(y_{1,\alpha}^t, y_{1,\beta}^t)$ are of $\mathcal{O}(1)$.*
   - *if $AA^\top$ is singular, $\mathrm{Cov}(X_{1,\alpha}^t, X_{1,\beta}^t)$ is of $\Theta(t^2)$, $\mathrm{Cov}(y_{1,\alpha}^t, y_{1,\beta}^t)$ is of $\mathcal{O}(1)$.*

Theorem 5.1 distinguishes clearly between the covariance evolution on Euler discretization and Symplectic discretization: Euler discretization exhibits an exponential growth

rate, while Symplectic discretization maintains a quadratic growth rate that matches the growth rate of the primary continuous dynamics. Therefore, an observer can achieve higher prediction accuracy when players use Symplectic discretization. The proof of Theorem 5.1 are deferred to Appendix C, and experimental results are presented in Section 6.

The dependence on singularity of $AA^\top$ can be intuitively explained in terms of the covariance evolution of the average strategies (i.e., X^t/t). In an unconstrained zero-sum game, the average strategies of (AltGDA) will converge to some equilibrium (Gidel et al., 2019a). When $AA^\top$ is non-singular, the only equilibrium is the all-zero vector. The average strategy converges to this point for all initial points, resulting in a small covariance of average strategy and implying slow cumulative strategy growth. However, when $AA^\top$ is singular, there exist multiple distinct equilibria and the average strategy will converge towards these different equilibrium. Consequently, throughout this process, the covariance of average strategy remains substantial which leads to a large covariance of cumulative strategy.

Theorem 5.1 also implies the following corollary regarding to the covariance evolution of primal space. The proof is directly as in this case $x_i^t = y_i^t$ holds. See Appendix A.4.

Corollary 5.2. *Under the same conditions as in Theorem 5.1, the covariance of the actual strategy x_1^t evolves as following : for all $\alpha, \beta \in [n_1]$*

1. *In Euler discretization, it holds that $\mathrm{Cov}(x_{1,\alpha}^t, x_{1,\beta}^t)$ is of $\Theta\left((1+\gamma\eta^2)^{2t}\right)$.*
2. *In continuous time and symplectic discretization, it holds that $\mathrm{Cov}(x_{1,\alpha}^t, x_{1,\beta}^t)$ is of $\mathcal{O}(1)$.*

An immediate application of Theorem 5.1 is to utilize Chebyshev's inequality for predicting the quantitative measure of the risk associated with random variables that deviate significantly from the expected value.

Corollary 5.3. *Let k be any constant that is greater than 1, $X_{i,\alpha}(t)$ and $y_{i,\alpha}(t)$ are cumulative strategy and payoffs of the ith agent with respect to strategy α at time t. Then we have*

$$Pr\{|X_{i,\alpha}(t) - \mu_X(t)| > k\sqrt{ct}\} < \frac{1}{k^2}$$
$$Pr\{|y_{i,\alpha}(t) - \mu_y(t)| > k\} < \frac{c}{k^2}$$

where c is a constant, and $\mu_X(t)$ and $\mu_y(t)$ are expectations of cumulative strategy and payoffs up to time t.

5.2. Covariance evolution with general regularizer.

So far we have left the evolution of covariance in continuous time FTRL with general regularizers unaddressed. The challenge comes from the non-linearity of Hamiltonian system induced by continuous time FTRL algorithm. Suppose the integral flow of Hamiltonian system is $\phi_t(X_i(t_0), y_i(t_0))$. In the most general setting, we are able to provide a lower bound for the product of standard deviation of $X_{i,\alpha}(t)$ and $y_{i,\alpha}(t)$, i.e., $\Delta X_{i,\alpha}(t)\Delta y_{i,\alpha}(t) \geq$ constant. We state the conditions and results formally in the following Theorem.

Theorem 5.4. *Let vector $X_i(t)$ and $y_i(t)$ be cumulative strategy and payoff, for $i \in [2]$ and $\alpha \in [n_i]$. Assume that the higher order differentials of $\phi_t(\cdot)$ are bounded by some constant K, and the standard deviations $\Delta X_{i,\alpha}(t_0)$ and $\Delta y_{i,\alpha}(t_0)$ at initial time t_0 are sufficient small,[2] then for $t > t_0$ it holds that*

$$\Delta X_{i,\alpha}(t)\Delta y_{i,\alpha}(t) \geq \frac{1}{\sqrt{2}}\frac{w_L(P(t_0))}{\pi},$$

where $w_L(P(t_0))$ is the linear Gromov width of the ellipsoid defined by the initial covariance matrix $P(t_0)$.

The definition of Gromov width can be found in Appendix D.2. Note that the inequality holds trivially when there is no uncertainty, i.e., $P(t_0) = 0$, since in this case $w_L^2(P(t_0)) = 0$. The necessary background and details of proof are left in Appendix D.

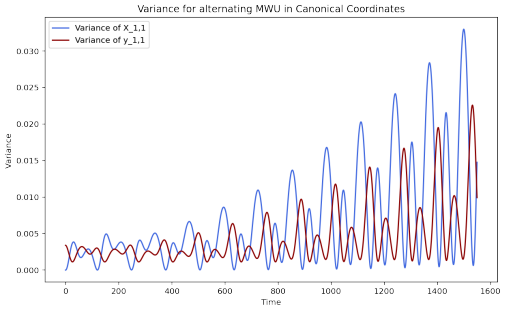


Figure 3: Covariance evolution of $\Delta(X_{1,1}^t)$ and $\Delta(y_{1,1}^t)$ when two players use (AltMWU) in a randomly generated game. Covariance is calculate based on sample variance of 500 randomly generated initial conditions.

The significance of Theorem 5.4 lies not only in providing a lower bound on the covariance evolution of general FTRL dynamics. Similar to the *Heisenberg Uncertainty Principle* in quantum mechanics, Theorem 5.4 implies that it is impossible for both $\Delta X_{i,\alpha}(t)$ and $\Delta y_{i,\alpha}(t)$ to be simultaneously small. A numerical experiment illustrating this point is presented in Figure 3. In this figure, it can be observed that when the curve representing $\Delta X_{i,\alpha}(t)$ is on an increasing stage, the curve representing $\Delta y_{i,\alpha}(t)$ is on an decreasing stage, and vice versa. Moreover, the occurrence

[2] "Sufficietly small" follows the convention in statistical modeling, e.g. p166 in (Benaroya et al., 2005), refer to Appendix D.3 for more details.

time of the local minima of $\Delta(y_{1,1}(t))$ coincides with the local maxima of $\Delta(X_{1,1}(t))$. This implies that $\Delta(X_{1,1}(t))$ and $\Delta(y_{1,1}(t))$ cannot be small at the same time.

It is interesting to ask whether similar phenomena as in Theorem 5.4 occur in the primal space $(x_1(t), x_2(t))$. In Figure 4, we demonstrate an experiment on the covariance evolution in the primal space, and observe that similar phenomena exist, at least when the number of pure strategies for players is small. Further discussions can be found in Appendix D.4.

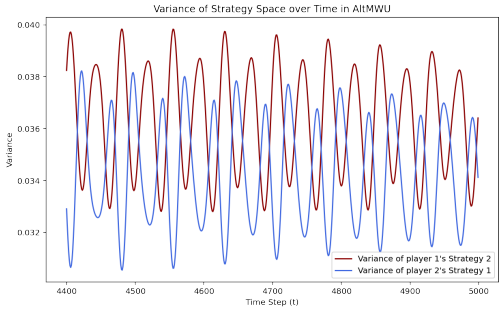


Figure 4: Covariance evolution on primal space. Calculate based on 100 random samples of the initial mixed strategies used by two players when they employ (AltMWU) in a randomly generated 3×3 game.

6. Experiments

In this section we provide numerical experiments illustrating the covariance evolution results proved for Euclidean norm regularized FTRL in Theorem 5.1. More numerical experiments on the non-singular cases are presented in Appendix E.

Continuous time FTRL. We illustrate how $\mathrm{Var}(X_{1,1}(t))$ and $\mathrm{Var}(y_{1,1}(t))$ evolve with continuous time FTRL with payoff matrices

$$\cdot A_1 = [[1,-1],[-1,1]],\ \cdot A_2 = [[1.2,-1.2],[-1,1]]$$

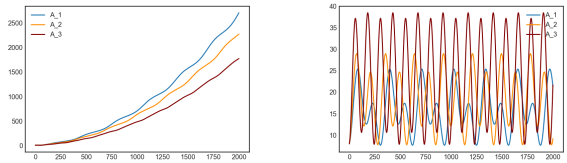
$$\cdot A_3 = [[1.5,-1.5],[-1,1]].$$

See Figure 5. In (a), the $\mathrm{Var}(X_{1,1}(t))$ has a quadratic growth rate, and in (b) $\mathrm{Var}(y_{1,1}(t))$ is bounded, which support results of continuous time part in Theorem 5.1.

Symplectic discretization. We illustrate how $\mathrm{Var}(X^t_{1,1})$ and $\mathrm{Var}(y^t_{1,1})$ evolves with symplectic discretization, the payoff matrices are given as follows:

$$\cdot B_1 = [[1,-1],[-1,1]],\ \cdot B_2 = [[1.2,-1.2],[-1,1]]$$

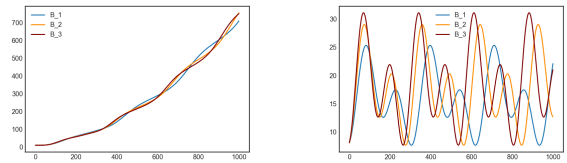
$$\cdot B_3 = [[1,-1.3],[-1,1.3]].$$



(a) $\mathrm{Var}(X_{1,1}(t))$ (b) $\mathrm{Var}(y_{1,1}(t))$

Figure 5: Variance evolution of continuous FTRL, singular cases.

See Figure 6. From the experimental results, we can see the variance behavior of symplectic discretization is same as continuous case, which support results of symplectic discretization part of Theorem 5.1.

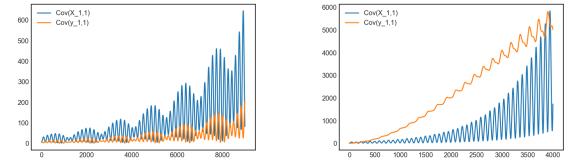


(a) $\mathrm{Var}(X^t_{1,1})$ (b) $\mathrm{Var}(y^t_{1,1})$

Figure 6: Variance evolution of Symplectic discretization.

Euler discretization. We show experimental results on $\mathrm{Var}(X^t_{1,1}), \mathrm{Var}(y^t_{1,1})$ where (X^t_1, y^t_1) evolve as Euler discretization and payoff matrices are given as follows:

- $C_1 = [[1,-1.31],[-1,1.31]]$ is singular.
- $C_2 = [[2,-1.7],[-1.7,1.5]]$ is non-singular.



(a) C_1, singular (b) C_2, non-singular

Figure 7: Variance evolution of Euler discretization.

In Figure 7 we can observe $\mathrm{Var}(X^t_{1,1})$ and $\mathrm{Var}(y^t_{1,1})$ exhibit an exponential growth rate which support the result of Euler discretization part in Theorem 5.1. As shown in Appendix 5.1, the function of the covariance evolution process contains polynomials combinations of trigonometric functions, which cause the oscillations in Figure 7.

7. Conclusion

In this paper we investigate the evolution of observer uncertainty in learning dynamics from a covariance perspective. We prove concrete rates of covariance evolution for different discretization schemes of FTRL dynamics and establish a Heisenberg-type uncertainty inequality that constrains the predictive ability of an observer. In our analysis, we leverage the techniques from symplectic geometry for analyzing the evolution of uncertainty, which to the best of our knowledge is the first of its kind. An interesting direction is to extend current results for different classes of games (e.g. potential games, multiplayer games).

Acknowledgments

Xiao Wang acknowledges Grant 202110458 from Shanghai University of Finance and Economics and support from the Shanghai Research Center for Data Science and Decision Technology.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abernethy, J. D., Hazan, E., and Rakhlin, A. Competing in the dark: An efficient algorithm for bandit linear optimization. 2009.

Arnold, V. I. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013.

Bailey, J. and Piliouras, G. Muti-agent learning in network zero-sum games is a hamiltonian system. In *AAMAS*, 2019.

Bailey, J. P. and Piliouras, G. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 321–338, 2018.

Bailey, J. P., Gidel, G., and Piliouras, G. Finite regret and cycles with fixed step-size via alternating gradient descent-ascent. In *Conference on Learning Theory*, pp. 391–407. PMLR, 2020.

Benaroya, H., Han, S. M., and Nagurka, M. *Probability models in engineering and science*, volume 192. CRC press, 2005.

Bronson, R. *Matrix methods: An introduction*. Gulf Professional Publishing, 1991.

Brown, N. and Sandholm, T. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Busch, P., Heinonen, T., and Lahti, P. Heisenberg's uncertainty principle. *Physics reports*, 452(6):155–176, 2007.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Cevher, V., Cutkosky, A., Kavis, A., Piliouras, G., Skoulakis, S., and Viano, L. Alternation makes the adversary weaker in two-player games. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Cheung, R. K., Piliouras, G., and Tao, Y. The evolution of uncertainty of learning in games. In *ICLR*, 2022.

Colonius, F. and Kliemann, W. *Dynamical systems and linear algebra*, volume 158. American Mathematical Society, 2014.

Conrad, B. Higher derivatives and taylor's formula via multilinear maps. In *http://math.stanford.edu/ conrad/diffgeomPage/handouts/taylor.pdf*.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *ICLR*, 2018.

Feutrill, A. and Roughan, M. A review of shannon and differential entropy rate estimation. *Entropy*, 23(8):1046, 2021.

Galla, T. and Farmer, J. D. Complex dynamics in learning complicated games. *Proceedings of the National Academy of Sciences*, 110(4):1232–1236, 2013.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *ICLR*, 2019a.

Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811. PMLR, 2019b.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Gromov, M. Pseudo holomorphic curves in symplectic manifolds. *Inventiones Mathematicae*, 82,307-347, 1985.

Haier, E., Lubich, C., and Wanner, G. *Geometric Numerical integration: structure-preserving algorithms for ordinary differential equations*. Springer, 2006.

Holmes, M. H. *Introduction to numerical methods in differential equations*. Springer, 2007.

Hong, Y. and Horn, R. A. The jordan cononical form of a product of a hermitian and a positive semidefinite matrix. *Linear Algebra and its Applications*, 147:373–386, 1991.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Hsiao, F.-Y. and Scheeres, D. J. Fundamental constraints on uncertainty evolution in hamiltonian systems. In *2006 American Control Conference*, pp. 6–pp. IEEE, 2006.

Hyvärinen, A. New approximations of differential entropy for independent component analysis and projection pursuit. *Advances in neural information processing systems*, 10, 1997.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

McDuff, D. and Salamon, D. *Introduction to symplectic topology*, volume 27. Oxford University Press, 2017.

Mertikopoulos, P. and Sandholm, W. H. Riemannian game dynamics. *Journal of Economic Theory*, 177:315–364, 2018.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Cycles in adversarial regularized learning. In *SODA*, 2018.

Nachbar, J. H. Prediction, optimization, and learning in repeated games. *Econometrica: Journal of the Econometric Society*, pp. 275–309, 1997.

Sato, Y., Akiyama, E., and Farmer, J. D. Chaos in learning a simple two-person game. *Proceedings of the National Academy of Sciences*, 99(7):4748–4751, 2002.

Shalev-Shwartz, S. and Singer, Y. Convex repeated games and fenchel duality. *Advances in neural information processing systems*, 19, 2006.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Vilone, D., Robledo, A., and Sánchez, A. Chaos and unpredictability in evolutionary dynamics in discrete time. *Physical review letters*, 107(3):038101, 2011.

Wang, D. Some aspects of hamiltonian systems and symplectic algorithms. *Physica D: Nonlinear Phenomena*, 73 (1-2):1–16, 1994.

Wibisono, A., Tao, M., and Piliouras, G. Alternating mirror descent for constrained min-max games. *Advances in Neural Information Processing Systems*, 35:35201–35212, 2022.

Yang, Y. and Wang, J. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

Zhou, X. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

A. Proof of Proposition 3.1

A.1. Hamiltonian formulation of FTRL

We first recall the Hamiltonian formulation of continuous FTRL in zero sum game from (Bailey & Piliouras, 2019).

The Hamiltonian function $H(X_1, y_1)$ for agent 1 is defined to be

$$H(X_1, y_1) = h_1^*(y_1(t)) + h_2^*(y_2(0) + A^{(21)}X_1(t)), \tag{9}$$

and the Hamiltonian function $H(X_2, y_2)$ for agent 2 is defined to be

$$H(X_2, y_2) = h_2^*(y_2(t)) + h_1^*(y_1(0) + A^{(12)}X_2(t)), \tag{10}$$

where h_i^* is the regularizer used by agent i, $i = 1, 2$.

Theorem 3.2 of (Bailey & Piliouras, 2019) shows the dynamical behaviors of $(X_i(t), y_i(t))$ are completely determined by these two Hamiltonian functions. A similar non-canonical Hamiltonian system formulation (also known as a Possion system) for continuous time mirror descent algorithms is also presented in (Wibisono et al., 2022). Theorem 5.1 of (Bailey & Piliouras, 2019) demonstrates that the Hamiltonian function defined here is inherently connected to the Bregman divergence, which is a commonly used concepts in optimization, plus a an additional term determined by the regularizers and equilibrium of the game.

More precisely, (Bailey & Piliouras, 2019) was shown that the cumulative strategies and payoffs of agent 1, (X_1, y_1), of continuous FTRL for agent 1 satisfies the following equations:

$$\frac{d}{dt}X_1(t) = \frac{\partial H}{\partial y_1}(X_1, y_1) = \nabla h_1^*(y_1(t)), \tag{11}$$

$$\frac{d}{dt}y_1(t) = -\frac{\partial H}{\partial X_1}(X_1, y_1) = A^{(12)}\nabla h_2^*(y_2(0) + A^{(21)}X_2(t)). \tag{12}$$

Similarly results also hold for agent 2, (X_2, y_2), of continuous FTRL for agent 2 satisfies the following equations:

$$\frac{d}{dt}X_2(t) = \frac{\partial H}{\partial y_2}(X_2, y_2) = \nabla h_2^*(y_2(t)), \tag{13}$$

$$\frac{d}{dt}y_2(t) = -\frac{\partial H}{\partial X_2}(X_2, y_2) = A^{(21)}\nabla h_1^*(y_1(0) + A^{(12)}X_1(t)). \tag{14}$$

The proof of Proposition 3.1 is divided into two parts :

- The proof of entropy regularizers is presented in Section A.3,
- The proof of Euclidean norm regularizers is presented in Section A.4,

and in Section A.2, we introduce the Euler and Symplectic discretization of FTRL.

A.2. Euler and Symplectic discretization of FTRL

Both in Euler and Symplectic, we denote the initial condition of the discrete equation on $(X_i^t, y_i^t),\ i = 1, 2$ to be $y_i^0 = y_i(0)$ and $X_i^0 = 0$.

