

A PROOF OF PROPOSITION 3.1

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

$$H(X_1, y_1) = h_1^*(y_1(t)) + h_2^*(y_2(0) + A^{(21)}X_1(t)), \quad (8)$$

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

$$H(X_2, y_2) = h_2^*(y_2(t)) + h_1^*(y_1(0) + A^{(12)}X_2(t)), \quad (9)$$

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.

More precisely, it 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:

$$\frac{d}{dt}X_1(t) = \frac{\partial H}{\partial y_1}(X_1, y_1) = \nabla h_1^*(y_1(t)), \quad (10)$$

$$\frac{d}{dt}y_1(t) = -\frac{\partial H}{\partial X_1}(X_1, y_1) = A^{(12)}\nabla h_2^*(y_2(0) + A^{(21)}X_1(t)). \quad (11)$$

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

$$\frac{d}{dt}X_2(t) = \frac{\partial H}{\partial y_2}(X_2, y_2) = \nabla h_2^*(y_2(t)), \quad (12)$$

$$\frac{d}{dt}y_2(t) = -\frac{\partial H}{\partial X_2}(X_2, y_2) = A^{(21)}\nabla h_1^*(y_1(0) + A^{(12)}X_2(t)). \quad (13)$$

The proof of Proposition 3.1 is divided into two parts :

- The proof of entropy regularizers is presented in Section A.2.
- The proof of Euclidean norm regularizers is presented in Section A.3.

and in Section A.1, we introduce the Euler and Symplectic discretization of FTRL.

A.1 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 (5) with Euler method for both agent $i = 1, 2$ gives*

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

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

and

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

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

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 (10) and (11). Similarly, (agent 2 Euler discretize equation) directly follows from applying Euler discretization to (12) and (13). \square

Lemma A.2 (Symplectic discretization of FTRL). *Discretizing (5) for $i = 1$ with I-type Euler symplectic method Type I method gives*

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

and discrete (5) for $i = 2$ with II-type Euler symplectic method Type II method gives

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

Proof. (agent 1 Symplectic discretize equation) directly follows from applying (Type I method) to equation 10 and 11. Similarly, (agent 2 Symplectic discretize equation) directly follows from applying (Type II method) to equation 12 and 13. \square

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

$$x_1^t = \frac{X_1^{t+1} - X_1^t}{\eta}, \quad x_2^t = \frac{X_2^{t+1} - X_2^t}{\eta}. \quad (14)$$

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

$$x_1^t = \nabla h_1^*(y_1^t), \quad x_2^t = \nabla h_2^*(y_2^t), \quad (15)$$

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

$$x_1^t = \nabla h_1^*(y_1^{t+1}), \quad x_2^t = \nabla h_2^*(y_2^t). \quad (16)$$

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*

$$y_1^n = y_1^0 + A^{(12)} X_2^n \quad (17)$$

$$y_2^n = y_2^0 + A^{(21)} X_1^n \quad (18)$$

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 $n = 0$, (17) and (18) are

$$y_1^0 = y_1^0 + A^{(12)} X_2^0 \quad (19)$$

$$y_2^0 = y_2^0 + A^{(21)} X_1^0, \quad (20)$$

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

Now assume (17) and (18) hold for n , i.e.,

$$y_1^n = y_1^0 + A^{(12)} X_2^n \quad (21)$$

$$y_2^n = y_2^0 + A^{(21)} X_1^n. \quad (22)$$

Then, we have

$$y_1^{n+1} = y_1^n + \eta A^{(12)} \nabla h_2^*(y_2^0 + A^{(21)} X_1^n) \quad (23)$$

$$\stackrel{(22)}{=} y_1^n + \eta A^{(12)} \nabla h_2^*(y_2^n) \quad (24)$$

$$= y_1^n + \eta A^{(12)} x_2^n \quad (25)$$

$$\stackrel{(21)}{=} y_1^0 + A^{(12)} X_2^n + \eta A^{(12)} x_2^n \quad (26)$$

$$= y_1^0 + A^{(12)} X_2^{n+1}. \quad (27)$$

Moreover, we have

$$y_2^{n+1} = y_2^n + \eta A^{(21)} \nabla h_1^*(y_1^0 + A^{(12)} X_2^{n+1}) \quad (28)$$

$$\stackrel{(27)}{=} y_2^n + \eta A^{(21)} \nabla h_1^*(y_1^{n+1}) \quad (29)$$

$$= y_2^n + \eta A^{(21)} x_1^n \quad (30)$$

$$\stackrel{(22)}{=} y_2^0 + A^{(21)} X_1^n + \eta A^{(21)} x_1^n \quad (31)$$

$$= y_2^0 + A^{(21)} X_1^{n+1} \quad (32)$$

This finish the proof. \square

A.2 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., $\Delta = \{x \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = 1, x_i \geq 0\}$, we have

$$\nabla h^*(y) = \left(\frac{e^{y_i}}{\sum_{s=1}^n e^{y_s}} \right)_{i=1}^n. \quad (33)$$

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

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

Since the gradient of f can be computed to be

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

the KKT condition becomes

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

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

$$\log x_i + 1 + \lambda = y_i \text{ for all } i \in [n], \sum_{i=1}^n x_i = 1.$$

This gives solution of x_i and λ :

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

thus we have completed the proof. \square

Lemma A.5. The (x_1^t, x_2^t) in (15) with entropy regularizer is the same as (MWU).

Proof. In (15), we have $x_1^t = \nabla h_1^*(y_1^t)$, thus

$$x_1^t = \nabla h_1^*(y_1^t) \quad (34)$$

$$(35)$$

$$\stackrel{(33)}{=} \left(\frac{e^{y_{1,s}^t}}{\sum_{j=1}^{n_1} e^{y_{1,j}^t}} \right)_{s=1}^{n_1} \quad (36)$$

$$(37)$$

$$\stackrel{(17)}{=} \left(\frac{e^{y_{1,s}^0 + (A^{(12)} X_2^t)_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^0 + (A^{(12)} X_2^t)_j}} \right)_{s=1}^{n_1} \quad (38)$$

$$(39)$$

$$= \left(\frac{e^{y_{1,s}^0 + \eta(A^{(12)} \sum_{k=1}^{t-1} x_2^k)_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^0 + \eta(A^{(12)} \sum_{k=1}^{t-1} x_2^k)_j}} \right)_{s=1}^{n_1} \quad (40)$$

$$(41)$$

$$= \left(\frac{x_{1,s}^{t-1} e^{\eta(A^{(12)} x_2^{t-1})_s}}{\sum_{j=1}^{n_1} x_{1,j}^{t-1} e^{\eta(A^{(12)} x_2^{t-1})_j}} \right)_{s=1}^{n_1}. \quad (42)$$

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

$$x_2^t = \left(\frac{x_{2,s}^{t-1} e^{\eta(A^{(21)} x_1^{t-1})_s}}{\sum_{j=1}^{n_2} x_{2,j}^{t-1} e^{\eta(A^{(21)} x_1^{t-1})_j}} \right)_{s=1}^{n_2}. \quad (43)$$

That is same as (MWU). \square

Lemma A.6. The (x_1^t, x_2^t) in (16) with entropy regularizer is the same as (AltMWU).

