Skew-symmetrically perturbed gradient flow for convex optimization

Futoshi Futami FUTOSHI.FUTAMI.UK@HCO.NTT.CO.JP
Tomoharu Iwata TOMOHARU.IWATA.GY@HCO.NTT.CO.JP
Naonori Ueda NAONORI.UEDA.FR@HCO.NTT.CO.JP
Communication Science Laboratories, NTT, KYOTO, JAPAN

Ikko Yamane IKKO.YAMANE@DAUPHINE.PSL.EU
LAMSADE, CNRS, Université Paris-Dauphine, PSL Research University, PARIS/ RIKEN AIP, TOKYO, JAPAN

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

Recently, many methods for optimization and sampling have been developed by designing continuous dynamics followed by discretization. The dynamics that have been used for optimization have their corresponding underlying functionals to be minimized. On the other hand, a wider class of dynamics have been studied for sampling, which is not necessarily limited to functional minimization. For example, dynamics perturbed with skew-symmetric matrices, which cannot be seen as minimization of functionals, have been widely used to reduce asymptotic variance. Following this success in sampling, exploring such perturbed dynamics in the context of optimization can open a new avenue to optimization algorithm design. In this work, we introduce a perturbation technique for sampling into optimization for strongly convex functions. We show that perturbation applied to the gradient flow yields rapid convergence in optimization for strongly convex functions. Based on this continuous dynamics, we propose an optimization algorithm for strongly convex functions with a novel discretization framework that combines the Euler method with the leapfrog method which is used in the Hamilton Monte Carlo method. Our numerical experiments show that the perturbation technique is useful for optimization.

Keywords: Convex optimization, skew-symmetric matrices, gradient flow, discretization

1. INTRODUCTION

Analysis of continuous dynamics and discretization methods has been a driving force in recent developments in optimization and sampling. In optimization, inspired by the relation between the gradient flow as continuous dynamics and the gradient descent as discretized dynamics (Scieur et al., 2017), acceleration methods such as Nesterov's scheme have been analyzed as second-order differential equations (Su et al., 2014). Recent analysis showed that the various first-order optimization methods are closely related to continuous dynamics and discretization methods (Scieur et al., 2017). Zhang et al. (2018) and Shi et al. (2019) showed that using the high-order discretization results in acceleration.

For sampling, Wibisono (2018) analyzed the Langevin dynamics (LD) as a gradient flow in the space of probability measures and proposed a method for discretizing continuous dynamics based on a technique used in optimization. Motivated by this connection, many useful optimization techniques have been introduced into sampling (e.g., Durmus and Majewski (2019)). In particular, Muehlebach and Jordan (2019) introduced the continuous dynamics, which is used in optimization, into sampling for acceleration. In this way, studying the continuous dynamics and discretization methods in optimization has brought significant advances in recent efficient sampling algorithms.

Most continuous dynamics designed for optimization have their corresponding underlying functionals to be minimized. For example, the gradient flow and the second-order differential equations for acceleration are derived through minimization of the Bregman Lagrangian (Wibisono et al., 2016). On the other hand, designing the dynamics in sampling is not limited to minimizing functionals. For example, a *perturbation* approach that adds a small perturbation composed of a skew-symmetric matrix to the original LD has been gathering attention (Hwang et al., 2005, 2015; Duncan et al., 2016, 2017a; Kaiser et al., 2017). This perturbation technique never changes the stationary distribution but reduces the asymptotic variance for sampling. An interesting point of this perturbed dynamics is that it cannot be seen as a minimization of a functional.

Based on this success of the perturbation technique in sampling, we expect that understanding perturbed dynamics in the context of optimization may pave a new avenue for designing optimization algorithms. In this paper, we show, for the first time to the best of our knowledge, that such perturbed continuous dynamics are also useful when optimizing strongly convex functions.

However, when we adopt such perturbed dynamics into optimization, two major challenges arise. First, the advantage of the perturbation in optimization is unclear although this technique reduces the *asymptotic variance* in sampling. In optimization, we adopt the final state of a parameter as a solution, and thus the asymptotic variance is not even defined. Second, it is not obvious what kind of discretization is preferable for such perturbed dynamics. Existing work on perturbed dynamics in sampling only focused on continuous dynamics since the obtained samples can be adjusted by Metropolis-Hasting steps (Bishop, 2006).

We address the above challenges and show that the convergence rate of perturbed dynamics is improved compared to un-perturbed dynamics in the continuous and discrete-time settings.

First, we present new continuous dynamics using skew-symmetric matrices that converge more rapidly than the gradient flow under mild conditions. To show faster convergence, we analyze the perturbed Hessian matrix. Since it is neither symmetric nor skew-symmetric, it remains unclear whether diagonalization is possible. We clarify what kind of perturbation preserves the diagonalization of the Hessian matrix and then show the largest and smallest eigenvalues are changed by perturbation. This leads to faster convergence of the continuous dynamics.

Second, we provide a novel discretization method for the proposed dynamics and analyze its convergence properties. We show that a simple Euler method cannot guarantee faster convergence. To achieve faster convergence, inspired by Hamilton Monte Carlo (Bishop, 2006), we propose a new discretization method that combines the Euler and leapfrog methods to effectively exploit the particular structure of skew-symmetric matrices. Finally, we present methods for tuning the hyper-parameters of our proposed method, including those of the skew-symmetric perturbations.

2. PRELIMINARIES

In this section, we briefly introduce the gradient flow, gradient descent, and the perturbation technique in sampling.

2.1. Gradient flow and gradient descent

Consider a strongly convex loss function $F(x)$ on $\mathbb{R}^d$. We assume that F is an m-strong and M-smooth function. To minimize $F(x)$, we consider the gradient flow:

$$\frac{dx(t)}{dt} = -\nabla_x F(x(t)), \quad x(0) = x_0, \tag{1}$$

for which one can show

$$\|x(t) - x^*\| \leq e^{-mt}\|x_0 - x^*\|, \tag{2}$$

which ensures the convergence to the optimal point $x^* = \underset{x\in\mathbb{R}^d}{\arg\min} F(x)$. In many cases, we approximate Eq. (1) with a discretization method since we cannot directly implement it due to its continous nature. A widely used method is the gradient descent (GD). We use x_k to express a candidate of a solution obtained from the k-th iterate of the GD. Then, the GD algorithm is given by recursion:

$$x_{k+1} = x_k - \eta\nabla_x F(x_k), \tag{3}$$

where $\eta > 0$ is the step size. The convergence behavior is characterized as follows. GD converges if the step size satisfies $\eta \in [0, \frac{2}{M})$. Furthermore, if $\eta = \frac{2}{M+m}$, we have

$$\|x_k - x^*\| \leq e^{-\frac{m}{M}k}\|x_0 - x^*\|. \tag{4}$$

Based on the definition of strong convexity and smoothness, we can regard m and M as an upper bound and a lower bound of eigenvalues of Hessian matrix $H_x = \nabla_x^2 F(x)$ for all x. Equivalently, $H_x \succeq mI$ and $MI \succeq H_x$ hold, where I is the $d \times d$ identity matrix. Hereinafter, we simply express H_x as H. Then, the convergence of the GD is charactderized by the ratio of the largest and smallest eigenvalue of Hessian matrix. Thus, analyzing the properties of Hessian matrix is important to understand the convergence behavior.

2.2. Perturbation techinque in sampling

A perturbation to the LD is used for sampling from Gibbs distribution $\pi(x) \propto e^{-U(x)}$, where $U : \mathbb{R}^d \to \mathbb{R}$ is a potential function. Let X_t denote a random variable on $\mathbb{R}^d$ and let W denote the Wiener process. Then the perturbation to the LD is given by

$$dX_t = -(I + J)\nabla U(X_t)dt + \sqrt{2}dW, \tag{5}$$

where J is a skew-symmetric matrices that satisfies and $J = -J^\top$, and I is the identity matrix. The stationary distribution of this dynamics is $\pi(x)$. Compared to the standard LD, which corresponds to $J = 0$, the perturbed dynamics shows smaller variance in the asymptotic limit. The perturbation of J changes the smallest real part of the eigenvalue of the infinitesimal generator of Eq. (5), which is larger than the standard LD. See the following works for details: Hwang et al. (2005, 2015); Duncan et al. (2016, 2017a); Kaiser et al. (2017); Futami et al. (2020, 2021). Intuitively, the change of the eigenvalue of the generator indicates that, if U is a strongly convex function, the smallest real part of the eigenvalue of $(I + J)\nabla^2 U$ is larger than that of $\nabla^2 U$. Lelièvre et al. (2013) showed that when ∇U is a linear function, the optimal J improves the smallest and largest real part of the eigenvalue of $(I + J)\nabla^2 U$ to $\mathrm{Tr}(\nabla^2 U)/d$.

3. PROPOSED METHOD

In this section, we first present continuous dynamics. Then, we propose two types of discretization methods and an algorithm to tune the hyper-parameters.

3.1. Theoretical properties of the perturbed Hessian matrix

Inspired by the perturbation in sampling, we incorporate a skew-symmetric matrix J to the gradient term in the gradient flow:

$$\frac{dx(t)}{dt} = -\nabla_x F(x(t)) - \alpha J \nabla_x F(x(t)), \tag{6}$$

where $\alpha \in \mathbb{R}$ expresses the strength of the perturbation and J satisfies

$$J^\top = -J, \qquad \|J\|_F \le d, \tag{7}$$

where $\|\cdot\|_F$ is the Frobenius norm. We call this dynamics *skew-symmetrically perturbed gradient flow*. Inspired by the perturbation to the LD, we expect that introducing the skew-symmetric matrix changes the eigenvalues of Hessian matrix and leads to rapid convergence. We denote the original Hessian matrix by $H = \nabla^2 F$ and the perturbed Hessian matrix by $H' = (I + \alpha J)H$. To analyze the skew-symmetrically perturbed gradient flow, we need to understand H' by elucidating the following three factors: 1) whether the stationary point of the perturbed dynamics is x^*, 2) the condition in which H' is diagonalizable, and 3) the condition in which the eigenvalues are improved. In this section, we assume that J is a general skew-symmetric matrix that satisfies Eq. (7). We discuss the concrete algorithm to generate J that has nice properties in Section 3.4.

Stationary point: First, we study question 1) by analyzing the stationary point of the perturbed dynamics. From Eq. (6) and the property of the optimal point $\nabla_x F(x^*) = 0$, it is clear that x^* is also the stationary point of the perturbed dynamics. Furthermore, we can show that x^* remains the unique stationary point that satisfies $(I + \alpha J)\nabla_x F(x^*) = 0$ since $(I + \alpha J)$ has an inverse matrix; see Appendix C.

Diagonalization: Next, we discuss the diagonalizability of H'. We emphasize that the diagonalizability is an important property for convergence analysis of the continuous dynamics (see Appendix D for details). Although without the diagonalizability, we can analyze the dynamics by Jordan decomposition, it produces unsatisfactory constants in the convergence bound. Since H is a real-valued symmetric matrix, it is always diagonalizable. On the other hand, since H' is not symmetric, the diagonalizability is not assured. Remarkably, the following proposition provides the practical guarantee for the diagonalization of J'.

Proposition 1 *Suppose that J is a random matrix whose upper triangular entries follow a probability distribution that is absolutely continuous with respect to the Lebesgue measure. Then, $(I + \alpha J)H$ is diagonalizable with probability 1.*

Improvement of eigenvalues: We discuss how the real parts of the eigenvalues of H' are changed from those of H. Denote the pairs of the eigenvectors and the eigenvalues of H' as $\{(v_i^\alpha(x), \lambda_i^\alpha(x))\}_{i=1}^d$. Order them as $\mathrm{Re}(\lambda_1^\alpha(x)) \le \cdots \le \mathrm{Re}(\lambda_d^\alpha(x))$. Thus, the eigenvectors and eigenvalues of H are expressed by $\{(v_i^0(x), \lambda_i^0(x))\}_{i=1}^d$. Let $m' := \inf_{x\in\mathbb{R}^d} \mathrm{Re}(\lambda_1^\alpha(x))$ and $M' := \sup_{x\in\mathbb{R}^d} \mathrm{Re}(\lambda_d^\alpha(x))$. These m' and M' can be regarded as the modified constants of (m, M) of the objective $F(x)$. The following proposition describes the relation between the eigenvalues:

Proposition 2 *For all x, the real parts of the eigenvalues of $(I+\alpha J)H$ satisfy*

$$\lambda_1^0(x) \le \mathrm{Re}\,(\lambda_1^\alpha(x)) \le \cdots \le \mathrm{Re}\,(\lambda_d^\alpha(x)) \le \lambda_d^0(x). \tag{8}$$

In addition, denote the set of the eigenvectors of eigenvalue $\lambda_1^0(x)$ as V_1^0. Let us denote the size of V_1^0 as $|V_1^0|$. If the following condtions are satisfied, then we have $\lambda_1^0(x) = \mathrm{Re}\,(\lambda_1^\alpha(x))$:

$$\begin{cases} |V_1^0| = 1, \text{ and } v \in V_1^0,\ Jv = 0, \\ |V_1^0| > 1, \text{ and for any } v, v' \in V_1^0,\ \lambda_1^0(x)\alpha Jv = (\mathrm{Im}\,(\lambda_1^\alpha(x)))v' \text{ and } \lambda_1^0(x)\alpha Jv' = -(\mathrm{Im}\,(\lambda_1^\alpha(x)))v. \end{cases} \tag{9}$$

We have similar sufficient conditions for $\lambda_d^0(x) = \mathrm{Re}\,(\lambda_d^\alpha(x))$. Furthermore, the following relation holds:

$$\mathrm{Re}\,(\lambda_1^\alpha(x)) \le \mathrm{Tr}H/d \le \mathrm{Re}\,(\lambda_d^\alpha(x)). \tag{10}$$

Thus, from the above proposition, we have $m \le m'$ and $M' \le M$ by definition. Moreover, if α is small enough, we can evaluate the change of the largest and smallest eigenvalues:

Proposition 3 *Suppose H has d distinct eigenvalues. With the same notation as in Proposition 2, for all x and for any $i \in \{1, \dots, d\}$, we have*

$$\mathrm{Re}\,(\lambda_i^\alpha(x)) = \lambda_i^0(x) + \alpha^2 \sum_{k=1,k\neq i}^{d} \frac{|v_k^0(x) J v_i^0(x)|^2}{\lambda_k^0(x) - \lambda_i^0(x)} + \mathcal{O}(\alpha^3). \tag{11}$$

The proof is shown in Appendix C.4. Note that the first-order term in α is zero owing to the skew-symmetric property of J. From this proposition, for example,

$$\mathrm{Re}\,(\lambda_1^\alpha(x)) = \lambda_1^0(x) + \alpha^2 \sum_{k=2}^{d} \frac{|v_k^0(x) J v_1^0(x)|^2}{\lambda_k^0(x) - \lambda_1^0(x)} + \mathcal{O}(\alpha^3) \tag{12}$$

holds up to the second order. Since for all $k \ge 1$, $\lambda_k^0(x) > \lambda_1^0(x)$ holds, the second term above is positive, indicating $\mathrm{Re}\,(\lambda_1^\alpha(x)) > \lambda_1^0(x)$ for any sufficiently small α. Similarly, $\mathrm{Re}(\lambda_d^\alpha(x)) < \lambda_d^0(x)$ holds for any sufficiently small α.

3.2. Continuous dynamics

Based on the above analysis, since the largest and smallest eigenvalues are improved by introducing the skew-symmetric matrix, we expect that it will improve the convergence speed of the dynamics. We present our first main theorem that describes the effect of the skew-symmetric gradient on convergence.

Proposition 4 *If $(I+\alpha J)H$ is diagonalizable, the convergence of Eq. (6) is*

$$\|x(t) - x^*\| \le e^{-m't}\|x_0 - x^*\|. \tag{13}$$

We outline the proof since it includes an important property of continuous dynamics.
Proof (Outline) Let $r(t) := x(t) - x^*$ and define a functional $\mathcal{L}(t) = r(t)^\top r(t)$. Then

$$\frac{d\mathcal{L}}{dt} = -2r(t)^\top (I+\alpha J)\left(\nabla F(x(t)) - \nabla F(x^*)\right) = -2\int_0^1 r(t)^\top (I+\alpha J) H(\bar{x}(\tau)) r(t) d\tau,$$

where $\bar{x}(\tau) := x^* + \tau(x(t) - x^*)$. We used the Taylor expansion and expressed its residual by the integral. Since $(I + \alpha J)H$ is diagonalizable, we can analyze the dynamics based on the eigenvalue decomposition. We assume that it has l real eigenvalues $\lambda'_1, \dots, \lambda'_l$ and $2m$ complex eigenvalues, $\mu_1 = \alpha_1 \pm i\beta_1, \dots, \mu_m = \alpha_m \pm i\beta_m$. Thus, $d = l + 2m$. We express the corresponding eigenvectors as $\{v_j\}_{j=1}^{l}$ for the real eigenvalues and $\{w_j = a_j + ib_j\}_{j=1}^{m}$ for the complex eigenvalues. We express the corresponding conjugate eigenvectors as $\{\bar{w}_j\}$. Define $d \times d$ matrix V as

$$V = [v_1, \dots, v_l, a_1, b_1, \dots, a_m, b_m]. \tag{14}$$

Then, we decompose $(I + \alpha J)H(x(\tau))$ into a block diagonal matrix known as the Jordan canonical form (Golub and Van Loan, 2012);

$$(I + \alpha J)H(x^* + \tau(x(t) - x^*)) = VDV^{-1}, \tag{15}$$

where

$$D = \underbrace{\begin{pmatrix} \lambda'_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda'_l & & & & & \\ & & & \alpha_1 & 0 & & & \\ & & & 0 & \alpha_1 & & & \\ & & & & & \ddots & & \\ & & & & & & \alpha_m & 0 \\ & & & & & & 0 & \alpha_m \end{pmatrix}}_{D_1} + \underbrace{\begin{pmatrix} 0 & & & & & & & \\ & \ddots & & & & & & \\ & & 0 & & & & & \\ & & & 0 & \beta_1 & & & \\ & & & -\beta_1 & 0 & & & \\ & & & & & \ddots & & \\ & & & & & & 0 & \beta_m \\ & & & & & & -\beta_m & 0 \end{pmatrix}}_{D_2}. \tag{16}$$

Then,

$$\begin{aligned} \frac{d\mathcal{L}}{dt} = -2\int_0^1 r(t)^\top VDV^{-1} r(t) d\tau &= -2\int_0^1 r(t)^\top V(D_1 + D_2)V^{-1} r(t) d\tau \\ &= -2\int_0^1 r(t)^\top VD_1V^{-1} r(t) d\tau \\ &\leq -2\mathrm{Re}(\lambda_1^\alpha(x(\tau))) r(t)^\top r(t), \end{aligned} \tag{17}$$

where $\mathrm{Re}(\lambda_1^\alpha(x(\tau))) = \min\{\lambda'_1, \dots, \lambda'_l, \alpha_1, \dots, \alpha_m\}$. We used the skew-symmetric property of D_2 and applied the Gronwall inequality to obtain the proposition. ■

Compared to the standard GD in Eq. (2), the acceleration is confirmed since $m' \geq m$ holds. This improvement can be quantified by Eq.(12). In the proof, the key factor is the skew-symmetric property of D_2, with which we can eliminate the imaginary part of the eigenvalue from the convergence rate. See Appendix D.3 for details.