Lemma A.1 (Euler discretization of FTRL)**.** *Discretizing equation (6) with Euler method for both agent $i = 1, 2$ gives*

$$X_1^{t+1} = X_1^t + \eta\frac{\partial H}{\partial y_1}(X_1^t, y_1^{t+1}) = X_1^t + \eta\nabla h_1^*(y_1^t), \qquad \text{(agent 1 Euler discretize equation)}$$

$$y_1^{t+1} = y_1^t - \eta\frac{\partial H}{\partial X_1}(X_1^t, y_1^t) = y_1^t + \eta A^{(12)}\nabla h_2^*(y_2^0 + A^{(21)}X_1^t),$$

and

$$X_2^{t+1} = X_2^t + \eta\frac{\partial H}{\partial y_2}(X_2^t, y_2^t) = X_2^t + \eta\nabla h_2^*(y_2^t), \qquad \text{(agent 2 Euler discretize equation)}$$

$$y_2^{t+1} = y_1^t - \eta\frac{\partial H}{\partial X_2}(X_2^{t+1}, y_2^t) = y_2^t + \eta A^{(21)}\nabla h_1^*(y_1^0 + A^{(12)}X_2^t).$$

Proof. Note that in Euler discretization, we use the derivative on point of t-th round to find the point of $t+1$ round. Thus (agent 1 Euler discretize equation) directly follows from applying Euler discretization to 11 and 12. Similarly, (agent 2 Euler discretize equation) directly follows from applying Euler discretization to 13 and 14. □

Lemma A.2 (Symplectic discretization of FTRL)**.** *Discretizing (6) for $i = 1$ with I-type Euler symplectic method (Type I method) gives*

$$y_1^{t+1} = y_1^t - \eta \frac{\partial H}{\partial X_1}(X_1^t, y_1^t) = y_1^t + \eta A^{(12)} \nabla h_2^*(y_2^0 + A^{(21)} X_1^t) \qquad \text{(agent 1 Symplectic discretize equation)}$$
$$X_1^{t+1} = X_1^t + \eta \frac{\partial H}{\partial y_1}(X_1^t, y_1^{t+1}) = X_1^t + \eta \nabla h_1^*(y_1^{t+1})$$

and discrete (6) for $i = 2$ with II-type Euler symplectic method (Type II method)gives

$$X_2^{t+1} = X_2^t + \eta \frac{\partial H}{\partial y_2}(X_2^t, y_2^t) = X_2^t + \eta \nabla h_2^*(y_2^t) \qquad \text{(agent 2 Symplectic discretize equation)}$$
$$y_2^{t+1} = y_1^t - \eta \frac{\partial H}{\partial X_2}(X_2^{t+1}, y_2^t) = y_2^t + \eta A^{(21)} \nabla h_1^*(y_1^0 + A^{(12)} X_2^{t+1})$$

Proof. (agent 1 Symplectic discretize equation) directly follows from applying (Type I method) to equation 11 and 12. Similarly, (agent 2 Symplectic discretize equation) directly follows from applying (Type II method) to equation 13 and 14. □

We define (x_1^t, x_2^t) to be

$$x_1^t = \frac{X_1^{t+1} - X_1^t}{\eta},\ x_2^t = \frac{X_2^{t+1} - X_2^t}{\eta}. \tag{15}$$

In the case of Euler discretization of FTRL (Lemma A.1), we have

$$x_1^t = \nabla h_1^*(y_1^t),\ x_2^t = \nabla h_2^*(y_2^t), \tag{16}$$

and in the case of Symplectic discretization of FTRL (Lemma A.2), we have

$$x_1^t = \nabla h_1^*(y_1^{t+1}),\ x_2^t = \nabla h_2^*(y_2^t). \tag{17}$$

Note that in Symplectic method, x_1^t is determined by y_1^{t+1}, but in Euler method, x_1^t is determined by y_1^t.

In the following, we will show (x_1^t, x_2^t) evolves as (MWU) under Euler method or (AltMWU) under Symplectic method on the strategy space if the regularizers $h_i(\cdot)$ are choose to be entropy functions, and the constrained sets $\mathcal{X}_i$ are chosen to be simplexes for $i = 1, 2$, this exactly the second part of Proposition 3.1.

Lemma A.3. *Both in Euler discretization of FTRL dynamics and Symplectic discretization of FTRL dynamics, the equalities*

$$y_1^n = y_1^0 + A^{(12)} X_2^n \tag{18}$$
$$y_2^n = y_2^0 + A^{(21)} X_1^n \tag{19}$$

hold for any $n \geq 0$.

Proof. Here we only prove the case of Symplectic discretization of FTRL dynamics, as the case of Euler discretization of FTRL dynamics is similar. We prove this by induction. For $n = 0$, (18) and (19) are

$$y_1^0 = y_1^0 + A^{(12)} X_2^0 \tag{20}$$
$$y_2^0 = y_2^0 + A^{(21)} X_1^0, \tag{21}$$

which hold trivially since by definition $X_1^0 = X_2^0 = 0$.

Now assume (18) and (19) hold for n, i.e,

$$y_1^n = y_1^0 + A^{(12)} X_2^n \tag{22}$$
$$y_2^n = y_2^0 + A^{(21)} X_1^n. \tag{23}$$

Then, we have

$$\begin{aligned} y_1^{n+1} &= y_1^n + \eta A^{(12)} \nabla h_2^*(y_2^0 + A^{(21)} X_1^n) && (24)\\ &\overset{(23)}{=} y_1^n + \eta A^{(12)} \nabla h_2^*(y_2^n) && (25)\\ &= y_1^n + \eta A^{(12)} x_2^n && (26)\\ &\overset{(22)}{=} y_1^0 + A^{(12)} X_2^n + \eta A^{(12)} x_2^n && (27)\\ &= y_1^0 + A^{(12)} X_2^{n+1}. && (28) \end{aligned}$$

Moreover, we have

$$\begin{aligned} y_2^{n+1} &= y_2^n + \eta A^{(21)} \nabla h_1^*(y_1^0 + A^{(12)} X_2^{n+1}) && (29)\\ &\overset{(28)}{=} y_2^n + \eta A^{(21)} \nabla h_1^*(y_1^{n+1}) && (30)\\ &= y_2^n + \eta A^{(21)} x_1^n && (31)\\ &\overset{(23)}{=} y_2^0 + A^{(21)} X_1^n + \eta A^{(21)} x_1^n && (32)\\ &= y_1^0 + A^{(21)} X_1^{n+1} && (33) \end{aligned}$$

This finish the proof. □

A.3. Proof of Entropy regularizers

Lemma A.4. *For entropy regularizer $h(x) = \sum_{i=1}^n x_i \ln x_i$ with simplex constrain, i.e., $\Delta = \{x \in \mathbb{R}^n | \sum_{i=1}^n x_i = 1, x_i \geq 0\}$, we have*

$$\nabla h^*(y) = \left(\frac{e^{y_i}}{\sum_{s=1}^n e^{y_s}} \right)_{i=1}^n. \tag{34}$$

Proof. By the definition, $h^*(y) = \max_{x \in \Delta}(\langle x, y \rangle - h(x))$, and $\nabla h^*(y) = \arg\max_{x \in \Delta}(\langle x, y \rangle - h(x))$.

Denote $f(x) = (\langle x, y \rangle - h(x))$, by the KKT condition, if x^* is the maximum of f, then there exist μ_i and λ such that

$$-\nabla f(x^*) + \sum_{i=1}^n \mu_i \nabla g_i(x^*) + \lambda \nabla h(x^*) = 0$$

and

$$\begin{aligned} &g_i(x^*) = -x_i^* \leq 0 \ \text{ for all } \ i \in [n],\\ &h(x^*) = \sum_{i=1}^n x_i^* - 1 = 0,\\ &\mu_i g_i(x^*) = 0 \ \text{ for all } \ i \in [n]. \end{aligned}$$

Since the gradient of f can be computed to be

$$\nabla f(x) = y - (\log x_1 + 1, ..., \log x_n + 1),$$

the KKT condition becomes

$$-y + (\log x_1 + 1, ..., \log x_n + 1) + \sum_{i=1}^{n} \mu_i (0, ..., -1, ..., 0) + \lambda(1, ..., 1) = 0.$$

Suppose the feasible x^* is interior point of Δ, i.e., $x_i > 0$, then we have for all $i \in [n]$, $\mu_i = 0$. Then the KKT condition is reduced to the following equations $\log x_i + 1 + \lambda = y_i$ for all $i \in [n], \sum_{i=1}^{n} x_i = 1$. This gives solution of x_i and λ:

$$x_i = \frac{e^{y_i}}{\sum_{s=1}^{n} e^{y_s}} \text{ for all } i \in [n], \quad \lambda = \log \left(\sum_{s=1}^{n} e^{y_s} \right) - 1,$$

thus we have completed the proof. □

Lemma A.5. *The (x_1^t, x_2^t) in (16) with entropy regularizer is the same as (MWU).*

Proof. In (16), we have $x_1^t = \nabla h_1^*(y_1^t)$, thus

$$x_1^t = \nabla h_1^*(y_1^t) \tag{35}$$

$$\tag{36}$$

$$\overset{(34)}{=} \left(\frac{e^{y_{1,s}^t}}{\sum_{j=1}^{n_1} e^{y_{1,j}^t}} \right)_{s=1}^{n_1} \tag{37}$$

$$\tag{38}$$

$$\overset{(18)}{=} \left(\frac{e^{y_{1,s}^0 + (A^{(12)} X_2^t)_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^0 + (A^{(12)} X_2^t)_j}} \right)_{s=1}^{n_1} \tag{39}$$

$$\tag{40}$$

$$= \left(\frac{e^{y_{1,s}^0 + \eta (A^{(12)} \sum_{k=1}^{t-1} x_2^k)_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^0 + \eta (A^{(12)} \sum_{k=1}^{t-1} x_2^k)_j}} \right)_{s=1}^{n_1} \tag{41}$$

$$\tag{42}$$

$$= \left(\frac{x_{1,s}^{t-1} e^{\eta (A^{(12)} x_2^{t-1})_s}}{\sum_{j=1}^{n_1} x_{1,j}^{t-1} e^{\eta (A^{(12)} x_2^{t-1})_j}} \right)_{s=1}^{n_1}. \tag{43}$$

The case of 2 agent is exactly same as 1 agent as they are symmetry, and we have

$$x_2^t = \left(\frac{x_{2,s}^{t-1} e^{\eta (A^{(21)} x_2^{t-1})_s}}{\sum_{j=1}^{n_2} x_{2,j}^{t-1} e^{\eta (A^{(21)} x_2^{t-1})_j}} \right)_{s=1}^{n_2}. \tag{44}$$

That is same as (MWU) . □

Lemma A.6. *The (x_1^t, x_2^t) in (17) with entropy regularizer is the same as (AltMWU).*

Proof. For 1 agent, from (17), we have $x_1^t = \nabla h_1^*(y_1^{t+1})$, thus

$$x_1^t = \nabla h_1^*(y_1^{t+1}) \tag{45}$$

$$\tag{46}$$

$$\overset{(34)}{=} \left(\frac{e^{y_{1,s}^{t+1}}}{\sum_{j=1}^{n_1} e^{y_{1,j}^{t+1}}} \right)_{s=1}^{n_1} \tag{47}$$

$$\tag{48}$$

$$\overset{(18)}{=} \left(\frac{e^{y_{1,s}^{0}+(A^{(12)}X_2^{t+1})_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^{0}+(A^{(12)}X_2^{t+1})_j}} \right)_{s=1}^{n_1} \tag{49}$$

$$\tag{50}$$

$$= \left(\frac{e^{y_{1,s}^{0}+\eta(A^{(12)}\sum_{k=1}^{t} x_2^k)_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^{0}+\eta(A^{(12)}\sum_{k=1}^{t} x_2^k)_j}} \right)_{s=1}^{n_1} \tag{51}$$

$$\tag{52}$$

$$= \left(\frac{x_{1,s}^{t-1} e^{\eta(A^{(12)}x_2^t)_s}}{\sum_{j=1}^{n_1} x_{1,j}^{t-1} e^{\eta(A^{(12)}x_2^t)_j}} \right)_{s=1}^{n_1}. \tag{53}$$

Note that the update rule of x_1^t use x_2^t, this is a characteristic of (AltMWU).

For 2 agent, from (17) we have $x_2^t = \nabla h_2^*(y_2^t)$, thus

$$x_2^t = \nabla h_2^*(y_2^t) \tag{54}$$

$$\tag{55}$$

$$= \left(\frac{e^{y_{2,s}^{t}}}{\sum_{j=1}^{n_2} e^{y_{2,j}^{t}}} \right)_{s=1}^{n_2} \tag{56}$$

$$\tag{57}$$

$$\overset{(19)}{=} \left(\frac{e^{y_{2,s}^{0}+(A^{(21)}X_1^{t})_s}}{\sum_{j=1}^{n_2} e^{y_{2,j}^{0}+(A^{(21)}X_1^{t})_j}} \right)_{s=1}^{n_2} \tag{58}$$

$$\tag{59}$$

$$= \left(\frac{e^{y_{2,s}^{0}+\eta(A^{(21)}\sum_{k=1}^{t-1} x_1^k)_s}}{\sum_{j=1}^{n_2} e^{y_{2,j}^{0}+\eta(A^{(21)}\sum_{k=1}^{t-1} x_1^k)_j}} \right)_{s=1}^{n_2} \tag{60}$$

$$\tag{61}$$

$$= \left(\frac{x_{2,s}^{t-1} e^{\eta(A^{(21)}x_1^{t-1})_s}}{\sum_{j=1}^{n_2} x_{2,j}^{t-1} e^{\eta(A^{(21)}x_1^{t-1})_j}} \right)_{s=1}^{n_2}. \tag{62}$$

Combine (62) and (53), we can see the update rule of (x_1^t, x_2^t) is same as (AltMWU) . □

Combine Lemma A.5 and Lemma A.6, we proved the second part of Proposition 3.1. The first part is very similar, except the regularizers are changed to Euclidian norm. However, this change will not affect the proof, so we omit it here.

A.4. Proof of Euclidean norm regularizers

Note that for Euclidean norm regularizers, i.e., $h_i(x) = \|x\|^2$, we have

$$\nabla h_i^*(y) = \arg\max_{x \in \mathbb{R}^n} \{\langle x, y\rangle - \|x\|^2\} = y. \tag{63}$$

Lemma A.7. *The (x_1^t, x_2^t) in (16) with Euclidean norm regularizer is the same as (GDA).*

Proof. For agent 1, we have

$$x_1^t \overset{(16)}{=} \nabla h_1^*(y_1^t) \overset{(63)}{=} y_1^t = x_1^{t-1} + \eta \cdot A^{(12)} x_2^{t-1}. \tag{64}$$

Agent 2 is exactly same as agent 1 since in Euler method, two agent are symmetry. Thus we have shown the update rule of (x_1^t, x_2^t) is same as (GDA). □

Lemma A.8. *The (x_1^t, x_2^t) in (17) with Euclidean norm regularizer is the same as (AltGDA).*

Proof. For agent 1, we have

$$x_1^t \overset{(17)}{=} \nabla h_1^*(y_1^{t+1}) \overset{(63)}{=} y_1^{t+1} = x_1^{t-1} + \eta \cdot A^{(12)} x_2^t. \tag{65}$$

For agent 2, we have

$$x_2^t \overset{(17)}{=} \nabla h_1^*(y_2^t) \overset{(63)}{=} y_2^t = x_2^{t-1} + \eta \cdot A^{(21)} x_2^{t-1}. \tag{66}$$

Thus we have shown the update rule of (x_1^t, x_2^t) is same as (AltGDA).

□

B. Proof of Section 4

In this appendix we prove results in Section 4. Proposition 4.2 is proved in Section B.3, and Proposition 4.1 is proved in Section B.4. In fact, we prove a more general result, which states that when two players choose arbitrary regularizers that satisfies strongly convex and Lipschitz gradient condition except a bounded region on the domain, then the differential entropy of Euler discretization has linear growth rate, while differential entropy of Symplectic discretization keeps constant. Note that both Euclidian norm regularizer and entropy regularizer satisfy these conditions, for example, entropy regularizer is 1-strongly convex on the interior points of simplex and has Lipschitz gradient except an arbitrary small neighbourhood of zero point.

The main technical lemma for proving Proposition 4.2 is Lemma B.6, which states for sufficient small step size, the update rule of Euler discretization of FTRL is an injective map. This injective property is necessary for calculating the evolution of differential entropy, see Lemma B.1. The proof of Proposition 4.1 is easier, as the symplectic discretization is naturally an injective map.

B.1. Evolution of differential entropy under diffeomorphism

The following result and its proof are informally stated in (Cheung et al., 2022), for convenience of applying their statement later, we formulate it into a lemma as follows.

Lemma B.1. *Let $X \in \mathbb{R}^d$ be a random vector with probability density function $g(x)$ and the support set of $g(x)$ is $\mathcal{X}$. Assume $f : \mathbb{R}^d \to \mathbb{R}^d$ be a diffeomorphism, thus $f(X)$ is a random vector. Then we have*

$$S(f(X)) = S(X) + \int_{\mathcal{X}} g(x) \log\left(\left|\det J_f(x)\right|\right) dx \tag{67}$$

where $J_f(x)$ is the Jacobian matrix of f at point $x \in \mathbb{R}^d$.

Proof. Denote $Y = f(X)$, and let $\widehat{g}(Y)$ represent the probability density function of Y, and $\mathcal{Y}$ be the support set of Y.

Then we have

$$S(Y) = S(f(X)) = -\int_{\mathcal{Y}} \widehat{g}(y) \cdot \log\left(\widehat{g}(y)\right) dy \tag{68}$$

$$= -\int_{\mathcal{Y}} g\left(f^{-1}(y)\right) |\det J_{f^{-1}}(y)| \cdot \log\left(g(f^{-1}(y))|\det J_{f^{-1}}(y)|\right) dy \tag{69}$$

$$= -\int_{\mathcal{X}} g(x)|\det J_{f^{-1}}\left(f(x)\right)| \cdot \log\left(g(x)|\det J_{f^{-1}}(f(x))|\right) \cdot |\det J_f(x)| dx \tag{70}$$

$$= -\int_{\mathcal{X}} g(x) \cdot \log\left(g(x)|\det J_{f^{-1}}(f(x))|\right) dx \tag{71}$$

$$= -\int_{\mathcal{X}} g(x) \log\left(g(x)\right) dx - \int_{\mathcal{X}} g(x) \log\left(|\det J_{f^{-1}}\left(f(x)\right)|\right) dx \tag{72}$$

$$= S(X) + \int_{\mathcal{X}} g(x) \log\left(|\det J_f(x)|\right) dx, \tag{73}$$

where (71) comes from the inverse function theorem, which states

$$J_{f^{-1}}\left(f(x)\right) = \left(J_f(x)\right)^{-1}. \tag{74}$$

□

B.2. Technical lemmas for Proposition 4.2

We first present several lemmas used later.

