GRADIENT-ADJUSTED UNDERDAMPED LANGEVIN DYNAMICS FOR SAMPLING*

XINZHE ZUO†, STANLEY OSHER†, AND WUCHEN LI‡

Abstract. Sampling from a target distribution is a fundamental problem with wide-ranging applications in scientific computing and machine learning. Traditional Markov chain Monte Carlo (MCMC) algorithms, such as the unadjusted Langevin algorithm (ULA), derived from the overdamped Langevin dynamics, have been extensively studied. From an optimization perspective, the Kolmogorov forward equation of the overdamped Langevin dynamics can be treated as the gradient flow of the relative entropy in the space of probability densities embedded with Wasserstein-2 metrics. Several efforts have also been devoted to including momentum-based methods, such as underdamped Langevin dynamics for faster convergence of sampling algorithms. Recent advances in optimizations have demonstrated the effectiveness of primal-dual damping and Hessian-driven damping dynamics for achieving faster convergence in solving optimization problems. Motivated by these developments, we introduce a class of stochastic differential equations (SDEs) called gradient-adjusted underdamped Langevin dynamics (GAUL), which add stochastic perturbations in primal-dual damping dynamics and Hessian-driven damping dynamics from optimization. We prove that GAUL admits the correct stationary distribution, whose marginal is the target distribution. The proposed method outperforms overdamped and underdamped Langevin dynamics regarding convergence speed in the total variation distance for Gaussian target distributions. Moreover, using the Euler-Maruyama discretization, we show that the mixing time towards a biased target distribution only depends on the square root of the condition number of the target covariance matrix. Numerical experiments for non-Gaussian target distributions, such as Bayesian regression problems and Bayesian neural networks, further illustrate the advantages of our approach over classical methods based on overdamped or underdamped Langevin dynamics.

Key words. Hessian-driven damping dynamics; Primal-dual damping dynamics; Nesterov's method; Langevin dynamics; Optimal convergence rate.

MSC codes. 37M25, 65C05, 82C31.

1. Introduction.

Sampling from a target distribution is a long-standing quest and has numerous applications in scientific computing, including Bayesian statistical inference [46, 53, 43, 31], Bayesian inverse problems [56, 35, 23, 29], as well as Bayesian neural networks [65, 2, 61, 36, 45, 51]. In this direction, various algorithms have been developed to sample a target distribution $\pi \propto \exp(-f)$ for a given function $f : \mathbb{R}^d \to \mathbb{R}$, where π is only known up to a normalization constant. In this area, a simple and popular algorithm is the unadjusted Langevin algorithm (ULA):

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k) + \sqrt{2h}\boldsymbol{z}_k \,, \tag{1.1}$$

where $\boldsymbol{x}_k \in \mathbb{R}^d$, k is the iteration number, f is assumed to be a differentiable function, $h > 0$ is a step size, and $\boldsymbol{z}_k$ is a d-dimensional random variable with independently and identically distributed (i.i.d) entries following standard Gaussian distributions. The ULA algorithm (1.1) comes from the forward Euler discretization of a stochastic differential equation (SDE) known as overdamped Langevin dynamics:

$$d\boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t)dt + \sqrt{2}d\boldsymbol{B}_t \,, \tag{1.2}$$

*Submitted to the editors DATE.
†Department of Mathematics, University of California, Los Angeles, CA, 90095. ({zxz,sjo}@math.ucla.edu).
‡Department of Mathematics, University of South Carolina, Columbia, SC, 29208. (wuchen@mailbox.sc.edu).

where $\boldsymbol{x}_t \in \mathbb{R}^d$ and $\boldsymbol{B}_t$ is a standard d-dimensional Brownian motion. Under some mild conditions on f, it has been shown that the SDE (2.15) has a unique strong solution $\{\boldsymbol{x}_t, t \geq 0\}$ that is a Markov process [54, 49]. Moreover, the distribution of $\boldsymbol{x}_t$ converges to the invariant distribution $\pi \propto \exp(-f)$ as $t \to \infty$. The asymptotic convergence guarantees of (1.1) have been established decades ago [59, 30, 48]. In more recent years, non-asymptotic behaviors of (1.1) have also been explored by several works [19, 20, 26, 21, 15, 63].