3.3. Discretization

We need to discretize the continuous dynamics to implement it. In this section, we first observe that the Euler discretization degrades the convergence rate and propose a discretization scheme that integrates the Euler and leapfrog methods to overcome this issue.

Euler discretization: First, the Euler discretization is given by

$$x_{k+1} = x_k - \eta(I + \alpha J)\nabla_x F(x_k), \tag{18}$$

where η is a stepsize. The convergence behavior is analyzed in the following way.

Proposition 5 *Define $r := \alpha \max_i \left(\sum_{j=1}^{d} |J_{ij}|\right)$. Suppose that in Eq.* (18)*, $(I + \alpha J)H$ is diagonalizable. Also suppose that α and η satisfy $r \leq m'$ and $\eta \in (0, \frac{2}{M'+r}]$. Then, x_k converges to x^* as $k \to \infty$ and the rate of convergence is*

$$\|x_k - x^*\| \leq e^{-\frac{m'-r}{M'}k} \|x_0 - x^*\|. \tag{19}$$

The proof shown in Appendix E is outlined here.

Proof (Outline) From the discretized dynamics Eq. (18), subtract x^* from both sides and define $r_k := x_k - x^*$. Then, we have

$$\|r_{k+1}\| = \|r_k - \eta(I + \alpha J)\nabla F(x_k)\|. \tag{20}$$

We define $h(x) := x - \eta(I + \alpha J)\nabla F(x)$. The above equation can be expressed:

$$\|r_{k+1}\| = \|h(x_k) - h(x^*)\|. \tag{21}$$

Then, we apply the mean-value theorem. There exists a point $\xi_k = (1-\beta)x_k + \beta x^*,\ \beta \in [0,1) \subset \mathbb{R}$ (expressed by $\xi_k \in [x_k, x^*) \subset \mathbb{R}^d$ for simplicity), such that

$$\|r_{k+1}\| \leq \| \left(I - \eta(I + \alpha J)H(\xi_k)\right) \| \|r_k\|. \tag{22}$$

Here we used the operator norm $\|\cdot\|$ defined by

$$\|M\| := \sup_{x \neq 0} \frac{\|Mx\|}{\|x\|} = \|M^\dagger M\|^{1/2} = s(M), \tag{23}$$

for any matrix M, where $s(M)$ is M's largest singular value.

To bound $\| \left(I - \eta(I + \alpha J)H\right) r_k\|$, we evaluate the singular value of $H' = I - \eta(I + \alpha J)H$. Note that from the Jordan canonical form,

$$H' = I - \eta V D V^{-1} = I - \eta V D_1 V^{-1} - \eta V D_2 V^{-1} \tag{24}$$

holds. We define $P = I - \eta V D_1 V^{-1}$ and $Q = -\eta V D_2 V^{-1}$. The largest singular value of H' (denoted by $s(H')$) is upper bounded by the largest singular values of P and Q (Bhatia, 2013) (denoted as $s(P)$ and $s(Q)$),

$$s(H') \leq s(P) + s(Q). \tag{25}$$

Note that $s(P)$ and $s(Q)$ depend on η. Thus, all we need is to bound each term. The remaining part of the proof is shown in Appendix E. ■

The key factor is that the convergence rate depends on the singular value. Given a matrix that has complex eigenvalues, its singular values depend on both the real and imaginary parts of the eigenvalues. Thus, unlike the continuous dynamics, discretized dynamics is characterized by both the real and imaginary parts of the eigenvalues. Since propositions 4 and 5 suggest a large gap between the Euler discretization and continuous dynamics, the convergence rate of Eq. (18) is not always improved compared to that of the GD.

We can intuitively understand why the Euler discretization does not work well by focuing on J. Consider the change in $F(x(t))$ in the continuous case:

$$\frac{dF(x(t))}{dt} = \nabla F(x(t))^\top \frac{dx(t)}{dt} = -\|\nabla F(x(t))\|^2 - \alpha \nabla F(x(t))^\top J \nabla F(x(t)) = -\|\nabla F(x(t))\|^2. \tag{26}$$

The value of F is preserved for J due to the skew-symmetric property. Although such preservation is a critical property, the Euler method does not take it into consideration.

Euler-leapfrog discretization: To exploit the preservation property of J, we propose a new discretization method that combines the Euler and the leapfrog methods, which is widely used in Hamilton Monte Carlo (Bishop, 2006). We split the dynamics into two parts. One is related to J, and we discretize it by the leapfrog method. The other part is unrelated to J, and we discretize it by the Euler method. To implement the leapfrog method, we introduce auxiliary variable $y_k \in \mathbb{R}^d$ and optimize augmented objective function $\tilde{F}(x, y) = F(x) + \frac{1}{2c}\|y\|^2$ where c is a positive constant, whose condition is described in Proposition 6. Then, we update $\{x_k\}$ and $\{y_k\}$ by

$$\begin{cases} x_{k+\frac{1}{2}} = x_k - \frac{\eta\alpha}{c} J y_k, \\ y_{k+\frac{1}{2}} = y_k - \eta\alpha J \nabla F(x_{k+\frac{1}{2}}), \end{cases} \tag{27}$$

$$\begin{cases} y_{k+1} = y_{k+\frac{1}{2}} - \frac{\eta}{c} y_{k+\frac{1}{2}}, \\ x_{k+1} = x_{k+\frac{1}{2}} - \eta \nabla F(x_{k+\frac{1}{2}}). \end{cases} \tag{28}$$

Eq. (27) corresponds to the leapfrog step, which discretizes the dynamics related to J. Eq. (28) corresponds to the Euler step, which discretizes the gradient flow. We call this the Euler-leapfrog (ELF) discretization. In Appendix F.4, we compared the ELF method with other discretization methods. Before we present a formal statement, we explain the intuition of our ELF method in matrix form:

$$\left\| \begin{pmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{pmatrix} \right\| \le \left\| \underbrace{\begin{pmatrix} -\eta H(\xi_{k+\frac{1}{2}}) & 0 \\ 0 & (1-\eta/c)I \end{pmatrix}}_{=\tilde{H}(c,\eta)} \underbrace{\begin{pmatrix} I & -\eta\alpha c^{-1} J \\ -\eta\alpha J H(\xi_{k+\frac{1}{2}}) & I + \eta^2\alpha^2 c^{-1} J H(\xi_{k+\frac{1}{2}}) J \end{pmatrix}}_{=L(\eta,c,\alpha,J)} \right\| \left\| \begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} \right\|, \tag{29}$$

where $\xi_{k+\frac{1}{2}} \in [x_{k+\frac{1}{2}}, x^*)$ is a constant in $\mathbb{R}^d$, specified by the mean-value theorem; see Appendix F.1 for details. In Eq. (29), $\tilde{H}$ corresponds to the Euler step of Eq. (28) and L corresponds to the leapfrog step of Eq. (27). If we appropriately select α, the singular values of L will be 1. This is the characteristic property of the leapfrog step. From the submultiplicativity of the matrix norm, $\|\tilde{H}L\| \le \|\tilde{H}\|\|L\| \le \|\tilde{H}\| = 1 - \frac{m}{M}$ under appropriate conditions for η and c. Furthermore, if we generate J following the rules described in Section 3.4, the ELF method will converge faster than the GD. Summarizing these results, we have the following theorem, whose proof is shown in Appendix F.1:

Proposition 6 *In Eqs. (27),(28), if η, c, and α satisfy $\eta \in (0, \frac{2}{M}]$, $c^{-1} \in (0, \frac{2}{\eta}]$, and $\alpha^2 \le 4c(\eta^2 M s_d^2)^{-1}$, where s_d is the largest singular value of J, x_k converges to x^* as $k \to \infty$. If we set $\eta = \frac{2}{m+M}$, $c^{-1} \in (m, M]$, and $\alpha < 2\sqrt{c(\eta^2 M s_d^2)^{-1}}$,*

$$\|x_k - x^*\| \le e^{-\kappa(\alpha,m,M,c,J)k} \|x_0 - x^*\| \tag{30}$$

holds for positive constant $\kappa(\alpha, m, M, c, J)$ that satisfies $\kappa(\alpha, m, M, c, J) \ge 2m/(m + M)$. If $m \ne M$ and $\ker J = \{0\}$ are satisfied, then $\kappa(\alpha, m, M, c, J) > 2m/(m + M)$ holds.

From above proposition, if we choose hyperparameters appropriately, the ELF method shows faster convergence than gradient descent.

3.4. Tuning hyper-parameters for the ELF method

Here, we present an algorithm to tune J, α, and c in the ELF method to satisfy conditions of Propositions 1 and 6. Detailed explanation of the algorithm is shown in Appendix F.2. We assume that $m \neq M$. Our proposed algorithm is summarized in Algorithm 1 and its theoretical property is shown in Theorem 7.

First, we discuss how to generate J. Lelièvre et al. (2013) obtained the optimal J when the drift function is linear under continuous time. However, to get the optimal J, we require $\mathrm{O}(d^3)$ time per iteration, which is computationally demanding. Such J may cause numerical instability for discretized dynamics shown in Section 5. Instead, we propose using a random matrix for J, as suggested from Proposition 1, and fix it during the optimization to reduce the computational cost. Although this choice is not optimal, it successfully alters the trajectory and improves the convergence rate but does not cause the numerical instability due to the large singular values. See Section 5 for details.

J needs to be generated to satisfy the assumption of Proposition 1. We also want to ensure the condition for $\kappa > 2m/(m+M)$ in Proposition 6 so that acceleration will occur. Also, from Proposition 6, $\mathrm{ker}J = \{0\}$ is a sufficient condition for that. To satisfy $\mathrm{ker}J = \{0\}$, we first generate matrix J' wherein the upper triangular entries $(i < j)$ are

$$J'_{ij} = \begin{cases} 1 + \rho_{ij}/d \quad \rho_{ij} \sim \mathcal{N}(0, \epsilon) & \text{if } i \text{ is odd and } i = j+1, \\ \rho_{ij}/d \quad \rho_{ij} \sim \mathcal{N}(0, \epsilon) & \text{otherwise,} \end{cases} \tag{31}$$

where $\mathcal{N}(0, \epsilon)$ denotes the zero-mean Gaussian distribution with small variance ϵ. For example, we set $\epsilon = 10^{-4}$ in numerical experiments. Finally, we set J as $J = J'^{\top} - J'$. This J is diagonalizable, and the eigenvalues are very close to $\pm i$ if d is even, which means that $\mathrm{ker}J = \{0\}$. If d is odd, however, eigenvalues are $\pm i$ and 0, which implies that $\mathrm{ker}J \neq \{0\}$. To resolve this problem with the case of odd d, a simple idea is to introduce dummy variable $\tilde{x}$ so that the dimension of the objective function will be even. Then, we optimize $F(x, \tilde{x}) = F(x) + \gamma\|\tilde{x}\|^2$ where γ is a positive constant. We can use $\gamma = \frac{1}{2c}$ so that the convergence rate of $F(x, \tilde{x})$ is dominated by the original $F(x)$.

Next, we set c as $c^{-1} = \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x-y\|}$, where x and y are arbitrary distinct points, e.g., those chosen from the initial point and its neighborhood. Condition $m \leq c^{-1} \leq M$ in Proposition 6 holds by definition.

Finally, α must satisfy $0 < \chi < 4$, where $\chi := \eta^2\alpha^2 c^{-1} s_d^2 M$, which is required for the ELF method to accelerate the convergence. We also empirically observed that the ELF method works well with χ around 1. Based on these insights, we set α so that $\alpha^2 = \frac{c}{2\eta s_d^2}$, which ensures $0 < \chi \leq 1 < 4$ since $\eta M \leq 2$. From the construction of J, the largest singular value s_d is upper bounded by $s_d^2 \leq \max_i(1 + \sum_{j \neq i} |J_{ij}|/d)^2$ from the Gerchgorin theorem (Golub and Van Loan, 2012). We use this as an estimate of s_d^2. In practice, setting η to a large value is advisable within condition $\eta M \leq 2$ so that χ will be close to 1. Summarizing the above discussions, we generate α and J by Algorithm 1, analyzed by the following proposition.

Proposition 7 *Suppose that d is even, $m \neq M$, and J and α are generated by Algorithm 1. Then, with high probability, the conditions of Proposition 6 are satisfied and $\kappa > \frac{2m}{m+M}$ holds.*

The detailed proof is shown in Appendix F.3. This proposition guarantees that the ELF shows better convergence than the GD with high probability. We confirm that matrix J obtained by Algorithm 1

Algorithm 1 Tuning hyperparameters α and J

```
Input: $\eta, c, \epsilon$ (e.g., $\epsilon = 10^{-4}$)
Output: $\alpha, J$
Make a random matrix $J'$ by Eq. (31).
Calculate $J = J'^{\top} - J'$
Calculate $s_d^2 = \max_i (1 + \sum_{j \neq i} |J_{ij}|/d)^2$
Set $\alpha = \sqrt{c(2\eta s_d^2)^{-1}}$
```

indeed shows better convergence behavior in numerical experiments. When d is odd, we solve $F(x, \tilde{x}) = F(x) + \frac{1}{2c}\|\tilde{x}\|^2$, where $\tilde{x}$ is a dummy variable so that we can apply Proposition 7.

Compared to the optimal J that requires $\mathcal{O}(d^3)$ (Lelièvre et al. (2013)), the calculation cost of Algorithm 2 is $\mathcal{O}(d)$ time. Our hyper-parameter tuning also works even in nonlinear dynamics, although the optimal J given by Lelièvre et al. (2013) can only be applicable to linear drift functions. In the numerical experiments in Section 5, we observed that the optimal J of Lelièvre et al. (2013) is unstable for discretized dynamics. When implementing the ELF discretization, we can re-use the gradient calculation in Eqs. (27) and (28), and thus, the computation cost of the ELF method is not much larger than that of the Euler discretization.

4. DISCUSSION AND RELATED WORK

In this section, we discuss the relationship between our proposed method, perturbation technique in sampling, and other optimization methods.

4.1. Relation to perturbation technique in sampling

Although our work is inspired by perturbation technique in sampling (Hwang et al., 2005, 2015; Duncan et al., 2016, 2017a; Kaiser et al., 2017), it is different in the sense that we focused on the property of J and discretizations. For the first time, our work propose using a random matrix for J and analyze the desirable conditions. We present a concrete algorithm to construct J in the ELF method. No previous work has considered the relation between J and discretization methods. Lelièvre et al. (2013) derived the optimal J, but it is limited to linear dynamics and is computationally demanding. Our numerical experiments in Section 5 also show that such an optimal J causes a numerical issue for discretized dynamics. Although Duncan et al. (2017b) worked on the splitting method, they focused on its asymptotic behavior with a general skew-symmetric matrix.

4.2. Relation to preconditioning methods

Our methods can be understood as preconditioning schemes. One of the most successful preconditioning methods is Newton's method and its approximations. These methods take metric information into consideration and multiply the inverse of Hessian matrix to the gradient. Thus, the gradient of each dimension is re-scaled, and the condition number of these dynamics becomes one in the re-scaled space. See Appendix G for details. However, since calculating such inverse matrices is computationally demanding, many variants of methods have been established.

Our proposed dynamics correlate different dimension by skew-symmetric matrices, and the perturbed Hessian matrix shows that the smallest real part of the eigenvalue is larger than that of the un-perturbed Hessian matrix. This results in a faster convergence compared to the un-perturbed dynamics and makes the trajectory smoother than the GD. See Section 5. As Lelièvre et al. (2013)

argued, for linear dynamics, we can construct optimal J, and the smallest and largest real parts of the eigenvalues become $\mathrm{Tr}H/d$, which means that the condition number becomes one, which is the same as Newton's method. Concerning the computational cost, our methods need an additional matrix and a vector product computation, which is usually much smaller than Newton's method.