Lemma B.2 (Corollary 2.2 and 2.3 of (Hong & Horn, 1991))**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetry and positive semidefinite. Then AB is diagonalizable and has nonnegative eigenvalues. Moreover, if A is positive definite, then the number of positive eigenvalues, negative eigenvalues, and 0 eigenvalues of AB are the same as B.*

Lemma B.3. *If $A \in \mathbb{R}^{n\times n}$ is a symmetry matrix, and λ is an eigenvalue of A, then λ^2 is an eigenvalue of A^2.*

Proof. Since A is a symmetry matrix, there is an invertible matrix P makes

$$P \cdot A \cdot P^{-1} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \tag{75}$$

where $\lambda_1, ..., \lambda_n$ are eigenvalues of A. Thus

$$P \cdot (A)^2 \cdot P^{-1} = \left(P \cdot A \cdot P^{-1}\right) \cdot \left(P \cdot A \cdot P^{-1}\right) \tag{76}$$

$$= \begin{bmatrix} (\lambda_1)^2 & & \\ & \ddots & \\ & & (\lambda_n)^2 \end{bmatrix}, \tag{77}$$

this implies $\{\lambda_i^2\}$ are eigenvalues of A^2. □

The following lemma is the standard Fenchel duality property, a proof can be found in Theorem 1 (Zhou, 2018).

Lemma B.4. *Let $h : \mathcal{X} \to \mathbb{R}$ be a μ-strongly convex function with L-Lipschitz continuous gradient, let*

$$h^*(y) = \max_{x\in\mathcal{X}}\{\langle x, y\rangle - h(x)\} \tag{78}$$

be the convex conjugate of h, then we have

(1) h^* is a $\frac{1}{L}$-strongly convex function.

(2) h^* has $\frac{1}{\mu}$-Lipschitz continuous gradient.

Lemma B.5. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable function on a convex set $U \subset \mathbb{R}^n$, and*

$$\|J_f(x) - I\| < 1 \tag{79}$$

for any $x \in U$, where $\|\cdot\|$ is the L^2-operator norm, then f is an injective map.

Proof. Let $g(x) = f(x) - x$. Then for any $x \neq y$, we have

$$\begin{aligned} \|f(x) - f(y) + y - x\| &= \|g(x) - g(y)\| && (80)\\ &= \|J_g(\zeta)(x - y)\| && (81)\\ &\leq \|J_g(\zeta)\| \cdot \|x - y\| && (82)\\ &< \|x - y\|, && (83) \end{aligned}$$

where (81) use the mean value theorem, and (83) is due to the fact that $\|J_g(\zeta)\| < 1$ for any $\zeta \in U$. Thus $f(x) \neq f(y)$. □

Lemma B.6. *If the step size $\eta < \min\{\mu_1, \mu_2/\|A^{(21)}\|^2\}$, then the iterate map*

$$\phi : (X_1^n, y_1^n) \to (X_1^{n+1}, y_1^{n+1}) \tag{84}$$

of Euler discretization of FTRL in Lemma A.1 is an injective function.

Proof. Recall the iterate map $\phi : (X_1^n, y_1^n) \to (X_1^{n+1}, y_1^{n+1})$ can be written as an Euler discretization with the following form

$$y_1^{n+1} = y_1^n - \eta \cdot \frac{\partial H}{\partial X_1}(X_1^n, y_1^n) \tag{85}$$

$$X_1^{n+1} = X_1^n + \eta \cdot \frac{\partial H}{\partial y_1}(X_1^n, y_1^n) \tag{86}$$

and the Hamiltonian function has form

$$H(X_1, y_1) = h_1^*(y_1) + h_2^*(y_2(0) + A^{(21)} X_1). \tag{87}$$

Note that $H_1(X_1, y_1)$ is separable, i.e., $h_1^*(\cdot)$ is independent with X_1 and $h_2^*(\cdot)$ is independent with y_1, thus we have

$$\frac{\partial^2 H}{\partial X_1 \partial y_1} = 0, \ \frac{\partial^2 H}{\partial y_1 \partial X_1} = 0. \tag{88}$$

Next we calculate the Jacobin matrix of ϕ,

$$J_\phi = \begin{bmatrix} \frac{\partial y_1^{n+1}}{\partial y_1^n} & \frac{\partial y_1^{n+1}}{\partial X_1^n} \\ \\ \frac{\partial X_1^{n+1}}{\partial y_1^n} & \frac{\partial X_1^{n+1}}{\partial X_1^n} \end{bmatrix} = \begin{bmatrix} I - \eta \frac{\partial^2 H}{\partial y_1 \partial X_1}(X_1^n, y_1^n) & -\eta \frac{\partial^2 H}{\partial^2 X_1}(X_1^n, y_1^n) \\ \\ \eta \frac{\partial^2 H}{\partial^2 y_1}(X_1^n, y_1^n) & I + \eta \frac{\partial^2 H}{\partial X_1 \partial y_1}(X_1^n, y_1^n) \end{bmatrix} \tag{89}$$

$$= \begin{bmatrix} I & -\eta (A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)} \\ \\ \eta \nabla^2 h_1^* & I \end{bmatrix}, \tag{90}$$

and

$$J_\phi - I = \begin{bmatrix} 0 & -\eta(A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)} \\ \\ \eta\nabla^2 h_1^* & 0 \end{bmatrix}. \tag{91}$$

Since h_i is μ_i-strongly convex, by Lemma B.4, h_i^* has $\frac{1}{\mu_i}$-Lipschitz continuous gradient, thus we have

$$\|\nabla^2 h_i^*\| \leq \frac{1}{\mu_i} \tag{92}$$

holds at arbitrary points within the domain of h_i^*.

Next we estimate the L^2-operator norm of the matrix $J_\phi - I$, since the L^2-operator norm is equivalent to the spectral norm, we have

$$\|J_\phi - I\| = \sqrt{\lambda_{\max}\left((J_\phi - I)^\top \cdot (J_\phi - I)\right)}, \tag{93}$$

and

$$(J_\phi - I)^\top \cdot (J_\phi - I) = \eta^2 \cdot \begin{bmatrix} \left(\nabla^2 h_1^*\right)^2 & 0 \\ \\ 0 & \left((A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)}\right)^2 \end{bmatrix}. \tag{94}$$

Since both $\nabla^2 h_1^*$ and $(A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)}$ are symmetry matrix, thus by Lemma B.3, eigenvalues of $(J_\phi - I)^\top \cdot (J_\phi - I)$ has form $\eta^2\lambda^2$, where λ is an eigenvalue of $\nabla^2 h_1^*$ or $(A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)}$.

Note that h_i^* is a function of X_1, y_1, thus λ is also a function of X_1, y_1, and it is not clear whether there is an upper bound on λ without more information on the Hessian matrix of h_i^*. However, as we have shown in (92), the L^2-operator norm of $\nabla^2 h_i^*$ has an upper bound $\frac{1}{\mu_i}$, thus we have

$$0 \leq \lambda < \max\left\{\frac{1}{\mu_1}, \frac{\|A^{(21)}\|^2}{\mu_2}\right\}. \tag{95}$$

Thus we can choose $\eta < \min\left\{\mu_1, \frac{\mu_2}{\|A^{(21)}\|^2}\right\}$ to make

$$\|J_\phi - I\| < 1, \tag{96}$$

and by Lemma B.5, ϕ is an injective map. □

B.3. Proof of Proposition 4.2

Proof. By Lemma B.6, the iterate map of simultaneous FTRL is an diffeomorphism, thus we can use Lemma B.1 to calculate the evolution of differential entropy in simultaneous FTRL. We firstly prove differential entropy is a non-decrease function.

Recall (9), the Hamiltonian function of FTRL is

$$H(X_1^t, y_1^t) = h_1^*(y_1^t) + h_2^*(y_2^0 + A^{(21)} X_1^t), \tag{97}$$

and the iterate map $\phi : (y_1^n, X_n^1) \to (y_1^{n+1}, X_1^{n+1})$ of simultaneous FTRL can be written as Euler discretization of continuous FTRL, i.e.,

$$y_1^{t+1} = y_1^t - \eta \cdot \frac{\partial H}{\partial X_1}(X_1^t, y_1^t) \tag{98}$$

$$X_1^{t+1} = X_1^t + \eta \cdot \frac{\partial H}{\partial y_1}(X_1^t, y_1^t). \tag{99}$$

Recall from (90), the jacobian map of ϕ is

$$J_\phi(X,y) = \begin{bmatrix} I & -\eta(A^{(21)})^\top \cdot \nabla^2 h_2^*(X,y) \cdot A^{(21)} \\ \\ \eta\nabla^2 h_1^*(X,y) & I \end{bmatrix}, \tag{100}$$

thus

$$\det\left(J_\phi(X,y)\right) = \det\left(I + \eta^2 \cdot \nabla^2 h_1^*(X,y) \cdot (A^{(21)})^\top \cdot \nabla^2 h_2^*(X,y) \cdot A^{(21)}\right). \tag{101}$$

Since $\nabla^2 h_1^*$ and $(A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)}$ are both symmetry and positive semidefinite matrix, by Lemma B.2, their product is diagonalizable and has non-negative eigenvalues. Thus we have

$$\det\left(J_\phi(X,y)\right) \geq 1. \tag{102}$$

Combine this with (67), we have

$$\begin{aligned} S(X_1^{t+1}, y_1^{t+1}) &= S\left(\phi(X_1^{t+1}, y_1^{t+1})\right) && (103)\\ &= S(X_1^t, y_1^t) + \int_{\mathcal{X}} g^t(X_1^t, y_1^t) \log\left(\left|\det J_\phi(X_1^t, y_1^t)\right|\right) dX_1^t dy_1^t && (104)\\ &\geq S(X_1^t, y_1^t) + \int_{\mathcal{X}} g^t(X_1^t, y_1^t) \log(1) dX_1^t dy_1^t && (105)\\ &= S(X_1^t, y_1^t), && (106) \end{aligned}$$

where $g^t(\cdot)$ is the probability density function of (X_1^t, y_1^t), and (105) comes from $\det\left(J_\phi(X,y)\right) \geq 1$. Thus $S(X_1^t, y_1^t)$ is a non-decreasing function.

Moreover, as $\log(1+x) > x/(1+x)$ for $x > 0$, thus from (104), to prove a linear growth rate of differential entropy, it is sufficient to prove a uniform lower bound on $\det\left(J_\phi(X,y)\right)$. With the assumption that $h_i(\cdot)$ has Lipschitz continuous gradient, by Lemma B.4, $h_i^*(\cdot)$ is μ_i-strongly convex, thus $\nabla^2 h_i^*(X,y)$ is positive definite, i.e.,

$$\nabla^2 h_i^*(X,y) \succcurlyeq \mu_i I \tag{107}$$

for some $\mu_i > 0$. From Lemma B.2, with $A = \nabla^2 h_1^*(X,y)$ and $B = (A^{(21)})^\top \cdot \nabla^2 h_2^*(X,y) \cdot A^{(21)}$, we have

$$\eta^2 \cdot \nabla^2 h_1^*(X,y) \cdot (A^{(21)})^\top \cdot \nabla^2 h_2^*(X,y) \cdot A^{(21)} \tag{108}$$

is a diagonalizable matrix, and there are all eigenvalues of $J_\phi(X,y)$ are real number and larger than 1. Moreover, these eigenvalues has a uniform lower bound $1+c$, where c is determined by the strongly convex coefficients μ_i and the payoff matrix $A^{(21)}$. □

B.4. Proof of Proposition 4.1

Lemma B.7. *The update map*

$$(X_1^t, y_1^t) \to (X_1^{t+1}, y_1^{t+1})$$

from (agent 1 Symplectic discretize equation) is an injective map.

Proof. Recall the update rule can be written as

$$y_1^{t+1} = y_1^t - \eta \frac{\partial H}{\partial X_1}(X_1^t, y_1^t) = y_1^t + \eta A^{(12)} \nabla h_2^*(y_2^0 + A^{(21)} X_1^t), \tag{109}$$

$$X_1^{t+1} = X_1^t + \eta \frac{\partial H}{\partial y_1}(X_1^t, y_1^{t+1}) = X_1^t + \eta \nabla h_1^*(y_1^{t+1}). \tag{110}$$

$$\tag{111}$$

Thus given (X_1^{t+1}, y_1^{t+1}), we can directly find an unique X_1^t from (110), i.e.,

$$X_1^t = X_1^{t+1} - \eta \nabla h_1^*(y_1^{t+1}).$$

Then use this X_1^t, we can also determine a unique y_1^t from (109), i.e.,

$$y_1^t = y_1^{t+1} - \eta A^{(12)} \nabla h_2^*(y_2^0 + A^{(21)} X_1^t).$$

This finish the proof. □

Now we are ready to prove Proposition 4.1.

Proof. Since the iterate map $\psi : (X_1^n, y_1^n) \to (X_1^{n+1}, y_1^{n+1})$ in Symplectic discretization of FTRL defined A.2 in is naturally an injective map from Lemma B.7 , we can directly use lemma B.1. We have

$$S(X_1^{n+1}, y_1^{n+1}) - S(X_1^n, y_1^n) = \int_{\mathcal{X}} g^n(X_1^n, y_1^n) \log\left(|\det J_\phi(X_1^n, y_1^n)|\right) dx \tag{112}$$

where $g^n(\cdot)$ is the probability density function of random vector (X_1^n, y_1^n).

Moreover, since ψ is a symplectomorphism, we have $|\det J_\psi(X_1^n, y_1^n)| = 1$. Thus the right hand side of (112) equals to 0, and this implies $S(X_1^{n+1}, y_1^{n+1}) = S(X_1^n, y_1^n)$. □

B.5. Numerical examples of Proposition Propositions 4.2 and 4.1

Although differential entropy plays important roles in several subjects, estimating the value of differential entropy under transformations is generally a challenging task. Even in the one-dimensional case, special methods need to be designed for calculating differential entropy (Hyvärinen, 1997). A recent review of this topic can be found in (Feutrill & Roughan, 2021).

However, for the case of gradient descent, it is possible to calculate the variation of differential entropy due to the linear structure of the algorithm and the equality

$$S(AX) - S(X) = \log(|\det(A)|). \tag{113}$$

In Figure 2, we present numerical experiments on the variation of differential entropy using game defined by $A = [[1, -1], [-1, 1]]$. Numerical results show differential entropy has a linear growth rate in simultaneous case and keeps invariant in alternating case, which support Propositions 4.2 and 4.1.

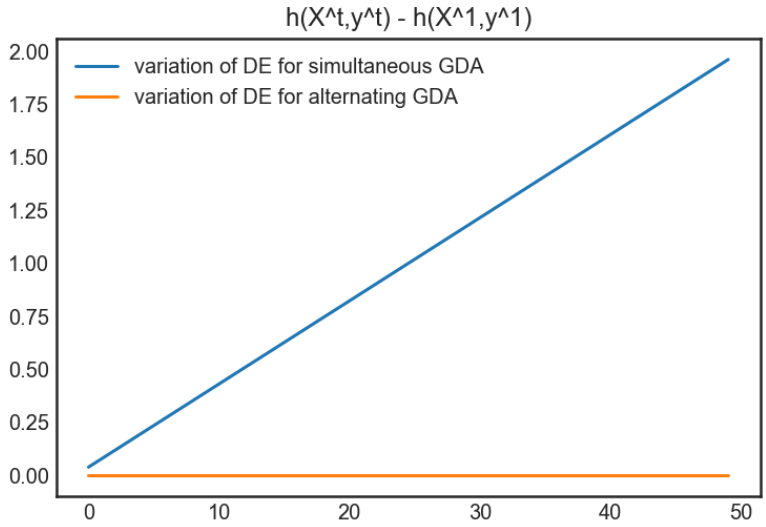


Figure 8: Variation of differential entropy for simultaneous and alternating GDA. The growth rate of the variation in DE is linear for simultaneous GDA, while it is 0 for alternating GDA.

C. Proof of Theorem 5.1

This appendix is divided into two parts. In Section C.1, we presented necessary backgrounds from linear algebra, differential equation, and difference equation. In Section C.2, we provide detailed prove of Theorem 5.1.

C.1. Additional Backgrounds

C.1.1. Complex Jordan normal form

We will consider the complex Jordan normal form of a real square matrix $A \in \mathbb{R}^{d\times d}$. Let $\mathrm{Spec}(A)$ be the set of eigenvalues of A. Consider A acts on vector space $\mathbb{C}^d$ as a linear operator.

Definition C.1 (Generalized Eigenvector). *A vector $v_m \in \mathbb{C}^d$ is called a generalized eigenvector of **type** m corresponding to the eigenvalue μ if*

$$(A-\lambda I)^m v_m = 0$$

but

$$(A-\mu I)^{m-1} v_m \neq 0.$$

Definition C.2 (Jordan Chain). *let v_m be a generalized eigenvector of type m corresponding to the matrix A and the eigenvalue μ. The Jordan chain generated by v_m is a set of m vectors $\{v_m, v_{m-1}, ..., v_1\}$ given by*

$$\begin{aligned} v_{m-1} &= (A-\mu I)v_m \\ v_{m-2} &= (A-\mu I)^2 v_m = (A-\mu I)v_{m-1} \\ &... \\ v_1 &= (A-\mu I)^{m-1} v_m = (A-\mu I)v_2 \end{aligned}$$

Remark C.3. *If $\mu \in \mathbb{R}$ is a real eigenvalue of A, then the generalized eigenvectors of μ are also vectors over real numbers, and the Jordan chain are also made up by vectors over real numbers.*

Proposition C.4. *A Jordan chain is a linearly independent set of vectors.*

Proof. Let $\{v_m, v_{m-1}, ..., v_1\}$ be a Jordan chain generated by a type m generalized eigenvector v_m corresponding to an eigenvalue λ of A, and consider the equation

$$c_m v_m + c_{m-1} v_{m-1} + ... + c_1 v_1 = 0 \tag{114}$$

We will show $c_m = c_{m-1} = ... = c_1 = 0$.

Multiply equation 114 by $(A-\mu I)^{m-1}$, and note that for $j \leq m-1$

$$\begin{aligned} (A-\mu I)^{m-1} c_j v_j &= c_j (A-\mu I)^{m-j-1} (A-\mu I)^j x_j \\ &= 0 \end{aligned}$$

Thus equation 114 becomes to be $c_m (A-\mu I)^{m-1} v_m = 0$. However, since v_m is a type m generalized eigenvector, we have

$$(A-\mu I)^{m-1} v_m \neq 0,$$

thus $c_m = 0$. Continuing this process, we will finally obtain $c_m = c_{m-1} = ... = c_1 = 0$. □

Proposition C.5 (page 366 of (Bronson, 1991))**.** *Every $d \times d$ matrix has d linearly independent generalized eigenvectors.*

Given a Jordan chain $\{v_1, v_2, ..., v_m\}$ of length m, by Proposition C.4, we will get a subspace spans by $\{v_1, v_2, ..., v_m\}$. The linear operator $A : \mathbb{C}^d \to \mathbb{C}^d$ can acts on vectors' set $\{v_1, v_2, ..., v_m\}$.

Denote $[v_1, v_2, ..., v_m]$ to be the matrix consists of column vectors $\{v_1, v_2, ..., v_m\}$, then A acts on $[v_1, v_2, ..., v_m]$ as

$$
\begin{aligned}
A[v_1, v_2, ..., v_m] &= [Av_1, Av_2, ..., Av_m] \\
&= [\mu v_1, v_1 + \mu v_2, ..., v_{m-1} + \mu v_m] \\
&= [v_1, v_2, ..., v_m] \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix}.
\end{aligned}
$$

Thus we have

$$
[v_1, v_2, ..., v_m]^{-1} A [v_1, v_2, ..., v_m] = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix} \tag{115}
$$

which is a $m \times m$ upper triangular matrix, with eigenvalue μ on diagonal and each non-zero off-diagonal entry equal to 1. Such an upper triangular matrix is called a size m **Jordan block** of A corresponding to eigenvalue μ.