Proof. For 1 agent, from (16), we have $x_1^t = \nabla h_1^*(y_1^{t+1})$, thus

$$x_1^t = \nabla h_1^*(y_1^{t+1}) \quad (44)$$

$$(45)$$

$$\stackrel{(33)}{=} \left(\frac{e^{y_{1,s}^{t+1}}}{\sum_{j=1}^{n_1} e^{y_{1,j}^{t+1}}} \right)_{s=1}^{n_1} \quad (46)$$

$$(47)$$

$$\stackrel{(17)}{=} \left(\frac{e^{y_{1,s}^0 + (A^{(12)} X_2^{t+1})_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^0 + (A^{(12)} X_2^{t+1})_j}} \right)_{s=1}^{n_1} \quad (48)$$

$$(49)$$

$$= \left(\frac{e^{y_{1,s}^0 + \eta(A^{(12)} \sum_{k=1}^t x_2^k)_s}}{\sum_{j=1}^{n_1} e^{y_{1,j}^0 + \eta(A^{(12)} \sum_{k=1}^t x_2^k)_j}} \right)_{s=1}^{n_1} \quad (50)$$

$$(51)$$

$$= \left(\frac{x_{1,s}^{t-1} e^{\eta(A^{(12)} x_2^t)_s}}{\sum_{j=1}^{n_1} x_{1,j}^{t-1} e^{\eta(A^{(12)} x_2^t)_j}} \right)_{s=1}^{n_1}. \quad (52)$$

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

For 2 agent, from (16) we have $x_2^t = \nabla h_2^*(y_2^t)$, thus

$$x_2^t = \nabla h_2^*(y_2^t) \quad (53)$$

$$(54)$$

$$= \left(\frac{e^{y_{2,s}^t}}{\sum_{j=1}^{n_2} e^{y_{2,j}^t}} \right)_{s=1}^{n_2} \quad (55)$$

$$(56)$$

$$\stackrel{(18)}{=} \left(\frac{e^{y_{2,s}^0 + (A^{(21)} X_1^t)_s}}{\sum_{j=1}^{n_2} e^{y_{2,j}^0 + (A^{(21)} X_1^t)_j}} \right)_{s=1}^{n_2} \quad (57)$$

$$(58)$$

$$= \left(\frac{e^{y_{2,s}^0 + \eta(A^{(21)} \sum_{k=1}^{t-1} x_1^k)_s}}{\sum_{j=1}^{n_2} e^{y_{2,j}^0 + \eta(A^{(21)} \sum_{k=1}^{t-1} x_1^k)_j}} \right)_{s=1}^{n_2} \quad (59)$$

$$(60)$$

$$= \left(\frac{x_{2,s}^{t-1} e^{\eta(A^{(21)} x_1^t)_s}}{\sum_{j=1}^{n_2} x_{2,j}^{t-1} e^{\eta(A^{(21)} x_1^t)_j}} \right)_{s=1}^{n_2}. \quad (61)$$

Combine (61) and (52), we can see the update rule of (x_1^t, x_2^t) is same as (AltMWU). \square

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 Euclidean norm. However, this change will not affect the proof, so we omit it here.

A.3 PROOF OF EUCLIDEAN NORM REGULARIZERS

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

$$\nabla h_i^*(y) = \arg \max_{x \in \mathbb{R}^n} \{ \langle x, y \rangle - \|x\|^2 \} = y. \quad (62)$$

Lemma A.7. The (x_1^t, x_2^t) in (15) with Euclidean norm regularizer is the same as (GDA).

Proof. For agent 1, we have

$$x_1^t \stackrel{(15)}{=} \nabla h_1^*(y_1^t) \quad (63)$$

$$\stackrel{(62)}{=} y_1^t \quad (64)$$

$$= y_1^0 + A^{(12)} X_2^t \quad (65)$$

$$= y_1^0 + \eta \cdot A^{(12)} \sum_{i=0}^{t-1} x_2^i \quad (66)$$

$$= x_1^{t-1} + \eta \cdot A^{(12)} x_2^{t-1}. \quad (67)$$

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). \square

Lemma A.8. The (x_1^t, x_2^t) in (16) with Euclidean norm regularizer is the same as (AltGDA).

Proof. For agent 1, we have

$$x_1^t \stackrel{(16)}{=} \nabla h_1^*(y_1^{t+1}) \quad (68)$$

$$\stackrel{(62)}{=} y_1^{t+1} \quad (69)$$

$$= y_1^0 + A^{(12)} X_2^{t+1} \quad (70)$$

$$= y_1^0 + \eta \cdot A^{(12)} \sum_{i=0}^t x_2^i \quad (71)$$

$$= x_1^{t-1} + \eta \cdot A^{(12)} x_2^t. \quad (72)$$

For agent 2, we have

$$x_2^t \stackrel{(16)}{=} \nabla h_1^*(y_2^t) \quad (73)$$

$$\stackrel{(62)}{=} y_2^t \quad (74)$$

$$= y_2^0 + A^{(21)} X_2^t \quad (75)$$

$$= y_2^0 + \eta \cdot A^{(21)} \sum_{i=0}^{t-1} x_2^i \quad (76)$$

$$= x_2^{t-1} + \eta \cdot A^{(21)} x_2^{t-1}. \quad (77)$$

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

B PROOF OF SECTION 4

In this appendix we prove results in Section 4. Proposition 4.1 is proved in Section B.3, and Proposition 4.2 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 4.1 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 4.2 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a diffeomorphism, thus $f(X)$ is a random vector. Then we have*

$$S(f(X)) = S(X) + \int_{\mathcal{X}} g(x) \log(|\det J_f(x)|) dx \quad (78)$$

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

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

Then we have

$$S(Y) = S(f(X)) = - \int_{\mathcal{Y}} \hat{g}(y) \cdot \log(\hat{g}(y)) dy \quad (79)$$

$$= - \int_{\mathcal{Y}} g(f^{-1}(y)) |\det J_{f^{-1}}(y)| \cdot \log(g(f^{-1}(y)) |\det J_{f^{-1}}(y)|) dy \quad (80)$$

$$= - \int_{\mathcal{X}} g(x) |\det J_{f^{-1}}(f(x))| \cdot \log(g(x) |\det J_{f^{-1}}(f(x))|) \cdot |\det J_f(x)| dx \quad (81)$$

$$= - \int_{\mathcal{X}} g(x) \cdot \log(g(x) |\det J_{f^{-1}}(f(x))|) dx \quad (82)$$

$$= - \int_{\mathcal{X}} g(x) \log(g(x)) dx - \int_{\mathcal{X}} g(x) \log(|\det J_{f^{-1}}(f(x))|) dx \quad (83)$$

$$= S(X) + \int_{\mathcal{X}} g(x) \log(|\det J_f(x)|) dx, \quad (84)$$

where (82) comes from the inverse function theorem, which states

$$J_{f^{-1}}(f(x)) = (J_f(x))^{-1}. \quad (85)$$

□

B.2 TECHNICAL LEMMAS FOR PROPOSITION 4.1

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

$$P \cdot A \cdot P^{-1} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \quad (86)$$

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

$$P \cdot (A)^2 \cdot P^{-1} = (P \cdot A \cdot P^{-1}) \cdot (P \cdot A \cdot P^{-1}) \quad (87)$$

$$= \begin{bmatrix} (\lambda_1)^2 & & \\ & \ddots & \\ & & (\lambda_n)^2 \end{bmatrix}, \quad (88)$$

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 $h : \mathcal{X} \rightarrow \mathbb{R}$ be a μ -strongly convex function with L -Lipschitz continuous gradient, let*

$$h^*(y) = \max_{x \in \mathcal{X}} \{\langle x, y \rangle - h(x)\} \quad (89)$$

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 $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a differentiable function on a convex set $U \subset \mathbb{R}^n$, and*

$$\|J_f(x) - I\| < 1 \quad (90)$$

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

$$\|f(x) - f(y) + y - x\| = \|g(x) - g(y)\| \quad (91)$$

$$= \|J_g(\zeta)(x - y)\| \quad (92)$$

$$\leq \|J_g(\zeta)\| \cdot \|x - y\| \quad (93)$$

$$< \|x - y\|, \quad (94)$$

where [\(92\)](#) use the mean value theorem, and [\(94\)](#) is due to the fact that $\|J_g(\zeta)\| < 1$ for any $\zeta \in U$. Thus $f(x) \neq f(y)$. \square

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

$$\phi : (X_1^n, y_1^n) \rightarrow (X_1^{n+1}, y_1^{n+1}) \quad (95)$$

of Euler discretization of FTRL in Lemma [A.1](#) is an injective function.