4.3. Condition number and ℓ_2 regularization

Our method and ℓ_2 regularization are similar in the sense that the smallest and largest eigenvalues of the Hessian matrix change. If $\gamma \in \mathbb{R}^+$ is a regularization parameter, then objective function $F(x) + \gamma\|x\|^2$ is a $(m+\gamma)$-strong convex and a $(M+\gamma)$-smooth function. Thus, the condition number becomes $\frac{M+\gamma}{m+\gamma}$. This indicates that the convergence rate improved. However, the obtained solution is biased. On the other hand, our method improves the convergence rate without biasing the solution.

4.4. Relation to other continuous dynamics for optimization

Studying optimization algorithms through continuous dynamics has become an important approach. For example, Scieur et al. (2017) recently described the relation between the gradient flow and several discretization methods with a variety of optimization methods, including accelerated optimization methods such as the Nesterov method. Since our proposed dynamics is a perturbed gradient flow, we can combine more sophisticated higher-order discretization methods to ours following by Scieur et al. (2017). We note that the continuous dynamics of Nesterov's scheme is known as a second-order differential equation (Wibisono et al., 2016), while our continuous dynamics are first-order differential equations. Future work might introduce perturbation to that second-order equation.

5. NUMERICAL EXPERIMENTS

We confirmed our theoretical findings through numerical experiments. First, we confirmed the acceleration of continuous dynamics. Then, we observed the convergence behavior of two different discretization methods: the Euler and Euler-leapfrog (ELF) methods. We also show additional numerical experiments in Appendix I.

5.1. Least square experiments

We considered $F(x) = \frac{1}{N}\sum_{i=1}^{N}(A_{i*}x - y_i)^2$ where $y = (y_1, \dots y_N)^\top$ and A_{i*} denotes the ith row of $A \in \mathbb{R}^{N\times d}$. We generated design matrix A with entries following $\mathcal{N}(0,1)$. y was generated by $y = Az + \epsilon$ where $z \sim \mathcal{N}(0, I_{d\times d})$ and $\epsilon \sim \mathcal{N}(0, I_{N\times N})$. Then, Hessian matrix is $H = A^\top A$. Since the properties of J depend on whether d is odd or even, we considered $(d,N) = (400,600)$ and $(401,600)$. For $d = 401$, we introduced a dummy variable and solved the problem with $d = 402$. Detailed experimental settings and further discussions are presented in Appendix H.

First, we compared the continuous dynamics of the gradient flow (GF) and the perturbed dynamics under three types of J. One is a completely random matrix of which each entry follows the standard Gaussian (Random-J); the second is obtained by Algorithm 1 (Alg-J); the third is the optimal matrix obtained by the method in Lelièvre et al. (2013) (Opt-J) (its algorithm is shown in Appendix H). The results are shown in Fig. 1. For the perturbed dynamics, the results are the averages of ten repetitions for different realizations of random perturbation. Table 1 shows how the largest and the smallest real part of the eigenvalues of the Hessian matrix are changed by the perturbation. As shown in Proposition 4, the larger the smallest real part of the eigenvalue is, the faster convergence we have.

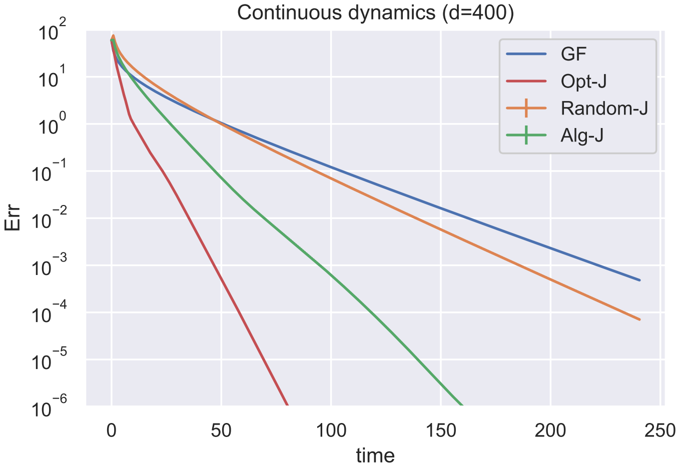


Figure 1: Convergence behavior of continuous dynamics

Table 1: Smallest and largest real parts of eigenvalues of Hessian matrix $A^\top A$

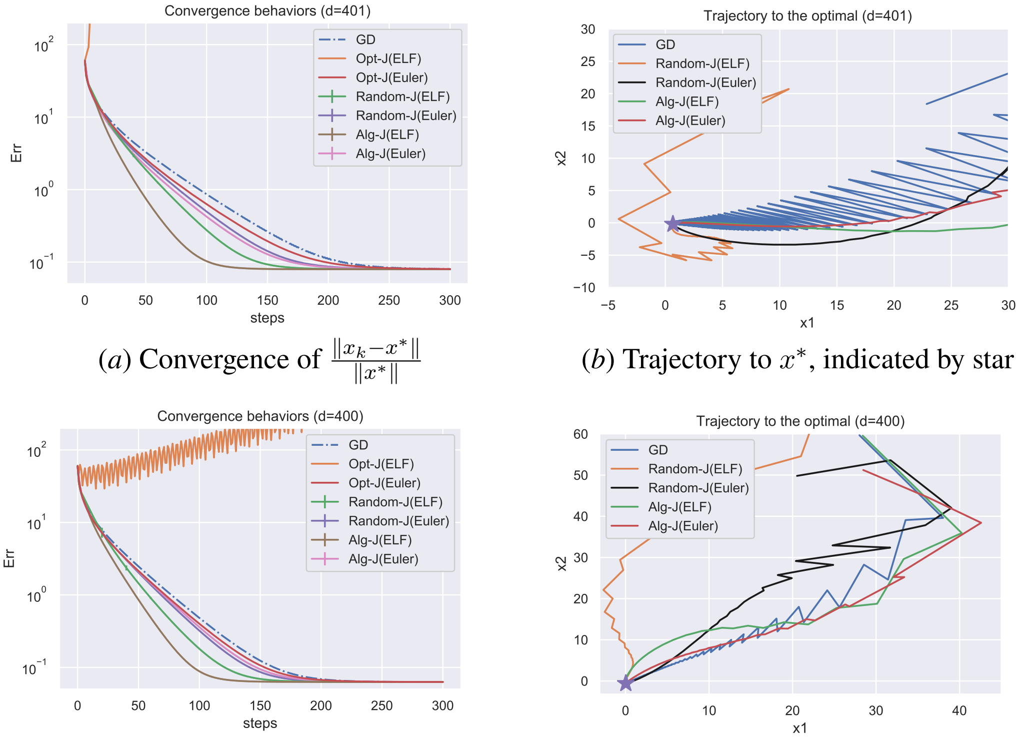
	$\mathrm{Re}\lambda_1$	$\mathrm{Re}\lambda_d$
GF	0.03	3.26
Random-J	0.05 $\pm$0	2.64 $\pm$ 0.003
Alg1-J	0.10 $\pm$0	2.52 $\pm$ 0
Opt-J	0.20	1.74

The optimal choice of J shows the best performance. We can confirm that completely random J still remains useful for acceleration. We also confirmed that for each different J, all the eigenvalues of the perturbed Hessian matrix are distinct, meaning that the perturbed Hessian matrix is diagonalizable. We also show the histogram of the sigular values for each J in Appendix 7.

Next, we compared the discretization methods and different choices of J. The choice of J is identical as the continuous settings. We used optimal step sizes. For the ELF, we tuned α following Algorithm 1. For the Euler method, since it was sensitive to the choice of α, we reported the best result among those obtained with several different αs. The results are shown in Fig. 2. As shown in Propositions 5 and 6, although the ELF method shows faster convergence than the GD, the Euler discretization does not. We also found that the optimal choice of J by Lelièvre et al. (2013) is unstable with the ELF method. This is because its singular values are significantly large, and it does not satisfy the conditions of the ELF method. Figs. 2(b) and 2(d) show the trajectories of the GD and the proposed perturbed dynamics. Those of the perturbed dynamics are smoother. This figure suggests that the proposed method achieved rapid convergence.

5.2. Logistic regression experiments

We considered learning parameters of logistic regression for binary classification. Let the input and output pairs of data $\{(z_i, y_i)\}_{i=1}^N$, where $z_i \in \mathbb{R}^d$ and $y_i \in \{1, -1\}$. Let $\tilde{z}_i := (z_i, 1)^\top \in \mathbb{R}^{d+1}$. The objective function is given as $F(x) = \frac{1}{N}\sum_{i=1}^N \ln \sigma(y_i x^\top \tilde{z}_i)$, where $\sigma(x) = (1 + \exp(-x))^{-1}$ is the logistic function and $x \in \mathbb{R}^{d+1}$ is the parameter that we optimized. We compared the convergence



(a) Convergence of $\frac{\|x_k - x^*\|}{\|x^*\|}$

(b) Trajectory to x^*, indicated by star

(c) Convergence of $\frac{\|x_k - x^*\|}{\|x^*\|}$

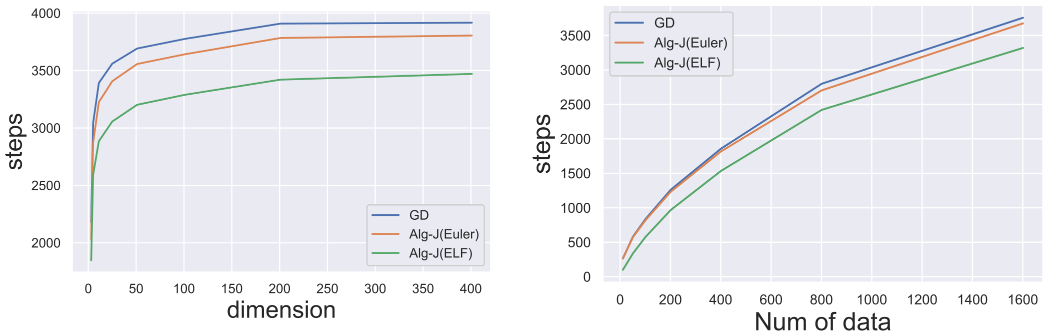
(d) Trajectory to x^*, indicated by star

Figure 2: Comparisons of different discretization. $d = 401$ for (a) and (b). $d = 400$ for (c) and (d)

speed of the GD and our proposed algorithm with skew-symmetric matrices generated by Algorithm 1 (Alg-J) and used the discretizations of the ELF method and the Euler method. Note that in logistic regression, using optimal skew-symmetric matrices is computationally demanding since Hessian matrices depend on the current position of x_k. This means the optimal J changes during the optimization, and thus we need to calculate the optimal J at each step. We also found that using the completely random skew-symmetric matrices with each entry following the standard Gaussian does not accelerate the convergence.

First, we considered toy data experiments to observe the convergence behavior of the GD and the Euler and Euler-leapfrog (ELF) methods of our proposed methods. To generate toy data, we drew each dimension of z from the uniform distribution between -1 and 1 and generated each dimension of the true parameter x from the uniform distribution between -5 and 5. The result is shown in Fig. 3. In Fig. 3(a), we fixed $N = 5000$, changed d, and measured the number of steps required for convergence. In Fig. 3(b), we fixed $d = 501$ and changed N. In both experiments, we confirmed that our proposed algorithm consistently accelerated the convergence in both large sample and large dimension settings.

Next, we used a real dataset to confirm that our proposed algorithm is useful in practice. We used four datasets in the UCI machine learning repository (Dheeru and Karra Taniskidou, 2017), and the result is shown in Fig. 4. Our proposed algorithm using the ELF method consistently accelerated the convergence. We found that using the Euler discretization did not always accelerate the convergence, which is consistent with our Proposition 5.



(a) Change d and fix $N = 5000$

(b) Change N and fix $d = 501$

Figure 3: Convergence behaviors of logistic regression under different d and N

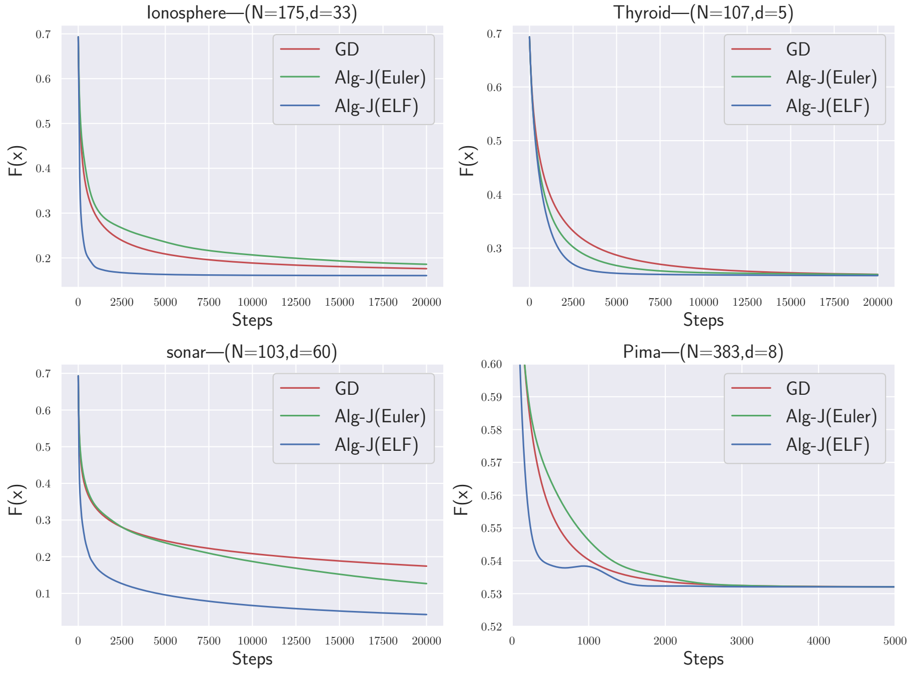


Figure 4: Convergence behavior of logistic regression: N is amount of data points and d is input dimensions.

6. CONCLUSION

We proposed a new continuous dynamics, which was obtained by perturbing the gradient flow by a random skew-symmetric matrix. By analyzing the perturbed Hessian matrix, we proved that perturbed dynamics shows rapid convergence. We presented a new discretization method that combines the Euler and leapfrog methods. It preserved the faster convergence property better than the gradient descent. We also presented an effective algorithm to select hyper-parameters.

An important conclusion of our work is that the perturbation technique in sampling is also useful for optimization. Our result suggests that perturbing the underlying dynamics is different from the standard scheme of minimizing a functional, it is a promising approach for designing optimization algorithms.

Our work can be extended in various ways. In this paper, we focused on the perturbation of a skew-symmetric matrix although there are other types of perturbations in sampling such as the one proposed by Maragliano and Vanden-Eijnden (2006). Incorporating such techniques into optimization would be an interesting research direction. Combining our technique with Nesterov's second-order dynamics scheme is also promising. In sampling, applying our discretization technique to existing Langevin-based sampling may provide potential improvements.

Acknowledgments

FF was supported by JST ACT-X Grant Number JPMJAX190R. IY acknowledges the support of the ANR as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute).

References

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

Dua Dheeru and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Andrew B. Duncan, Tony Lelièvre, and Grigorios. A. Pavliotis. Variance reduction using nonreversible langevin samplers. *Journal of Statistical Physics*, 163(3):457–491, May 2016.

Andrew B. Duncan, Nikolas. Nüsken, and Grigorios. A. Pavliotis. Using perturbed underdamped langevin dynamics to efficiently sample from probability distributions. *Journal of Statistical Physics*, 169(6):1098–1131, Dec 2017a.

Andrew B. Duncan, Grigorios. A. Pavliotis, and Konstantinos. C. Zygalakis. Nonreversible langevin samplers: Splitting schemes, analysis and implementation. *arXiv preprint arXiv:1701.04247*, 2017b.

Alain Durmus and Szymon Majewski. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

Futoshi Futami, Issei Sato, and Masashi Sugiyama. Accelerating the diffusion-based ensemble sampling by non-reversible dynamics. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3337–3347. PMLR, 13–18 Jul 2020.

Futoshi Futami, Tomoharu Iwata, Naonori Ueda, and Issei Sato. Accelerated diffusion-based sampling by the non-reversible dynamics with skew-symmetric matrices. *Entropy*, 23(8), 2021. ISSN 1099-4300. doi: 10.3390/e23080993.

Gene H Golub and Charles F. Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

Chii-Ruey Hwang, Shu-Yin Hwang-Ma, Shuenn-Jyi Sheu, et al. Accelerating diffusions. *The Annals of Applied Probability*, 15(2):1433–1444, 2005.

Chii-Ruey Hwang, Raoul Normand, and Sheng-Jhih Wu. Variance reduction for diffusions. *Stochastic Processes and their Applications*, 125(9):3522–3540, 2015.

Marcus Kaiser, Robert L. Jack, and Johannes Zimmer. Acceleration of convergence to equilibrium in markov chains by breaking detailed balance. *Journal of Statistical Physics*, 168(2):259–287, Jul 2017.

Tony Lelièvre, Francis Nier, and Grigorios A Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013.

Luca Maragliano and Eric Vanden-Eijnden. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chemical physics letters*, 426(1-3):168–175, 2006.

Michael Muehlebach and Michael Jordan. A dynamical systems perspective on nesterov acceleration. In *International Conference on Machine Learning*, pages 4656–4662, 2019.

Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *Ann. Statist.*, 1(4):763–765, 07 1973. doi: 10.1214/aos/1176342472.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Damien Scieur, Vincent Roulet, Francis Bach, and Alexandre d'Aspremont. Integration methods and optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 1109–1118, 2017.

Bin Shi, Simon S Du, Weijie Su, and Michael I Jordan. Acceleration via symplectic discretization of high-resolution differential equations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

Andre Wibisono. Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem. In *Conference On Learning Theory*, pages 2093–3027, 2018.