An important result by [37] states that the Kolmogorov forward equation of Langevin dynamics corresponds to the gradient flow of the relative entropy functional in the space of probability density functions with the Wasserstein-2 metric. This observation serves as a bridge between the sampling community and the optimization community by studying optimization problems in Wasserstein-2 space. In the field of optimization, Nesterov's accelerated gradient [52] is a first order algorithm for finding the minimum of a convex/strongly convex objective function f. The intuition is that Nesterov's method incorporates momentum into the updates. It is much faster than the traditional gradient descent method, in the sense that the convergence speed for convex functions is $\mathcal{O}(\frac{1}{k^2})$ where k is the number of iterations compared to $\mathcal{O}(\frac{1}{k})$ for gradient descent. The convergence speed of Nesterov's method for L-smooth, m-strongly convex functions is $\mathcal{O}\big(\exp(-k/\sqrt{\kappa})\big)$ where $\kappa = L/m$ is the condition number of f compared to $\mathcal{O}\big(\exp(-k/\kappa)\big)$ for gradient descent. By taking the step size to 0, one obtains a second-order ODE for Nesterov's method called the Nesterov's accelerated gradient flow or Nesterov's ODE [57, 5]. In recent years, one extends the gradient flow of the relative entropy into Nesterov's accelerated gradient flow [57], which is explored in [64, 58, 44] from different perspectives. For the optimization in Wasserstein-2 space perspective, [64, 58, 13] study a class of accelerated dynamics with depending on the score function, i.e., the gradient of logarithm of density function. This results in the approximation of a non-linear partial differential equation, known as the damped Euler equation [10]. In this case, the optimal choices of parameters for sampling a target distribution share similarities with the classical Nesterov's accelerated gradient flow. On the other hand, from a stochastic dynamics perspective, a line of research has been devoted to study the accelerated version of Langevin dynamics, known as the underdamped Langevin dynamics [9, 16, 44, 66]. As explained later in Subsection 2.2, the underdamped Langevin dynamics consists of a deterministic component and a stochastic component. The deterministic component exactly corresponds to the Nesterov's accelerated gradient flow. The marginal of invariant distribution in x-axis satisfies the target distribution. However, the optimal choice of parameters in underdamped Langevin dynamics might not directly follow the classical Nesterov's method [16].

Recently, [67] proposed to use the primal-dual hybrid gradient (PDHG) method [12, 62] to solve unconstrained optimization problems. The original PDHG method is designed for optimization problem with linear constraints. [67] formulated the optimality condition $\nabla f(\boldsymbol{x}) = 0$ of a strongly convex function f into the solution of a saddle point problem

$$\inf_{\boldsymbol{x}\in\mathbb{R}^d} \sup_{\boldsymbol{p}\in\mathbb{R}^d} \quad \langle \nabla f(\boldsymbol{x}), \boldsymbol{p}\rangle - \frac{\gamma}{2}\|\boldsymbol{p}\|^2 \,,$$

where $\gamma > 0$ is a selected regularization parameter. They proceed by using the PDHG algorithm with appropriate preconditioners to solve the above saddle point problem. By taking the limit as the step size goes to zero, their algorithm yields a continuous-time flow, which is a second-order ordinary differential equation (ODE) called the primal-dual damping (PDD) dynamics. In particular, the PDD dynamic contains Nesterov's ODE [57]. In other words, Nesterov's ODE is a special case of PDD dynamics. The PDD dynamics also shares similarities with the Hessian–driven damping dynamics that has been studied in recent years [5, 3, 4]. The main difference between the PDD dynamics and the Nesterov's ODE is a second-order term $\nabla^2 f(\boldsymbol{x})\dot{\boldsymbol{x}}$ that appears in the former. This term is also presented in the Hessian driven damping dynamics. It has been observed that the PDD dynamics and the Hessian driven damping dynamics yield faster convergence towards the global minimum than the traditional gradient flow and Nesterov's ODE. Therefore, it is natural to extend the PDD dynamics and Hessian driven damping dynamics to SDEs for sampling a target distribution.