Proposition C.6 (Page 367 of (Bronson, 1991))**.** *Every $d \times d$ matrix A has a set of $d \times d$ linearly independent generalized eigenvectors composed entirely of Jordan chains, such a set of generalized eigenvectors is a called a **canonical bases** of A.*

Definition C.7 (**Generalized modal matrix**)**.** *Let A be an $d \times d$ matrix. A generalized modal matrix M for A is a $d \times d$ matrix whose columns, considered as vectors, form a canonical basis for A and appear in M according to the following rules:*

(1) All vectors of the same chain appear together in adjacent columns of M.

(2) Each chain appears in M in order of increasing type.

Remark C.8. *The Jordan chain corresponding to a real eigenvalue $\mu \in \mathbb{R}$ of A will composed by vectors in $\mathbb{R}^d$, but if $\mu \in \mathbb{C}$ is a complex eigenvalue of A, then the Jordan chain corresponding to μ will contain vectors in $\mathbb{C}^d$.*

Combine equation 115 and Proposition C.6, we have following proposition.

Proposition C.9. *Any matrix $A \in \mathbb{R}^{d \times d}$ is similar to a matrix in Jordan normal form under the similarity transformation of a generalized modal matrix M of A, i.e.,*

$$
J^{\mathbb{C}} = M^{-1} A M
$$

has block diagonal form $J^{\mathbb{C}} = \oplus_i J_i$ with Jordan blocks J_i given with $\mu \in \mathrm{Spec}(A)$ by

$$
J_i = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix}.
$$

The Jordan normal form is unique up to the order of the Jordan blocks.

We have the following proposition that determines the number of a particular size of Jordan blocks in Jordan normal form corresponding to an eigenvalue λ .

Proposition C.10 (Page 368 of (Bronson, 1991)). *Let λ be an eigenvalue of A, and denote*

$$m = \max\{i|\ \ker(A-\mu I)^i \supsetneq \ker(A-\mu I)^{i-1}\}.$$

Denote ρ_k as the number of linear independent generalized eigenvectors of type k corresponding to the eigenvalue μ that appear in a canonical basis for A, then

$$\rho_k = \dim\ker(A-\mu I)^k - \dim\ker(A-\mu I)^{k-1}\ (k=1,2,...,m).$$

Since every Jordan chains of length k in a canonical basis gives a size k Jordan block, ρ_k is also the number of size k Jordan blocks in the complex Jordan normal form corresponding to λ.

There is another characterization on the size of the largest Jordan block corresponding eigenvalue based on the minimal polynomial of A:

Proposition C.11 (Theorem 3.3.6 in (Horn & Johnson, 2012)). *Let m be the size of the largest Jordan block corresponding to the eigenvalue μ of a matrix A, then m equals to the degree of the factor $(x-\mu)$ in the minimal polynomial of A.*

C.1.2. Real Jordan normal form

The real Jordan normal is important for computing exponential function of a matrix $A \in \mathbb{R}^{d\times d}$. In the complex Jordan form, Jordan blocks may contain elements in $\mathbb{C}-\mathbb{R}$. Thus for a matrix $A \in \mathbb{R}^{d\times d}$, we cannot use its complex Jordan normal form to calculate exponential functions of A directly. We need to define a standard form of A, which should only contain real numbers and keep the shape as a diagonal block matrix. This motive the definition of real Jordan normal form.

Let $\mu = \mathrm{Re}(\mu) + i\mathrm{Im}(\mu), \mathrm{Im}(\mu) \neq 0$ be an eigenvalue of A, and $v_m \in \mathbb{C}^d$ is a generalized eigenvalue of type m corresponding to μ. Then the complex conjugate of μ, denote by $\bar{\mu}$, is also an eigenvalue of A, and the complex conjugate of v_m, denote by $\bar{v}_m \in \mathbb{C}^d$ is a generalized eigenvalue of type m corresponding to $\bar{\mu}$. Let

$$\{v_{m-i}|\ v_{m-i} = (A-\mu I)^i v_m,\ i=0,1,...,m-1\}$$

be a Jordan chain of length m corresponding to λ, then it gives a complex Jordan block of size m in the complex Jordan normal form of A as in equation 115.

The $2m$ vectors $\{\mathrm{Re}(v_1), \mathrm{Im}(v_1), \mathrm{Re}(v_2), \mathrm{Im}(v_2), ..., \mathrm{Re}(v_m), \mathrm{Im}(v_m)\} \subset \mathbb{R}^d$ will play the role of complex generalized vectors of eigenvalues $\mu, \bar{\mu}$. It is directly to check A acts on $[\mathrm{Re}(v_1), \mathrm{Im}(v_1), \mathrm{Re}(v_2), \mathrm{Im}(v_2), ..., \mathrm{Re}(v_m), \mathrm{Im}(v_m)]$ gives the following matrix representation :

$$\begin{aligned}
&A[\mathrm{Re}(v_1), \mathrm{Im}(v_1), \mathrm{Re}(v_2), \mathrm{Im}(v_2), ..., \mathrm{Re}(v_m), \mathrm{Im}(v_m)] \\
&= [A\mathrm{Re}(v_1), A\mathrm{Im}(v_1), A\mathrm{Re}(v_2), A\mathrm{Im}(v_2), ..., A\mathrm{Re}(v_m), A\mathrm{Im}(v_m)] \\
&= [\mathrm{Re}(\mu v_1), \mathrm{Im}(\mu v_1), \mathrm{Re}(v_1)+\mathrm{Re}(\mu v_2), \mathrm{Im}(v_1)+\mathrm{Im}(\mu v_2), ..., \mathrm{Re}(v_{m-1})+\mathrm{Re}(\mu v_m), \mathrm{Im}(v_{m-1})+\mathrm{Im}(\mu v_m)] \\
&= [\mathrm{Re}(v_1), \mathrm{Im}(v_1), \mathrm{Re}(v_2), \mathrm{Im}(v_2), ..., \mathrm{Re}(v_m), \mathrm{Im}(v_m)] \begin{bmatrix} D & I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & I \\ & & & & D \end{bmatrix}
\end{aligned}$$

where $D = \begin{bmatrix} \mathrm{Re}(\mu) & \mathrm{Im}(\mu) \\ -\mathrm{Im}(\mu) & \mathrm{Re}(\mu) \end{bmatrix}$ and $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

The matrix

$$\begin{bmatrix} D & I & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ & & & & I \\ & & & & D \end{bmatrix} \tag{116}$$

is called the real Jordan blocks corresponding to a conjugate pair of image eigenvalues $\mu, \bar{\mu}$.

As in the case of complex Jordan normal form, we need to define the real generalized modal matrix, and under the similar transformation by real generalized modal matrix, the original matrix will be transformed into a real Jordan normal form.

Definition C.12 (Real generalized modal matrix). *Let A be an $d \times d$ matrix. A real generalized modal matrix M for A is a $d \times d$ matrix whose columns, considered as vectors, are real or image parts of a complex generalized eigenvectors in a canonical basis for A and appear in M according to the following rules:*

(1) Real and image parts of a generalized eigenvectors corresponding to same image eigenvalues appear in the first columns of M

(2) If $\mathrm{Re}(v), \mathrm{Im}(v)$ appear in the real generalized modal matrix, $(\mathrm{Re}(v), \mathrm{Im}(v))$ appear in adjacent columns of M

(3) All vectors of the same chain appear together in adjacent columns of M.

(4) Each chain appears in M in order of increasing type.

Note that comparing to definition C.7, the new requirements are (1) and (2). (1) will make the Jordan blocks as in equation 116 appear firstly in the real Jordan normal form, and the necessary of (2) can be seen from the derivation of equation 116.

Proposition C.13 (Theorem 1.2.3 in (Colonius & Kliemann, 2014)). *Any matrix $A \in \mathbb{R}^{d\times d}$ is similar to a matrix in the real Jordan normal from via a similarity transformation by A's real generalized modal matrix. That is, $\exists M \in \mathbb{R}^{d\times d}$ be A's real generalized modal matrix to make*

$$J^{\mathbb{R}} = M^{-1}AM$$

where $J^{\mathbb{R}} = (J_{\mu_1,\bar{\mu}_1} \oplus ... \oplus J_{\mu_k,\bar{\mu}_k}) \oplus_{i=1}^{l} (J_{\mu_{k+i}})$ with real Jordan blocks given for $\mu \in \mathrm{Spec}(\mathcal{L}) \cap \mathbb{R}$ by

$$J_\mu = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix}$$

and for $\mu, \bar{\mu} = \lambda \pm i\nu \in \mathrm{Spec}(\mathcal{L}), \nu > 0$, by

$$J_{\mu,\bar{\mu}} = \begin{bmatrix} D & I_2 & \cdot & \cdot & 0 \\ 0 & D & & & \cdot \\ \cdot & \ddots & \ddots & & \\ \cdot & & & D & I_2 \\ 0 & & & 0 & D \end{bmatrix}$$

where $D = \begin{bmatrix} \lambda & -\nu \\ \nu & \lambda \end{bmatrix}$ and $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

Note that the difference between complex and real Jordan form, complex and real generalized modal matrix are only in the Jordan blocks corresponding to a eigenvalue whose image part is not 0. Thus the number of a given size Jordan blocks corresponding to a real eigenvalue is same in complex and real Jordan form. For a pair of conjugate complex eigenvalues with nonzero image part, every real Jordan block is a combination of two complex Jordan blocks.

C.1.3. SOLUTION FORMULA FOR LINEAR DIFFERENTIAL EQUATION

For a linear differential equation with constant coefficients $A \in \mathbb{R}^{d\times d}$

$$\frac{dx}{dt}(t) = Ax$$

with initial condition $x(0)$, it's solution formula is given by $x(t) = e^{tA}x(0)$. Thus to understand the dynamic behavior of $x(t)$, it is necessary to calculate the matrix exponential matrix e^{tA}. This can be done by using the real Jordan normal form of A. Let J_A be A's real Jordan normal form, and $J_A = MAM^{-1}$, then $e^{tA} = Me^{tJ_A}M^{-1}$. Thus calculation of e^{tA} can be reduced to calculation of e^{tJ_A} and the real generalized modal matrix of A. The real generalized modal matrix of A is a combination of real and image parts of A's generalized eigenvectors, and in this section we consider calculation of e^{tJ_A}.

Proposition C.14. *(page 12 in (Colonius & Kliemann, 2014)) Let J_μ be a real Jordan block of size $m \times m$ associated with the real eigenvalue μ of a matrix $A \in \mathbb{R}^{d\times d}$. Then*

$$J_\mu = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix} \in \mathbb{R}^{m\times m} \tag{117}$$

and

$$e^{J_\mu t} = e^{\mu t}\begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdot & \cdot & \frac{t^{m-1}}{(m-1)!} \\ & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \frac{t^2}{2!} \\ & & & & \cdot & t \\ & & & & & 1 \end{bmatrix} \in \mathbb{R}^{m\times m}. \tag{118}$$

Let $J_{\mu,\bar{\mu}}$ be a real Jordan block of size $2m \times 2m$ associated with the real eigenvalue $\mu, \bar{\mu} = \lambda \pm i\nu, \nu > 0$ of a matrix $A \in \mathbb{R}^{d\times d}$. With

$$D = \begin{bmatrix} \lambda & -\nu \\ \nu & \lambda \end{bmatrix},\ R := R(t) = \begin{bmatrix} \cos(\nu t) & -\sin(\nu t) \\ \sin(\nu t) & \cos(\nu t) \end{bmatrix},\ I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

one obtains for

$$J_{\mu,\bar{\mu}} = \begin{bmatrix} D & I_2 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & I_2 \\ & & & & D \end{bmatrix} \in \mathbb{R}^{2m\times 2m}, \tag{119}$$

and

$$e^{J_{\mu,\bar{\mu}}t} = e^{\lambda t}\begin{bmatrix} R & tR & \frac{t^2}{2!}R & \cdot & \cdot & \frac{t^{m-1}}{(m-1)!}R \\ & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \frac{t^2}{2!}R \\ & & & & \cdot & tR \\ & & & & & R \end{bmatrix} \in \mathbb{R}^{2m\times 2m}. \tag{120}$$

C.1.4. Solution formula for linear difference equation

For a matrix $A \in \mathbb{R}^{d\times d}$, a linear difference equation with coefficient matrix A has form

$$x_{n+1} = Ax_n \tag{121}$$

By induction, equation 122 has solution formula

$$x_n = A^n x_0,\ x_0 \in \mathbb{R}^d. \tag{122}$$

If $J_A = M^{-1}AM$ be the real Jordan normal form of A, then $x_n = M(J_A)^n M^{-1}x_0$. Thus to solve equation 122, it is sufficient to know the formula for $(J_A)^n$ and the real generalized modal matrix M. Moreover, if $J_A = \oplus_i J_i$, then $(J_A)^n = \oplus_i (J_i)^n$, thus we only need to consider the power of real Jordan blocks.

Proposition C.15 (Page 19 in (Colonius & Kliemann, 2014))**.** *Let J be a real Jordan block of size $m\times m$ associated with a real eigenvalue μ of $A \in \mathbb{R}^{d\times d}$. Then*

$$J_\mu = \begin{bmatrix} \mu & & & & \\ & \mu & & & \\ & & \ddots & & \\ & & & \mu & \\ & & & & \mu \end{bmatrix} + \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \tag{123}$$

$$= \mu I + N \tag{124}$$

with $N^m = 0$. Thus we have

$$J_\mu^n = (\mu I + N)^n = \sum_{i=0}^{m-1} \binom{n}{i} \mu^{n-i} N^i. \tag{125}$$

Note that if $\mu > 1$, elements in J_μ^n will have an exponential growth rate as μ^n. If $\mu = 1$, elements in J_μ^n have a polynomial growth rate. If $\mu < 1$, elements in J_μ^n tends to 0 as n growth.

Proposition C.16 (Page 20 in (Colonius & Kliemann, 2014))**.** *Let J be a real Jordan block of size $2m\times 2m$ associated with a pair of conjugate complex eigenvalue $\mu = \lambda + i\nu, \bar{\mu} = \lambda - i\nu$ of $A \in \mathbb{R}^{d\times d}$. With $D = \begin{bmatrix} \lambda & -\nu \\ \nu & \lambda \end{bmatrix}$ and $I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, one obtains that*

$$J_{\mu,\bar{\mu}} = \begin{bmatrix} D & 0 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & 0 \\ & & & & D \end{bmatrix} + \begin{bmatrix} 0 & I_2 & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & I_2 \\ & & & & 0 \end{bmatrix} \tag{126}$$

$$= \tilde{D} + N \tag{127}$$

with $N^m = 0$. Moreover, since $\mu = |\mu|e^{i\phi}$ for some $\phi \in [0, 2\pi)$, one can write

$$D = \begin{bmatrix} \lambda & -\nu \\ \nu & \lambda \end{bmatrix} = |\mu| R,\ R = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}. \tag{128}$$

Thus

$$J^n_{\mu,\bar{\mu}} = (\tilde{D} + N)^n = \sum_{i=0}^{m-1} \binom{n}{i} \tilde{D}^{n-i} N^i = \sum_{i=0}^{m-1} \binom{n}{i} |\mu|^{n-i} \tilde{R}^{n-i} N^i. \tag{129}$$

where $\tilde{R}$ is a block diagonal matrix with matrix block R. Note that if $|\mu| > 1$, elements in $J^n_{\mu,\bar{\mu}}$ have exponential growth rate as $\mathcal{O}(|\mu|^n)$. If $|\mu| = 1$, elements in $J^n_{\mu,\bar{\mu}}$ have polynomial growth rate. If $|\mu| < 1$, elements in $J^n_{\mu,\bar{\mu}}$ tends to 0 as n growth.

C.2. Proof of Theorem 5.1

Now we are ready to prove Theorem 5.1. The proof is divided into three parts :

- covariance evolution of continuous time equation, proved in C.2.1.
- covariance evolution of Euler discretization, proved in C.2.2.
- covariance evolution of Symplectic discretization, proved in C.2.3.

C.2.1. Covariance evolution of continuous equation

We firstly give a summary of the proof. The continuous time equation of FTRL with Euclidean norm for agent 1 is written as:

$$\frac{dy_1}{dt} = -\frac{\partial H(X_1, y_1)}{\partial X_1} = -AA^\top X_1(t) + Ay_2(0) \tag{130}$$

$$\frac{dX_1}{dt} = \frac{\partial H(X_1, y_1)}{\partial y_1} = y_1(t). \tag{131}$$

Similarly for agent 2 we have

$$\frac{dy_2}{dt} = -\frac{\partial H(X_2, y_2)}{\partial X_2} = -A^\top AX_2(t) - A^\top y_1(0) \tag{132}$$

$$\frac{dX_2}{dt} = \frac{\partial H(X_2, y_2)}{\partial y_2} = y_2(t). \tag{133}$$

In the following, we will focus on the viewpoint of agent 1, thus we will consider equation (130) and (131).

Lemma C.17. *The solution of the linear differential system consisted by equation (130) and (131) and initial condition $(y_1(t_0), X_1(t_0))$ can be written as*

$$\begin{bmatrix} y_1(t+t_0) \\ X_1(t+t_0) \end{bmatrix} = e^{\mathcal{L}t} \cdot \begin{bmatrix} y_1(t_0) \\ X_1(t_0) \end{bmatrix} + e^{\mathcal{L}t} \cdot \int_{t_0}^{t+t_0} e^{-\mathcal{L}s} Ay_2(0) ds \tag{134}$$

Proof. It is directly to verify the derivate of right hind side satisfies equation (130) and (131), and the solution satisfies initial condition $(y_1(t_0), X_1(t_0))$. Due to the uniqueness of the solution of linear differential equation, we can conclude this is the solution of equation (130) and (131). □

Form (134) we can also see that if uncertainty are introduced to the system at some initial time t_0, i.e., $(y_1(t_0), X_1(t_0))$ is a random variable, then the evolution of covariance of random variable $(y_1(t+t_0), X_1(t+t_0))$ will not be affected by the term $\int_{t_0}^{t+t_0} e^{-\mathcal{L}s} Ay_2(0) ds$ since this is a determined quantity, thus will only affect the expectation of $(y_1(t+t_0), X_1(t+t_0))$. Therefore, without loss of generality, we will let $y_2(0) = 0$ in the following.

Thus the continuous equation of agent 1 is

$$\begin{bmatrix} \frac{dy_1}{dt} \\ \\ \frac{dX_1}{dt} \end{bmatrix} = \mathcal{L} \begin{bmatrix} y_1(t) \\ \\ X_1(t) \end{bmatrix} \tag{135}$$

where

$$\mathcal{L} = \begin{bmatrix} 0 & -AA^\top \\ \\ I & 0 \end{bmatrix} \in \mathbb{R}^{2n \times 2n}.$$

Since the solution of (135) is $\begin{bmatrix} y_1(t) \\ \\ X_1(t) \end{bmatrix} = e^{t\mathcal{L}} \begin{bmatrix} y_1(0) \\ \\ X_1(0) \end{bmatrix}$, we have

$$P(t+t_0) = e^{t\mathcal{L}} P(t_0)(e^{t\mathcal{L}})^\top.$$

Thus to analysis the behavior of $\mathrm{Var}(X_1(t))$, $\mathrm{Var}(y_1(t))$, and $\mathrm{Cov}(X_1(t), y_1(t))$, we need to understand the behavior of $e^{t\mathcal{L}}$. A standard method to calculate the matrix exponential of a matrix with elements in $\mathbb{R}$ is though the matrix's real Jordan normal form and real generalized modal matrix, see Proposition C.14. For our purpose, the most important question about the Jordan form of $\mathcal{L}$ is :

What is the size of the largest Jordan blocks corresponding to $\mathcal{L}'s$ eigenvalues ?