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

$$y_1^{n+1} = y_1^n - \eta \cdot \frac{\partial H}{\partial X_1}(X_1^n, y_1^n) \quad (96)$$

$$X_1^{n+1} = X_1^n + \eta \cdot \frac{\partial H}{\partial y_1}(X_1^n, y_1^n) \quad (97)$$

and the Hamiltonian function has form

$$H(X_1, y_1) = h_1^*(y_1) + h_2^*(y_2(0) + A^{(21)}X_1). \quad (98)$$

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

$$\frac{\partial^2 H}{\partial X_1 \partial y_1} = 0, \quad \frac{\partial^2 H}{\partial y_1 \partial X_1} = 0. \quad (99)$$

Next we calculate the Jacobin matrix of ϕ ,

$$J_\phi = \begin{bmatrix} \frac{\partial y_1^{n+1}}{\partial y_1^n} & \frac{\partial y_1^{n+1}}{\partial X_1^n} \\ \frac{\partial X_1^{n+1}}{\partial y_1^n} & \frac{\partial X_1^{n+1}}{\partial X_1^n} \end{bmatrix} = \begin{bmatrix} I - \eta \frac{\partial^2 H}{\partial y_1 \partial X_1}(X_1^n, y_1^n) & -\eta \frac{\partial^2 H}{\partial^2 X_1}(X_1^n, y_1^n) \\ \eta \frac{\partial^2 H}{\partial^2 y_1}(X_1^n, y_1^n) & I + \eta \frac{\partial^2 H}{\partial X_1 \partial y_1}(X_1^n, y_1^n) \end{bmatrix} \quad (100)$$

$$= \begin{bmatrix} I & -\eta(A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)} \\ \eta \nabla^2 h_1^* & I \end{bmatrix}, \quad (101)$$

and

$$J_\phi - I = \begin{bmatrix} 0 & -\eta(A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)} \\ \eta \nabla^2 h_1^* & 0 \end{bmatrix}. \quad (102)$$

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

$$\|\nabla^2 h_i^*\| \leq \frac{1}{\mu_i} \quad (103)$$

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

$$\|J_\phi - I\| = \sqrt{\lambda_{\max}((J_\phi - I)^\top \cdot (J_\phi - I))}, \quad (104)$$

and

$$(J_\phi - I)^\top \cdot (J_\phi - I) = \eta^2 \cdot \begin{bmatrix} (\nabla^2 h_1^*)^2 & 0 \\ 0 & ((A^{(21)})^\top \cdot \nabla^2 h_2^* \cdot A^{(21)})^2 \end{bmatrix}. \quad (105)$$

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 (103), the L^2 -operator norm of $\nabla^2 h_i^*$ has an upper bound $\frac{1}{\mu_i}$, thus we have

$$0 \leq \lambda < \max \left\{ \frac{1}{\mu_1}, \frac{\|A^{(21)}\|^2}{\mu_2} \right\}. \quad (106)$$

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

$$\|J_\phi - I\| < 1, \quad (107)$$

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

B.3 PROOF OF PROPOSITION 4.1

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 (8), the Hamiltonian function of FTRL is

$$H(X_1^t, y_1^t) = h_1^*(y_1^t) + h_2^*(y_2^0 + A^{(21)} X_1^t), \quad (108)$$

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

$$y_1^{t+1} = y_1^t - \eta \cdot \frac{\partial H}{\partial X_1}(X_1^t, y_1^t) \quad (109)$$

$$X_1^{t+1} = X_1^t + \eta \cdot \frac{\partial H}{\partial y_1}(X_1^t, y_1^t). \quad (110)$$

Recall from (101), the jacobian map of ϕ is

$$J_\phi(X, y) = \begin{bmatrix} I & -\eta(A^{(21)})^\top \cdot \nabla^2 h_2^*(X, y) \cdot A^{(21)} \\ \eta \nabla^2 h_1^*(X, y) & I \end{bmatrix}, \quad (111)$$

thus

$$\det(J_\phi(X, y)) = \det\left(I + \eta^2 \cdot \nabla^2 h_1^*(X, y) \cdot (A^{(21)})^\top \cdot \nabla^2 h_2^*(X, y) \cdot A^{(21)}\right). \quad (112)$$

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

$$\det(J_\phi(X, y)) \geq 1. \quad (113)$$

Combine this with (78), we have

$$S(X_1^{t+1}, y_1^{t+1}) = S(\phi(X_1^t, y_1^t)) \quad (114)$$

$$= S(X_1^t, y_1^t) + \int_{\mathcal{X}} g^t(X_1^t, y_1^t) \log(|\det J_\phi(X_1^t, y_1^t)|) dX_1^t dy_1^t \quad (115)$$

$$\geq S(X_1^t, y_1^t) + \int_{\mathcal{X}} g^t(X_1^t, y_1^t) \log(1) dX_1^t dy_1^t \quad (116)$$

$$= S(X_1^t, y_1^t), \quad (117)$$

where $g^t(\cdot)$ is the probability density function of (X_1^t, y_1^t) , and (116) comes from $\det(J_\phi(X, y)) \geq 1$. 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 (115), to prove a linear growth rate of differential entropy, it is sufficient to prove a uniform lower bound on $\det(J_\phi(X, y))$. 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.,

$$\nabla^2 h_i^*(X, y) \succcurlyeq \mu_i I \quad (118)$$

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

$$\eta^2 \cdot \nabla^2 h_1^*(X, y) \cdot (A^{(21)})^\top \cdot \nabla^2 h_2^*(X, y) \cdot A^{(21)} \quad (119)$$

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 $A^{(21)}$. \square

B.4 PROOF OF PROPOSITION 4.2

Lemma B.7. *The update map*

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

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

Proof. Recall the update rule can be written as

$$y_1^{t+1} = y_1^t - \eta \frac{\partial H}{\partial X_1}(X_1^t, y_1^t) = y_1^t + \eta A^{(12)} \nabla h_2^*(y_2^0 + A^{(21)} X_1^t), \quad (120)$$

$$X_1^{t+1} = X_1^t + \eta \frac{\partial H}{\partial y_1}(X_1^t, y_1^{t+1}) = X_1^t + \eta \nabla h_1^*(y_1^{t+1}). \quad (121)$$

$$(122)$$

Thus given (X_1^{t+1}, y_1^{t+1}) , we can directly find a unique X_1^t from (121), 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 (120), 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. \square

Now we are ready to prove Proposition 4.2

Proof. Since the iterate map $\psi : (X_1^n, y_1^n) \rightarrow (X_1^{n+1}, y_1^{n+1})$ 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

$$S(X_1^{n+1}, y_1^{n+1}) - S(X_1^n, y_1^n) = \int_{\mathcal{X}} g^n(X_1^n, y_1^n) \log(|\det J_\psi(X_1^n, y_1^n)|) dx \quad (123)$$

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

$$|\det J_\psi(X_1^n, y_1^n)| = 1. \quad (124)$$

Thus the right hand side of (123) equals to 0, and this implies $S(X_1^{n+1}, y_1^{n+1}) = S(X_1^n, y_1^n)$. \square

B.5 NUMERICAL EXAMPLES OF PROPOSITION PROPOSITIONS 4.1 AND 4.2

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

$$S(AX) - S(X) = \log(|\det(A)|). \quad (125)$$

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 4.1 and 4.2.