In this paper, we take inspirations from [67, 3] to design a system of SDE called gradient-adjusted underdamped Langevin dynamics (GAUL) that resembles the primal-dual damping dynamics and the Hessian driven damping dynamics. Consider

$$\begin{pmatrix} d\boldsymbol{x}_t \\ d\boldsymbol{p}_t \end{pmatrix} = \begin{pmatrix} -a\boldsymbol{C}\nabla f(\boldsymbol{x}_t)dt + \boldsymbol{C}\boldsymbol{p}_t dt \\ -\nabla f(\boldsymbol{x}_t)dt - \gamma \boldsymbol{p}_t dt \end{pmatrix} + \sqrt{\begin{pmatrix} 2a\boldsymbol{C} & \mathbf{I} - \boldsymbol{C} \\ \mathbf{I} - \boldsymbol{C} & 2\gamma \mathbf{I} \end{pmatrix}} \begin{pmatrix} d\boldsymbol{B}_t^{(1)} \\ d\boldsymbol{B}_t^{(2)} \end{pmatrix}, \tag{1.3}$$

for some constants $a, \gamma > 0$, whose detailed choices will be explained later. $\boldsymbol{C}$ is a preconditioner such that the diffusion matrix in front of the Brownian motion term is well-defined and positive semidefinite. And $\boldsymbol{B}_t^{(i)}$ is a standard Brownian motion in $\mathbb{R}^d$ for $i = 1, 2$. The supercript on $\boldsymbol{B}_t$ indicates that $\boldsymbol{B}_t^{(1)}$ and $\boldsymbol{B}_t^{(2)}$ are independent. We show that the stationary distribution GAUL (1.3) is the desired target distribution of the form $\frac{1}{Z}\exp(-f(\boldsymbol{x}) - \|\boldsymbol{p}\|^2/2)$. Noticeably, the $\boldsymbol{x}$-marginal distribution is the target distribution π. Additionally, we demonstrate that for a quadratic function f, GAUL achieves the exponential convergence and outperforms both overdamped and underdamped Langevin dynamics. A series of numerical examples are provided to demonstrate the advantage of the proposed method.

To illustrate the main idea, we summarize main theoretical results into the following informal theorem.

THEOREM 1.1 (Informal). *Suppose that $f : \mathbb{R}^d \to \mathbb{R}^d$ is given by $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T \Lambda \boldsymbol{x}$ with a symmetric positive definite matrix $\Lambda \in \mathbb{R}^{d\times d}$ with eigenvalues $s_1 \geq s_2 \geq \ldots \geq s_d > 0$. Let $\kappa = s_1/s_d$ be the condition number of matrix Λ. And let $\boldsymbol{C} = \boldsymbol{I}$.*

(1) Denote by $\rho_x(\boldsymbol{x}, t)$ the law of $\boldsymbol{x}_t$ driven by (1.3), and $\pi(\boldsymbol{x}) \propto \exp(-f(\boldsymbol{x}))$ the target distribution. Let $a > 0$, $\gamma = as_d + 2\sqrt{s_d}$. Then it takes at most $t = \mathcal{O}(\log(d/\delta))/(as_d + 2\sqrt{s_d})$ for the total variation distance between $\rho_x(\boldsymbol{x}, t)$ and $\pi(\boldsymbol{x})$ to decrease to δ.

(2) Denote by $\tilde{\rho}_x(\boldsymbol{x}, k)$ the law of $\boldsymbol{x}$ after k iterations of the Euler-Maruyama discretization of (1.3). Suppose $\sqrt{s_1} - \sqrt{s_d} \geq 2$, $a = 1$, $\gamma = s_d + 2\sqrt{s_d}$ and consider the Euler-Maruyama discretization of (1.3) with step size $h = 1/5s_1$. Then it takes at most $N = \mathcal{O}(\log(d/\delta)/(\kappa^{-1} + (\kappa s_1)^{-1/2})$ iterations for the total variation distance between $\tilde{\rho}_x(\boldsymbol{x}, k)$ and $\tilde{\pi}(\boldsymbol{x})$ to decrease to δ, where $\tilde{\pi}(\boldsymbol{x})$ is a biased target distribution given by Equation (B.24).