Andre Wibisono, Ashia C Wilson, and Michael I Jordan. A variational perspective on accelerated methods in optimization. *proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

Jingzhao Zhang, Aryan Mokhtari, Suvrit Sra, and Ali Jadbabaie. Direct runge-kutta discretization achieves acceleration. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Appendix A. Motivating example

Before going to the detailed analysis, let us observe a simple example. Let $x, y \in \mathbb{R}$ and $F(x,y) := x^2 + y^2 + xy$. Then, the hessian matrix is

$$\nabla^2 F = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \tag{32}$$

and its eigenvalues are

$$\lambda = 3, 1. \tag{33}$$

Let us add the skew-symmtric matrix:

$$(I+J)\nabla^2 F = \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \right) \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & 3 \\ -1 & 1 \end{pmatrix}. \tag{34}$$

Then, the eigenvalues are

$$\lambda = 2 \pm i\sqrt{2}. \tag{35}$$

Thus, the eigenvalues are changed, while the trace is preserved. If the convergence is dominated the real part of the largest and the smallest eigenvalues, then this skew-matrix can accelerate the convergence.

Appendix B. Motivation of the assumption

In the main paper, we assumed $\|J\|_F \leq d$. This assumption is motivated by the fact that $\|I\|_F = d$.

Appendix C. Properties of a skew-symmetric matrix

C.1. Proof of the existence of an inverse matrix $I + \alpha J$

Proof A matrix has an inverse matrix if its determinant is not 0. Then, we will prove $\det(I+J) \neq 0$. This holds because, since eigenvalues of a skew-symmetric matrix J is 0 or purely imaginary (see Petersen et al. (2008)). Thus, $J + I$ has 1 or complex values as eigenvalues thus it cannot have 0 as an eigenvalue. Thus, $\det(I+J) \neq 0$ holds. ■

From this analysis, $(I+\alpha J)\nabla F(x) = 0$ indicates $\nabla F(x) = 0$. Since x^* is the unique stationary point that satisfies $\nabla F(x^*) = 0$, the perturbed dynamics has the same stationary point with the un-perturbed dynamics.

C.2. Proof of Proposition 1

Next, we discuss diagonalization of $(I+\alpha J)H$. Before that, we state a useful property for J:

Lemma 1 *$(I+\alpha J)H$ and $H + \alpha H^{1/2} J H^{1/2}$ are similar, i.e., have common eigenvalues.*

$H^{1/2}JH^{1/2}$ is skew-symmetric while the original JH is not. Thus $H^{1/2}JH^{1/2}$ is much easier to analyze than JH.

Proof First, observe that

$$H^{1/2}((I+J)H)H^{-1/2} = H + H^{1/2}JH^{1/2}. \tag{36}$$

This means $(I+J)H$ and $H + H^{1/2}JH^{1/2}$ are similar. Since the matrices which are similar with each other have the same eigenvalues, $(I+J)H$ and $H + H^{1/2}JH^{1/2}$ have the same eigenvalues. ■

Proof [Proof of Proposition 1] The proof is almost the same as that of the main theorem in Okamoto (1973). As shown in Okamoto (1973), we only need to prove that the discriminant of the characteristic polynomial of $H + \alpha JH$ is not identically 0. In this proof, we use the lemma in Okamoto (1973);

Lemma 2 *(Okamoto (1973)) If $f(x_1, \dots, x_m)$ is a polynomial in real variables $x_1, \dots, x_m$, which is not identically zero, i.e., there exists $(x_1, \dots, x_m)$ such that $f(x_1, \dots, x_m) \neq 0$, then the subset $N_m = \{(x_1, \dots, x_m) | f(x_1, \dots, x_m) = 0\}$ of the Euclidean m-space $\mathbb{R}^m$ has the Lebesgue measure zero.*

We consider that f in the above lemma corresponds to the discriminant of the characteristic polynomial of $H + \alpha JH$. That is, the characteristic polynomial is given as

$$f(J_{1,2}, \dots, J_{d-1,d}) = |\lambda I_d - H'|. \tag{37}$$

Then, if the discriminant of this polynomial is not equal to 0, then H' has distinct eigenvalues.

From the above lemma, if the discriminant of the characteristic polynomial of $H + \alpha JH$ is not identically 0, then the probability that the discriminant of the characteristic polynomial is 0 with probability 0. That means the probability that $H + \alpha JH$ has distinct eigenvalues is 1. This means that $H + \alpha JH$ is diagonalizable with probability 1.

Thus our goal here is to prove that the discriminant of the characteristic polynomial of $H + \alpha JH$ is not identically 0. The outline is that given x, we have H_x. For any random J generated by Alg. 1, we prove that there exists a $\tilde{J}$ that is arbitrarily close to J, for which the discriminant of the characteristic polynomial of $H + \alpha \tilde{J} H$ is not 0. If such $\tilde{J}$ exists, it is clear that the characteristic polynomial of $H + \alpha JH$ is not identically 0.

Hereafter, for simplicity, we set $\alpha = 1$. Also from lemma 1, we only need to consider the eigenvalues of $H + H^{1/2}JH^{1/2}$. First, let us express the Jordan canonical form of $H + H^{1/2}JH^{1/2}$ as follows: $H + H^{1/2}JH^{1/2}$ has l real eigenvalues $\lambda_1, \dots, \lambda_l$ and $2m$ complex eigenvalues, $\mu_1 = \alpha_1 \pm i\beta_1, \dots, \mu_m = \alpha_m \pm i\beta_m$. Thus, $d = l + 2m$. We denote the corresponding generalized eigenvectors as $\{v_j\}_{j=1}^{l}$ for real eigenvalues. Here we assumed the generalized eigenvectors since $H + H^{1/2}JH^{1/2}$ is not always diagonalizable. We also denote the generalized eigenvectors as $\{w_j = a_j + ib_j\}_{j=1}^{m}$ for complex eigenvalues $\{\mu_j\}_{j=1}^{m}$, we denote their conjugate as $\{\bar{w}_j\}$, which are the generalized eigenvectors of the conjugate eigenvalues. Then define a matrix as

$$V = (v_1, \dots, v_l, a_1, b_1, \dots, a_m, b_m). \tag{38}$$

From the definition of the generalized eigenvectors, we have

$$H + H^{1/2}JH^{1/2} = V\Lambda V^{-1} \tag{39}$$

$$\Lambda := \begin{pmatrix} \Lambda_1 & & \\ & \ddots & \\ & & \Lambda_r \end{pmatrix}, \tag{40}$$

where each elementary Jordan block Λ_i are expressed as

$$\Lambda_i^{\mathrm{R}} := \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ & & \ddots & & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}, \tag{41}$$

for the real eigenvalues and

$$\Lambda_i^{\mathrm{C}} := \begin{pmatrix} A & I_2 & 0 & \dots & 0 \\ 0 & A & I_2 & \dots & 0 \\ & & \ddots & & \\ & & & A & I_2 \\ & & & & A \end{pmatrix}, \quad A := \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}, \quad I_2 := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{42}$$

for the complex eigenvalues. We will construct a specific skew-symmetric matrix J' such that perturbing Λ with J' will yield a matrix with all eigenvalues distinct from each other, and thus the discriminant characteristic polynomial will be nonzero.

We perturb Λ_i by a skew-symmetric matrix J_i' so that the perturbed Λ_i has distinct eigenvalues. For Λ_i^{R}, if its size is an even number $2m'$, take the following skew-symmetric matrix

$$J_i'^{\mathrm{R}} := \begin{pmatrix} 0 & c_1 & 0 & \dots & 0 \\ -c_1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & c_2 \dots & 0 \\ 0 & 0 & -c_2 & 0 \dots & 0 \\ & & \ddots & & \\ & & & 0 & c_{m'} \\ & & & -c_{m'} & 0 \end{pmatrix}, \tag{43}$$

where $\{c_j\}_{j=1}^{m'}$ are the distinct positive real numbers, and let $\Lambda_i' = \Lambda_i^{\mathrm{R}} + J_i'^{\mathrm{R}}$ then, its eigenvalues $\omega \in \mathbb{C}$ are the solution of

$$0 = |\Lambda_i' - \omega I| = \prod_{j=1}^{m'} \left((\lambda - \omega)^2 + c_j(1 + c_j) \right), \tag{44}$$

which is derived by the formula of determinant of the block diagonal matrix as

$$\det \begin{pmatrix} A & B \\ 0 & D \end{pmatrix} = \det(A)\det(D); \tag{45}$$

see Petersen et al. (2008). Thus, it is easy to find that ws are distinct m pairs of conjugate complex values since all c_j are distinct from each other. If the size of Λ_i^{R} is an odd number $2m'+1$, then we just need to prepare the skew-symmetric matrix

$$J_i'^{\mathrm{R}} := \left(\begin{array}{cccccc} 0 & c_1 & 0 & \dots & & 0 \\ -c_1 & 0 & 0 & \dots & & 0 \\ 0 & 0 & 0 & c_2 & \dots & 0 \\ 0 & 0 & -c_2 & 0 & \dots & 0 \\ & & \ddots & & & \\ & & & 0 & c_{m'} & 0 \\ & & & -c_{m'} & 0 & 0 \\ & & & 0 & 0 & 0 \end{array}\right). \tag{46}$$

Then, in the same way as the above discussion, we get the m' pairs of distinct complex eigenvalues and one real eigenvalue. In conclusion, we get the distinct eigenvalues by adding the skew-symmetric matrix.

Next, we consider Λ_i^{C}. This is almost similar to the case of Λ_i^{C}. If its size is an even number $2m'$, by preparing the following skew matrix

$$J_i'^{\mathrm{C}} := \left(\begin{array}{ccccc} 0 & c_1 I_2 & 0 & \dots & 0 \\ -c_1 I_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & c_2 I_2, \dots & 0 \\ 0 & 0 & -c_2 I_2 & 0, \dots & 0 \\ & & \ddots & & \\ & & & 0 & c_{m'} I_2 \\ & & & -c_{m'} I_2 & 0 \end{array}\right), \tag{47}$$

where $\{c_j\}_{j=1^{2m'}}$ are the distinct positive real numbers, and let $\Lambda_i' = \Lambda_i^{\mathrm{C}} + J_i'^{\mathrm{C}}$. Define its eigenvalues $\omega \in \mathbb{C}$, which are the solution of

$$0 = |\Lambda_i' - \omega I| = \prod_{j=1}^{m'} \left(|A - \omega I|^2 + |c_j I_2||(1+c_j)I_2|\right) = \prod_{j=1}^{m'} \left((\alpha - \omega)^2 + \beta^2 + (1+c_j)c_j\right). \tag{48}$$

Thus, it is easy to find that ws are distinct m pairs of conjugate complex values. For the odd size matrix, the argument is the same as the case of real eigenvalues.

Finally, we collect the above perturbations and define

$$J' := \begin{pmatrix} J_1'^{\mathrm{R}} & \dots & 0 \\ & \ddots & \\ 0 & & J_m'^{\mathrm{C}} \end{pmatrix}, \tag{49}$$

and $\Lambda + J'$ has d distinct eigenvalues. Since

$$V^{-1}(H + H^{1/2} J H^{1/2})V = \Lambda, \tag{50}$$

we have

$$V^{-1}(H + H^{1/2}JH^{1/2})V + J' = \Lambda + J'. \tag{51}$$

Thus

$$H + H^{1/2}JH^{1/2} + VJ'V^{-1} = V(\Lambda + J')V^{-1}. \tag{52}$$

and since $\Lambda + J'$ and $V(\Lambda + J')V^{-1}$ are similar, $H + H^{1/2}JH^{1/2} + VJ'V^{-1}$ have distinct eigenvalues.

Then if we prepare

$$J'' = H^{-1/2}VJ'V^{-1}H^{-1/2}, \tag{53}$$

$H + H^{1/2}(J + J'')H^{1/2}$ has distinct eigenvalues. Then, it is possible to generate such a random matrix $J + J''$. This means that the discriminant of the characteristic polynomial of $H + H^{1/2}(J + J'')H^{1/2}$ is not identically zero. This concludes the proof. ■

C.3. Proof of Proposition 2

Proof Since we assume that the Hessian matrix $H := \nabla^2 F(x)$ is a diagonalizable matrix, $(I + J)H$ and $H + H^{1/2}JH^{1/2}$ have the same eigenvalues since they are similar. Thus, to study the eigenvalues of $(I + J)H$, we will study those of $H + H^{1/2}JH^{1/2}$ instead.

We set $A = H + H^{1/2}JH^{1/2}$, that is, non-symmetric matrix A has $d(= 2m)$ complex eigenvalues and eigenvectors,

$$Aw_j = \mu_j w_j \Leftrightarrow A(a_j + ib_j) = (\alpha_j + i\beta_j)(a_j + ib_j). \tag{54}$$

We denote the eigenvalues and eigenvectors of H as $\{\lambda_j, v_j\}_{kj=1}^{d}$ and v_js are linearly independent. And we assume that $\lambda_1 \leq, \dots, \lambda_d$. We assume that the lengths of all eigenvectors are normalized to 1. We assume that d is even and all the eigenvalues are complex. It is straightforward to extend the proof in this section to the case when d is odd and there exists real value eigenvalues.

From the above definition, by checking the real parts and complex parts, the following relations are derived

$$Aa_j = \alpha_j a_j - \beta b_j, \tag{55}$$
$$Ab_j = \alpha_j b_j + \beta a_j. \tag{56}$$

thus, by the skew-symmetric property

$$\begin{aligned} a_j^\top A a_j + b_j^\top A b_j &= \alpha_j(\|a_j\|^2 + \|b_j\|^2) = \alpha_j \quad (57) \\ &= a_j^\top H a_j + b_j^\top H b_j, \quad (58) \end{aligned}$$

and in the third equality, we used the property

$$a_j^\top H^{1/2}JH^{1/2}a_j = b_j^\top H^{1/2}JH^{1/2}b_j = 0, \tag{59}$$

since $H^{1/2}JH^{1/2}$ is a skew symmetric matrix.

Then, we expand a_j and b_j by v_j as

$$a_k = \sum_{j=1}^{d} a_k^\top v_j v_j \tag{60}$$

$$b_k = \sum_{j=1}^{d} b_k^\top v_j v_j, \tag{61}$$

and substitute this into Eq.(58). Then, we have

$$\begin{aligned}\alpha_k &= \sum_{j=1}^{d} \lambda_j (a_k^\top v_j)^2 + \sum_{j=1}^{d} \lambda_j (b_k^\top v_j)^2 &(62)\\ &\geq \lambda_1 \sum_{j=1}^{d} ((a_k^\top v_j)^2 + (b_k^\top v_j)^2) &(63)\\ &= \lambda_1. &(64)\end{aligned}$$

This means that any real part of the eigenvalue of A is larger than λ_1 which is the smallest eigenvalue of H. Thus, if the α_1 is the smallest real part of the eigenvalue of A, that is larger than the smallest eigenvalue of H. This concludes the proof.

In the same way,

$$\begin{aligned}\alpha_k &= \sum_{j=1}^{d} \lambda_j (a_k^\top v_j)^2 + \sum_{j=1}^{d} \lambda_j (b_k^\top v_j)^2 &(65)\\ &\leq \lambda_d \sum_{j=1}^{d} ((a_k^\top v_j)^2 + (b_k^\top v_j)^2) &(66)\\ &= \lambda_d, &(67)\end{aligned}$$

which means any real part of the eigenvalues of A is smaller than the largest eigenvalue of H. Thus, if α is the largest real part of the eigenvalues of A, it is smaller than the largest eigenvalue of H.

Next, we discuss when the equality holds for α_1 and λ_1. First, we assume that eigenvalues of H are distinct. Later, we discuss if eigenvalues are not distinct. From Eq. (62), we have

$$\begin{aligned}\alpha_1 &= \sum_{j=1}^{d} \lambda_j (a_1^\top v_j)^2 + \sum_{j=1}^{d} \lambda_j (b_1^\top v_j)^2 &(68)\\ &\geq \lambda_1 \sum_{j=1}^{d} ((a_1^\top v_j)^2 + (b_1^\top v_j)^2) &(69)\\ &= \lambda_1, &(70)\end{aligned}$$

in general. To study when the equality holds, let us assume that a_1 and b_1 corresponds to $a_1 = cv_k$ and $b_1 = c'v_{k'}$ where $c^2 + c'^2 = 1$. Note that if

$$a_1, b_1 \propto v_1, \tag{71}$$

does not hold, for example $a_1 = c_1 v_1 + c_1 v_2$), where $c_1^2 + c_2^2 = c^2$ and $b_2 = c' v_1$, then, we have

$$\alpha_1 = \sum_{j=1}^{d} \lambda_j (a_1^\top v_j)^2 + \sum_{j=1}^{d} \lambda_j (b_1^\top v_j)^2 \tag{72}$$

$$= c_1^2 \lambda_1 + c_2^2 \lambda_2 + c'^2 \lambda_1 \tag{73}$$

$$> \lambda_1. \tag{74}$$

Thus, $a_1, b_1 \propto v_1$ is required to have $\alpha_1 = \lambda_1$. Moreover, a_1 and b_1 are linearly independent from the basic property of the skew matrices (Bhatia, 2013), thus we only have either i) $v_1 = a_1$ or ii) $v_1 = ib_1$ as the condition.

If i) is satisfied, then $Av_1 = (\lambda_1 \pm i\mathrm{Im}(\lambda_1))v_1$ holds. Since Av_1 is a real-valued vector, we can see that $\mathrm{Im}(\lambda_1) = 0$ holds. Thus, if i) is satisfied, $Av_1 = \lambda_1 v_1$ holds.