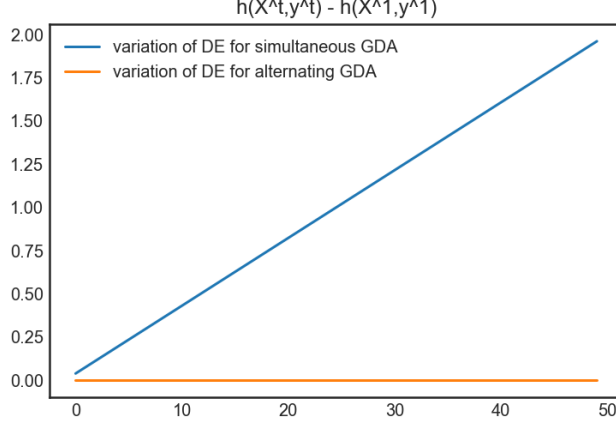


Figure 5: Variation of differential entropy for simultaneous and alternating GDA. The growth rate of the variation in DE is linear for simultaneous GDA, while it is 0 for alternating GDA.

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 $\text{Spec}(A)$ 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 $\{v_m, v_{m-1}, \dots, v_1\}$ given by

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

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

$$c_m v_m + c_{m-1} v_{m-1} + \dots + c_1 v_1 = 0 \quad (126)$$

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

Multiply equation [126](#) 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 [126](#) 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 $c_m = c_{m-1} = \dots = c_1 = 0$. \square

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

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

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

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

Thus we have

$$[v_1, v_2, \dots, v_m]^{-1} A [v_1, v_2, \dots, v_m] = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix} \quad (127)$$

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 [127](#) 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.,

$$J^C = M^{-1}AM$$

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

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

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

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

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 $\mu = \text{Re}(\mu) + i\text{Im}(\mu)$, $\text{Im}(\mu) \neq 0$ 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

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

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 [I27](#)

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

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

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

The matrix

$$\begin{bmatrix} D & I & & & \\ & \cdot & \cdot & & \\ & & \cdot & \cdot & \\ & & & \cdot & \cdot \\ & & & & I \\ & & & & & D \end{bmatrix} \quad (128)$$

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 $\operatorname{Re}(v), \operatorname{Im}(v)$ appear in the real generalized modal matrix, $(\operatorname{Re}(v), \operatorname{Im}(v))$ 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 128 appear firstly in the real Jordan normal form, and the necessary of (2) can be seen from the derivation of equation 128.

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

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

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

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

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

$$J_{\mu} = \begin{bmatrix} \mu & 1 & & & \\ & \mu & 1 & & \\ & & \ddots & \ddots & \\ & & & \mu & 1 \\ & & & & \mu \end{bmatrix} \in \mathbb{R}^{m \times m} \quad (129)$$

and

$$e^{J_{\mu}t} = e^{\mu t} \begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdot & \cdot & \frac{t^{m-1}}{(m-1)!} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \in \mathbb{R}^{m \times m}. \quad (130)$$

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

$$J_{\mu, \bar{\mu}} = \begin{bmatrix} D & I_2 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & I_2 \\ & & & & D \end{bmatrix} \in \mathbb{R}^{2m \times 2m}, \quad (131)$$

and

$$e^{J_{\mu, \bar{\mu}} t} = e^{\lambda t} \begin{bmatrix} R & tR & \frac{t^2}{2!}R & \cdot & \cdot & \frac{t^{m-1}}{(m-1)!}R \\ & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \frac{t^2}{2!}R \\ & & & & \cdot & tR \\ & & & & & R \end{bmatrix} \in \mathbb{R}^{2m \times 2m}. \quad (132)$$

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

$$x_{n+1} = Ax_n \quad (133)$$

By induction, equation (134) has solution formula

$$x_n = A^n x_0, \quad x_0 \in \mathbb{R}^d. \quad (134)$$

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 (134), 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*

$$J_\mu = \begin{bmatrix} \mu & & & \\ & \mu & & \\ & & \ddots & \\ & & & \mu \end{bmatrix} + \begin{bmatrix} 0 & 1 & & \\ & 0 & 1 & \\ & & \ddots & \ddots \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \quad (135)$$

$$= \mu I + N \quad (136)$$

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

$$J_\mu^n = (\mu I + N)^n = \sum_{i=0}^{m-1} \binom{n}{i} \mu^{n-i} N^i. \quad (137)$$

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 [Colonijs & 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*

$$J_{\mu, \bar{\mu}} = \begin{bmatrix} D & 0 & & \\ & \ddots & \ddots & \\ & & \ddots & \\ & & & 0 \\ & & & & D \end{bmatrix} + \begin{bmatrix} 0 & I_2 & & \\ & \ddots & \ddots & \\ & & \ddots & \\ & & & I_2 \\ & & & & 0 \end{bmatrix} \quad (138)$$

$$= \tilde{D} + N \quad (139)$$

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

$$D = \begin{bmatrix} \lambda & -\nu \\ \nu & \lambda \end{bmatrix} = |\mu|R, \quad R = \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix}. \quad (140)$$

Thus

$$J_{\mu, \bar{\mu}}^n = (\tilde{D} + N)^n = \sum_{i=0}^{n-1} \binom{n}{i} \tilde{D}^{n-i} N^i = \sum_{i=0}^{n-1} \binom{n}{i} |\mu|^{n-i} \tilde{R}^{n-i} N^i. \quad (141)$$

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

$$\frac{dy_1}{dt} = -\frac{\partial H(X_1, y_1)}{\partial X_1} = -AA^\top X_1(t) + Ay_2(0) \quad (142)$$

$$\frac{dX_1}{dt} = \frac{\partial H(X_1, y_1)}{\partial y_1} = y_1(t). \quad (143)$$

Similarly for agent 2 we have

$$\frac{dy_2}{dt} = -\frac{\partial H(X_2, y_2)}{\partial X_2} = -A^\top AX_2(t) - A^\top y_1(0) \quad (144)$$

$$\frac{dX_2}{dt} = \frac{\partial H(X_2, y_2)}{\partial y_2} = y_2(t). \quad (145)$$

In the following, we will focus on the viewpoint of agent 1, thus we will consider equation [\(142\)](#) and [\(143\)](#).

Lemma C.17. *The solution of the linear differential system consisted by equation [\(142\)](#) and [\(143\)](#) and initial condition $(y_1(t_0), X_1(t_0))$ can be written as*

$$\begin{bmatrix} y_1(t+t_0) \\ X_1(t+t_0) \end{bmatrix} = e^{\mathcal{L}t} \cdot \begin{bmatrix} y_1(t_0) \\ X_1(t_0) \end{bmatrix} + e^{\mathcal{L}t} \cdot \int_{t_0}^{t+t_0} e^{-\mathcal{L}s} Ay_2(0) ds \quad (146)$$

Proof. It is directly to verify the derivate of right hind side satisfies equation (142) and (143), 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 (142) and (143). \square

Form (146) 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 $\int_{t_0}^{t+t_0} e^{-\mathcal{L}s} A y_2(0) ds$ 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

$$\begin{bmatrix} \frac{dy_1}{dt} \\ \frac{dX_1}{dt} \end{bmatrix} = \mathcal{L} \begin{bmatrix} y_1(t) \\ X_1(t) \end{bmatrix} \quad (147)$$

where

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

Since the solution of (147) 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 $\text{Var}(X_1(t))$, $\text{Var}(y_1(t))$, and $\text{Cov}(X_1(t), y_1(t))$, 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 $\text{Var}(X_1(t))$ and $\text{Var}(y_1(t))$, $\text{Var}(X_1(t))$ may have quadratic growth and $\text{Var}(y_1(t))$ 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 $\text{Var}(y_1(t))$ 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 (147), 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

$$f(x) = \det(xI_{n \times n} + AA^\top) \quad (148)$$

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

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

since $\langle v, v \rangle \geq 0$, thus we have $\mu \leq 0$. This implies the zeros of (148) are 0 or negative.