(3) When taking $a = \frac{2}{\sqrt{s_1} - \sqrt{s_d}}$, $\gamma = as_d + 2\sqrt{s_d}$ and $h = \frac{1}{2(as_1 + \gamma)}$, we can improve the number of iterations in (2) to $N = \mathcal{O}(\sqrt{\kappa}\log(d/\delta))$.

The detailed version of Theorem 1.1 is given in Theorem 3.9, Theorem 3.15 and Theorem 3.16. It is worth noting that GAUL (1.3) reduces to underdamped Langevin dynamics when $a = 0$ and $\boldsymbol{C} = \mathbf{I}$. Our theorem implies that in the Gaussian case, GAUL converges to the target measure faster than underdamped Langvein dynamics. In particular, we demonstrate that the Euler-Maruyama discretization admits a mixing time proportional to the square root of the condition number of covariance matrix. While this work primarily focuses on Gaussian distributions, our numerical experiments also explore non-log-concave target distributions in Bayesian linear regressions and Bayesian neural networks, which demonstrate potential advantages of GAUL over overdamped and underdamped Langevin dynamics. Extending these results to more general distributions and discretization schemes is an important future research direction. The choice of preconditoner $\boldsymbol{C}$ is tricky as one needs to guarantee that the diffusion matrix in (1.3) is positive semidefinite. Therefore, we mainly focus on the case when $\boldsymbol{C} = \mathbf{I}$. We address on our results for $\boldsymbol{C} \neq \mathbf{I}$ in Remark 3.10 and Remark 3.19. For $\boldsymbol{C} = \mathbf{I}$, [42] also explored dynamics (1.3), which they called Hessian-Free High-Resolution (HFHR) dynamics. For this closely related work, we provide some comparisons later in Remark 2.4.

This paper is organized as follows. In Section 2, we review the connection between optimization methods and sampling dynamics, which leads to the construction of our proposed SDE called gradient-adjusted underdamped Langevin dynamics (GAUL). Our main results are presented in Section 3, where we prove the exponential convergence of GAUL to the target distribution when the target measure follows a Gaussian distribution. We also study the Euler-Maruyama discretization of GAUL and prove its linear convergence to a biased target distribution. Lastly, in Section 4, we present several numerical examples to compare GAUL with both overdamped and underdamped Langevin dynamics.

2. Preliminaries.

In this section, we briefly review the relation among Euclidean gradient flows, overdamped Langevin dynamics and Wasserstein gradient flows. We then draw the connection between the underdamped Langevin dynamics and Nesterov's ODEs. We next review primal-dual damping (PDD) flows [67] and Hessian driven damping dynamics. Finally, we introduce a new SDE called gradient-adjusted underdamped Langevin dynamics (GAUL) for sampling, which resembles the PDD flow and the Hessian-driven damping dynamics with designed stochastic perturbations in terms of Brownian motions.

2.1. Gradient descent, unadjusted Langevin algorithms, and optimal transport gradient flows.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function with L-Lipschitz gradient. The classical gradient descent algorithm for finding the global minimum of $f(\boldsymbol{x})$ is an iterative algorithm that reads:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k)\,, \tag{2.1}$$

where $h > 0$ is the step size. When f is convex and the step size is not too large, this algorithm converges at a rate of $\mathcal{O}(k^{-1})$. When f is m-strongly convex, the same algorithm can be shown to converge at a rate of $\mathcal{O}\big((1 - m/L)^k\big)$, if the step size is chosen appropriately. The gradient descent algorithm (2.1) can be understood as the forward Euler time discretization of the gradient flow

$$\dot{\boldsymbol{x}}(t) = -\nabla f(\boldsymbol{x}(t))\,, \tag{2.2}$$

where $\boldsymbol{x}(t)$ describes a trajectory in $\mathbb{R}^d$ that travels in the direction of the steepest descent. Similar convergence results can be obtained for the gradient flow (2.2). When f is convex, the gradient flow (2.2) converges at a rate of $\mathcal{O}(t^{-1})$. When f is assumed to be m-strongly convex, the gradient flow (2.2) converges at a rate of $\mathcal{O}\big(\exp(-mt)\big)$.