Because this number will determine the growth rate of elements in $e^{t\mathcal{L}}$. In Proposition C.20, we will determine the minimal polynomial of $\mathcal{L}$, combine this result with Proposition C.11, we can get elements in $e^{t\mathcal{L}}$ that will at most have a linear growth rate. However as we see in Theorem 5.1 there is a difference between $\mathrm{Var}(X_1(t))$ and $\mathrm{Var}(y_1(t))$, $\mathrm{Var}(X_1(t))$ may have quadratic growth and $\mathrm{Var}(y_1(t))$ will always be bounded, only the growth rate of $e^{t\mathcal{L}}$ is not sufficient for one to explain this difference. To show that $\mathrm{Var}(y_1(t))$ is always bounded, we need a more detailed analysis of the real generalized modal matrix of $\mathcal{L}$. We will see that the first n rows of the real generalized modal matrix has many 0 elements, and these 0 elements will make the first n rows of $e^{t\mathcal{L}}$ not to have linear growth rate. This will be shown in Proposition C.23.

Lemma C.18. *Let*

$$\mathcal{L} = \begin{bmatrix} 0 & -AA^\top \\ \\ I & 0 \end{bmatrix}$$

be the coefficient matrix of (135), then the eigenvalues of $\mathcal{L}$ are pure imaginary numbers or 0. Moreover, if 0 is an eigenvalue of $\mathcal{L}$, then its multiplicity is an even number.

Proof. Let $f(x)$ be the character polynomial of $-AA^\top$, thus

$$f(x) = \det(xI_{n\times n} + AA^\top) \tag{136}$$

Firstly, every eigenvalue of $-AA^\top$ is a negative number or 0. That is because if μ is an eigenvalue of $-AA^\top$ and v is an eigenvector of μ, then

$$\begin{aligned} \mu\langle v, v\rangle &= \langle -AA^\top v, v\rangle \\ &= -\langle A^\top v, A^\top v\rangle \\ &= -\left\|A^\top v\right\|^2 \leq 0, \end{aligned}$$

since $\langle v, v\rangle \geq 0$, thus we have $\mu \leq 0$. This implies the zeros of (136) are 0 or negative.

The character polynomial of $\mathcal{L}$ is

$$\det(\lambda I_{2n\times 2n} - \mathcal{L}) = \det(\lambda^2 I_{n\times n} + AA^\top) \tag{137}$$

$$= f(\lambda^2) \tag{138}$$

Thus the eigenvalues of $\mathcal{L}$ are square roots of zeros of (136). Since the zeros of (136) are 0 or negative, thus the eigenvalues of $\mathcal{L}$ are 0 or pure imaginary numbers. Moreover, since (137) is an even polynomial, it means that the polynomial only have terms of even degree, thus if 0 is a zero of (137), then its multiplicity is at least 2. □

Next we will calculate the minimal polynomial of $\mathcal{L}$. Combining the calculated results with Proposition C.11, we will get the size of the biggest Jordan blocks in the Jordan normal form of $\mathcal{L}$. Before that, we need the following lemma.

Lemma C.19. *(Corollary 3.3.10. in (Horn & Johnson, 2012))Let $M \in \mathbb{R}^{d\times d}$ and $f_M(x)$ be its minimal polynomial. Then M is diagonalizable if and only if every eigenvalue of M has multiplicity 1 as a root of $f_M(x) = 0$.*

Proposition C.20 (Minimal polynomial of $\mathcal{L}$)**.** *Let $\lambda_1 \neq \lambda_2 \neq ... \neq \lambda_l$ be the distinct eigenvalues of $-AA^\top$, thus $\forall i \in [l], \lambda_i \in \mathbb{R},\ \lambda_i \leq 0$. Then the minimal polynomial of $\mathcal{L}$ is :*

$$f_{\mathcal{L}}(x) = (x - \sqrt{\lambda_1}i)(x + \sqrt{\lambda_1}i)...(x - \sqrt{\lambda_l}i)(x + \sqrt{\lambda_l}i)$$

Note that this implies that eigenvalues of $\mathcal{L}$ are purely imaginary or 0, moreover, if $\lambda_i = 0$ for some i, then the factor x in $f_{\mathcal{L}}(x)$ has degree 2.

Proof. Let $f_{-AA^\top}(x)$ be the minimal polynomial of $-AA^\top$. Since $-AA^\top$ is symmetric, thus diagonalizable, by Lemma C.19, $f_{-AA^\top}(x)$ only contains linear factor of $(x - \lambda_i)$,where λ_i is an eigenvalue of $-AA^\top$. Thus we have

$$f_{-AA^\top}(x) = \prod_i (x - \lambda_i)$$

We claim $f_{-AA^\top}(\mathcal{L}^2) = 0$, since:

$$\mathcal{L}^2 = \begin{bmatrix} -AA^\top & 0 \\ 0 & -AA^\top \end{bmatrix},$$

therefore

$$f_{-AA^\top}(\mathcal{L}^2) = \begin{bmatrix} f_{-AA^\top}(-AA^\top) & 0 \\ 0 & f_{-AA^\top}(-AA^\top) \end{bmatrix} = 0.$$

Thus if $f_{\mathcal{L}}(x)$ is the minimal polynomial of $\mathcal{L}$, we have

$$f_{\mathcal{L}}(x) \mid f_{-AA^\top}(x^2) = \prod_j (x - \sqrt{\lambda_j}i)(x + \sqrt{\lambda_j}i) \tag{139}$$

Moreover, since every eigenvalue of $\mathcal{L}$ is also a root of $f_{\mathcal{L}}(x)$, from (139), we have :

(1) If $\pm\sqrt{\lambda_j}i \neq 0$ is an eigenvalue of $\mathcal{L}$, then the degree of $(x \pm \sqrt{\lambda_j}i)$ in $f_{\mathcal{L}}(x)$ must be 1.

(2) If $\lambda_j = 0$ is an eigenvalue of $\mathcal{L}$, then the degree of x in $f_{\mathcal{L}}(x)$ must be 1 or 2.

In the following arguments, we will prove that there is at least a real Jordan block corresponding to eigenvalue 0 has size 2. We have

$$\mathcal{L} = \begin{bmatrix} 0 & -AA^\top \\ I & 0 \end{bmatrix},\ \mathcal{L}^2 = \begin{bmatrix} -AA^\top & 0 \\ 0 & -AA^\top \end{bmatrix}.$$

Thus for some $x = (x_1, x_2) \in \mathrm{Ker}(\mathcal{L}^2)$,where $x_1, x_2 \in \mathbb{R}^n$. Then we have

$$\mathcal{L}^2 \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -AA^\top \cdot x_1 \\ -AA^\top \cdot x_2 \end{bmatrix} = 0.$$

This implies $x_1, x_2 \in \text{Ker} - AA^\top$. Since $-AA^\top$ has eigenvalue 0, x_1, x_2 may not equal to 0. But

$$\mathcal{L} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -AA^\top \cdot x_2 \\ x_1 \end{bmatrix} \quad ,$$

so if $(x_1, x_2) \in \text{Ker}\mathcal{L}$, x_1 must be 0. This completes the proof of $\text{Ker}(\mathcal{L}) \neq \text{Ker}(\mathcal{L}^2)$. By Proposition C.10, there exists Jordan block of size 2. □

Lemma C.21. *Let J_0 be the largest Jordan blocks of $\mathcal{L}$ corresponding to 0 eigenvalues, then elements in $e^{J_0 t}$ are of $\mathcal{O}(t)$. Let $J_{\mu,\bar{\mu}}$ be the largest Jordan blocks of $\mathcal{L}$ corresponding to a pair of conjugate purely imaginary eigenvalues $(\mu, \bar{\mu})$, then elements in $e^{J_{\mu,\bar{\mu}} t} \in \mathcal{O}(1)$.*

Proof. From Proposition C.20, the size of the largest Jordan blocks of $\mathcal{L}$ corresponding to the 0 eigenvalues are 2. Thus the corollary follows from (118) with $\mu = 0$ and $m = 2$. From Proposition C.20, the size of the largest Jordan blocks of $\mathcal{L}$ corresponding to a pair of conjugate imaginary eigenvalues $(\mu, \bar{\mu})$ are 2. So the corollary follows from (120) with $\lambda = 0$ and $m = 1$. □

Next we consider the set of real generalized vectors of $\mathcal{L}$. Let $S_\mathcal{L}$ be the set of vectors that appear in the real generalized modal matrix $M_\mathcal{L}$. Then as shown in Proposition C.20, these generalized eigenvectors corresponds to an imaginary eigenvalues or 0 eigenvalues. If $v \in S_\mathcal{L}$ corresponds to 0, v may in a Jordan chain with length 1 or length 2. If $v \in S_\mathcal{L}$ corresponds to a purely imaginary number, then there exists some w be the eigenvector of an imaginary eigenvalue and $v = \text{Re}(w)$ or $v = \text{Im}(w)$.

Thus we have

$$S_\mathcal{L} = (\cup_{(\mu,\bar{\mu})} S_\mathcal{L}(\mu, \bar{\mu})) \cup S_\mathcal{L}(0)$$

where

$$S_\mathcal{L}(\mu, \bar{\mu}) = \{v \in \mathbb{R}^{2n} | \ \exists\ w \text{ be an eigenvector for } \mu \text{ or } \bar{\mu} \in \mathbb{C} - \mathbb{R}\ , v = \text{Re}(w) \text{ or } v = \text{Im}(w)\},$$

and

$$S_\mathcal{L}(0) = \{v \in \mathbb{R}^{2n} | \ v \text{ is a generalized eigenvector for } 0\}.$$

So as in definition C.12, $M_\mathcal{L}$ has the following form:

$$M_\mathcal{L} = \big[\overbrace{v_1, \quad \dots \quad, v_{m_1}}^{m_1 \text{ columns}}, \ \overbrace{v_{1+m_1}, \quad \dots \quad, v_{m_1+m_2}}^{m_2 \text{ columns}}, \ \overbrace{v_{m_1+m_2+1}, \quad \dots \quad, v_{m_1+m_2+m_3}}^{m_3 \text{ columns}}\big] \tag{140}$$

where

- $\{v_1, ..., v_{m_1}\} \subset \cup_{(\lambda,\bar{\lambda})} S_\mathcal{L}(\mu, \bar{\mu})$.
- $\{v_{1+m_1}, ..., v_{m_1+m_2}\} \subset S_\mathcal{L}(0)$, and each each v_i is a Jordan chain of length 1.
- $\{v_{m_1+m_2+1}, ..., v_{m_1+m_2+m_3}\} \subset S_\mathcal{L}(0)$, (v_i, v_{i+1}) is a Jordan chain of length 2, and $v_i = \mathcal{L} v_{i+1}$.
- $m_1 + m_2 + m_3 = 2n$.

Lemma C.22. *If $w = (w_1, w_2, ..., w_{2n})^\top \neq 0$ is a type 1 generalized eigenvectors of $\mathcal{L}$ corresponding to 0 eigenvalue , then the first n components $(w_1, w_2, ..., w_n)^\top = 0$, and the the last n components $(w_{n+1}, ..., w_{2n})^\top \in \ker(-AA^\top)$.*

If $s = (s_1, s_2, ..., s_{2n})^\top \neq 0$ is a type 2 generalized eigenvectors of $\mathcal{L}$ corresponding to 0 eigenvalue , then $(s_1, ..., s_n)^\top \neq 0$ and $(s_1, ..., s_n)^\top, (s_{n+1}, ..., s_{2n})^\top \in \ker(-AA^\top)$.

Proof. Firstly, note that type 1 generalized eigenvectors of $\mathcal{L}$ corresponding to 0 are just vectors in $\ker(\mathcal{L})$. Thus if w is a type 1 generalized eigenvectors of $\mathcal{L}$ corresponding to 0, we have

$$\begin{aligned}\mathcal{L}\cdot w &= \begin{bmatrix} 0 & -AA^\top \\ I & 0 \end{bmatrix}\cdot (w_1, w_2, ..., w_{2n})^\top \\ &= \begin{bmatrix} -AA^\top \cdot (w_{n+1}, ..., w_{2n})^\top \\ \\ (w_1, ..., w_n)^\top \end{bmatrix} = 0\end{aligned}$$

Thus $(w_1, w_2, ..., w_n)^\top = 0$ and $(w_{n+1}, ..., w_{2n})^\top \in \ker(-AA^\top)$.

Secondly, if $s = (s_1, s_2, ..., s_{2n})^\top \neq 0$ is a type 2 generalized eigenvectors of $\mathcal{L}$ corresponding to 0 eigenvalue, then we have

$$\mathcal{L}\cdot s \neq 0 \text{ and } \mathcal{L}^2\cdot s = 0.$$

Since

$$\begin{aligned}\mathcal{L}^2\cdot s &= \begin{bmatrix} -AA^\top & 0 \\ \\ 0 & -AA^\top \end{bmatrix}\cdot (s_1, s_2, ..., s_{2n})^\top \\ &= \begin{bmatrix} -AA^\top \cdot (s_{n+1}, ..., s_{2n})^\top \\ \\ -AA^\top (s_1, ..., s_n)^\top \end{bmatrix} = 0\end{aligned}$$

This implies $(s_1, ..., s_n)^\top, (s_{n+1}, ..., s_{2n})^\top \in \ker(-AA^\top)$. Moreover, since s is a type 2 generalized eigenvector, we have $(s_1, ..., s_n)^\top \neq 0$. □

From the correspondence between real generalized vectors and real Jordan blocks as described in Proposition C.13, the real Jordan form $J_\mathcal{L}$ of $\mathcal{L}$ under a similar transformation by real generalized modal matrix $M_\mathcal{L}$ has form

$$J_\mathcal{L} = \begin{bmatrix} \overbrace{\begin{matrix} J_{\mu_1,\bar{\mu}_1} & & \\ & \ddots & \\ & & J_{\mu_{m_1/2},\bar{\mu}_{m_1/2}} \end{matrix}}^{m_1 \text{ columns}} & & & \\ & \overbrace{0}^{m_2 \text{ columns}} & & \\ & & \ddots & \\ & & & \overbrace{\begin{matrix} 0 & & \\ & J_0^2 & \\ & & \ddots \\ & & & J_0^2 \end{matrix}}^{m_3 \text{ columns}} \end{bmatrix} \tag{141}$$

$J_{\mu_k,\bar{\mu}_k} = \begin{bmatrix} \mathrm{Re}(\mu_k) & -\mathrm{Im}(\mu_k) \\ \\ \mathrm{Im}(\mu_k) & \mathrm{Re}(\mu_k) \end{bmatrix}$ are real Jordan blocks corresponding to conjugate eigenvalue $(\mu_k, \bar{\mu}_k)$ and $J_0^2 = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is the real Jordan block corresponding to the type 2 generalized eigenvectors of 0 eigenvalue. Thus we have

$$e^{tJ_\mathcal{L}} = \oplus_{i=1}^{m_1}(e^{tJ_{\mu_i,\bar{\mu}_i}}) \oplus_{i=1}^{m_2} I_1 \oplus_{i=1}^{m_3} e^{tJ_0^2} \tag{142}$$

where

- $e^{tJ_{\mu_i,\bar{\mu}_i}}$ is defined in (120) with $\lambda = 0,\ m = 1$. Thus elements in $e^{tJ_{\mu_i,\bar{\mu}_i}}$ are in $\mathcal{O}(1)$.

- $I_1 = [1] \in \mathbb{R}^{1\times 1}$.
- $e^{tJ_0^2}$ is defined in (118) with $\mu = 0, m = 2$. Matrix elements in $e^{tJ_0^2}$ are in $\mathcal{O}(t)$.

More precisely, we have

$$e^{tJ_{\mathcal{L}}} = \begin{bmatrix} \overbrace{\begin{matrix} \ddots & & \\ & \ddots & \end{matrix}}^{m_1+m_2} & & \\ & \overbrace{\begin{matrix} 1 & t & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & t \\ & & & & 1 \end{matrix}}^{m_3 \text{ columns}} \end{bmatrix} \tag{143}$$

and only elements in the upper diagonal of the last m_3 columns belongs to $\mathcal{O}(t)$, other elements are all bounded function of t.

Lemma C.23. *Let $(e^{t\mathcal{L}})_i$ denote the i-th row of $e^{t\mathcal{L}}$. Then we have*

- *If $AA^\top$ has 0 eigenvalues, then for $i \in \{1, 2, ..., n\}$, elements in $(e^{t\mathcal{L}})_i$ are bounded and for $i \in \{n+1, n+2, ..., 2n\}$, $(e^{t\mathcal{L}})_i$ belongs to $\mathcal{O}(t)$.*
- *If $AA^\top$ doesn't have 0 eigenvalues, all elements in $e^{t\mathcal{L}}$ are bounded.*

Proof. Let $(M_{\mathcal{L}})^{-1} = [p_1, p_2, ..., p_{2n}],\ p_i \in \mathbb{R}^{2n}$ be the inverse of the real generalized modal matrix of $\mathcal{L}$, then we have

$$e^{t\mathcal{L}} = M_{\mathcal{L}}(e^{tJ_{\mathcal{L}}}[w_1, w_2, ..., w_{2n}]). \tag{144}$$

If $AA^\top$ has 0 eigenvalues, from (143), we have $e^{tJ_{\mathcal{L}}}w_i$ has form

$$\begin{bmatrix} \vdots \\ \vdots \\ \vdots \\ \hline t \\ \mathcal{O}(1) \\ \vdots \\ t \\ \mathcal{O}(1) \end{bmatrix} \begin{matrix} \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} \in \mathcal{O}(1),\ m_1+m_2 \text{ rows} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} m_3 \text{ rows} \end{matrix} . \tag{145}$$

From Lemma C.22, the first n rows of $M_{\mathcal{L}}$ has form

$$\overbrace{\begin{bmatrix} & \end{bmatrix}}^{m_1+m_2} \Bigg| \overbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{bmatrix}}^{m_3 \text{ columns}} \Bigg\} \text{n rows}$$

$$\begin{bmatrix} \cdots & \cdots & \vdots & v_{m+2} & \vdots & \cdots & \vdots & v_{2n} \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 0 \\ \vdots & \vdots & \vdots \end{bmatrix} \Bigg\} \text{n rows} \quad . \tag{146}$$

The first n rows of $M_{\mathcal{L}} e^{tJ_{\mathcal{L}}} p_i$ is the matrix-vector product of (146) and (145), we can see the term t in (145) will product with term 0 in (146), thus the first n rows of $M_{\mathcal{L}} e^{tJ_{\mathcal{L}}} p_i$ belong to $\mathcal{O}(1)$. The last n rows of $M_{\mathcal{L}} e^{tJ_{\mathcal{L}}} p_i$ belong to $\mathcal{O}(t)$ because there may exist nonzero elements in the last n rows of $M_{\mathcal{L}}$ that can product with t in (145).