The character polynomial of \mathcal{L} is

$$\det(\lambda I_{2n \times 2n} - \mathcal{L}) = \det(\lambda^2 I_{n \times n} + AA^\top) \quad (149)$$

$$= f(\lambda^2) \quad (150)$$

Thus the eigenvalues of \mathcal{L} are square roots of zeros of (148). Since the zeros of (148) are 0 or negative, thus the eigenvalues of \mathcal{L} are 0 or pure imaginary numbers. Moreover, since (149) is an even polynomial, it means that the polynomial only have terms of even degree, thus if 0 is a zero of (149), then its multiplicity is at least 2. \square

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 $\lambda_1 \neq \lambda_2 \neq \dots \neq \lambda_l$ 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 :

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

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

$$f_{\mathcal{L}}(x) \mid f_{-AA^\top}(x^2) = \prod_j (x - \sqrt{\lambda_j}i)(x + \sqrt{\lambda_j}i) \quad (151)$$

Moreover, since every eigenvalue of \mathcal{L} is also a root of $f_{\mathcal{L}}(x)$, from (151), 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

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

Thus for some $x = (x_1, x_2) \in \text{Ker}(\mathcal{L}^2)$, 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

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

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. \square

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 (130) 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 (132) with $\lambda = 0$ and $m = 1$. \square

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

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

and

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

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

$$M_{\mathcal{L}} = \left[\overbrace{v_1, \dots, v_{m_1}}^{m_1 \text{ columns}}, \overbrace{v_{1+m_1}, \dots, v_{1+m_1+m_2}}^{m_2 \text{ columns}}, \overbrace{v_{1+m_1+m_2+1}, \dots, v_{1+m_1+m_2+m_3}}^{m_3 \text{ columns}} \right] \quad (152)$$

where

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

$$\bullet m_1 + m_2 + m_3 = 2n.$$

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

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

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

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

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

Secondly, if $s = (s_1, s_2, \dots, s_{2n})^\top \neq 0$ 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

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

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

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

$$J_{\mathcal{L}} = \begin{bmatrix} \overbrace{J_{\mu_1, \bar{\mu}_1}}^{m_1 \text{ columns}} & & & \\ & \ddots & & \\ & & J_{\mu_{m_1/2}, \bar{\mu}_{m_1/2}} & \overbrace{0}^{m_2 \text{ columns}} \\ & & & \ddots \\ & & & & 0 & \overbrace{J_0^2}^{m_3 \text{ columns}} \\ & & & & & \ddots \\ & & & & & & J_0^2 \end{bmatrix} \quad (153)$$

$J_{\mu_k, \bar{\mu}_k} = \begin{bmatrix} \text{Re}(\mu_k) & -\text{Im}(\mu_k) \\ \text{Im}(\mu_k) & \text{Re}(\mu_k) \end{bmatrix}$ 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 eigenvector.

tors of 0 eigenvalue. Thus we have

$$e^{tJ_{\mathcal{L}}} = \oplus_{i=1}^{m_1} (e^{tJ_{\mu_i, \bar{\mu}_i}}) \oplus_{i=1}^{m_2} I_1 \oplus_{i=1}^{m_3} e^{tJ_0^2} \quad (154)$$

where

- $e^{tJ_{\mu_i, \bar{\mu}_i}}$ is defined in (132) 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 (130) with $\mu = 0$, $m = 2$. Matrix elements in $e^{tJ_0^2}$ are in $\mathcal{O}(t)$.

More precisely, we have

$$e^{tJ_{\mathcal{L}}} = \begin{bmatrix} \overbrace{\begin{matrix} \ddots & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \ddots \end{matrix}}^{m_1 + m_2} & & & \\ & \ddots & & \\ & & \overbrace{\begin{matrix} 1 & t \\ & 1 \end{matrix}}^{m_3 \text{ columns}} & \\ & & & \ddots & \\ & & & & \overbrace{\begin{matrix} 1 & t \\ & 1 \end{matrix}} \end{bmatrix} \quad (155)$$

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

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

$$e^{t\mathcal{L}} = M_{\mathcal{L}}(e^{tJ_{\mathcal{L}}}[w_1, w_2, \dots, w_{2n}]). \quad (156)$$

If AA^\top has 0 eigenvalues, from (155), we have $e^{tJ_{\mathcal{L}}}w_i$ has form

$$\left[\begin{array}{c} \vdots \\ \vdots \\ \vdots \\ \vdots \\ \hline t \\ \mathcal{O}(1) \\ \vdots \\ t \\ \mathcal{O}(1) \end{array} \right] \left. \vphantom{\begin{array}{c} \vdots \\ \vdots \\ \vdots \\ \vdots \\ \hline t \\ \mathcal{O}(1) \\ \vdots \\ t \\ \mathcal{O}(1) \end{array}} \right\} \begin{array}{l} \in \mathcal{O}(1), m_1 + m_2 \text{ rows} \\ \\ m_3 \text{ rows} \end{array} \quad (157)$$

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

$$\left[\begin{array}{c|ccc} \overbrace{\quad\quad\quad}^{m_1 + m_2} & & & & \overbrace{\quad\quad\quad}^{m_3 \text{ columns}} \\ \hline & 0 & 0 & 0 & \\ & 0 & 0 & 0 & \\ & 0 & 0 & 0 & \\ & \vdots & \vdots & \vdots & \\ & 0 & 0 & 0 & \\ \cdots & \vdots & \vdots & \vdots & \\ \cdots & \vdots & v_{m+2} & \vdots & \cdots & \vdots & v_{2n} \\ & 0 & 0 & 0 & \\ & \vdots & \vdots & \vdots & \end{array} \right] \left. \vphantom{\begin{array}{c|ccc} \overbrace{\quad\quad\quad}^{m_1 + m_2} \\ \hline \\ \\ \\ \\ \cdots \\ \cdots \end{array}} \right\} n \text{ rows} \quad (158)$$

The first n rows of $M_{\mathcal{L}}e^{tJ_{\mathcal{L}}}p_i$ is the matrix-vector product of (158) and (157), we can see the term t in (157) will product with term 0 in (158), 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 (157).

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. \square

The covariance evolution directly follows from above calculation.

Proof of covariance in continuous time equation. Denote

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

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

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

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 $\text{Var}(y_1(t_0 + t)) \in \mathcal{O}(1)$, $\text{Cov}(y_1(t_0), X_1(t_0 + t)) \in \Theta(t)$, $\text{Var}(X_1(t_0 + t)) \in \Theta(t^2)$. 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. \square

C.2.2 COVARIANCE EVOLUTION OF EULER DISCRETIZATION

Proof. The Euler discretization of (147) with step size η can be written as

$$\begin{bmatrix} y_1^t \\ X_1^t \end{bmatrix} = \begin{bmatrix} y_1^{t-1} \\ X_1^{t-1} \end{bmatrix} + \begin{bmatrix} 0 & -\eta AA^\top \\ \eta I & 0 \end{bmatrix} \cdot \begin{bmatrix} y_1^{t-1} \\ X_1^{t-1} \end{bmatrix} = \begin{bmatrix} I & -\eta AA^\top \\ \eta I & I \end{bmatrix} \cdot \begin{bmatrix} y_1^{t-1} \\ X_1^{t-1} \end{bmatrix}. \quad (159)$$