While the goal of optimization is to find the global minimum of f, the goal of sampling algorithm is to sample from a distribution of the form $\frac{1}{Z_1}\exp(-f(\boldsymbol{x}))$, where the normalization constant $Z_1 > 0$ is assumed to be finite, i.e.,

$$Z_1 = \int_{\mathbb{R}^d} e^{-f(x)} dx < +\infty.$$

The classical unadjusted Langevin algorithm (ULA) given in (1.1) is a simple modification to the gradient descent method. Recall that ULA is given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - h\nabla f(\boldsymbol{x}_k) + \sqrt{2h}\boldsymbol{z}_k\,, \tag{2.3}$$

where $\boldsymbol{z}_k$ is a d-dimensional standard Gaussian random variable and h is the step size. We obtain (2.3) from (2.1) by adding a Gaussian noise term $\boldsymbol{z}_k$ scaled by $\sqrt{2h}$. Similar to how (2.1) can be viewed as the Euler discretization of (2.2), ULA (2.3) represents the forward Euler discretization of the overdamped Langevin dynamics:

$$d\boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t)dt + \sqrt{2}d\boldsymbol{B}_t\,, \tag{2.4}$$

where $\boldsymbol{B}_t$ is a standard d-dimensional Brownian motion. Denote by $\rho(\boldsymbol{x}, t)$ the probability density function for $\boldsymbol{x}_t$. Then the Kolmogorov forward equation (also known as the Fokker-Planck equation) of the overdamped Langevin dynamics (2.4) is given as

$$\frac{\partial \rho}{\partial t} = \nabla\cdot(\rho\nabla f) + \Delta\rho\,. \tag{2.5}$$

Clearly, $\pi(\boldsymbol{x}) = \frac{1}{Z_1}\exp(-f(\boldsymbol{x}))$ is a stationary solution of the Fokker-Planck equation (2.5). In other words, note that $\nabla\pi = -\pi\nabla f$, then

$$0 = \partial_t\pi = \nabla\cdot(\pi\nabla f) + \Delta\pi = \nabla\cdot((\pi\nabla f + \nabla\pi))\,.$$

In the literature, one can also study the gradient drift Fokker-Planck equation (2.5) from a gradient flow point of view. This means that equation (2.5) is a gradient flow in the probability space embedded with a Wasserstein-2 metric. We review some facts on a formal manner; see rigorous treatment in [1].

Define the probability space on $\mathbb{R}^d$ with finite second-order moment:

$$\mathcal{P}(\mathbb{R}^d) = \left\{\rho(\cdot)\in C^\infty:\ \int_{\mathbb{R}^d}\rho(\boldsymbol{x})d\boldsymbol{x} = 1,\ \int_{\mathbb{R}^d}|\boldsymbol{x}|^2\rho(\boldsymbol{x})\ d\boldsymbol{x} < \infty, \quad \rho(\cdot)\geq 0\right\}.$$

We note that $\mathcal{P}(\mathbb{R}^d)$ can be equipped with the L_2–Wasserstein metric g_W at each $\rho\in\mathcal{P}(\mathbb{R}^d)$ to form a Riemannian manifold $(\mathcal{P}(\mathbb{R}^d), g_W)$. Let $\mathcal{F}:\mathcal{P}(\mathbb{R}^d)\to\mathbb{R}$ be an energy functional on $\mathcal{P}(\mathbb{R}^d)$. To be more precise, denote the Wassertein gradient operator of functional $\mathcal{F}(\rho)$ at the density function $\rho\in\mathcal{P}(\mathbb{R}^d)$, such that

$$\mathrm{grad}_W\mathcal{F}(\rho) := -\nabla\cdot\left(\rho\nabla\frac{\delta}{\delta\rho}\mathcal{F}(\rho)\right),$$

where $\frac{\delta}{\delta\rho}$ is the L_2–first variation with respect to ρ. This yields that the gradient descent flow in the Wasserstein-2 space satsifies

$$\frac{\partial\rho}{\partial t} = -\mathrm{grad}_W\mathcal{F}(\rho) = \nabla\cdot\left(\rho\nabla\frac{\delta}{\delta\rho}\mathcal{F}(\rho)\right).$$

The above PDE is also named the *Wasserstein gradient descent flow*, in short Wasserstein gradient flows, which depend on the choices of the energy functionals $\mathcal{F}(\rho)$.