If $AA^\top$ doesn't have 0 eigenvalues, from Lemma C.21, all elements in $e^{J_{\mathcal{L}} t}$ are bounded and since $e^{t\mathcal{L}} = M_{\mathcal{L}} e^{tJ_{\mathcal{L}}} (M_{\mathcal{L}})^{-1}$, thus all elements in $e^{t\mathcal{L}}$ are bounded. □

The covariance evolution directly follows from above calculation.

Proof of covariance in continuous time equation. Denote

$$P(t_0) = \begin{bmatrix} \mathrm{Var}(y_1(t_0)) & \mathrm{Cov}(y_1(t_0), X_1(t_0)) \\ \mathrm{Cov}(y_1(t_0), X_1(t_0)) & \mathrm{Var}(X_1(t_0)) \end{bmatrix}$$

and $e^{t\mathcal{L}} = \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix}$. Since $P(t+t_0) = e^{t\mathcal{L}} P(t_0) (e^{t\mathcal{L}})^\top$, we have

$$P(t+t_0) = \begin{bmatrix} A(t) & B(t) \\ C(t) & D(t) \end{bmatrix} \begin{bmatrix} \mathrm{Var}(y_1(t_0)) & \mathrm{Cov}(y_1(t_0), X_1(t_0)) \\ \mathrm{Cov}(y_1(t_0), X_1(t_0)) & \mathrm{Var}(X_1(t_0)) \end{bmatrix} \begin{bmatrix} (A(t))^\top & (C(t))^\top \\ (B(t))^\top & (D(t))^\top \end{bmatrix}$$

$$= \begin{bmatrix} A(t)\mathrm{Var}(y_1(t_0)) + B(t)\mathrm{Cov}(y_1(t_0), X_1(t_0)) & A(t)\mathrm{Cov}(y_1(t_0), X_1(t_0)) + B(t)\mathrm{Var}(X_1(t_0)) \\ C(t)\mathrm{Var}(y_1(t_0)) + D(t)\mathrm{Cov}(y_1(t_0), X_1(t_0)) & C(t)\mathrm{Cov}(y_1(t_0), X_1(t_0)) + D(t)\mathrm{Var}(X_1(t_0)) \end{bmatrix} \begin{bmatrix} (A(t))^\top & (C(t))^\top \\ (B(t))^\top & (D(t))^\top \end{bmatrix}$$

$$= \begin{bmatrix} (A\mathrm{Var}(y_1) + B\mathrm{Cov})A^\top + (A\mathrm{Cov} + B\mathrm{Var}(X_1))B^\top & (A\mathrm{Var}(y_1) + B\mathrm{Cov})C^\top + (A\mathrm{Cov} + B\mathrm{Var}(X_1))D^\top \\ (C\mathrm{Var}(y_1) + D\mathrm{Cov})A^\top + (C\mathrm{Cov} + D\mathrm{Var}(X_1))B^\top & (C\mathrm{Var}(y_1) + D\mathrm{Cov})C^\top + (C\mathrm{Cov} + D\mathrm{Var}(X_1))D^\top \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Var}(y_1(t_0+t)) & \mathrm{Cov}(y_1(t_0+t), X_1(t_0+t)) \\ \mathrm{Cov}(y_1(t_0+t), X_1(t_0+t)) & \mathrm{Var}(X_1(t_0+t)) \end{bmatrix}$$

From Lemma C.23, if $AA^\top$ has 0 eigenvalue, elements in $A(t), B(t)$ are bounded and elements in $C(t), D(t)$ are in $\mathcal{O}(t)$ and there exist elements in $C(t), D(t)$ belongs to $\Theta(t)$, thus from above equation we have $\mathrm{Var}(y_1(t_0+t)) \in \mathcal{O}(1)$, $\mathrm{Cov}(y_1(t_0), X_1(t_0+t)) \in \Theta(t)$,$\mathrm{Var}(X_1(t_0+t)) \in \Theta(t^2)$. Moreover, if $AA^\top$ doesn't have 0 eigenvalue, Lemma C.23 implies all elements in $A(t), B(t), C(t), D(t)$ are bounded, thus all elements in $P(t+t_0)$ are bounded. □

C.2.2. Covariance evolution of Euler discretization

Proof. The Euler discretization of (135) with step size η can be written as

$$\begin{bmatrix} y_1^t \\ X_1^t \end{bmatrix} = \begin{bmatrix} y_1^{t-1} \\ X_1^{t-1} \end{bmatrix} + \begin{bmatrix} 0 & -\eta AA^T \\ \eta I & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1^{t-1} \\ X_1^{t-1} \end{bmatrix} \qquad = \begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix} \cdot \begin{bmatrix} y_1^{t-1} \\ X_1^{t-1} \end{bmatrix}. \tag{147}$$

We want to prove if (y_1^t, X_1^t) evolve as this equation, then $\mathrm{Cov}(y_1^t), \mathrm{Cov}(X_1^t)$ have exponential growth rate. This can be done by calculating the real Jordan normal form of $\begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}$. Since $AA^\top \in \mathbb{R}^{n\times n}$ is a diagonalizable matrix, there exists a matrix P, such that

$$PAA^\top P^{-1} = \begin{bmatrix} \gamma_1 & & & \\ & \gamma_2 & & \\ & & \ddots & \\ & & & \gamma_n \end{bmatrix} \tag{148}$$

where $\gamma_1, ..., \gamma_n$ are eigenvalues of $AA^\top$. Moreover, since $AA^\top$ is a positive semidefinite matrix, we have $\gamma_i \geq 0, i \in [n]$.

Then we have

$$\begin{bmatrix} P & \\ & P \end{bmatrix} \begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix} \begin{bmatrix} P^{-1} & \\ & P^{-1} \end{bmatrix} = \left[\begin{array}{cccc|cccc} 1 & & & & -\gamma_1\eta & & & \\ & 1 & & & & -\gamma_2\eta & & \\ & & \ddots & & & & \ddots & \\ & & & 1 & & & & -\gamma_n\eta \\ \hline \eta & & & & 1 & & & \\ & \eta & & & & 1 & & \\ & & \ddots & & & & \ddots & \\ & & & \eta & & & & 1 \end{array}\right]. \tag{149}$$

This matrix is similar to $\begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}$, thus they have same character polynomial defined by

$$\det(\mu I - \left[\begin{array}{ccc|ccc} 1 & & & -\gamma_1\eta & & \\ & 1 & & & -\gamma_2\eta & \\ & & \ddots & & & \ddots \\ & & 1 & & & -\gamma_n\eta \\ \hline \eta & & & 1 & & \\ & \eta & & & 1 & \\ & & \ddots & & & \ddots \\ & & \eta & & & 1 \end{array}\right]) \tag{150}$$

$$= \det(\left[\begin{array}{ccc|ccc} \mu-1 & & & \gamma_1\eta & & \\ & \mu-1 & & & \gamma_2\eta & \\ & & \ddots & & & \ddots \\ & & \mu-1 & & & -\gamma_n\eta \\ \hline -\eta & & & \mu-1 & & \\ & -\eta & & & \mu-1 & \\ & & \ddots & & & \ddots \\ & & -\eta & & & \mu-1 \end{array}\right]) \tag{151}$$

$$= \det(\begin{bmatrix} (\mu-1)^2+\gamma_1\eta^2 & & & \\ & (\mu-1)^2+\gamma_2\eta^2 & & \\ & & \ddots & \\ & & & (\mu-1)^2+\gamma_n\eta^2 \end{bmatrix}) \tag{152}$$

$$= \prod_{i=1}^{n}((\mu-1)^2+\gamma_i\eta^2) \tag{153}$$

Recall we have $\gamma_i \geq 0, i \in [n]$, thus the root of character polynomial is $\mu_j, \bar{\mu}_j = 1 \pm i\sqrt{\gamma_j}\eta,\ j \in [n]$. Moreover, we have $|\mu_j| = 1+\gamma_j\eta^2 \geq 1$. Thus by Proposition C.16, elements in $\begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}^t$ have an exponential growth rate as $\Theta(|\mu|^t)$, where μ is the eigenvalue of $AA^\top$ which has largest norm.

Since

$$P(t) = \begin{bmatrix} \mathrm{Var}(y_1^t) & \mathrm{Cov}(y_1^t, X_1^t) \\ \mathrm{Cov}(X_1^t, y_1^t) & \mathrm{Var}(X_1^t) \end{bmatrix} \tag{154}$$

$$= \begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}^{t-t_0} \begin{bmatrix} \mathrm{Var}(y_1^{t_0}) & \mathrm{Cov}(y_1^{t_0}, X_1^{t_0}) \\ \mathrm{Cov}(X_1^{t_0}, y_1^{t_0}) & \mathrm{Var}(X_1^{t_0}) \end{bmatrix} (\begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}^{t-t_0})^\top \tag{155}$$

Since elements in matrix $\begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}^{t-t_0}$ has growth rate $\Theta(|\mu|^t)$, thus elements in $P(t)$ has growth rate $\Theta(|\mu|^{2t})$. □

C.2.3. Covariance evolution of Symplectic discretization

We firstly determine the Symplectic discretization of continuous FTRL (135) with Euclidian norm regularizer.

Lemma C.24. *Discrete continuous FTRL (135) with (Type I method), we get*

$$\begin{bmatrix} y_1^{n+1} \\ X_1^{n+1} \end{bmatrix} = \begin{bmatrix} I & -\eta AA^\top \\ \eta & I - \eta^2 AA^\top \end{bmatrix} \cdot \begin{bmatrix} y_1^n \\ X_1^n \end{bmatrix}. \tag{156}$$

Proof. Directly calculate gives

$$y_1^{t+1} = y_1^t - \eta AA^\top X_1^t, \tag{157}$$
$$X_1^{t+1} = X_1^t + \eta y_1^{t+1} \tag{158}$$
$$= X_1^t + \eta y_1^t - \eta^2 AA^\top X_1^t. \tag{159}$$

Combine above gives

$$\begin{bmatrix} y_1^{n+1} \\ X_1^{n+1} \end{bmatrix} = \begin{bmatrix} I & -\eta AA^\top \\ \eta & I - \eta^2 AA^\top \end{bmatrix} \cdot \begin{bmatrix} y_1^n \\ X_1^n \end{bmatrix}. \tag{160}$$

This finish the proof. □

Let $\mathcal{M}_\eta = \begin{bmatrix} I & -AA^\top \cdot \eta \\ \eta & I - AA^\top \cdot \eta^2 \end{bmatrix}$. Since $AA^\top$ is diagonalizable, and $AA^\top$'s eigenvalues are all non-negative, there exists a matrix $P \in \mathbb{R}^{n\times n}$ and P invertible, such that

$$P^{-1} AA^\top P = \begin{bmatrix} \gamma_1 & & & \\ & \gamma_2 & & \\ & & \ddots & \\ & & & \gamma_n \end{bmatrix} \tag{161}$$

where $\gamma_1, ..., \gamma_n \geq 0$ are eigenvalues of $AA^\top$.

Lemma C.25. *$\mathcal{M}_\eta$ has real eigenvalues if and only if $AA^\top$ has 0 eigenvalue, and in this case the only real eigenvalue of $\mathcal{M}_\eta$ equals to 1. Moreover, every image eigenvalue of $\mathcal{M}_\eta$ has norm equals to 1.*

Proof. With (161), we have

$$\det(\mu I - \mathcal{M}_\eta) = \det(\begin{bmatrix} \mu - I & AA^\top \eta \\ -\eta & \mu - I + AA^\top \eta^2 \end{bmatrix}) \tag{162}$$
$$= \det((\mu - I)(\mu - I + AA^\top \eta^2) + AA^\top \eta^2) \tag{163}$$
$$= \det(\mu^2 - 2\mu + I + \mu AA^\top \eta^2) \tag{164}$$
$$\overset{(161)}{=} \det(\begin{bmatrix} \mu^2 - 2\mu + 1 + \mu\eta^2\gamma_1 & & & \\ & \mu^2 - 2\mu + 1 + \mu\eta^2\gamma_2 & & \\ & & \ddots & \\ & & & \mu^2 - 2\mu + 1 + \mu\eta^2\gamma_n \end{bmatrix}) \tag{165}$$
$$= \prod_{i=1}^n (\mu^2 - 2\mu + 1 + \mu\eta^2\gamma_i) \tag{166}$$

From above, we can see if μ makes $\det(\mu I - \mathcal{M}_\eta) = 0$, then there exists some $i \in [n]$, such that

$$\mu^2 - 2\mu + 1 + \mu\eta^2\gamma_i = 0. \tag{167}$$

(167) is a quadratic function about μ, and has solution $\mu = \frac{(2-\eta^2\gamma_i) \pm \sqrt{(\eta^2\gamma_i-2)^2-4}}{2}$.

To make $\mu \in \mathbb{R}$, we need $(\eta^2\gamma_i - 2)^2 - 4 \geq 0$. For sufficient small η ($\eta^2\gamma_i \leq 2$), that can only happen when $\gamma_i = 0$, and in this case we have $\mu = 1$. Moreover, if $(\eta^2\gamma_i - 2)^2 - 4 < 0$, we have

$$\|\mu\|^2 = \|\frac{(2-\eta^2\gamma_i) \pm i\sqrt{4-(\eta^2\gamma_i-2)^2}}{2}\|^2 \tag{168}$$

$$= \frac{(2-\eta^2\gamma_i)^2 + 4 - (\eta^2\gamma_i - 2)^2}{4} \tag{169}$$

$$= 1 \tag{170}$$

Thus we have

- $\mathcal{M}_\eta$ has a real eigenvalue if and only if $\gamma = 0$ is an eigenvalue of $AA^\top$, and in this case the real eigenvalue of $\mathcal{M}_\eta$ equals to 1.
- Every image eigenvalue of $\mathcal{M}_\eta$ has norm equals to 1.

□

Lemma C.26. *The largest Jordan blocks corresponding to $\mu = 1$ have size 2.*

Proof. As in proposition C.10, the number of size k Jordan blocks corresponding to eigenvalue 1 is determined by the dimension of linear space $\ker(\mathcal{M}_\eta - I)^k$ and $\ker(\mathcal{M}_\eta - I)^{k-1}$. Since similar matrixes have same minimal polynomial and same kernel space, thus we can change $\mathcal{M}_\eta$ to any similar matrix. Thus by (161), we only need to consider

$$\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \mathcal{M}_\eta \begin{bmatrix} P^{-1} & 0 \\ 0 & P^{-1} \end{bmatrix} - I = \left[\begin{array}{cccc|cccc} 0 & & & & -\gamma_1\eta & & & \\ & 0 & & & & -\gamma_2\eta & & \\ & & \ddots & & & & \ddots & \\ & & & 0 & & & & -\gamma_n\eta \\ \hline \eta & & & & -\gamma_1\eta^2 & & & \\ & \eta & & & & -\gamma_2\eta^2 & & \\ & & \ddots & & & & \ddots & \\ & & & \eta & & & & -\gamma_n\eta^2 \end{array}\right] \tag{171}$$

Our claim is : $k = 3$ is the smallest number to make $\rho_k = 0$ in proposition C.10, thus by proposition C.10 the largest Jordan block corresponding to eigenvalue 1 have size 2 .

As shown in lemma C.25, 1 is an eigenvalue of $\mathcal{M}_\eta$ if and only if 0 is an eigenvalue of $AA^\top$. Assume 0 is an eigenvalue of $AA^\top$ with multiplicity m, then it is easy to see (171) has rank $2n - m$. A direct calculate show the square of (171) equals to

$$\left[\begin{array}{cccc|cccc} -\gamma_1\eta^2 & & & & -\gamma_1\eta^3 & & & \\ & -\gamma_2\eta^2 & & & & -\gamma_2\eta^3 & & \\ & & \ddots & & & & \ddots & \\ & & & -\gamma_n\eta^2 & & & & -\gamma_n\eta^3 \\ \hline -\gamma_1\eta^3 & & & & -\gamma_1\eta^2+\gamma_1^2\eta^4 & & & \\ & -\gamma_2\eta^3 & & & & -\gamma_2\eta^2+\gamma_2^2\eta^4 & & \\ & & \ddots & & & & \ddots & \\ & & & -\gamma_n\eta^3 & & & & -\gamma_n\eta^2+\gamma_n^2\eta^4 \end{array}\right] \tag{172}$$

Thus if 0 is an eigenvalue of $AA^\top$ with multiplicity m, then there are $2m$ rows in (172) be 0, thus (172) has rank $2n - 2m$, this shows the ρ_2 in proposition C.10 is not zero, which implies there exists Jordan blocks with size 2×2 corresponding to eigenvalue 1.

Moreover, the cubic of (171) equals to

$$\left[\begin{array}{cccc|cccc} \gamma_1^2\eta^4 & & & & \gamma_1\eta^3-\gamma_1^3\eta^5 & & & \\ & \gamma_2^2\eta^4 & & & & \gamma_2\eta^3-\gamma_2^3\eta^5 & & \\ & & \ddots & & & & \ddots & \\ & & & \gamma_n^2\eta^4 & & & & \gamma_n\eta^3-\gamma_n^3\eta^5 \\ \hline -\gamma_1\eta^3 & & & & 2\gamma_1^2\eta^4-\gamma_1^3\eta^6 & & & \\ & -\gamma_2\eta^3 & & & & 2\gamma_2^2\eta^4-\gamma_2^3\eta^6 & & \\ & & \ddots & & & & \ddots & \\ & & & -\gamma_n\eta^3 & & & & 2\gamma_n^2\eta^4-\gamma_n^3\eta^6 \end{array}\right] \tag{173}$$

We can see the rank of (173) is also $2n-2m$, thus ρ_3 in proposition C.10 is zero. This finish the proof of our claim. □

The following lemma determines $\mathcal{M}_\eta$'s type 1 and type 2 generalized eigenvectors of $\mathcal{M}_\eta$ corresponding to eigenvalue $\mu=1$. Recall $v\in\mathbb{R}^{2n}$ is a type 2 generalized eigenvectors of $\mathcal{M}_\eta$ corresponding to eigenvalue $\mu=1$ if $v\in\ker(\mathcal{M}_\eta-I)^2$ but $v\notin\ker(\mathcal{M}_\eta-I)$.

Lemma C.27. *Let $x,y\in\mathbb{R}^n$, then*

(1) If $(x,y)\in\mathbb{R}^{2n}$ is a type 2 generalized eigenvectors of $\mathcal{M}_\eta$ corresponding to eigenvalue $\mu=1$, then $x,y\in\ker(AA^\top)$ and $x\neq 0$.

(2) If $(x,y)\in\mathbb{R}^{2n}$ is a type 1 generalized eigenvectors of $\mathcal{M}_\eta$ corresponding to eigenvalue $\mu=1$, then $y\in\ker(AA^\top)$ and $x=0$.

Proof. We have

$$(\mathcal{M}_\eta-I)^2=\begin{bmatrix}-AA^\top\eta^2 & (AA^\top)^2\eta^3\\ -AA^\top\eta^3 & -AA^\top\eta^2+(AA^\top)^2\eta^4\end{bmatrix}. \tag{174}$$

Thus if $(x,y)\in\ker(\mathcal{M}_\eta-I)^2$, we have

$$-AA^\top\eta^2x+(AA^\top)^2\eta^3y=0 \tag{175}$$

$$-AA^\top\eta^3x+(-AA^\top\eta^2+(AA^\top)^2\eta^4)y=0 \tag{176}$$

(176) $-\eta\cdot$ (175) gives

$$-AA^\top\eta^2y=0. \tag{177}$$

Thus we have $y\in\ker(AA^\top)$, and take this back to (176), we get $x\in\ker(AA^\top)$.