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

$$PAA^\top P^{-1} = \begin{bmatrix} \gamma_1 & & & \\ & \gamma_2 & & \\ & & \ddots & \\ & & & \gamma_n \end{bmatrix} \quad (160)$$

where $\gamma_1, \dots, \gamma_n$ 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

$$\begin{bmatrix} P & \\ & P \end{bmatrix} \begin{bmatrix} I & -\eta AA^\top \\ \eta I & I \end{bmatrix} \begin{bmatrix} P^{-1} & \\ & P^{-1} \end{bmatrix} = \left[\begin{array}{ccc|ccc} 1 & & & -\gamma_1\eta & & \\ & 1 & & -\gamma_2\eta & & \\ & & \ddots & & \ddots & \\ & & & 1 & & -\gamma_n\eta \\ \hline \eta & & & 1 & & \\ & \eta & & & 1 & \\ & & \ddots & & & \ddots \\ & & & \eta & & 1 \end{array} \right]. \quad (161)$$

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

$$\det(\mu I - \begin{bmatrix} 1 & & & -\gamma_1\eta \\ & 1 & & -\gamma_2\eta \\ & & \ddots & \\ & & & 1 \\ \hline \eta & & & -\gamma_n\eta \\ & \eta & & \\ & & \ddots & \\ & & & \eta \end{bmatrix}) \quad (162)$$

$$= \det(\begin{bmatrix} \mu-1 & & & \gamma_1\eta \\ & \mu-1 & & \gamma_2\eta \\ & & \ddots & \\ & & & \mu-1 \\ \hline -\eta & & & -\gamma_n\eta \\ & -\eta & & \\ & & \ddots & \\ & & & -\eta \end{bmatrix}) \quad (163)$$

$$= \det(\begin{bmatrix} (\mu-1)^2 + \gamma_1\eta^2 & & & \\ & (\mu-1)^2 + \gamma_2\eta^2 & & \\ & & \ddots & \\ & & & (\mu-1)^2 + \gamma_n\eta^2 \end{bmatrix}) \quad (164)$$

$$= \prod_{i=1}^n ((\mu-1)^2 + \gamma_i\eta^2) \quad (165)$$

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

$$P(t) = \begin{bmatrix} \text{Var}(y_1^t) & \text{Cov}(y_1^t, X_1^t) \\ \text{Cov}(X_1^t, y_1^t) & \text{Var}(X_1^t) \end{bmatrix} \quad (166)$$

$$= \begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}^{t-t_0} \begin{bmatrix} \text{Var}(y_1^{t_0}) & \text{Cov}(y_1^{t_0}, X_1^{t_0}) \\ \text{Cov}(X_1^{t_0}, y_1^{t_0}) & \text{Var}(X_1^{t_0}) \end{bmatrix} \left(\begin{bmatrix} I & -\eta AA^T \\ \eta I & I \end{bmatrix}^{t-t_0} \right)^\top \quad (167)$$

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 $\Theta(|\mu|^{2t})$. \square

C.2.3 COVARIANCE EVOLUTION OF SYMPLECTIC DISCRETIZATION

We firstly determine the Symplectic discretization of continuous FTRL (147) with Euclidian norm regularizer.

Lemma C.24. Discrete continuous FTRL (147) with (Type I method), we get

$$\begin{bmatrix} y_1^{n+1} \\ X_1^{n+1} \end{bmatrix} = \begin{bmatrix} I & -\eta AA^\top \\ \eta & I - \eta^2 AA^\top \end{bmatrix} \cdot \begin{bmatrix} y_1^n \\ X_1^n \end{bmatrix}. \quad (168)$$

Proof. Directly calculate gives

$$y_1^{t+1} = y_1^t - \eta AA^\top X_1^t, \quad (169)$$

$$X_1^{t+1} = X_1^t + \eta y_1^{t+1} \quad (170)$$

$$= X_1^t + \eta y_1^t - \eta^2 AA^\top X_1^t. \quad (171)$$

Combine above gives

$$\begin{bmatrix} y_1^{n+1} \\ X_1^{n+1} \end{bmatrix} = \begin{bmatrix} I & -\eta AA^\top \\ \eta & I - \eta^2 AA^\top \end{bmatrix} \cdot \begin{bmatrix} y_1^n \\ X_1^n \end{bmatrix}. \quad (172)$$

This finish the proof. \square

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

$$P^{-1}AA^\top P = \begin{bmatrix} \gamma_1 & & & \\ & \gamma_2 & & \\ & & \ddots & \\ & & & \gamma_n \end{bmatrix} \quad (173)$$

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

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

Proof. With (173), we have

$$\det(\mu I - \mathcal{M}_\eta) = \det\left(\begin{bmatrix} \mu - I & AA^\top \eta \\ -\eta & \mu - I + AA^\top \eta^2 \end{bmatrix}\right) \quad (174)$$

$$= \det((\mu - I)(\mu - I + AA^\top \eta^2) + AA^\top \eta^2) \quad (175)$$

$$= \det(\mu^2 - 2\mu + I + \mu AA^\top \eta^2) \quad (176)$$

$$\stackrel{(173)}{=} \det\left(\begin{bmatrix} \mu^2 - 2\mu + 1 + \mu\eta^2\gamma_1 & & & \\ & \mu^2 - 2\mu + 1 + \mu\eta^2\gamma_2 & & \\ & & \ddots & \\ & & & \mu^2 - 2\mu + 1 + \mu\eta^2\gamma_n \end{bmatrix}\right) \quad (177)$$

$$= \prod_{i=1}^n (\mu^2 - 2\mu + 1 + \mu\eta^2\gamma_i) \quad (178)$$

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

$$\mu^2 - 2\mu + 1 + \mu\eta^2\gamma_i = 0. \quad (179)$$

(179) 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

$$\|\mu\|^2 = \left\| \frac{(2 - \eta^2\gamma_i) \pm i\sqrt{4 - (\eta^2\gamma_i - 2)^2}}{2} \right\|^2 \quad (180)$$

$$= \frac{(2 - \eta^2\gamma_i)^2 + 4 - (\eta^2\gamma_i - 2)^2}{4} \quad (181)$$

$$= 1 \quad (182)$$

Thus we have

- \mathcal{M}_η 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}_η equals to 1.
- Every image eigenvalue of \mathcal{M}_η 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}_η to any similar matrix. Thus by (173), we only need to consider

$$\begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \mathcal{M}_\eta \begin{bmatrix} P^{-1} & 0 \\ 0 & P^{-1} \end{bmatrix} - I = \left[\begin{array}{ccc|ccc} 0 & & & -\gamma_1\eta & & \\ & 0 & & & -\gamma_2\eta & \\ & & \ddots & & & \ddots \\ & & & 0 & & -\gamma_n\eta \\ \hline \eta & & & -\gamma_1\eta^2 & & \\ & \eta & & & -\gamma_2\eta^2 & \\ & & \ddots & & & \ddots \\ & & & \eta & & -\gamma_n\eta^2 \end{array} \right] \quad (183)$$

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}_η 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 (183) has rank $2n - m$. A direct calculate show the square of (183) equals to

$$\left[\begin{array}{ccc|ccc} -\gamma_1\eta^2 & & & -\gamma_1\eta^3 & & \\ & -\gamma_2\eta^2 & & & -\gamma_2\eta^3 & \\ & & \ddots & & & \ddots \\ & & & -\gamma_n\eta^2 & & -\gamma_n\eta^3 \\ \hline -\gamma_1\eta^3 & & & -\gamma_1\eta^2 + \gamma_1^2\eta^4 & & \\ & -\gamma_2\eta^3 & & & -\gamma_2\eta^2 + \gamma_2^2\eta^4 & \\ & & \ddots & & & \ddots \\ & & & -\gamma_n\eta^3 & & -\gamma_n\eta^2 + \gamma_n^2\eta^4 \end{array} \right] \quad (184)$$