An important example observed by [37] is as follows. Consider the relative entropy functional, also named Kullback–Leibler(KL) divergence

$$\mathcal{F}(\rho) := \mathrm{D}_{\mathrm{KL}}(\rho\|\pi) = \int_{\mathbb{R}^d} \rho(\boldsymbol{x}) \log\big(\frac{\rho(\boldsymbol{x})}{\pi(\boldsymbol{x})}\big) d\boldsymbol{x}\,.$$

One can show that the Fokker-Planck equation (2.5) is the gradient flow of the relative entropy in $(\mathcal{P}(\mathbb{R}^d), g_W)$. Upon recognizing $\frac{\delta}{\delta\rho}\mathrm{D}_{\mathrm{KL}}(\rho\|\pi) = \log\big(\frac{\rho}{\pi}\big) + 1$, we obtain that (2.5) can be expressed as

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= - \operatorname{grad}_W \mathrm{D}_{\mathrm{KL}}(\rho\|\pi) = \nabla\cdot\Big(\rho\nabla\log\big(\frac{\rho}{\pi}\big)\Big) \\ &= \nabla\cdot(\rho\nabla\log\rho) - \nabla\cdot(\rho\nabla\log\pi) \\ &= \Delta\rho + \nabla\cdot(\rho\nabla f), \end{aligned} \tag{2.6}$$

where we use facts that $\rho\nabla\log\rho = \nabla\rho$ and $\nabla\log\pi = -\nabla f$.

We note that the gradient of the logarithm of the density function, i.e. $\nabla\log\rho$, is often called the score function. The analysis of score functions are essential in understanding the convergence behavior of the Fokker-Planck equation (2.5) toward its invariant distribution; see related analytical studies in [28].

2.2. Nesterov's ODEs and underdamped Langevin dynamics. Consider the problem of minimizing $f : \mathbb{R}^d \to \mathbb{R}$ for some convex function f with L-Lipschitz gradient. [52] proposed the following iterations:

$$\boldsymbol{x}_{k+1} = \boldsymbol{p}_k - h\nabla f(\boldsymbol{p}_k) \tag{2.7a}$$

$$\boldsymbol{p}_{k+1} = \boldsymbol{x}_{k+1} + \gamma_k(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\,, \tag{2.7b}$$

where $\gamma_k = (k-1)/(k-2)$. [52] showed that the above method converges at a rate of $\mathcal{O}(k^{-2})$ instead of $\mathcal{O}(k^{-1})$ which is the convergence rate of the classical gradient descent method. If f is further assumed to be m-strongly convex, then taking $h = 1/L$ and $\gamma_k = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$ where $\kappa = L/m$, yields a convergence rate of $\mathcal{O}\big(\exp(-k/\sqrt{\kappa})\big)$. This is also considerably faster than gradient descent, which is $\mathcal{O}\big((1-\kappa^{-1})^k\big)$. [57] showed that the continuous-time limit of Nesterov's accelerated gradient method [52] satisfies a second order ODE:

$$\ddot{\boldsymbol{x}} + \gamma_t\dot{\boldsymbol{x}} + \nabla f(\boldsymbol{x}) = 0\,. \tag{2.8}$$

If f is a convex function, then $\gamma_t = 3/t$; if f is a m-strongly convex function, then $\gamma_t = \gamma = 2\sqrt{m}$. As observed in [47], (2.8) can be formulated as a damped Hamiltonian system:

$$\begin{pmatrix}\dot{\boldsymbol{x}}\\ \dot{\boldsymbol{p}}\end{pmatrix} = \begin{pmatrix}0\\ -\gamma_t\boldsymbol{p}\end{pmatrix} + \begin{pmatrix}0 & \mathbf{I}\\ -\mathbf{I} & 0\end{pmatrix}\begin{pmatrix}\nabla_x H(\boldsymbol{x},\boldsymbol{p})\\ \nabla_p H(\boldsymbol{x},\boldsymbol{p})\end{pmatrix} = \begin{pmatrix}0 & \mathbf{I}\\ -\mathbf{I} & -\gamma_t\mathbf{I}\end{pmatrix}\begin{pmatrix}\nabla_x H(\boldsymbol{x},\boldsymbol{p})\\ \nabla_p H(\boldsymbol{x},\boldsymbol{p})\end{pmatrix}, \tag{2.9}$$