Next, if ii) is satisfied, then $Av_1 = A(\pm ib_1) = (\lambda_1 \pm i\mathrm{Im}(\lambda_1)) \pm ib_1$ holds. Since Av_1 is a purely imaginary-valued vector, we can see that $\mathrm{Im}(\lambda_1) = 0$ holds. Thus, if ii) is satisfied, then $Av_1 = \lambda_1 v_1$ holds. In conclusion, in both i) and ii), these are equivalent to the condition of $Av_1 = \lambda_1 v_1$.

Then, from the definition of the eigenvalue, we obtain the following relation

$$\lambda_1 v_1 = Av_1 = (H + \alpha H^{1/2} J H^{1/2}) v_1 = Hv_1 + \alpha H^{1/2} J H^{1/2} v_1 = \lambda_1 v_1 + \alpha \lambda_1 H^{1/2} J v_1. \tag{75}$$

This indicates

$$\alpha H^{1/2} J v_1 = 0. \tag{76}$$

From the definition of H, $H^{1/2}$ has an inverse matrix. By multiplying it to the above condition, the above condition is equivalent to $Jv_1 = 0$. This is the condition that $\lambda_1 = \alpha_1$ holds. The same relation can be derived for $\lambda_d = \alpha_d$.

Next, we assume that eigenvalues of H are not distinct and if the multiplicity of eigenvalue λ_1 is greater than 1. We denote the set of eigenvectors of H whose eigenvalues are 1, as V_1^0

To study when equality $\alpha_1 = \lambda_1$ holds, from the similar discussion with the case when H are distinct, we obtain the condition that

$$a_1, b_1 \in V_1^0, \tag{77}$$

Based on this, let us assume that $w_1 = ca_1 + ic'b_1$ where $c^2 + c'^2 = 1$. We consider the case $a_1 \neq b_1$. Then

$$H^{-1/2} A(ca_1 + ic'b_1) = \lambda_1^{-1/2} (\lambda_1 + i\beta_1)(ca_1 + ic'b_1)$$
$$H^{-1/2} (H + \alpha H^{1/2} J H^{1/2})(ca_1 + ic'b_1) = \lambda_1^{1/2} c(I + \alpha J) a_1 + i\lambda_1^{1/2} c'(I + \alpha J) b_1, \tag{78}$$

then we obtain the condition

$$\lambda_1 c \alpha J a_1 = -\beta_1 c' b_1, \tag{79}$$

$$\lambda_1 c' \alpha J b_1 = \beta_1 c a_1. \tag{80}$$

The same discussion can be made for α_d and λ_d.

Finally we discuss the trace bound. Note that

$$\mathrm{Tr}(H + \alpha J) = \mathrm{Tr}(H). \tag{81}$$

On the other hand

$$\mathrm{Tr}(H + \alpha J) = \sum_{i=1}^{d/2} \mathrm{Re}\lambda_i(\alpha) \geq d\mathrm{Re}\lambda_1(\alpha). \tag{82}$$

Here we assumed that d is even and only complex eigenvalues appear. The situation when d is odd or real eigenvalues exists is treated almost similar way. Then

$$\mathrm{Tr}(H) \geq d\mathrm{Re}\lambda_1(\alpha), \tag{83}$$

holds. The discussion for $\mathrm{Re}\lambda_d(\alpha)$ is almost the same. ■

C.4. Proof of Proposistion 3

In this section, we express $(I + \alpha J)H$ by H'. First, we introduce the notations. Since H' is almost surely diagonalizable, we will use the eigen decomposition in the proof. For simplicity, we assume that all the eigenvalues of H' are imaginary. Note that if the real eigenvalues are exist, following derivation can be used. First, the eigenvalues of H are expressed by $\lambda_1, \dots, \lambda_d$ and corresponding eigenvectors are $v_1, \dots, v_d$. As for H', there are $2m$ complex eigenvalues, $\mu_1 = \alpha_1 \pm i\beta_1, \dots, \mu_m = \alpha_m \pm i\beta_m$. Thus, $d = 2m$. Note that is d is odd, then there are real eigenvalues. We denote the corresponding eigenvectors as $\{w_j = a_j + ib_j\}_{j=1}^m$ for complex eigenvalues and $\{\bar{w}_j\}$ for corresponding conjugate eigenvalues.

Here we introduce the notation

$$w_j = v_j + \delta v_j, \tag{84}$$

$$\mu_j = \lambda_j + \delta\lambda_j, \tag{85}$$

Then from the definition

$$H'w_j = Hw_j + \alpha V w_j = \mu w_j = (\lambda_j + \delta\lambda_j)(v_j + \delta v_j), \tag{86}$$

where $V := H^{1/2}JH^{1/2}$. Note that V is similar to JH. Then we have

$$Hv_j + H\delta v_j + \alpha V v_j + \alpha V \delta v_j = \lambda_j v_j + \delta\lambda_j v_j + \lambda_j \delta v_j + \delta\lambda_j \delta v_j. \tag{87}$$

First, we consider the first order expansion. Thus, we consider

$$Hv_j + H\delta v_j + \alpha V v_j = \lambda_j v_j + \delta\lambda_j v_j + \lambda_j \delta v_j. \tag{88}$$

Since $\{v_1, \dots, v_k\}$ are an orthogonal basis, we expand δvs by this basis.

$$\delta v_j = \sum_{k=1}^{d} c_{jk} v_k, \tag{89}$$

where $c_{jk} = \delta v_j^\top v_k$.

By multiplying V_j to Eq.(92) from the left handside, we have

$$\lambda_j + \lambda_j v_j^\top \delta v_j + \alpha v_j^\top V v_j = \lambda_j + \delta\lambda_j + \lambda_j v_j^\top \delta v_j, \tag{90}$$

since $v_j^\top V v_j = 0$ due to the skew-symmetric property of V. Thus we have

$$\delta\lambda_j = 0, \tag{91}$$

up to the first-order expansion.

Then we substitute this into Eq.(92) and multiplying v_i where $i \neq j$, we have

$$\lambda_i c_{ji} + \alpha v_i^\top V v_j = \lambda_j c_{ji}. \tag{92}$$

Then we have

$$c_{ji} = \frac{\alpha v_i^\top V v_j}{\lambda_j - \lambda_i}. \tag{93}$$

As for c_{jj}, since $w_j^\top w_j = 1$, we have

$$c_{jj} = 0. \tag{94}$$

Then we get

$$\delta v_j = \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i. \tag{95}$$

We substitute this into Eq.(97), and multiplying $v_j^\top$, we have

$$\begin{aligned} &v_j^\top H\alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i + \alpha v_j^\top V v_j + \alpha v_j^\top V \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i \\ &= \delta\lambda_j v_j^\top v_j + \lambda_j v_j^\top \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i + \delta\lambda_j v_j^\top \alpha \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_i. \end{aligned} \tag{96}$$

Since $v_j^\top V v_j = 0$ and $v_j^\top v_i = 0$ and $v_j^\top v_j = 1$, we have

$$\alpha^2 \sum_{i\neq j}^{d} \frac{v_i^\top V v_j}{\lambda_j - \lambda_i} v_j^\top V v_i = \delta\lambda_j. \tag{97}$$

Thus, we have

$$\mu_j - \lambda_j = \alpha_j + i\beta_j - \lambda_j = -\alpha^2 \sum_{i\neq j}^{d} \frac{(v_i^\top V v_j)^2}{\lambda_j - \lambda_i}. \tag{98}$$

Thus by taking the real part, and note that $\mathrm{Re}\lambda_j(\alpha) = \alpha_j$, we have

$$\mathrm{Re}\lambda_j(\alpha) - \lambda_j = \alpha^2 \mathrm{Re} \sum_{i\neq j}^{d} \frac{(v_i^\top V v_j)^2}{\lambda_i - \lambda_j} + \mathcal{O}(\alpha^3). \tag{99}$$

Appendix D. Analysis of the continuous dynamics

In this section, first we review the linear ODE with asymmetric matrix. After that we analyze proposed continuous dynamics.

D.1. Basic examples

Given a matrix $A \in \mathbb{R}^{d\times d}$, let us consider a linear differential equations

$$\dot{x} = Ax, \tag{100}$$

where $\dot{x} = \frac{dx}{dt}$. If A is a normal symmetric matrix, we can diagonalize it and can define the exponential matrix e^A. Then given the initial condition $x(0) = x_0$, we can solve the above ODEs as

$$x(t) = e^{tA}x_0. \tag{101}$$

Our next interest is the case in which A is not symmetric, that is, which can have complex eigenvalues and eigenvectors. Given a complex number $v = x + iy$, we denote its complex conjugate by $\bar{v} = x - iy$. Then

$$Av = \begin{pmatrix} a_{11} & \dots & a_{1d} \\ \vdots & & \vdots \\ a_{d1} & \dots & a_{dd} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_d \end{pmatrix} = A(x + iy) = Ax + iAy. \tag{102}$$

This implies

$$\bar{Av} = A\bar{v}. \tag{103}$$

Thus, if we assume that λ is a complex eigenvalue of A and v is the corresponding complex eigenvector, $Av = \lambda v$,

$$A\bar{v} = \bar{\lambda}\bar{v}, \tag{104}$$

holds. Furthermore, the Euler's formula is useful to consider the matrix exponential,

$$e^{\lambda} = e^{\alpha + i\beta} = e^{\alpha}(\cos\beta + i\sin\beta). \tag{105}$$

Having these relations in mind, let us consider following case,

$$A = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix}. \tag{106}$$

Then, its eigenvalues are $\lambda = \alpha + i\beta$ and $\bar{\lambda} = \alpha - i\beta$ and corresponding eigenvectors are $v = (1, i)^\top = (1, 0)^\top + i(0, 1)^\top$ and $\bar{v} = (1, -i)^\top = (1, 0)^\top - i(0, 1)^\top$. Next, let us consider the matrix exponential e^{tA}. Note that

$$A = \alpha I + \beta J, \quad I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad J = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}. \tag{107}$$

Then J satisfies $J^2 = -I$, $J^3 = -J$, $J^4 = I, \dots$, thus

$$\begin{aligned} e^{tJ} &= I + tJ + \frac{t^2}{2!}J^2 + \frac{t^3}{3!}J^3 + \dots && (108)\\ &= I + tJ - \frac{t^2}{2!}I - \frac{t^3}{3!}J + \dots && (109)\\ &= \left(1 - \frac{t^2}{2!} + \frac{t^4}{4!} - \dots\right) I + \left(t - \frac{t^3}{3!} + \frac{t^5}{5!} - \dots\right) J && (110)\\ &= \cos(t)I + \sin(t)J && (111)\\ &= \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}. && (112) \end{aligned}$$

(Eigenvalues of J is $\pm i$, thus above calculation is done via Euler's formula.) Thus,

$$e^{tA} = e^{t\alpha I + t\beta J} = e^{\alpha t} \begin{pmatrix} \cos(\beta t) & \sin(\beta t) \\ -\sin(\beta t) & \cos(\beta t) \end{pmatrix}. \tag{113}$$

D.2. General case

Based on the previous simple example, let us consider $d \times d$ skew-symmetric matrix. Furthermore, assume that $d \times d$ skew-symmetric matrix A has l real eigenvalues $\lambda_1, \dots, \lambda_l$ and $2m$ complex eigenvalues, $\mu_1 = \alpha_1 \pm i\beta_1, \dots, \mu_m = \alpha_m \pm i\beta_m$. Thus, $d = l + 2m$. We denote the corresponding eigenvectors as $\{v_j\}_{j=1}^l$ for real eigenvalues and $\{w_j = a_j + ib_j\}_{j=1}^m$ for complex eigenvalues $\{\mu_j\}_{j=1}^m$ and $\{\bar{w}_j\}$ for corresponding conjugate eigenvalues.

Then, let us define a $d \times d$ matrix V as

$$V = [v_1, \dots, v_l, a_1, b_1, \dots, a_m, b_m]. \tag{114}$$

Then, we can decompose A into a block diagonal matrix known as the Jordan canonical form Golub and Van Loan (2012);

$$AV = VD \tag{115}$$

$$D := \begin{pmatrix} \lambda_1 &&&&&&& \\ & \ddots &&&&&& \\ && \lambda_l &&&&& \\ &&& \alpha_1 & \beta_1 &&& \\ &&& -\beta_1 & \alpha_1 &&& \\ &&&&& \ddots && \\ &&&&&& \alpha_m & \beta_m \\ &&&&&& -\beta_m & \alpha_m \end{pmatrix}. \tag{116}$$

Then, we can calculate the matrix exponential by $A = VDV^{-1}$

$$e^{tA} = e^{VDV^{-1}} = Ve^{tD}V^{-1} \tag{117}$$

$$e^{tD} = \begin{pmatrix} e^{t\lambda_1} & & & & & & & \\ & \ddots & & & & & & \\ & & e^{t\lambda_l} & & & & & \\ & & & e^{t\alpha_1}\cos(\beta_1 t) & e^{t\alpha_1}\sin(\beta_1 t) & & & \\ & & & -e^{t\alpha_1}\sin(\beta_1 t) & e^{t\alpha_1}\cos(\beta_1 t) & & & \\ & & & & & \ddots & & \\ & & & & & & e^{t\alpha_m}\cos(\beta_m t) & e^{t\alpha_m}\sin(\beta_m t) \\ & & & & & & -e^{t\alpha_m}\sin(\beta_m t) & e^{t\alpha_m}\cos(\beta_m t) \end{pmatrix} \tag{118}$$

Based on this, let us consider the problem $\dot{x} = Ax$. To solve this, following theorem is useful;

Lemma 3 *(Linear independence)*
i)Vectors $(v_1, \ldots, v_l, a_1, b_1, \ldots, a_m, b_m)$ are linearly independent.
ii)When given a real vector $x \in \mathbb{R}^d$, it is written as a form

$$\begin{aligned} x &= c_1 v_1 + \cdots + c_l v_l + \frac{1}{2}c_1' w_1 + \frac{1}{2}\bar{c}_1' \bar{w}_1 + \cdots + \frac{1}{2}c_m' w_m + \frac{1}{2}\bar{c}_m' \bar{w}_m & (119) \\ &= c_1 v_1 + \cdots + c_l v_l + \mathrm{Re}\left(c_1' w_1 + \cdots + c_m' w_m\right), & (120) \end{aligned}$$

where $c_j{}_{j=1}^{l}$ are real values and $c_j'{}_{j=1}^{m}$ are complex values.

Proof i) is a direct consequence of the diagonalizable property of A. As for ii), since $(v_1, \ldots, v_l, a_1, b_1, \ldots, a_m, b_m)$ span d-dimensional linear space, thus we use this set as a basis and apply the eigen decomposition. ■

With this linear dependency in mind, we can solve $\dot{x} = Ax$ as

Proposition 8 *(Linear ODEs by a skew-symmtric matrix)*
The initial condition is given as $x(0) = c_1 v_1 + \cdots + c_l v_l + \mathrm{Re}\,(c_1' w_1 + \cdots + c_m' w_m)$, the solution of $\dot{x} = Ax$ is given by

$$\begin{aligned} x(t) &= c_1 e^{\lambda_1 t} v_1 + \cdots + c_l e^{\lambda_l t} v_l + \mathrm{Re}\left(c_1' e^{\mu_1 t} w_1 + \cdots + c_m' e^{\mu_m t} w_m\right) & (121) \\ &= c_1 e^{\lambda_1 t} v_1 + \cdots + c_l e^{\lambda_l t} v_l + e^{\alpha_1 t}\mathrm{Re}\left(c_1' e^{i\beta_1 t} w_1\right) + \cdots + e^{\alpha_m t}\mathrm{Re}\left(c_m' e^{i\beta_m t} w_m\right). & (122) \end{aligned}$$

Proof We can easily confirm each element of a set $(e^{\lambda_1 t} v_1, \ldots, e^{\lambda_l t} v_l, e^{\mu_1 t} w_1, e^{\bar{\mu}_1 t} \bar{w}_1, \ldots, e^{\mu_m t} w_m, e^{\bar{\mu}_m t} \bar{w}_m)$ satisfies the given ODE. Thus they are the solutions. Since they are linearly independent, we use them as a basis for the decomposition, then we get the proposition. ■

As we can see that the real parts of the eigenvalues determine the exponential convergence or divergence of the solution.

Note: In the above analysis, we assumed that A is diagonalizable. This is a strong assumption in general because A is neither symmetric nor skew-symmetric or normal. To show the diagonalization, the algebraic and geometric multiplicities of eigenvalues must coincide. To show this, one strategy is to show that all the eigenvalues are distinct. As we had seen in the previous appendix, this is not difficult for a random matrix in general.

D.3. Proof of Proposition 4

Proof Recall that our proposed dynamics is given as

$$\frac{dx(t)}{dt} = -(I + \alpha J)\nabla F(x). \tag{123}$$

Let us define a functional as

$$\mathcal{L} = (x(t) - x^*)^\top (x(t) - x^*). \tag{124}$$

Then

$$\frac{d\mathcal{L}}{dt} = -2(x(t) - x^*)^\top (I + \alpha J) \left(\nabla F(x(t)) - \nabla F(x^*)\right). \tag{125}$$

Then, from the taylor expansion and expressing its residual by integral, we have

$$\begin{aligned} \frac{d\mathcal{L}}{dt} &= -2(x(t) - x^*)^\top \left(\int_0^1 (I + \alpha J) H(x^* + \tau(x(t) - x^*)) \left(x(t) - x^*\right) dt \right) \\ &= -2 \left(\int_0^1 (x(t) - x^*)^\top (I + \alpha J) H(x^* + \tau(x(t) - x^*)) \left(x(t) - x^*\right) dt \right). \end{aligned} \tag{126}$$

Since $(I + \alpha J)H$ is almost surely diagonalizable, we will analyze the dynamics based on the eigen decomposition. Followings are the notation of the eigenvalues and vectors: l real eigenvalues $\lambda_1, \ldots, \lambda_l$ and $2m$ complex eigenvalues, $\mu_1 = \alpha_1 \pm i\beta_1, \ldots, \mu_m = \alpha_m \pm i\beta_m$. Thus, $d = l + 2m$. We denote the corresponding eigenvectors as $\{v_j\}_{j=1}^l$ for real eigenvalues and $\{w_j = a_j + ib_j\}_{j=1}^m$ for complex eigenvalues $\{\mu_j\}_{j=1}^m$ and $\{\bar{w}_j\}$ for corresponding conjugate eigenvalues.