Thus if 0 is an eigenvalue of AA^\top with multiplicity m , then there are $2m$ rows in (184) be 0, thus (184) 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 (183) equals to

$$\left[\begin{array}{ccc|ccc} \gamma_1^2 \eta^4 & & & \gamma_1 \eta^3 - \gamma_1^3 \eta^5 & & \\ & \gamma_2^2 \eta^4 & & & \gamma_2 \eta^3 - \gamma_2^3 \eta^5 & \\ & & \ddots & & & \ddots \\ & & & \gamma_n^2 \eta^4 & & \gamma_n \eta^3 - \gamma_n^3 \eta^5 \\ \hline -\gamma_1 \eta^3 & & & 2\gamma_1^2 \eta^4 - \gamma_1^3 \eta^6 & & \\ & -\gamma_2 \eta^3 & & & 2\gamma_2^2 \eta^4 - \gamma_2^3 \eta^6 & \\ & & \ddots & & & \ddots \\ & & & -\gamma_n \eta^3 & & 2\gamma_n^2 \eta^4 - \gamma_n^3 \eta^6 \end{array} \right] \quad (185)$$

We can see the rank of (185) is also $2n - 2m$, thus ρ_3 in proposition C.10 is zero. This finish the proof of our claim. \square

The following lemma determines \mathcal{M}_η 's type 1 and type 2 generalized eigenvectors of \mathcal{M}_η corresponding to eigenvalue $\mu = 1$. Recall $v \in \mathbb{R}^{2n}$ is a type 2 generalized eigenvectors of \mathcal{M}_η 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}_η 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}_η corresponding to eigenvalue $\mu = 1$, then $y \in \ker(AA^\top)$ and $x = 0$.*

Proof. We have

$$(\mathcal{M}_\eta - I)^2 = \begin{bmatrix} -AA^\top \eta^2 & (AA^\top)^2 \eta^3 \\ -AA^\top \eta^3 & -AA^\top \eta^2 + (AA^\top)^2 \eta^4 \end{bmatrix}. \quad (186)$$

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

$$-AA^\top \eta^2 x + (AA^\top)^2 \eta^3 y = 0 \quad (187)$$

$$-AA^\top \eta^3 x + (-AA^\top \eta^2 + (AA^\top)^2 \eta^4) y = 0 \quad (188)$$

(188) $- \eta \cdot$ (187) gives

$$-AA^\top \eta^2 y = 0. \quad (189)$$

Thus we have $y \in \ker(AA^\top)$, and take this back to (188), 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 $y \in \ker(AA^\top)$. \square

Lemma C.28. *For an image eigenvalue $\mu \neq 1$ of \mathcal{M}_η , 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}_η , by (179), there exists an eigenvalue γ_i of AA^\top to make

$$\mu^2 - 2\mu + 1 + \mu \eta^2 \gamma_i = 0. \quad (190)$$

Moreover, the algebraic multiplicity of μ as an eigenvalue of \mathcal{M}_η 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

$$\mu - \begin{bmatrix} P & 0 \\ 0 & P \end{bmatrix} \mathcal{M}_\eta \begin{bmatrix} P^{-1} & 0 \\ 0 & P^{-1} \end{bmatrix} = \left[\begin{array}{cccc|cccc} \mu-1 & & & & \gamma_1\eta & & & \\ & \mu-1 & & & & \gamma_2\eta & & \\ & & \ddots & & & & \ddots & \\ & & & \mu-1 & & & & \gamma_n\eta \\ \hline -\eta & & & & \mu-1+\gamma_1\eta^2 & & & \\ & -\eta & & & & \mu-1+\gamma_2\eta^2 & & \\ & & \ddots & & & & \ddots & \\ & & & -\eta & & & & \mu-1+\gamma_n\eta^2 \end{array} \right] \quad (191)$$

It is easy to see if

$$\mu^2 - 2\mu + 1 + \mu\eta^2\gamma_i = 0 \quad (192)$$

the i -th and $n+i$ -th row of (191) are linearly dependent. Moreover, 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 (191) linearly dependent on m rows in the last n rows of (191). Thus if μ as rank m , (191) has rank $2n - m$.

A directly calculate shows the square of (191) equals to

$$\left[\begin{array}{cccc|cccc} (\mu-1)^2 - \gamma_1\eta^2 & & & & \gamma_1\eta(2\mu-2+\gamma_1\eta^2) & & & \\ & \ddots & & & & \ddots & & \\ & & (\mu-1)^2 - \gamma_n\eta^2 & & & & \gamma_n\eta(2\mu-2+\gamma_n\eta^2) & \\ \hline -\eta & & & & -\gamma_1\eta^2 + (\mu-1+\gamma_1\eta)^2 & & & \\ & \ddots & & & & \ddots & & \\ & & & -\eta & & & -\gamma_n\eta^2 + (\mu-1+\gamma_n\eta)^2 & \end{array} \right] \quad (193)$$

Moreover, if μ and γ_i satisfy (192), the i -th row and $n+i$ -th row are linearly dependent, thus same as matrix in (191), has rank $2n - m$ if μ is an eigenvalue of \mathcal{M}_η 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. \square

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}_η^t for $i = 1, 2, \dots, 2n$. We have

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

Proof. Denote the real generalized modal matrix of \mathcal{M}_η by M , then we have $\mathcal{M}_\eta^t = MJ_{\mathcal{M}_\eta}^t M^{-1}$, where $J_{\mathcal{M}_\eta}$ is the real Jordan normal form of \mathcal{M}_η . 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}_η , and the largest size real Jordan block have size 2. Moreover in this case, from Lemma C.27 the real generalized modal matrix of \mathcal{M}_η has same structure as the real generalized modal matrix of \mathcal{L} as in (158), 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}_η are image numbers and from Lemma C.25 these eigenvalues have norm 1. From Lemma C.28 the largest size real Jordan blocks of \mathcal{M}_η 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}_η^t are bounded. \square

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

$$P(t + t_0) = \begin{bmatrix} \text{Var}(y_1^{t+t_0}) & \text{Cov}(y_1^{t+t_0}, X_1^{t+t_0}) \\ \text{Cov}(X_1^{t+t_0}, y_1^{t+t_0}) & \text{Var}(X_1^{t+t_0}) \end{bmatrix} \quad (194)$$

$$= \mathcal{M}_\eta^t \begin{bmatrix} \text{Var}(y_1^{t_0}) & \text{Cov}(y_1^{t_0}, X_1^{t_0}) \\ \text{Cov}(X_1^{t_0}, y_1^{t_0}) & \text{Var}(X_1^{t_0}) \end{bmatrix} (\mathcal{M}_\eta^t)^\top \quad (195)$$

$$= \begin{bmatrix} A^t & B^t \\ C^t & D^t \end{bmatrix} \begin{bmatrix} \text{Var}(y_1^{t_0}) & \text{Cov}(y_1^{t_0}, X_1^{t_0}) \\ \text{Cov}(X_1^{t_0}, y_1^{t_0}) & \text{Var}(X_1^{t_0}) \end{bmatrix} \left(\begin{bmatrix} A^t & B^t \\ C^t & D^t \end{bmatrix} \right)^\top \quad (196)$$

$$= \begin{bmatrix} A^t \text{Var}(y_1^{t_0}) + B^t \text{Cov}(X_1^{t_0}, y_1^{t_0}) & A^t \text{Cov}(y_1^{t_0}, X_1^{t_0}) + B^t \text{Var}(X_1^{t_0}) \\ C^t \text{Var}(y_1^{t_0}) + D^t \text{Cov}(X_1^{t_0}, y_1^{t_0}) & C^t \text{Cov}(y_1^{t_0}, X_1^{t_0}) + D^t \text{Var}(X_1^{t_0}) \end{bmatrix} \begin{pmatrix} (A^t)^\top & (C^t)^\top \\ (B^t)^\top & (D^t)^\top \end{pmatrix}^\top \quad (197)$$