where the Hamiltonian function is defined as $H(\boldsymbol{x},\boldsymbol{p}) = f(\boldsymbol{x}) + \|\boldsymbol{p}\|^2/2$, $\boldsymbol{p}\in\mathbb{R}^d$. On the other hand, the underdamped Langevin dynamics for sampling $\Pi(\boldsymbol{x},\boldsymbol{p}) \propto \exp(-f(\boldsymbol{x}) - \|\boldsymbol{p}\|^2/2)$ is given by the system of SDE:

$$\begin{aligned} d\boldsymbol{x}_t &= \boldsymbol{p}_t dt, \\ d\boldsymbol{p}_t &= -\nabla f(\boldsymbol{x}_t)dt - \gamma_t\boldsymbol{p}_t dt + \sqrt{2\gamma_t}\,d\boldsymbol{B}_t, \end{aligned}$$

where γ_t is some damping parameter, and $\boldsymbol{B}_t$ is a d-dimensional standard Brownian motion. This can be reformulated as

$$(2.10)\qquad \begin{pmatrix} d\boldsymbol{x}_t \\ d\boldsymbol{p}_t \end{pmatrix} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & -\gamma_t \mathbf{I} \end{pmatrix} \begin{pmatrix} \nabla_x H(\boldsymbol{x},\boldsymbol{p}) \\ \nabla_p H(\boldsymbol{x},\boldsymbol{p}) \end{pmatrix} dt + \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{2\gamma_t}\mathbf{I} \end{pmatrix} d\boldsymbol{B}_t \,,$$

where $\boldsymbol{B}_t$ is a $2d$-dimensional standard Brownian motion. Observe that by adding a suitable Brownian motion term (the last term on the right hand side of (2.10)) to (2.9), Nesterov's accelerated gradient method for convex optimization becomes an algorithm for sampling $\Pi(\boldsymbol{x},\boldsymbol{p}) = \frac{1}{Z}\exp(-f(\boldsymbol{x}) - \|\boldsymbol{p}\|^2/2)$, where $Z := \int_{\mathbb{R}^{2d}} \exp(-f(\boldsymbol{x}) - \|\boldsymbol{p}\|^2/2) d\boldsymbol{x} d\boldsymbol{p} < +\infty$ is a noramlization constant. Moreover, the $\boldsymbol{x}$-marginal of $\Pi(\boldsymbol{x},\boldsymbol{p})$ is simply $\pi(\boldsymbol{x}) = \frac{1}{Z_1}\exp(-f(\boldsymbol{x}))$ up to a normalizing constant $Z_1 := \int_{\mathbb{R}^{2d}} \exp(-f(\boldsymbol{x}) - \|\boldsymbol{p}\|^2/2) d\boldsymbol{x} d\boldsymbol{p} < +\infty$. Therefore, (2.10) can be used to sample distributions of the form $\exp(-f(\boldsymbol{x}))/Z_1$. We postpone the proofs in terms of Fokker-Planck equations and there invariant distributions in Proposition 2.1 and 2.2.

2.3. Primal-dual damping dynamics and Hessian driven damping dynamics. Recently, [67] proposed to solve an unconstrained strongly convex optimization problem using the PDHG method by considering the saddle point problem

$$\inf_{\boldsymbol{x}\in\mathbb{R}^d} \sup_{\boldsymbol{p}\in\mathbb{R}^d} \quad \langle \nabla f(\boldsymbol{x}), \boldsymbol{p}\rangle - \frac{\gamma}{2}\|\boldsymbol{p}\|^2 \,,$$

where γ is a damping parameter, and $f : \mathbb{R}^d \to \mathbb{R}$ is m-strongly convex. Note that the saddle point $(\boldsymbol{x}^*, \boldsymbol{p}^*)$ for the above inf-sup problem satisfies $\nabla f(\boldsymbol{x}^*) = \boldsymbol{p}^* = 0$. Then the primal-dual damping (PDD) algorithm [67] admits the following iterations