Then, let us define a $d \times d$ matrix V as

$$V = [v_1, \ldots, v_l, a_1, b_1, \ldots, a_m, b_m]. \tag{127}$$

Then, we can decompose $(I + \alpha J)H(\xi)$ into a block diagonal matrix known as the Jordan canonical form Golub and Van Loan (2012);

$$(I + \alpha J)H(x^* + \tau(x(t) - x^*)) = VDV^{-1}, \tag{128}$$

where

$$\begin{aligned}
D :=& \begin{pmatrix}
\lambda_1 & & & & & & & \\
& \ddots & & & & & & \\
& & \lambda_l & & & & & \\
& & & \alpha_1 & \beta_1 & & & \\
& & & -\beta_1 & \alpha_1 & & & \\
& & & & & \ddots & & \\
& & & & & & \alpha_m & \beta_m \\
& & & & & & -\beta_m & \alpha_m
\end{pmatrix} \\
=& \begin{pmatrix}
\lambda_1 & & & & & & & \\
& \ddots & & & & & & \\
& & \lambda_l & & & & & \\
& & & \alpha_1 & 0 & & & \\
& & & 0 & \alpha_1 & & & \\
& & & & & \ddots & & \\
& & & & & & \alpha_m & 0 \\
& & & & & & 0 & \alpha_m
\end{pmatrix} + \begin{pmatrix}
0 & & & & & & & \\
& \ddots & & & & & & \\
& & 0 & & & & & \\
& & & 0 & \beta_1 & & & \\
& & & -\beta_1 & 0 & & & \\
& & & & & \ddots & & \\
& & & & & & 0 & \beta_m \\
& & & & & & -\beta_m & 0
\end{pmatrix} \\
:=& A + B. \qquad (129)
\end{aligned}$$

Then,

$$\begin{aligned}
\frac{d\mathcal{L}}{dt} &= -2\left(\int_0^1 VDV^{-1}\left(x(t)-x^*\right)dt\right) \\
&= -2\left(\int_0^1 V(A+B)V^{-1}\left(x(t)-x^*\right)dt\right) \\
&= -2\left(\int_0^1 VAV^{-1}\left(x(t)-x^*\right)dt\right) \\
&\leq -2\mathrm{Re}(\lambda_1^\alpha(x(t)))\|x(t)-x^*\|^2 \\
&\leq -2m'\|x(t)-x^*\|^2 \qquad (130)
\end{aligned}$$

Then, from the Gronwall inequality, we get the proposition. ■

Appendix E. Analysis of the Euler discretization (Proof of Proposition 5)

Proof From the discretized dynamics Eq.(18), subtract x^* from both sides and set it as $r_k = x_k - x^*$, we have

$$\|r_{k+1}\| = \|r_k - \eta(I + \alpha J)\nabla F(x_k)\|. \tag{131}$$

We define $h(x) := x - \eta(I + \alpha J)\left(\nabla F(x)\right)$, then above equation can be expressed as

$$\|r_{k+1}\| = \|h(x_k) - h(x^*)\|. \tag{132}$$

Then, we apply the mean-value theorem. There exists a point $\xi_k = (1-t)x_k + tx^*,\ t \in [0,1) \subset \mathbb{R}$ (we express this $\xi_k \in [x_k, x^*) \subset \mathbb{R}^d$ for simplycity), such that

$$\|r_{k+1}\| \leq \|\left(I - \eta(I + \alpha J)\nabla^2 F(\xi_k)\right)\|\|r_k\|. \tag{133}$$

Here, the operator norm is used. Given a matrix M, it is defined as,

$$\|M\| := \sup_{x \neq 0} \frac{\|Mx\|}{\|x\|}, \tag{134}$$

and it is also characterized by the largest singular value $s(M)$;

$$\|M\| = \|M^\dagger M\|^{1/2} = s(M), \tag{135}$$

To bound $\|\left(I - \eta(I + \alpha J)H\right)r_k\|$, we evaluate the singular value of $H' = I - \eta(I + \alpha J)H$. We use the same notation in Appendix D.3. Note that from the Jordan canonical form in Appendix D.3, we have,

$$H' = I - \eta VDV^{-1} = I - \eta VAV^{-1} - \eta VBV^{-1}, \tag{136}$$

holds (A and B are the diagonal and skew matrices). We define $P = I - \eta VAV^{-1}$ and $Q = -\eta VBV^{-1}$. From III.6.4 in Bhatia (2013), the largest singular value of H' (we denote it $s(H')$) is upper bounded by the sum of the largest eigenvalues of P and Q (we denote them $s(P)$ and $s(Q)$).

$$s(H') \leq s(P) + s(Q). \tag{137}$$

$s(P)$ depends on η. Let $\lambda_S := \mathrm{Re}\lambda_1^\alpha(\xi_k) = \min\{\lambda_1, \ldots, \lambda_l, \alpha_1, \ldots, \alpha_m\}$ and $\lambda_L := \mathrm{Re}\lambda_d^\alpha(\xi_k) = \max\{\lambda_1, \ldots, \lambda_l, \alpha_1, \ldots, \alpha_m\}$. Then form Figure 5 where the vertical line is $s(P)$, and we can upperbound $s(P)$ by the bold line in Figure 5. The bold line can be analytically calculated by the definition of $s(P)$.

About $s(Q)$, it is calculated by the definition of B

$$s(Q) = \eta \max_l \beta_l, \tag{138}$$

where β is the imaginary part of the eigenvalues (see Appendix D.3). Then, from Theorem 8.2.1 in Golub and Van Loan (2012) (Gershgorin Circle Theorem), we can upper bound this as follows,

$$\max_l \beta_l \leq \alpha \max_j \left(\sum_{i=1}^{d} |J_{ij}|\right). \tag{139}$$

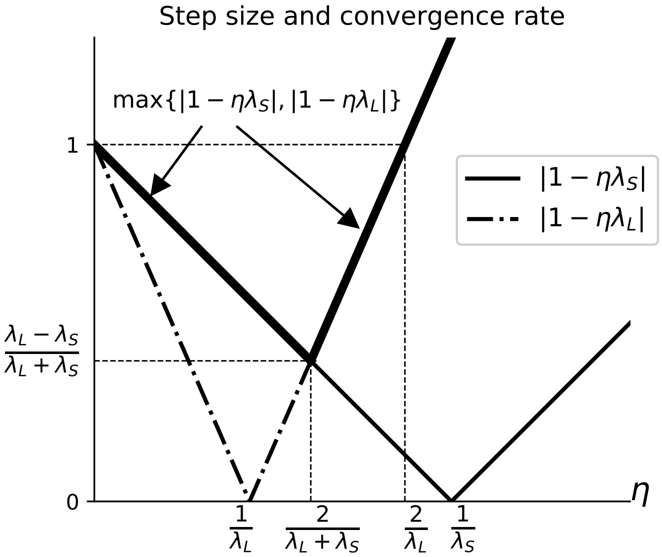


Figure 5: Conceptual figure of the relation between the convergence rate and the step size

Let us define

$$r := \alpha \max_j \left(\sum_{i=1}^{d} |J_{ij}| \right), \tag{140}$$

then, we obtain

$$s(Q) \leq \eta r. \tag{141}$$

Based on these facts, we obtain

$$s(H') \leq s(P) + s(Q). \tag{142}$$

We upper bound $s(P) + s(Q)$ using the relation shown in Figure 6 where the vertical line is $s(P) + s(Q)$. From the figure we obtain the optimal step size as $\eta = \frac{2}{\lambda_L + \lambda_S}$.

Then, by substituting the optimal step size, we obtain

$$s(H') \leq \frac{\lambda_L - \lambda_S + 2r}{\lambda_L + \lambda_S}. \tag{143}$$

Using this upper-bound of the singular value, we can upper bound the residual of r_{k+1} as

$$\|r_{k+1}\| \leq \left(1 - \frac{(\lambda_S - s)}{\lambda_L}\right) \|r_k\| \leq \left(1 - \frac{(\lambda_S - r)}{\lambda_L}\right) \|r_k\|. \tag{144}$$

Finally from definition, $m' \leq \lambda_S$ and $\lambda_L \leq M'$, we obtain

$$\|r_{k+1}\| \leq \left(1 - \frac{(m' - r)}{M'}\right) \|r_k\|. \tag{145}$$

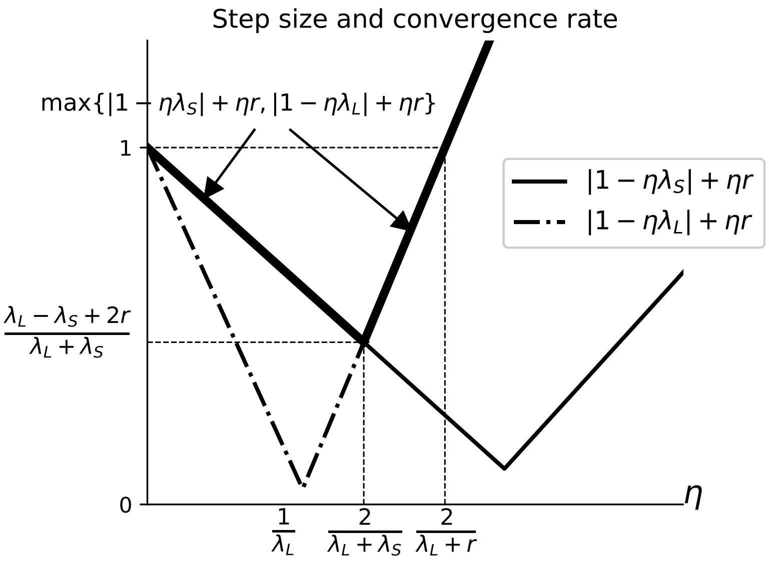


Figure 6: Conceptual figure of the relation between the convergence rate and the step size

Thus, combined with the above bound, we get

$$\|r_k\| \leq e^{-\frac{m'-r}{M'}k}\|r_0\|. \tag{146}$$

Also, from Figure 6, if $\eta \leq \frac{2}{\mathrm{Re}\lambda_d(\alpha)+s}$ and $\mathrm{Re}\lambda_1(\alpha) > r$ is satisfied, then the algorithm converges at r_k. Thus the condition is $\eta \leq \frac{2}{M'+r}$ and $m' > r$ holds.

■

Appendix F. Analysis of the Euler-leapfrog discretization

F.1. Proof of Proposition 6

Proof First, we express the Euler-leapfrog method in a matrix form. In the same way as the proof of Proposition 5, we first change the gradient to the Hessian matrix. First, subtract $(x^*, y^*)^\top$, where $y^* = 0$, from the update equation, then we obtain

$$\begin{aligned} \begin{pmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{pmatrix} &= \begin{pmatrix} x_k - \frac{\eta\alpha}{c} J y_k - \eta \nabla F(x_k - \frac{\eta\alpha}{c} J y_k) - x^* \\ (1 - \eta c^{-1})(y_k - \eta\alpha J \nabla F(x_k - \frac{\eta\alpha}{c} J y_k)) - y^* \end{pmatrix} \\ &= \begin{pmatrix} h_1(x_k, y_k) - x^* \\ h_2(x_k, y_k) - y^* \end{pmatrix} \\ &= \begin{pmatrix} h_1(x_k, y_k) - h_1(x^*, y^*) \\ h_2(x_k, y_k) - h_2(x^*, y^*) \end{pmatrix}, \end{aligned} \tag{147}$$

where we defined the function $h_1, h_2 : \mathbb{R}^d \to \mathbb{R}^d$, as

$$h_1(x_k, y_k) := x_k - \frac{\eta\alpha}{c} J y_k - \eta \nabla F(x_k - \frac{\eta\alpha}{c} J y_k) \tag{148}$$

$$h_2(x_k, y_k) := (1 - \eta c^{-1})(y_k - \eta\alpha J \nabla F(x_k - \frac{\eta\alpha}{c} J y_k)), \tag{149}$$

and from the definition

$$h_1(x^*, y^*) = x^* \tag{150}$$

$$h_2(x^*, y^*) = y^* = 0. \tag{151}$$

Then by applying the mean value theorem

$$\left\| \begin{pmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{pmatrix} \right\| \leq \left\| \begin{pmatrix} \nabla_x h_1(\xi_k, \zeta_k) & \nabla_y h_1(\xi_k, \zeta_k) \\ \nabla_x h_2(\xi_k, \zeta_k) & \nabla_y h_2(\xi_k, \zeta_k) \end{pmatrix} \right\| \left\| \begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} \right\|, \tag{152}$$

where $\xi_k = (1-\beta)x_k + \beta x^*,\ \beta \in [0,1) \subset \mathbb{R}$ and $\zeta_k = (1-\beta')y_k + \beta' y^*,\ \beta' \in [0,1) \subset \mathbb{R}$ (we express them $\xi_k \in [x_k, x^*)$ and $\zeta_k \in [y_k, y^*)$ for simplycity), are some constants in $\mathbb{R}^d$ which is specified by the mean-value theorem. And the definition of the matrix norm is that given a matrix A,

$$\|A\| = \sup_x \frac{\|Ax\|}{\|x\|}, \tag{153}$$

and

$$\|A\| = \|A^\dagger A\|^{1/2} = s_d(A), \tag{154}$$

where $s_d(A)$ is the largest singular value.

By calculating the Hessian matrix, we have

$$\left\| \begin{pmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{pmatrix} \right\| \leq \left\| \underbrace{\begin{pmatrix} I - \eta H(\xi_{k+\frac{1}{2}}) & 0 \\ 0 & (1 - \eta c^{-1}) I \end{pmatrix}}_{=\tilde{H}(\eta)} \underbrace{\begin{pmatrix} I & -\eta\alpha c^{-1} J \\ -\eta\alpha J H(\xi_{k+\frac{1}{2}}) & I + c^{-1}\eta^2\alpha^2 J H(\xi_{k+\frac{1}{2}}) J \end{pmatrix}}_{=L(\eta,\alpha,J)} \right\| \left\| \begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} \right\|, \tag{155}$$

where $\xi_{k+\frac{1}{2}}$ is specified by $\xi_k - \frac{\eta\alpha}{c}J\zeta_k$. Since ξ_k and ζ_k are specified by the mean-value theorem, $\xi_{k+\frac{1}{2}}$ is also a some constant in $\mathbb{R}^d$ which is specified by the mean-value theorem.

Next we analyze $\tilde{H}L$. From the definition of the matrix norm, we need to evaluate the singular value of $\tilde{H}L$. From the submultiplicativity of the matrix norm, that is, $\|\tilde{H}L\| \leq \|\tilde{H}\|\|L\|$ holds. Thus, we need to evaluate the norm of $\tilde{H}$ and L separately. Later, we consider when $\|\tilde{H}L\| = \|\tilde{H}\|\|L\|$ holds.