The final equation (197) equals to

$$\begin{bmatrix} P_1^t & P_2^t \\ P_3^t & P_4^t \end{bmatrix}, \quad (198)$$

where

$$P_1^t = (A^t \text{Var}(y_1^{t_0}) + B^t \text{Cov})(A^t)^\top + (A^t \text{Cov} + B^t \text{Var}(X_1^{t_0}))(B^t)^\top, \quad (199)$$

$$P_2^t = (A^t \text{Var}(y_1^{t_0}) + B^t \text{Cov})(C^t)^\top + (A^t \text{Cov} + B^t \text{Var}(X_1^{t_0}))(D^t)^\top, \quad (200)$$

$$P_3^t = (C^t \text{Var}(y_1^{t_0}) + D^t \text{Cov})(A^t)^\top + (C^t \text{Cov} + D^t \text{Var}(X_1^{t_0}))(B^t)^\top, \quad (201)$$

$$P_4^t = (C^t \text{Var}(y_1^{t_0}) + D^t \text{Cov})(C^t)^\top + (C^t \text{Cov} + D^t \text{Var}(X_1^{t_0}))(D^t)^\top. \quad (202)$$

From Lemma C.29, when AA^\top is non-singular, all elements in \mathcal{M}_η^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 $\text{Var}(y^t) \in \mathcal{O}(1)$, $\text{Cov}(X^t, y^t) \in \Theta(t)$ and $\text{Cov}(X_i^t, X_j^t) \in \Theta(t^2)$.

D DETAILS OF SECTION "LAST"

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 intrinsic 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

$$\dot{x} = \arg \max_{z \in T_x \mathcal{X}} \{G^v(x; z) - C(x; z)\}. \quad (203)$$

The dynamics equation (203) 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 $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ called the Euclidean structure. A symplectic form on an even dimensional space \mathbb{R}^n is a *skew-symmetric* bilinear map $\omega(\cdot, \cdot) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, 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 $\{u_1, \dots, u_n, v_1, \dots, v_n\}$ 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, ω) and (\mathbb{R}^n, ω') is a linear isomorphism $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$ such that $\varphi^* \omega' = \omega$, where $(\varphi^* \omega')(u, v) := \omega'(\varphi(u), \varphi(v))$. More generally, let $f : (\mathbb{R}^n, \omega) \rightarrow (\mathbb{R}^n, \omega')$ 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 $\varphi : \mathbb{R}^{2n} \rightarrow \mathbb{R}^{2n}$ 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 always 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:

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

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

D.3 PROOF OF THEOREM [5.2](#)

The first ingredient in proving Proposition [5.2](#) is based on standard results of Taylor series method in [Benaroya & Han \(2005\)](#). 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

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

and the expectation is given by

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

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

$$\sigma_Y^2 = E[Y^2] - \mu_Y^2 \simeq g^2(\mu_X) + \sigma_X^2 ([g'(\mu_X)]^2 + g(\mu_X)g''(\mu_X)) - \left(g(\mu_X) + \frac{\sigma_X^2}{2} g''(\mu_X)\right)^2. \quad (204)$$

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

$$\sigma_Y^2 \simeq \sigma_X^2 (g'(\mu_X))^2.$$

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 $Y = g(X_1, \dots, X_n)$, the Taylor series expansion about the mean value of each variable yield

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

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

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 $X = (q_1, \dots, q_n, p_1, \dots, p_n)$ 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

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

The third ingredient is the Taylor expansion of differential mapping $f : V \rightarrow W$ 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 $f_i : U \rightarrow \mathbb{R}$ denote the i th component of f , so f is described as a map $f = (f_1, \dots, f_m)$. 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 $\text{Hom}(V, W)$ be the space of linear mappings from V to W . Then the higher derivative $D^p f$ is a multi-linear mapping from $V^p \rightarrow W$, i.e.,

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

where $V^p = V \times \dots \times V$ 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,

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

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

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

which is a vector in W . The Taylor formula for differentiable mapping $f : V \rightarrow W$ 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

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

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,

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

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., $\text{Cov}(\phi_t^i(X), \phi_t^j(X))$. The fundamental property of covariance implies that for each pair (i, j) , $\text{Cov}(\phi_t^i(X), \phi_t^j(X))$ 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 $\text{Cov}(\phi_t^i(X), \phi_t^j(X))$ is of the higher power of the entries of $X - \mu$. Formally we have

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

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

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

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

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

The proof completes. \square

E EXPERIMENTS FOR NON-SINGULAR CASES.

In this section we provide more numerical experiments on the non-singular cases to complete the verification of Theorem 5.1. We provide experimental results on the evolution of $\text{Var}(X_{1,1})$, $\text{Var}(y_{1,1})$, the first components of X_1 and y_1 , where (X_1, y_1) evolve as continuous FTRL equation, Symplectic discretization. In all experiments, we assume that the payoff matrix is in $\mathbb{R}^{2 \times 2}$, thus $X_1, y_1 \in \mathbb{R}^2$, and at initial time the covariance matrix $\text{Cov}(y_1, X_1)$ is $[[8, 2, 1, 3], [2, 13, 7, 9], [1, 7, 9, 2], [3, 9, 2, 10]] \in \mathbb{R}^{4 \times 4}$, which is the same as the setting mentioned in the main text.

Continuous time FTRL. We illustrate how $\text{Var}(X_{1,1}(t))$ and $\text{Var}(y_{1,1}(t))$ 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 6. In (a) the $\text{Var}(X_{1,1}(t))$ is bounded, and in (b) $\text{Var}(y_{1,1}(t))$ is bounded, which support results of continuous time part in Theorem 5.1 for the non-singular cases.

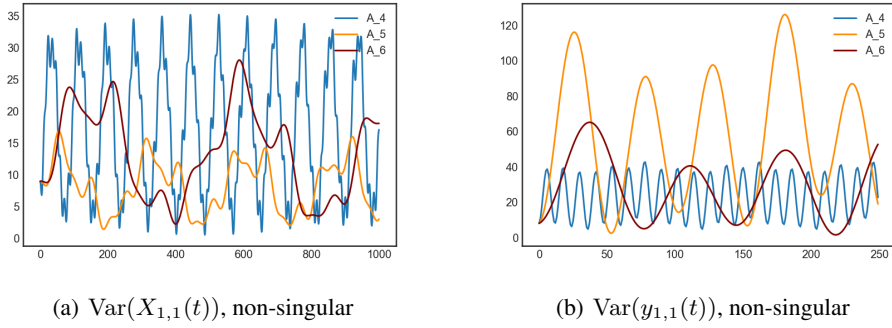


Figure 6: Variance evolution of continuous FTRL

Symplectic discretization. We illustrate how $\text{Var}(X_{1,1}^t)$ and $\text{Var}(y_{1,1}^t)$ 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 7. 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.

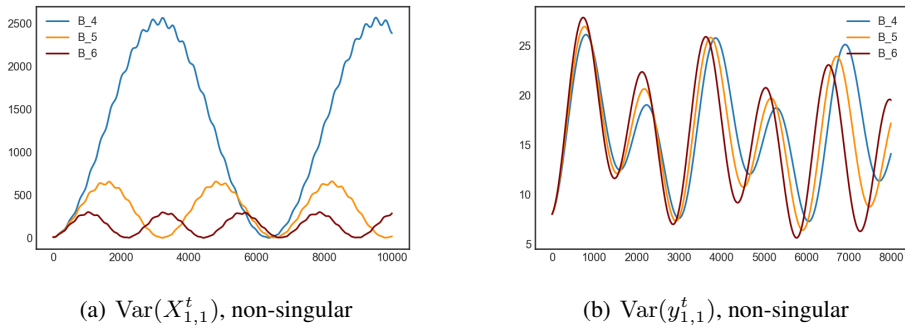


Figure 7: Variance evolution of Symplectic discretization