Moreover, it is directly to verify if $(x,y)\in\ker(\mathcal{M}_\eta-I)$, then $x=0$ and $y\in\ker(AA^\top)$. Thus if (x,y) is a type 2 generalized eigenvector, $x\in\ker(AA^\top)$, $x\neq 0$ and $y\in\ker(AA^\top)$.

□

Lemma C.28. *For an image eigenvalue $\mu\neq 1$ of $\mathcal{M}_\eta$, the largest real Jordan blocks corresponding to μ have size 2.*

Proof. The proof is similar to lemma C.26. Let $\mu\neq 1$ be an image eigenvalue of $\mathcal{M}_\eta$, by (167), there exists an eigenvalue γ_i of $AA^\top$ to make

$$\mu^2-2\mu+1+\mu\eta^2\gamma_i=0. \tag{178}$$

Moreover, the algebraic multiplicity of μ as an eigenvalue of $\mathcal{M}_\eta$ is same as the algebraic multiplicity of γ as an eigenvalue of $AA^\top$. We assume $\mu\neq 1$ has algebraic multiplicity m.

Consider the matrix

$$\mu - \begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \mathcal{M}_\eta \begin{bmatrix} P^{-1} & 0 \\ 0 & P^{-1} \end{bmatrix} = \left[\begin{array}{ccc|ccc} \mu-1 & & & \gamma_1\eta & & \\ & \ddots & & & \ddots & \\ & & \mu-1 & & & \gamma_n\eta \\ \hline -\eta & & & \mu-1+\gamma_1\eta^2 & & \\ & \ddots & & & \ddots & \\ & & -\eta & & & \mu-1+\gamma_n\eta^2 \end{array}\right] \tag{179}$$

It is easy to see if

$$\mu^2 - 2\mu + 1 + \mu\eta^2\gamma_i = 0 \tag{180}$$

the i-th and $n+i$-th row of (179) are linearly dependent. Moreover, if if μ as rank m, the multiplicity of γ_i as an eigenvalue of $AA^\top$ is also m. So there are m rows in the first n rows of (179) linearly dependent on m rows in the last n rows of (179). Thus if μ as rank m, (179) has rank $2n-m$.

A directly calculate shows the square of (179) equals to

$$\left[\begin{array}{ccc|ccc} (\mu-1)^2-\gamma_1\eta^2 & & & \gamma_1\eta(2\mu-2+\gamma_1\eta^2) & & \\ & \ddots & & & \ddots & \\ & & (\mu-1)^2-\gamma_n\eta^2 & & & \gamma_n\eta(2\mu-2+\gamma_n\eta^2) \\ \hline -\eta & & & -\gamma_1\eta^2+(\mu-1+\gamma_1\eta)^2 & & \\ & \ddots & & & \ddots & \\ & & -\eta & & & -\gamma_n\eta^2+(\mu-1+\gamma_n\eta)^2 \end{array}\right] \tag{181}$$

Moreover, if μ and γ_i satisfy (180), the i-th row and $n+i$-th row are linearly dependent, thus same as matrix in (179), has rank $2n-m$ if μ is an eigenvalue of $\mathcal{M}_\eta$ with multiplicity m.

Thus we have $\rho_2 = 0$ in proposition C.10, this implies the real Jordan blocks of a pair of conjugate eigenvalues $(\mu, \bar{\mu})$ have size 2. □

Lemma C.29. *Let t be a positive integer, and $(\mathcal{M}_\eta^t)_i$ be the i-th row of the matrix $\mathcal{M}_\eta^t$ for $i = 1, 2, ..., 2n$. We have*

- *If $AA^\top$ is singular, then for $1 \le i \le n$, elements in $(\mathcal{M}_\eta^t)_i$ is bounded, for $n+1 \le i \le 2n$, elements in $(\mathcal{M}_\eta^t)_i$ have growth rate $\mathcal{O}(t)$.*
- *If $AA^\top$ is non-singular, then for all $1 \le i \le 2n$, elements in $(\mathcal{M}_\eta^t)_i$ is bounded.*

Proof. Denote the real generalized modal matrix of $\mathcal{M}_\eta$ by M, then we have $\mathcal{M}_\eta^t = M J_{\mathcal{M}_\eta}^t M^{-1}$, where $J_{\mathcal{M}_\eta}$ is the real Jordan normal form of $\mathcal{M}_\eta$. If $AA^\top$ has 0 eigenvalue, then from Lemma C.25 and Lemma C.26, 1 is the only possible real eigenvalue of $\mathcal{M}_\eta$, and the largest size real Jordan block have size 2. Moreover in this case, form Lemma C.27 the real generalized modal matrix of $\mathcal{M}_\eta$ has same structure as the real generalized modal matrix of $\mathcal{L}$ as in (146), thus the proof follows from a same argument as in Proposition C.23.

If $AA^\top$ doesn't have 0 eigenvalue, all eigenvalues of $\mathcal{M}_\eta$ are image numbers and form Lemma C.25 these eigenvalues have norm 1. From Lemma C.28 the largest size real Jordan blocks of $\mathcal{M}_\eta$ is 2×2, thus from Proposition C.16 with $m = 1$, all elements in $J_{\mathcal{M}_\eta}^t$ are bounded, thus all elements in $\mathcal{M}_\eta^t$ are bounded. □

The covariance evolution of Symplectic discretization directly follows from above calculations. Denote the covariance matrix at time $t+t_0$ by $P(t+t_0)$ and $\mathcal{M}_\eta^t = \begin{bmatrix} A^t & B^t \\ C^t & D^t \end{bmatrix}$, then we have

$$P(t+t_0) = \begin{bmatrix} \mathrm{Var}(y_1^{t+t_0}) & \mathrm{Cov}(y_1^{t+t_0}, X_1^{t+t_0}) \\ \mathrm{Cov}(X_1^{t+t_0}, y_1^{t+t_0}) & \mathrm{Var}(X_1^{t+t_0}) \end{bmatrix} \tag{182}$$

$$= \mathcal{M}_\eta^t \begin{bmatrix} \mathrm{Var}(y_1^{t_0}) & \mathrm{Cov}(y_1^{t_0}, X_1^{t_0}) \\ \mathrm{Cov}(X_1^{t_0}, y_1^{t_0}) & \mathrm{Var}(X_1^{t_0}) \end{bmatrix} (\mathcal{M}_\eta^t)^\top \tag{183}$$

$$= \begin{bmatrix} A^t & B^t \\ C^t & D^t \end{bmatrix} \begin{bmatrix} \mathrm{Var}(y_1^{t_0}) & \mathrm{Cov}(y_1^{t_0}, X_1^{t_0}) \\ \mathrm{Cov}(X_1^{t_0}, y_1^{t_0}) & \mathrm{Var}(X_1^{t_0}) \end{bmatrix} (\begin{bmatrix} A^t & B^t \\ C^t & D^t \end{bmatrix})^\top \tag{184}$$

$$= \begin{bmatrix} A^t\mathrm{Var}(y_1^{t_0}) + B^t\mathrm{Cov}(X_1^{t_0}, y_1^{t_0}) & A^t\mathrm{Cov}(y_1^{t_0}, X_1^{t_0}) + B^t\mathrm{Var}(X_1^{t_0}) \\ C^t\mathrm{Var}(y_1^{t_0}) + D^t\mathrm{Cov}(X_1^{t_0}, y_1^{t_0}) & C^t\mathrm{Cov}(y_1^{t_0}, X_1^{t_0}) + D^t\mathrm{Var}(X_1^{t_0}) \end{bmatrix} (\begin{bmatrix} (A^t)^\top & (C^t)^\top \\ (B^t)^\top & (D^t)^\top \end{bmatrix})^\top \tag{185}$$

The finial equation (185) equals to

$$\begin{bmatrix} P_1^t & P_2^t \\ P_3^t & P_4^t \end{bmatrix}, \tag{186}$$

where

$$P_1^t = (A^t\mathrm{Var}(y_1^{t_0}) + B^t\mathrm{Cov})(A^t)^\top + (A^t\mathrm{Cov} + B^t\mathrm{Var}(X_1^{t_0}))(B^t)^\top, \tag{187}$$

$$P_2^t = (A^t\mathrm{Var}(y_1^{t_0}) + B^t\mathrm{Cov})(C^t)^\top + (A^t\mathrm{Cov} + B^t\mathrm{Var}(X_1^{t_0}))(D^t)^\top, \tag{188}$$

$$P_3^t = (C^t\mathrm{Var}(y_1^{t_0}) + D^t\mathrm{Cov})(A^t)^\top + (C^t\mathrm{Cov} + D^t\mathrm{Var}(X_1^{t_0}))(B^t)^\top, \tag{189}$$

$$P_4^t = (C^t\mathrm{Var}(y_1^{t_0}) + D^t\mathrm{Cov})(C^t)^\top + (C^t\mathrm{Cov} + D^t\mathrm{Var}(X_1^{t_0}))(D^t)^\top. \tag{190}$$

From Lemma C.29, when $AA^\top$ is non-singular, all elements in $\mathcal{M}_\eta^t$ are bounded, thus elements in $P(t+t_0)$ is bounded. When $AA^\top$ is singular, elements in A^t, B^t are bounded and elements in C^t, D^t has linear growth rate, thus $\mathrm{Var}(y^t) \in \mathcal{O}(1)$, $\mathrm{Cov}(X^t, y^t) \in \Theta(t)$ and $\mathrm{Cov}(X_i^t, X_j^t) \in \Theta(t^2)$.

D. Proof of Section 5

D.1. Riemannian Game Dynamics

We collect minimum amount of terminologies on Riemannian game dynamics, for a complete treatment on this topic, we refer to (Mertikopoulos & Sandholm, 2018). For the case of population game $\mathcal{G}(\mathcal{A}, v)$ where $\mathcal{A}$ is the strategy set and v is the set of utilities, the *gain from motion* from state $x \in \mathcal{X}$ along $z \in \mathbb{R}^{\mathcal{A}}$ is defined as

$$G^v(x; z) = \sum_{\alpha \in \mathcal{A}} v_\alpha(x) z_\alpha.$$

The *cost of motion* $C(x; z)$ represents the intrisic difficulty of moving from state x along a given displacement vector z, and it is defined to be

$$C(x; z) = \frac{1}{2} g_x(z, z)$$

where g is a smooth assignment of symmetric positive definite matrices g_x to each state $x \in \mathcal{X}$. The vector of motion from state x is required to maximize the difference between the gain of motion $G^v(x; z)$ and the cost of motion $C(x; z)$ subject to

$$\dot{x} = \arg\max_{z \in T_x\mathcal{X}} \{G^v(x; z) - C(x; z)\}. \tag{191}$$

The dynamics equation 191 is called *Riemannian game dynamics*.

D.2. Symplectic Geometry

Symplectic form. In order to present Gromov's non-squeezing theorem and its implication in uncertainty principle, we need some terminology of Symplectic Geometry. Roughly speaking, symplectic geometry studies the geometry of the space (of even dimension) equipped with the symplectic form. Take Euclidean geometry for example, it studies the vector space $\mathbb{R}^n$ with an inner product structure $\langle , \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ called the Euclidean structure. A symplectic form on a even dimensional space $\mathbb{R}^n$ is a *skew-symmetric* bilinear map $\omega(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, satisfying

- $\omega(u, v) = -\omega(v, u)$ for all $u, v \in \mathbb{R}^n$.
- $\omega(v, v) = 0$ for all $v \in \mathbb{R}^n$.
- $\omega(u, v) = 0$ for all $v \in \mathbb{R}^n$ implies that $u = 0$.

A typical symplectic form is the bilinear map defined by matrix $J = \begin{pmatrix} 0 & I_n \\ -I_n & 0 \end{pmatrix}$, where I_n denotes the identity matrix on $\mathbb{R}^n$. A basis $\{u_1, ..., u_n, v_1, ..., v_n\}$ of $\mathbb{R}^{2n}$ is called ω-standard if $\omega(u_j, u_k) = -\omega(v_j, v_k) = 0$ and $\omega(u_j, v_k) = \delta_{jk}$.

Symplectomorphism. A symplectomorphism φ between symplectic vector spaces $(\mathbb{R}^n, \omega)$ and $(\mathbb{R}^n, \omega')$ is a linear isomorphism $\varphi : \mathbb{R}^n \to \mathbb{R}^n$ such that $\varphi^*\omega' = \omega$, where $(\varphi^*\omega')(u, v) := \omega'(\varphi(u), \varphi(v))$. More generally, let $f : (\mathbb{R}^n, \omega) \to (\mathbb{R}^n, \omega')$ be a diffeomorphism. Then f is a symplectomorphism if $f^*\omega' = \omega$. These definitions generalize to manifold settings, but further discussion is beyond the scope of current paper. In the plane, symplectic form represents area, so a symplectic mapping is equivalent to an area preserving mapping.

Linear symplectic width. The main technique in the analysis of general regularizers is to leverage the power of symplectic geometry, especially a classic work of Gromov in 1980's (Gromov, 1985), to obtain a lower bound for the covariance of the conjugate coordinates. It is known as "Gromov's Non-squeezing Theorem",

Theorem D.1 (Gromov, 1985.)**.** *If $R < r$, there does not exist Hamiltonian map $\varphi : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ such that $\varphi(B(r)) \subset Z(R)$, where $B(r) = \{(x, y) \in \mathbb{R}^{2n} : \|x\|^2 + \|y\|^2 \leq r^2\}$ and $Z(R) = \{(x, y) \in \mathbb{R}^{2n} : x_i^2 + y_i^2 \leq R^2\}$ for any $i \in [n]$.*

Gromov's non-squeezing theorem asserts the following fact: Let $B(r)$ be a ball in the phase space $\mathbb{R}^n \times \mathbb{R}^n$, with center (a, b) and radius r: $B(r) : \|x - a\|^2 + \|y - b\|^2 \leq r^2$. The orthogonal projection of this ball on any plane of coordinates alway contains a disc of radius r. Now suppose that the ball is moved by a Hamiltonian flow $\varphi(t, \cdot)$, i.e., each point of $B(r)$ serves as an initial condition of a system of Hamiltonian equations. By Liouville's Theorem, the image $\varphi(t, B(r))$ at any moment t has the volume the same as the initial shape ball $B(r)$, but the shape is distorted. If we pair the conjugate coordinates x_i and y_i, then the projection of the deformed ball on any (x_i, y_i)-plane will never decrease below its original value πr^2. The FTRL algorithm induces a linear Hamiltonian system whose solution is a linear symplectic mapping on phase space $\mathbb{R}^{n_1} \times \mathbb{R}^{n_1}$. It suffices to consider a narrowed concept called "Linear symplectic width" which is defined below.

Definition D.2 (Linear symplectic width, (Hsiao & Scheeres, 2006))**.** The linear symplectic width of an arbitrary subset $A \subset \mathbb{R}^{2n}$, denoted as $w_L(A)$, is defined as:

$$w_L(A) = \sup_{r \in \mathbb{R}^+} \left\{ \pi r^2 : \phi(B^{2n}(r)) \subset A \text{ for some } \phi \in AS_p(\mathbb{R}^{2n}) \right\},$$

where $AS_p(\mathbb{R}^{2n})$ denotes the group of affine symplectomorphisms, i.e., linear map followed by translation.

D.3. Proof of Theorem 5.4

The first ingredient in proving Proposition 5.4 is based on standard results of Taylor series method in (Benaroya et al., 2005). Suppose $Y = g(X)$ where X is a random variable with μ_X the mean value. A full Taylor expansion of $Y = g(X)$ about the mean value yields

$$Y = g(X)|_{X=\mu_X} + (X - \mu_X)\frac{dg}{dx}|_{X=\mu_X} + \frac{1}{2!}(X - \mu_X)^2\frac{d^2g}{dx^2}|_{X=\mu_X} + ...$$

and the expectation is given by

$$\mu_Y = g(\mu_X) + \frac{\sigma_X^2}{2}g''(\mu_X) + ...$$

which holds due to $E[X-\mu_X]$=0. Furthermore, the variance of Y can be estimated as follows.

$$\sigma_Y^2 = E[Y^2] - \mu_X^2 \backsimeq g^2(\mu_X) + \sigma_X^2\left([g'(\mu_X)]^2 + g(\mu_X)g''(\mu_X)\right) - \left(g(\mu_X) + \frac{\sigma_X^2}{2}g''(\mu_X)\right)^2. \tag{192}$$

Therefore, if we assume that $\sigma_X^4 \ll \sigma_X^2$ which can be implied by $\sigma_X \ll 1$, the approximation to order σ_X^2 for the variance is

$$\sigma_Y^2 \backsimeq \sigma_X^2\left(g'(\mu_X)\right)^2.$$

This estimate enables one to focus only on the linearization of a general differentiable map on the multivariable case, with inevitably some extra assumption on higher order derivatives. In general, suppose $Y = g(X_1, ..., X_n)$, the Taylor series expansion about the mean value of each variable yield

$$\mathrm{Var}[Y] = \sum_{i=1}^n \left(\frac{\partial g(\mu_1, ..., \mu_n)}{\partial X_i}\right)^2 \sigma_i^2 + \sum_{i=1}^n \sum_{j=1, j\neq i}^n \left(\frac{\partial g(\mu_1, ..., \mu_n)}{\partial X_i}\right)\left(\frac{\partial g(\mu_1, ..., \mu_n)}{\partial X_j}\right)\rho_{ij}\sigma_i\sigma_j + ...$$

where $\rho_{ij} = \frac{\mathrm{Cov}(X_i, X_j)}{\sigma_i\sigma_j}$.