We first analyze L. We evaluate its eigenvalues. We denote its eigenvalue as l. It is derived by solving the following characteristic equation:

$$0 = \det\left(lI - L\right). \tag{156}$$

We use the fourmula of determinant of the block diagonal matrix as

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det A \det(D - BA^{-1}C), \tag{157}$$

see Petersen et al. (2008). Then,

$$\begin{aligned} 0 &= \det\left(lI - L\right) \\ &= \det\left((1-l)((1-l)I + \eta^2\alpha^2c^{-1}JHJ) - c^{-1}\eta^2\alpha^2JHJ\right) \\ &= \det\left(l^2I - l(2I + \eta^2\alpha^2c^{-1}JHJ) + I\right). \end{aligned} \tag{158}$$

Note that since $H^{1/2}J = H^{1/4}(H^{1/4}JH^{1/4})H^{-1/4}$, $H^{1/2}J$, $H^{1/4}JH^{1/4}$ are similar, thus they have common eigenvalues. Since $H^{1/4}JH^{1/4}$ is a skew-symmetric matrix, it has purely imaginary eigenvalues or 0 as eigenvalues. Thus let us denote the eigenvalues of $H^{1/2}J$ as $\pm i\omega_i$, where $\omega_i \geq 0$. Let us define the unitary matrix U (since a skew-symmetric is normal, thus it is diagonalizable by a unitary matrix), which is the set of eigenvectors of $H^{1/2}J$. Then

$$\begin{aligned} 0 &= \det\left(l^2I - l(2I + \eta^2\alpha^2c^{-1}JHJ) + I\right) \\ &= \det\left(U^\dagger\left(l^2I - l(2I + \eta^2\alpha^2c^{-1}JHJ) + I\right)U\right) \\ &= \prod_i\left(l^2 - l(2 - \eta^2\alpha^2c^{-1}\omega_i^2) + 1\right). \end{aligned} \tag{159}$$

Thus we get

$$l_i = \frac{2 - \eta^2\alpha^2c^{-1}\omega_i^2}{2} \pm \frac{1}{2}\sqrt{(2 - \eta^2\alpha^2c^{-1}\omega_i^2)^2 - 4}. \tag{160}$$

If $(2 - \eta^2\alpha^2c^{-1}\omega_i^2)^2 - 4 \leq 0$, then

$$l_i = \frac{2 - \eta^2\alpha^2c^{-1}\omega_i^2}{2} \pm i\frac{1}{2}\sqrt{4 - (2 - \eta^2\alpha^2c^{-1}\omega_i^2)^2}, \tag{161}$$

and

$$\|l_i\| = 1 \tag{162}$$

holds. The condition

$$(2-\eta^2\alpha^2 c^{-1}\omega_i^2)^2 - 4 \leq 0 \tag{163}$$

can also be expressed as

$$\eta^2\alpha^2 c^{-1}\omega_i^2 \leq 4, \tag{164}$$

and if we express the largest singular value of J as s_d, from Bhatia (2013), $\max_i \omega_i^2 \leq s_d^2 M$. Thus, the above condition is satisfied if

$$\eta^2\alpha^2 s_d^2 c^{-1} M \leq 4 \tag{165}$$

holds. Thus, if we set α sufficeintly small, this conditin will be satisfied. Then

$$\begin{aligned}\left\|\begin{pmatrix} x_{k+1}-x^* \\ y_{k+1}-y^* \end{pmatrix}\right\|_2 &= \left\|\tilde{H}(\eta)L(\eta,\alpha,J)\begin{pmatrix} x_k-x^* \\ y_k-y^* \end{pmatrix}\right\| \\ &\leq \|\tilde{H}\|\|L(\eta,\alpha,J)\|\left\|\begin{pmatrix} x_k-x^* \\ y_k-y^* \end{pmatrix}\right\| \\ &\leq \|\tilde{H}\|\left\|\begin{pmatrix} x_k-x^* \\ y_k-y^* \end{pmatrix}\right\|, \end{aligned} \tag{166}$$

where we used the submultiplicativity of the matrix norm. Next we consider $\|\tilde{H}\|$. We treat it as the direct sum:

$$\tilde{H} = (I-\eta c^{-1}I) \oplus (I-\eta H). \tag{167}$$

Thus from Bhatia (2013), we have

$$\|\tilde{H}\| = \max\{\|(I-\eta c^{-1}I)\|, \|(I-\eta H)\|\} \tag{168}$$

Thus, for $\|(I-\eta H)\|$, from the analysis of the GD for strongly convex function, setting $\eta = \frac{2}{m+M}$ is the optimal and it is bounded as $\|(I-\eta H)\| \leq 1-\frac{2m}{m+M}$. Thus, if we set $M \geq c^{-1} \geq m$,

$$\|\tilde{H}\| \leq 1-\frac{2m}{m+M} \tag{169}$$

holds. Thus,

$$\begin{aligned}\left\|\begin{pmatrix} x_{k+1}-x^* \\ y_{k+1}-y^* \end{pmatrix}\right\|_2 &\leq \left(1-\frac{2m}{m+M}\right)\left\|\begin{pmatrix} x_k-x^* \\ y_k-y^* \end{pmatrix}\right\|_2 \\ &\leq e^{-\frac{m}{M}k}\left\|\begin{pmatrix} x_0-x^* \\ y_0-y^* \end{pmatrix}\right\|_2. \end{aligned} \tag{170}$$

For the convergence,

$$\|1-\eta c^{-1}\| < 1 \tag{171}$$

must be satisfied and this is equivalent to $0 < c^{-1} \leq \frac{2}{\eta}$.

Next, we study the condition for α. As we confirmed, following condition must hold

$$\eta^2\alpha^2 s_d^2 c^{-1} M \leq 4. \tag{172}$$

Then,

$$\alpha^2 \leq 4c(\eta^2 M s_d^2)^{-1} \tag{173}$$

is required for the convergence.

Equality condition: Next, we consider when the equality holds for the submultiplicativity, that is $\|\tilde{H}L\| = \|\tilde{H}\|\|L\|$. If condition is satisfied then we have

$$\left\| \begin{pmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{pmatrix} \right\|_2 \leq \left(1 - \frac{2m}{m+M}\right) \left\| \begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} \right\|_2 \tag{174}$$

and if $\|\tilde{H}L\| < \|\tilde{H}\|\|L\|$ is satisfied then we have

$$\left\| \begin{pmatrix} x_{k+1} - x^* \\ y_{k+1} - y^* \end{pmatrix} \right\|_2 \leq s \left\| \begin{pmatrix} x_k - x^* \\ y_k - y^* \end{pmatrix} \right\|_2 . \tag{175}$$

where s the largest singular value of $\|\tilde{H}L\|$ and satisfies $s < 1 - \frac{2m}{m+M}$. Thus, if $\|\tilde{H}L\| = \|\tilde{H}\|\|L\|$ is not satisfied, then we have the faster convergence. Since $\|L\| = 1$ as we had proved so far, this condition is equivalent to $\|\tilde{H}L\| = \|\tilde{H}\|$.

From the definition of the matrix norm, we focus on the largest singular value of $\tilde{H}L$ and $\tilde{H}$. Note that the singular value of L is one as we proved so far. Also, note that eigenvalues and eigenvectors of $\tilde{H}$ are real values, and those of $\tilde{H}L$ can be complex values.

Let us express a eigenvector of L as $w = a + ib$ and define the corresponding eigenvalue as μ, of which norm is $|\mu| = 1$. We also express the pairs of eigenvalues and vectors of $\tilde{H}$ as $\{(v_k, \lambda_k)\}_{k=1}^d$ and $\lambda_1 \leq \cdots \leq \lambda_d$. Assume that ws and $\{v_k\}$s are normalized to 1.

If a, b corresponds to some $\{v_k\}$s, that is

$$w = a + ib = c_1 v_k + i c_2 v_{k'}, \tag{176}$$

where c_1 and c_2 are real values and $c_1^2 + c_2^2 = 1$ holds and since a and b is always linearly independent thus $k \neq k'$. Here we assumed that the eigenvalues of $\tilde{H}$ are distinct. Then

$$\|\tilde{H}Lw\| = |\mu| \|\tilde{H}(c_1 v_k + i c_2 v_{k'})\| = \|c_1 \lambda_k v_k + i c_2 \lambda_{k'} v_k'\| = \sqrt{c_1^2 \lambda_k^2 + c_2^2 \lambda_{k'}^2}. \tag{177}$$

If $\lambda_k^2 \geq \lambda_{k'}^2$, then

$$\|\tilde{H}Jw\| = \sqrt{c_1^2 \lambda_k^2 + c_2^2 \lambda_{k'}^2} \leq |\lambda_{k'}| \leq \lambda_d = \|\tilde{H}\|. \tag{178}$$

From above discussion, if the equality in Eq.(178) holds, $\|\tilde{H}J\| = \|H\|$ is satisfied. This means that if $w = v_d$, then $\|HJw\| = |\lambda_d| = \|H\|$ holds. In the same way, if $w = iv_d$, then $\|HJw\| = |\lambda_d| = \|H\|$ holds. If a and b does not corresponds to vs, $\|HJ\| < \|H\|\|J\|$ holds. The above discussion can be applied when there is a multiplicity for eigenvalues. Thus if the eigenvector w is v_d or $\pm iv_d$, the equality can be satisfied. Here $w = v_d$ means that the eigenvalue of L is real, that means, $l_i = \pm 1$ and $\eta^2 \alpha^2 c^{-1} w_i^2 = 0$ or 4. $w = \pm iv_d$ means that the eigenvalue of L is real thus $l_i = \pm 1$.

In the above, if $\lambda_1 = \lambda_2 = \cdots = \lambda_d$, then for any $k \neq k'$, we expand a, b by v

$$a = \sum_{j=1}^{d} a^\top v_j v_j \tag{179}$$

$$b = \sum_{j=1}^{d} b^\top v_j v_j . \tag{180}$$

Then we have

$$\|HLw\| = \|HJ(\sum_{j=1}^{d} a^\top v_j v_j + i\sum_{j=1}^{d} b^\top v_j v_j)\| = |\mu||\lambda_d|\|\sum_{j=1}^{d} a^\top v_j v_j + i\sum_{j=1}^{d} b^\top v_j v_j\| = |\lambda_d|. \tag{181}$$

Thus, if all the eigenvalues of $\tilde{H}$ are the same, the equality always holds. This condition is equivalent to $m = M$.

Thus, $\|\tilde{H}L\| = \|\tilde{H}\|$ holds if $m = M$ or when $m \neq M$, the eigenvalue of L is ± 1 and its eigenvector is corresponds to v_d or $\pm i v_d$.

In conclusion, if

$$\ker(\tilde{H} - \lambda_d I) \cap (\ker(L - I) \cup \ker(L + I)) = \{0\} \tag{182}$$

is satisfied, and $m \neq M$ is satisfied, $\|\tilde{H}L\| < \|\tilde{H}\|\|L\|$ holds.

Next, we simply the condition

$$\ker(\tilde{H} - \lambda_d I) \cap (\ker(L - I) \cup \ker(L + I)) = \{0\}. \tag{183}$$

From the definition, if $\ker(L - I) = \{0\}$ and $\ker(L + I) = \{0\}$ holds, then the above condition will be satisfied.

To investigate $\ker(L-I)($ or $\ker(L+I))$, we study the singular values of $L-I($ or $(L+I))$. This is because if all the singular values are larger than 0, then $\ker(L-I) = \{0\}($ or $\ker(L-I) = \{0\})$. We first discuss $L - I$. The discussion for $L + I$ is the same.

For that purpose, first, let us decompose $L - I$ as

$$\begin{aligned} L - I &= \begin{pmatrix} 0 & -\eta\alpha c^{-1} J \\ -\eta\alpha J H(\xi_k) & \eta^2\alpha^2 J H(\xi_k) J \end{pmatrix} \\ &= \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & \eta^2\alpha^2 J H(\xi_k) J \end{pmatrix}}_{A} + \underbrace{\begin{pmatrix} J & 0 \\ 0 & J \end{pmatrix}\begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}\begin{pmatrix} -\eta\alpha H(\xi_k) & 0 \\ 0 & -\eta\alpha c^{-1} I \end{pmatrix}}_{B}, \end{aligned} \tag{184}$$

where A is symmetric and B is similar to a skew-symmetric (Note that JH is not a skew symmetric, but it is similar to $H^{1/2}JH^{1/2}$. Thus they have common eigenvalues and eigenvectors. Thus JH has purely imaginary eigenvalues.). Thus, the eigenvalues of A are real and B are purely imaginary.

Let $s(B)$ and $s(A)$ denote the smallest singular values of B and A. From III.13 of Bhatia (2013), the smallest singular value of $L - I$ is lower bounded by $\sqrt{s(B)^2 + s(A)^2}$. Thus if $\sqrt{s(B)^2 + s(A)^2} > 0$ holds, $\ker(L - I) = \{0\}$. From the definition, $s(A) = 0$. Thus we need to study when $s(B) > 0$ is satisfied.

Let us decompose B by $B = \tilde{L}\tilde{B}$ where

$$\tilde{B} = \begin{pmatrix} 0 & I \\ I & 0 \end{pmatrix}\begin{pmatrix} -\eta\alpha H(\xi_k) & 0 \\ 0 & -\eta\alpha c^{-1} I \end{pmatrix}, \tag{185}$$

and

$$\tilde{L} = \begin{pmatrix} J & 0 \\ 0 & J \end{pmatrix}. \tag{186}$$

Suppose $s(\tilde{L})$ and $s(\tilde{B})$ denote the smallest singular values of $\tilde{L}$ and $\tilde{B}$. Then from III.6.14 of Bhatia (2013)),

$$s(B) \geq s(\tilde{L})s(\tilde{B}). \tag{187}$$

holds. Thus, if

$$s(\tilde{L})s(\tilde{B}) > 0, \tag{188}$$

holds, $s(B) > 0$ will be satisfied. From the definition of $\tilde{B}$, it is clear that B does not have 0 as eigenvalues. Thus, $s(\tilde{B}) > 0$ holds. Then if

$$s(\tilde{L}) > 0 \tag{189}$$

holds, $s(B) > 0$ will be satisfied. From the definition of the singular value, if $\ker\tilde{L} = \{0\}$, $s(\tilde{L}) > 0$ holds.

Therefore, in conclusion, if $\ker\tilde{L} = \{0\}$, then $s(B) > 0$ holds. This indicates $\ker(L - I) = \{0\}$. Then since $\tilde{L} = J \oplus J$, $\ker\tilde{L} = \{0\}$ is equivalent to $\ker J = \{0\}$. ■

F.2. Derivation of Algorithm 1

Here we present the algorithm to tune J and α in the ELF.

Tuning method for c: First of all, we discuss how to set c. This is conducted by roughly estimating m, M. For example, we can use the following relation.

$$\|\nabla f(x) - \nabla f(y)\| = \|\nabla^2 F(\xi)(x - y)\| \leq \|H\|\|x - y\|. \tag{190}$$

Thus,

$$m \leq \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} \leq M \tag{191}$$

holds for any x, y. For example, we calculate $\frac{\|\nabla f(x_1) - \nabla f(x_0)\|}{\|x_1 - x_0\|}$ and set it as c, then the condition will be satisfied. x_0 is the initial point and x_1 is an arbitrary nearby point.

Construction of J: From now on, we will consider how to generate J such that $\tilde{J}$ does not have 0 as eigenvalues. Thus, by constructing J carefully, 0 will not appear as eigenvalues. (Note that J is $d \times d$ matrix, of which dimension can be odd. If the dimension is odd, 0 always appears as eigenvalues.)

To assure these conditions, we propose to construct J as follows:

$$J' = \begin{pmatrix} 0 & 1 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 1\dots & 0 \\ 0 & 0 & 0 & 0\dots & 0 \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix}, \tag{192}$$

and then $J = J'^{\top} - J'$. Then J has eigenvalues $\pm i$ from the definition of the Jordan form of the skew-symmetric matrix.

The reason that we prepared the matrix that has eigenvalues $\pm i$ is that we need to control the largest and the smallest singular values of J due to the following reasons. Recall that the convergence of ELF is characterized by $\|\tilde{H}L\|$, and $\|\tilde{H}L\| = \|\tilde{H}\|\|L\|$ is satisfied if the eigenvalue of L is real, that is, in general, the eigenvalues of L is expressed by

$$l_i = \frac{2 - \eta^2\alpha^2c^{-1}\omega_i^2}{2} \pm i\frac{1}{2}\sqrt{4 - (2 - \eta^2\alpha^2c^{-1}\omega_i^2)^2}, \tag{193}$$

and if the above imaginary part disappears, that means $\|\tilde{H}L\| = \|\tilde{H}\|\|L\|$. Thus, we need to satisfy

$$0 \neq 4 - (2 - \eta^2\alpha^2c^{-1}\omega_i^2)^2 \tag{194}$$

and

$$\eta^2\alpha^2c^{-1}s_d^2M \leq 4, \tag{195}$$

by definition. Here, ω_i is the eigenvalues of $H^{1/2}L$. Note that if $\omega_i = 0$, the condition of Eq. (194) never satisfied. Thus, ω_i must be lower bounded by some positive constant. From the property of the eigenvalue, $|\omega_i| \geq m^{1/2}s_1$ holds, where s_1 denotes the smallest singular value of J. Thus, if J has 0 as an eigenvalue, the condition of Eq.(194) is never satisfied. Also, too small $\omega_i = 0$ results in a small imaginary part for the eigenvalue of L, which leads to a small acceleration effect. Thus, we need to control the smallest singular value about J. Also, a too large singular value of J violates the condition of Eq.(195). Thus, we need to control the singular values of J so that singular values will not become too large or too small. Based on these observations, making J by Eq. (192) results in J having $\pm i$ for all the eigenvalues, which is desirable.

However, since making J by Eq.(192) is not making a random matrix, the condition of Proposition 1 is not satisfied. Thus, we propose to add a very small Gaussian noise to generate the matrix. Then it satisfies the condition and becomes diagonalizable. Moreover, we need to control the singular value. This is achived by Theorem 8.2.1 in Golub and Van Loan (2012) (Gershgorin Circle Theorem). Following the main paper, its singular value is upper bounded by $s_d^2 \leq \max_i(1 + \sum_{j\neq i}|J_{ij}|/d)^2$.

Tuning for α: Next, we consider the condition of α. Since the eigenvalues of L are

$$l_i = \frac{2 - \eta^2\alpha^2c^{-1}\omega_i^2}{2} \pm i\frac{1}{2}\sqrt{4 - (2 - \eta^2\alpha^2c^{-1}\omega_i^2)^2}, \tag{196}$$

and

$$\eta^2\alpha^2c^{-1}\omega_i^2 \leq 4 \tag{197}$$

need to hold for the convergence. This is equivalent to

$$\eta^2\alpha^2s_d^2c^{-1}M \leq 4. \tag{198}$$

So, as for α, we set it $\eta^2\alpha^2c^{-1}s_d^2M \approx 2$ for example. This is because if $\eta^2\alpha^2c^{-1}s_d^2M$ are close to 0 or 4, then eigenvalues will be very close to ± 1 and no acceleration occurs.

From the construction of J, the largest singular value is estimated by

$$s_d^2 \leq \max_i (1 + \sum_{j \neq i} |J_{ij}|/d)^2 \tag{199}$$

and lower bounded by

$$s_d^2 \geq \min_i (1 - \sum_{j \neq i} |J_{ij}|/d)^2, \tag{200}$$

where J_{ij} is the realization of the element of J. If we choose ϵ very small, then s_d are very close to 1 and step size is set so that $\eta M \leq 2$ holds. So we have to set $\alpha^2 \approx \frac{c}{2\eta(1+\rho_{\max})^2}$, then acceleration occurs with high probability. The probability that s_d becomes 0 is estimated by the next proposition.

F.3. Proof of Proposition 7

Proof We prove that when we generate J randomly, the probability that has 0 as an eigenvalue is very small. From the definition of J and the Gershgorin circle theorem, the smallest eigenvalue of J appears inside the circle, centered at 1 with the radius $\sum_{j\neq i} |J_{ij}|/d$. Thus, if $\sum_{j\neq i} |J_{ij}|/d$ is smaller than 1, J will not have 0 as an eigenvalue.

Since J_{ij} follows Gaussian $N(0, \epsilon)$, its absolute values follows the folded Gaussian distribution. By definition, the folded Gaussian distribution of $N(0, \epsilon)$ has the mean of $\epsilon\sqrt{\frac{2}{\pi}}$ and the variance of ϵ. Since each J_{ij} is generated independently with each other, the mean of $\sum_{j\neq i} |J_{ij}|/d$ is $(d-1)/d\epsilon\sqrt{\frac{2}{\pi}}$. Then we need to estimate the probability that $(d-1)/d\epsilon\sqrt{\frac{2}{\pi}}$ is smaller than 1. This is easily estimated, for example, by Markov inequality. We have that with probability $1-(d-1)/d\epsilon\sqrt{\frac{2}{\pi}}$, $(d-1)/d\epsilon\sqrt{\frac{2}{\pi}}$ is smaller than 1. This indicates that the eigenvalue of 0 does not appear with high probability since we assume that ϵ is very small. ■

F.4. Other implementation for the leapfrog discretization

Copied objective function: We can also consider different leapfrog schemes. To implement the leap frog method, we prepare two sequence of parameters as $\{x_k\}$ and $\{y_k\}$. Then the dynamics is

$$x_{k+\frac{1}{2}} = x_k - \eta\alpha J \nabla F(y_k), \tag{201}$$

$$y_{k+\frac{1}{2}} = y_k - \eta\alpha J \nabla F(x_{k+\frac{1}{2}}), \tag{202}$$

$$x_{k+1} = x_{k+\frac{1}{2}} - \eta \nabla F(x_{k+\frac{1}{2}}), \tag{203}$$

$$y_{k+1} = y_{k+\frac{1}{2}} - \eta \nabla F(y_{k+\frac{1}{2}}), \tag{204}$$

where we copied the original objective function $F(x)$, and we optimize $\tilde{F}(x, y) = F(x) + F(y)$. The advantage of this approach is that we do not need the hyper-parameter c, which was appeared in our main paper. On the other hand, we need to evaluate $\nabla F(y_k)$ in addition. Thus, we need twice as much time for the gradient evaluations as in our main paper. As we had seen, c is tuned easily, thus we considered this copying approach is computationally heavy and computationally inefficient compared to the approach in the main paper.

Different step sizes: We can consider the discretization, in which different step sizes are used for the Euler discretization and leap-frog discretizations as:

$$\begin{cases} x_{k+\frac{1}{2}} = x_k - \frac{\eta'\alpha}{c} J y_k, \\ y_{k+\frac{1}{2}} = y_k - \eta'\alpha J \nabla F(x_{k+\frac{1}{2}}), \end{cases} \tag{205}$$

$$\begin{cases} y_{k+1} = y_{k+\frac{1}{2}} - \frac{\eta}{c} y_{k+\frac{1}{2}}, \\ x_{k+1} = x_{k+\frac{1}{2}} - \eta \nabla F(x_{k+\frac{1}{2}}). \end{cases} \tag{206}$$

This is more flexible but difficult to tune two step sizes in practice. Another promising way is that we incorporate the step size of the leapfrog step as

$$\begin{cases} x_{k+\frac{1}{2}} = x_k - \frac{\alpha}{c} J y_k, \\ y_{k+\frac{1}{2}} = y_k - \alpha J \nabla F(x_{k+\frac{1}{2}}), \end{cases} \tag{207}$$

$$\begin{cases} y_{k+1} = y_{k+\frac{1}{2}} - \frac{\eta}{c} y_{k+\frac{1}{2}}, \\ x_{k+1} = x_{k+\frac{1}{2}} - \eta \nabla F(x_{k+\frac{1}{2}}). \end{cases} \tag{208}$$

To control the convergence condition of the leapfrog method, this approach requires tuning only α. However, we also need to tune η in the whole, thus the difficulty in tuning η and α is the same as the method in the main paper.

Other discretizations: Other than the leapfrog method, the promising discretization method is the kind of backward discretization method. In optimization, they are known as the proximal steps. If the dynamics are linear, then we can implement the backward step by calculating the inverse matrix of the drift function. Thus, it is computationally demanding. Another way to implement the backward discretization is to use the proximal step for a given convex objective function. However, since the perturbation entails the skew-symmetric term, it is unclear how to incorporate it into the proximal step calculation. Thus, we leave the approach of proximal step to future work.

Appendix G. Eigenvalues of Newton's method

In this section, we observe the property of Newton's method. Newton's method selects the next point in an optimal way,

$$x_{k+1} = x_k - \eta d_k, \tag{209}$$

where the direction is chosen as a solution to the system

$$\nabla^2 F(x_k) d_t = \nabla F(x_k). \tag{210}$$

This is equivalent to minimizing the quadratic approximation:

$$F(y) \approx F(x_k) + \nabla F(x_k)^\top (y - X_k) + \frac{1}{2\alpha_t}(y - x_k)^\top \nabla^2 F(x_k)(y - x_k). \tag{211}$$

We can understand this Newton method in terms of condition number as follows: let us introduce a change of variables $x = Uy$. Then we consider minimizing $\tilde{F}(y) \equiv F(Uy)$ by a gradient descent. Since $\nabla \tilde{F}(y) = U \nabla F(x)$. Thus,

$$y_{k+1} = y_k - \eta \nabla \tilde{F}(y_k). \tag{212}$$

Then, multiply U, we get

$$U y_{k+1} = U y_k - \eta U \nabla \tilde{F}(y_k), \tag{213}$$

and this is equivalent to

$$x_{k+1} = x_k - \eta U^2 \nabla F(x_k). \tag{214}$$

Since

$$\nabla^2 \tilde{F}(y) = U \nabla F(x) U^\top. \tag{215}$$

Thus, if we choose $U^2 := \nabla^2 F(x)^{-1}$, it will be an best choice in terms of the condition number. Here all the eigenvalues of $\nabla^2 \tilde{F}(y)$ will be 1, thus condition number will be 1.

Next, we will observe the relationship between the Newton method and a skew-symmetric matrix. Let us denote the spectral decomposition of $\nabla^2 F(x) = V^\top \Lambda V$ where V is the orthonormal matrix and Λ is the diagonal matrix whose entries are eigenvalues. Then we can set $U = V \Lambda^{-1/2}$. Note that there exists a skew-symmetric matrix whose matrix exponential is equal to this V. This is the property of the Lie group. Then, we express it as $V = e^J = I + \sum_{n=1}^{\infty} J$. Here, let us consider rough approximation $V \approx I + J$. Then,

$$U^2 \approx (I + J)^\top \Lambda^{-1} (I + J) \approx (I - J^2), \tag{216}$$

here for simplicity, we assumed that all the eigenvalues are similar. Thus, in this sense Newton's method is related to a skew-symmetric matrix, however, it is quite different from our acceleration.

Table 2: Maximum and minimum singular values

	Random	Alg-1	optimal
Max	0.0002	1.00002	3.1488×10^4
Min	5.1×10^{-7}	0.9999	5.64

Appendix H. Experimental settings and discussions

First, we discuss how prepared J in numerical experiments and show their singular values which plays an important role.

About the random J, we first generated the upper triangular matrix with each entry following standard Gaussian distribution. Then we divide all the entries by d so that the assumption $\|J\|_F \leq$ will be satisfied.

Next, J of our proposed algorithm is generated at $\epsilon = 1e - 4$.

Finally, about the optimal J, we present the algorithm 2 of Lelièvre et al. (2013).

Algorithm 2 Generating optimal J

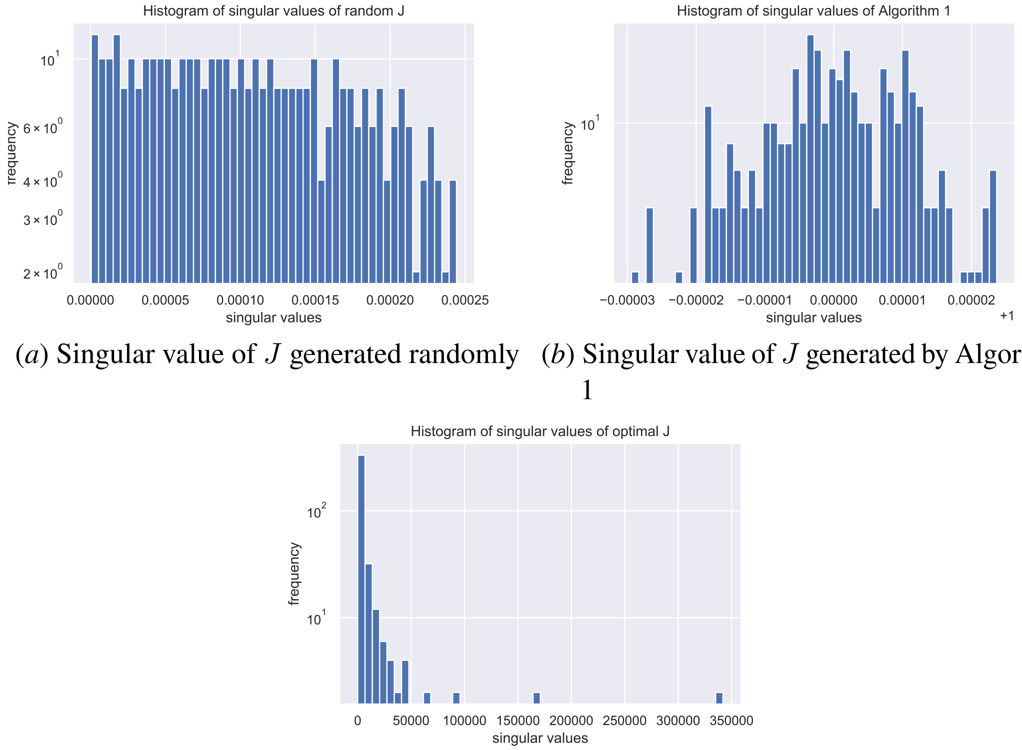
```
1: Input: Prepare an arbitrary orthonormal basis $\{\psi_i\}_{i=1}^d$
2: for $n = 1 \ldots d-1$ do
3:     Make a permutation of $(\psi_n \ldots \psi_d)$ so that
           $$\psi_n^\top H \psi_n = \max_{k=n,\ldots,d} \psi_k^\top H \psi_k > \mathrm{Tr}H/d \quad (217)$$
       and
           $$\psi_{n+1}^\top H \psi_{n+1} = \max_{k=n,\ldots,d} \psi_k^\top H \psi_k < \mathrm{Tr}H/d \quad (218)$$
4:     Compute $n^*$ such that $\psi_{n^*} = \cos n^* \psi_n + \sin n^* \psi_{n+1}$ satisfies $\psi_{n^*}^\top H \psi_{n^*} = \mathrm{Tr}H/d$
5:     By using a Gram-Schmidt procedure, change a set of $(\psi_{n^*}, \psi_{n+1}, \ldots, \psi_d)$ to an orthonormal basis $(\psi_{n^*}, \tilde{\psi}_{n+1}, \ldots, \tilde{\psi}_d)$
6: end for
7: Calculate the eigenvalues of $Q = [\psi_1, \ldots, \psi_d]$. Set them as $\{\lambda_i\}_{i=1}^d$
8: Solve $-\frac{\lambda_k+\lambda_j}{\lambda_k-\lambda_j} \psi_j^\top H \psi_k = \psi_j^\top J \psi_k$ if $k \neq j$ and $-\frac{\lambda_k+\lambda_j}{\lambda_k-\lambda_j} \psi_j^\top H \psi_k = 0$ if $k = j$.
9: Output: $J$
```

As for this optimal J, we found that it cannot satisfy $\|J\|_F$ numerically.

Singular value of J : Here we plot the singular values of different Js because singular values play important roles for the convergence analysis.

First, the singular values of J, which are generated randomly are very small and as shown in Table 2, the maximum and minimum singular values are very different. As for J generated by our algorithm shows that all the eigenvalues are concentrated near 1. Finally, as for optimal J, which shows very large singular values and its histogram shows the long tail to the right. This distribution of the singular value of optimal J makes the discretization difficult.

Setting α in continuous dynamics: As we have seen in Section D, the imaginary parts of the eigenvalues of $H + \alpha JH$ will never affect the convergence behavior. There is no condition for α



(a) Singular value of J generated randomly (b) Singular value of J generated by Algorithm 1

(c) Singular value of J generated by optimal method

Figure 7: Comparison of singular values of different J

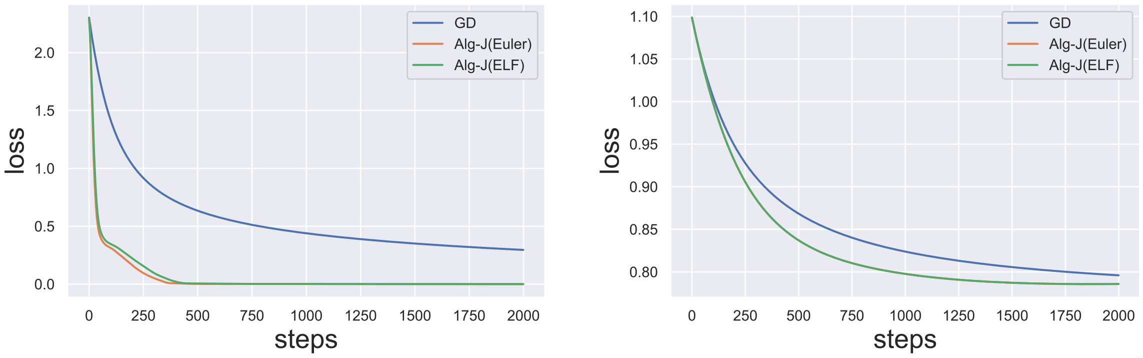
in general. Thus, in our experiments, we set α as the maximum singular value of J for randomly generated Js. As for optimal J, we set $\alpha = 1$.

Appendix I. Additional numerical experiments

Here we present the additional numerical experiments.

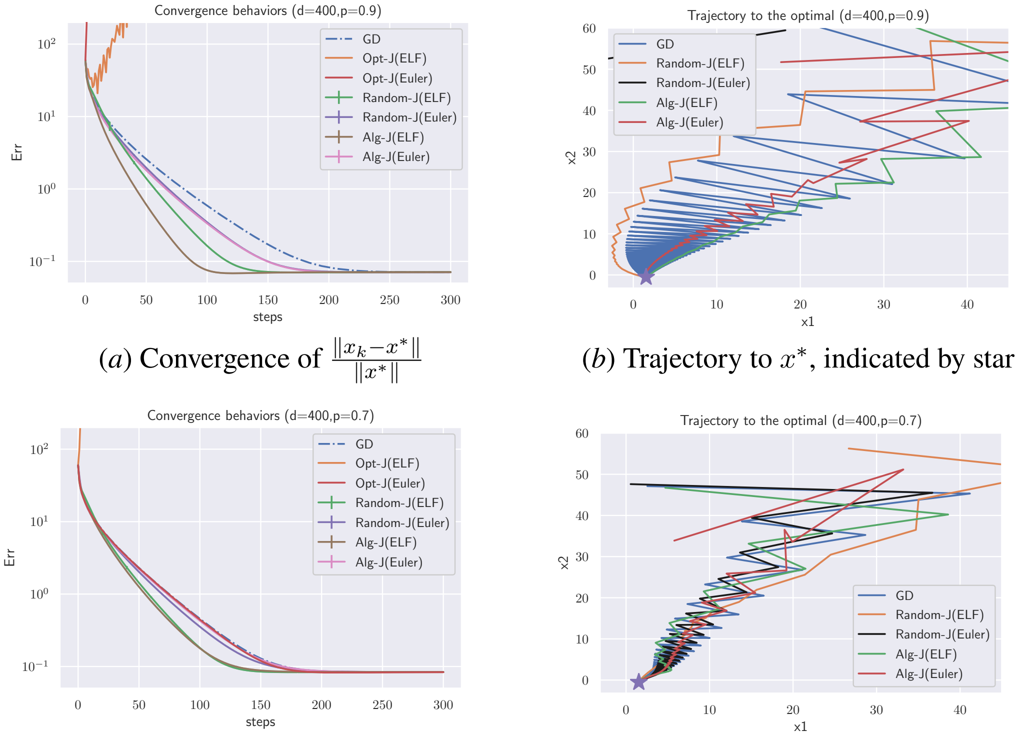
First, we present the softmax linear regression optimization for multiclass classification with L2 regularization. Figure 8 is the results of the softmax linear regression using MNIST and UCI (wine) datasets. We found that the proposed ELF method consistently outperformed the baseline GD.

Next, we present a least-square optimization using a sparse design matrix. The experimental setting is almost the same as the discretized experiments in Section 5.1. The difference is the design matrix A. We generated design matrix A with entries following $\mathcal{N}(0, 1)$. Then we replace each entry with 0 with probability $1-p$. Here we consider $p = 0.9$ and 0.7. We consider $N = 600$ and $d = 400$. The results are shown in Figure 9. We found that our proposed algorithm outperformed other methods in all experiments. We found that Euler did not improve the convergence in the experiment with $p = 0.7$.



(a) Softmax linear regression on MNIST dataset (b) Softmax linear regression on UCI (wine) dataset

Figure 8: Convergence behaviors of softmax linear regression



(a) Convergence of $\frac{\|x_k - x^*\|}{\|x^*\|}$

(b) Trajectory to x^*, indicated by star

(c) Convergence of $\frac{\|x_k - x^*\|}{\|x^*\|}$

(d) Trajectory to x^*, indicated by star

Figure 9: Comparisons of different discretization. $p = 0.9$ for (a) and (b). $p = 0.7$ for (c) and (d)