The other ingredient we need is the linear symplectic width evolving under a time-dependent linear Hamiltonian flow, which is the result of (Hsiao & Scheeres, 2006). In the setting of classic mechanics, the position and momentum $(\mathbf{q}, \mathbf{p})$ in a Hamiltonian system, the covariance matrix of $X = (q_1, ..., q_n, p_1, ..., p_n)$ is given by

$$P = E[XX^\top]$$

where we assume the system is zero-mean for convenience. If the system is linear

$$\dot{X} = A(t)X$$

with $X(t) = \Phi(t, t_0)X_0$ its solution, then the covariance is mapped as $P = \Phi P_0 \Phi^\top$. Furthermore, if we partition P into blocks such that

$$P = \begin{bmatrix} P_{11} & \dots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \dots & P_{nn} \end{bmatrix}$$

where

$$P_{ij} = \begin{bmatrix} E[q_i q_j^\top] & E[q_i p_j^\top] \\ \\ E[p_i q_j^\top] & E[p_i p_j^\top] \end{bmatrix}.$$

Then Theorem 3 of (Hsiao & Scheeres, 2006) asserts that

$$|P_{ii}(t)| \geq \left(\frac{w_L(P_0)}{\pi}\right)^2 \quad \text{for all } \ i = 1, ..., n.$$

The third ingredient is the Taylor expansion of differential mapping $f : V \to W$ between higher dimensional spaces (Conrad). In general suppose $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$. Let $U \subset V$ be open and let $f_i : U \to \mathbb{R}$ denote the ith component of f, so f is described as a map $f = (f_1, ..., f_m)$. Let $p \geq 0$ be a non-negative integer. Then f is C^p map if and only if all p-fold iterated partial derivatives of the f_i's exist and are continuous on U. Suppose $\mathrm{Hom}(V, W)$ be the space of linear mappings from V to W. Then the higher derivative $D^p f$ is a multi-linear mapping from $V^p \to W$, i.e.,

$$D^p f : U \to \text{Mult}(V^p, W),$$

where $V^p = V \times ... \times V$ is the p-th fold of Cartesian product of V. Choose $a \in U$ and $r > 0$ such that a small neighborhood $b_r(a) \subset U$ for a choice of norm on V. Choose $h = \sum h_j e_j \in V$ with $\|h\| < r$. For non-negative integer $k \leq p$, the higher order derivative as multi-linear mapping acting on $T_a U$ is given by the following expression,

$$\frac{(D^k f)(a)}{k!}(h^{(k)}) = \sum_{i_1+...+i_n=k} \frac{1}{i_1!...i_n!} h_1^{i_n}...h_n^{i_n} \frac{\partial^k f}{\partial x_1^{i_1}...\partial x_n^{i_n}}(a)$$

where $h^{(k)} = (h, ..., h) \in V^k$ and the sum is taken overa ll ordered n-tuples $(i_1, ..., i_n)$ of non-negative integer whose sum is k. To be more concrete

$$\frac{\partial^k f}{\partial x_1^{i_1} ... \partial x_n^{i_n}}(a) = \left(\frac{\partial^k f_1}{\partial x_1^{i_1} ... \partial x_n^{i_n}}(a), ..., \frac{\partial^k f_m}{\partial x_1^{i_1} ... \partial x_n^{i_n}}(a) \right)$$

which is a vector in W. The Taylor formula for differentiable mapping $f : V \to W$ is given as follows,

$$f(a+h) = \sum_{j=0}^{p} \frac{(D^j f)(a)}{j!}(h^{(j)}) + R_{p,a}(h)$$

in W, where

$$R_{p,a}(h) = \int_0^1 \frac{(1-t)^{p-1}}{(p-1)!}((D^p f)(a+th) - (D^p f)(a))(h^{(p)})dt$$

satisfies

$$\|R_{p,a}(h)\| \leq C_{p,h,a} \|h\|^p, \ \lim_{h \to 0} C_{p,h,a} = 0$$

with

$$C_{p,h,a} = \sup_{t \in [0,1]} \frac{\|(D^p f)(a+th) - (D^p f)(a)\|}{p!}.$$

With above settings, we are ready to prove the theorem.

Proof. Denote $X = (X_i^{t_0}, y_i^{t_0})$ for short, and let $\phi_t(X)$ be the flow of Hamiltonian system of X. The Taylor expansion with respect to mean of X, say μ, is computed as follows,

$$\phi_t(X) = \phi_t(\mu) + D\phi_t(\mu)(X - \mu) + \frac{1}{2} D^2 \phi_t(X)(X - \mu)^{(2)} + ...$$

Since the covariance matrix of the random vector given X as a random vector is encoded in the covariances of $\phi_t^i(X)$ and $\phi_t^j(X)$ for $i, j \in [n]$, i.e., $\mathrm{Cov}(\phi_t^i(X), \phi_t^j(X))$. The fundamental property of covariance implies that for each pair (i, j), $\mathrm{Cov}(\phi_t^i(X), \phi_t^j(X))$ is an infinite sum of covariances given by the Taylor formular. By assumption on the covariance of the initial input X such that the covariance is small, it suffices to use covariance matrix of $D\phi_t(\mu)(X - \mu)$ to approximate the covariance matrix of $\phi_t(X)$, since all the terms other than the linear ones in $\mathrm{Cov}(\phi_t^i(X), \phi_t^j(X))$ is of the higher power of the entries of $X - \mu$. Formally we have

$$\mathrm{Cov}(\phi_t(X)) \approx \mathrm{Cov}(D\phi_t(\mu)(X - \mu)).$$

Since we further assume that the Hamiltonian flow $\phi_t(\cdot)$ has Lipschitz derivatives of arbitrary order, uniformly, the remainder in the approximation is bounded by a constant multiplied by higher power of entries in the covariance matrix of X. Recall that we use notation $X = (X_i^{t_0}, y_i^{t_0})$, and apply Theorem 3 of (Hsiao & Scheeres, 2006) to the linear part $D\phi_t(\mu)(X - \mu)$ in the Taylor formula, we have

$$(\Delta X_{i,\alpha}^t \Delta y_{i,\alpha}^t)^2 - (\mathrm{Cov}(X_{i,\alpha}^t, y_{i,\alpha}^t))^2 \geq \frac{w_L^2(P_0)}{\pi^2}$$

provided the remainder in approximation is zero. Thus for any number strictly less than $\frac{w_L^2(P_0)}{\pi^2}$, say $\frac{1}{2}\frac{w_L^2(P_0)}{\pi^2}$, as long as the covariance entries are small enough, we can have

$$(\Delta X_{i,\alpha}^t \Delta y_{i,\alpha}^t)^2 - (\mathrm{Cov}(X_{i,\alpha}^t, y_{i,\alpha}^t))^2 \geq \frac{1}{2}\frac{w_L^2(P_0)}{\pi^2}.$$

The proof completes. □

D.4. Experiments on primal space

Note that primal space are closely related to canonical coordinates $(X_1(t), y_1(t))$ by

$$x_1(t) = \nabla h_1^*(y_1(t))$$
$$x_2(t) = \nabla h_2^*(A^{(21)}X_1(t) + y_2(0)).$$

It is hoped that Theorem 5.4 has its impact on primal space.

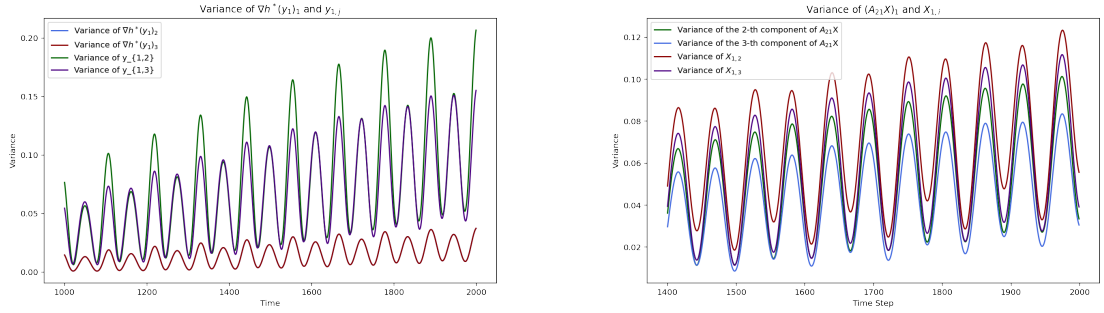
We firstly do experiments on how the operator $\nabla h^*(\cdot)$ and the matrix-vector $A^{(21)}X_1(t)$ can affect the wave shape of the canonical coordinates $(y_1(t), X_1(t))$. Our observation can be summerized as

Observation 1. For different $i_1, i_2 \in [n_1]$, the local minima of $\Delta(x_{1,i_1}(t))$ is close to the local minima of $\Delta(y_{1,i_2}(t))$. See (a) in Figure 9.

Observation 2. For different $i_1 \in [n_1]$ and $i_2 \in [n_2]$, if $\exists\ \epsilon > 0$ such that $x_{1,i_1}(t) > \epsilon$, then the local minima of $\Delta\left(X_{1,i_1}(t)\right)$, is close to local minima of $\Delta\left((A^{(21)}X_1(t) + y_2(0))_{i_2}\right)$. See (b) in Figure 9.

Observation 1 can be explained as viewing the function graph of $\Delta(x_{1,i_1}(t))$ and $\Delta(y_{1,i_2}(t))$ as waves, then the peaks(troughs) of wave of $\Delta(x_{1,i_1}(t))$ are close to the peaks(troughs) of wave of $\Delta(y_{1,i_2}(t))$. Observation 2 has the same meaning. Note that the condition $\exists\epsilon > 0$ such that $x_{1,i_1}(t), x_{1,i_2}(t) > \epsilon$ is necessary because otherwise $\lim_{t\to\infty} x_{1,i_1}(t) = \lim_{t\to\infty} x_{1,i_2}(t) = 0$, thus the player will not use these strategies as time process, and the variance of these two strategies will equal to 0, thus the uncertainty inequality will not hold for these strategies.

Although the above two observations are less rigid, however, combine with Theorem 5.4, they provides insights into the covariance evolution of primal space. Note that the second term implies that the positions of the local minima/maxima of $\Delta y_2(t) = \Delta(A^{(21)}X_1(t) + y_2(0))$ are close to those of $\Delta(X_1(t))$. Due to the symmetry between the two players, combining this with the first term allow us to conclude that the position of the local minima/maxima of $\Delta x_2(t)$ are close to those of $\Delta X_1(t)$. Furthermore, based on the first term, we can also infer that the position of the local minima/maxima of $\Delta x_1(t)$ are close those of $\Delta y_1(t)$. Therefore, by combining these two observations with Proposition 5.4, we can conclude that the positions of local minima in $\Delta x_1(t)$ are close to those of local maxima in $\Delta x_2(t)$, and vice versa.



(a) Functions graph of $\Delta(x_{1,i_1}(t))$ and $\Delta(y_{1,i_2}(t))$. (b) Functions graph of $\Delta(A^{(21)}X_1(t)+y_2(0))$ and $\Delta(X_1(t))$.

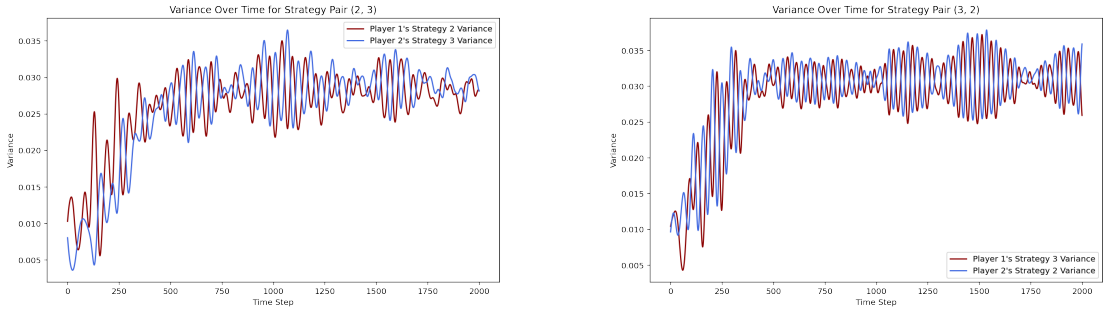
Figure 9: Variance evolution of classic Euler discretization

In Figure 9, experiments are presented in the case of MWU algorithms and a randomly generated $3 * 3$ game. The regularizer is chosen as $h(x) = \sum_{i=1}^{n_i} x_i \ln x_i$ and the constraint is chosen as simplex constrain $\mathcal{X} = \{x_i \mid \sum_{i=1}^{n} x_i = 1, x_i > 0\}$. Note that in this case, we have

$$\nabla h^*(y) = \left(\frac{e^{y_i}}{\sum_{s=1}^{n} e^{y_s}}\right)_{i=1}^{n}.$$

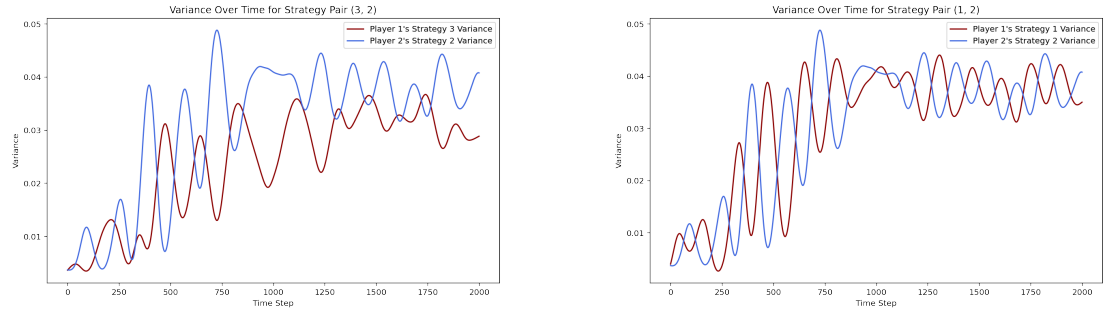
In Figure 9 (a), we can see although variance of $\nabla h^*(y_1)_1$ has smaller value than variance of $y_{1,i}$, but the positions of local minima/maxima of these curves are very close. Similarly, in Figure 9 (b), but the positions of local minima/maxima of these curves are also very close.

In the following, we provide additional experiments on the evolution of variance in the primal space. We observe that if the dimension of the game is low, for example, less than 10, we can see a similar covariance evolution in the primal space as shown in Figure 3. However, when the dimension becomes very high, this pattern can disappear.



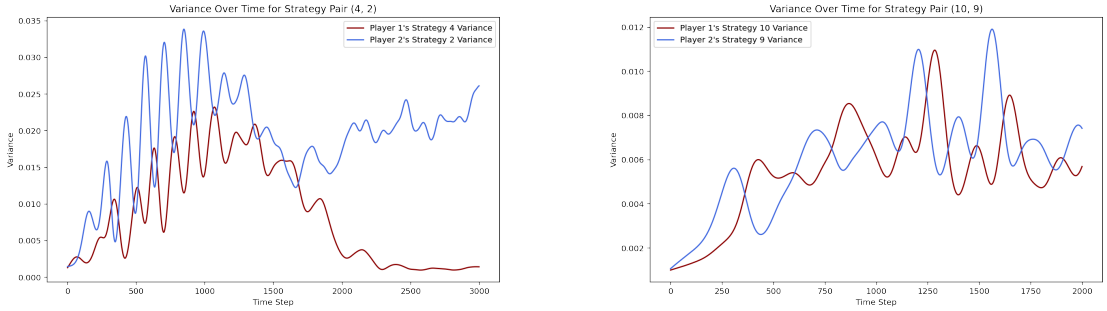
(a) (b)

Figure 10: Covariance evolution on a randomly generated 3*3 game



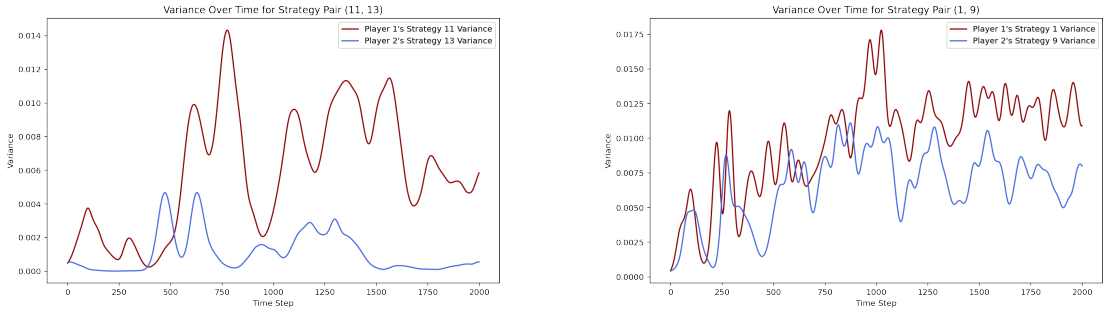
(a) (b)

Figure 11: Covariance evolution on a randomly generated 5*5 game



(a) (b)

Figure 12: Covariance evolution on a randomly generated 10*10 game



(a) (b)

Figure 13: Covariance evolution on a randomly generated 15*15 game

E. Experiments for non-singular cases.

Continuous time FTRL. We illustrate how $\mathrm{Var}(X_{1,1}(t))$ and $\mathrm{Var}(y_{1,1}(t))$ evolve with continuous time FTRL with payoff matrices

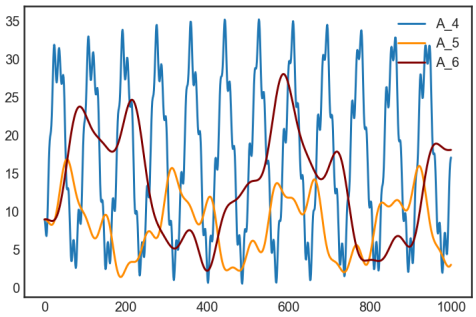
$$A_4 = [[1, -2], [-1, 1]], A_5 = [[2, -3], [-1, 5]],$$
$$A_6 = [[2, -1.5], [-2, 3]].$$

See (a)(b) Figure 14. In (a) the $\mathrm{Var}(X_{1,1}(t))$ is bounded, and in (b) $\mathrm{Var}(y_{1,1}(t))$ is bounded, which support results of continuous time part in Theorem 5.1 for the non-singular cases.

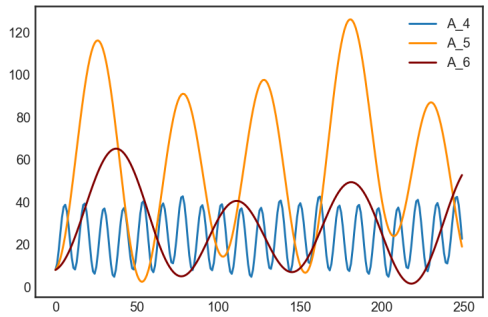
Symplectic discretization. We illustrate how $\mathrm{Var}(X^t_{1,1})$ and $\mathrm{Var}(y^t_{1,1})$ evolves with symplectic discretization, the payoff matrices are given as follows:

$$B_4 = [[1, -1.1], [-1, 1]], B_5 = [[1, -1.2], [-1, 1]],$$
$$B_6 = [[1, -1.3], [-1, 1]].$$

See Figure 15. From the experimental results, we can see the variance behavior of symplectic discretization is same as continuous case, which support results of symplectic discretization part of Theorem 5.1 for the non-singular cases.

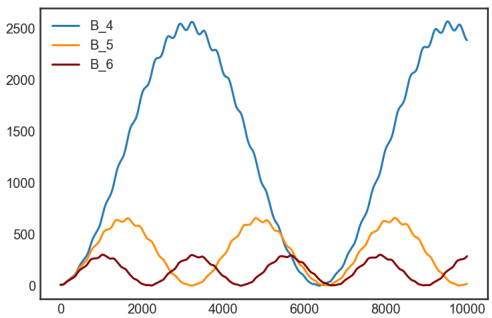


(a) $\mathrm{Var}(X_{1,1}(t))$, non-singular

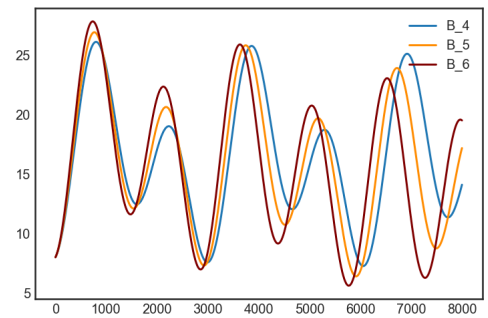


(b) $\mathrm{Var}(y_{1,1}(t))$, non-singular

Figure 14: Variance evolution of continuous FTRL



(a) $\mathrm{Var}(X^t_{1,1})$, non-singular



(b) $\mathrm{Var}(y^t_{1,1})$, non-singular

Figure 15: Variance evolution of Symplectic discretization