$$\begin{aligned} \boldsymbol{p}_{k+1} &= \frac{1}{1+\tau_1\gamma}\boldsymbol{p}_k + \frac{\tau_1}{1+\tau_1\gamma}\nabla f(\boldsymbol{x}_k)\,, \\ \tilde{\boldsymbol{p}}_{k+1} &= \boldsymbol{p}_{k+1} + \omega(\boldsymbol{p}_{k+1} - \boldsymbol{p}_k)\,, \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k - \tau_2 \boldsymbol{C}(\boldsymbol{x}_k)\tilde{\boldsymbol{p}}_{k+1}\,, \end{aligned}$$

where $\tau_1, \tau_2 > 0$ are dual and primal step sizes, $\omega > 0$ is an extrapolation parameter, and $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ is a preconditioning positive definite matrix that could depend on $\boldsymbol{x}_k$ and t. The continuous-time limit of the PDD algorithm can be obtained by letting $\tau_1, \tau_2 \to 0$ while keeping $\tau_1\omega \to a$ for some $a > 0$. This yields a second-order ODE called the PDD flow:

$$(2.11)\qquad \ddot{\boldsymbol{x}} + \left(\gamma + a\boldsymbol{C}\nabla^2 f(\boldsymbol{x}) - \dot{\boldsymbol{C}}\boldsymbol{C}^{-1}\right)\dot{\boldsymbol{x}} + \boldsymbol{C}\nabla f(\boldsymbol{x}) = 0\,.$$

In the case when $\boldsymbol{C}$ is constant, (2.11) reads

$$(2.12)\qquad \ddot{\boldsymbol{x}} + \left(\gamma + a\boldsymbol{C}\nabla^2 f(\boldsymbol{x})\right)\dot{\boldsymbol{x}} + \boldsymbol{C}\nabla f(\boldsymbol{x}) = 0\,.$$

And when $\boldsymbol{C} = \mathbf{I}$, the PDD flow simplifies to

$$(2.13)\qquad \ddot{\boldsymbol{x}} + \gamma\dot{\boldsymbol{x}} + a\nabla^2 f(\boldsymbol{x})\dot{\boldsymbol{x}} + \nabla f(\boldsymbol{x}) = 0\,.$$

This corresponds to the Hessian driven damping dynamic [3] when $\gamma = 2\sqrt{m}$. The terminology 'Hessian driven damping' comes from the Hessian term $\nabla^2 f(\boldsymbol{x})\dot{\boldsymbol{x}}$ in (2.13), which is controlled by a constant $a \geq 0$. When $a = 0$, equation (2.13) reduces to Nesterov's ODE (2.8). As in dynamics (2.9), we can express equation (2.11) as

$$(2.14)\qquad \begin{pmatrix} \dot{\boldsymbol{x}} \\ \dot{\boldsymbol{p}} \end{pmatrix} = \begin{pmatrix} -a\boldsymbol{C} & \boldsymbol{C} \\ (\gamma a - 1)\mathbf{I} & -\gamma\mathbf{I} \end{pmatrix} \begin{pmatrix} \nabla_x H(\boldsymbol{x},\boldsymbol{p}) \\ \nabla_p H(\boldsymbol{x},\boldsymbol{p}) \end{pmatrix} \,,$$

where as before the Hamiltonian function is $H(\boldsymbol{x},\boldsymbol{p}) = f(\boldsymbol{x}) + \|\boldsymbol{p}\|^2/2$. Note that one of the key differences between (2.9) and (2.14) is that the top left block of the preconditioner matrix is nonzero in (2.14), which gives rise to the Hessian damping term $\nabla^2 f(\boldsymbol{x})\dot{\boldsymbol{x}}$. Throughout this paper, we focus on the dynamical system (2.14).

2.4. Gradient-adjusted underdamped Langevin dynamics. We design a sampling dynamics that resembles the PDD flow and the Hessian driven damping with stochastic perturbations by Brownian motions. Our goal is still to sample a distribution proportional to $\exp(-f(\boldsymbol{x}))$ for some $f : \mathbb{R}^d \to \mathbb{R}$. Let $H(\boldsymbol{x},\boldsymbol{p}) = f(\boldsymbol{x}) + \|\boldsymbol{p}\|^2/2$. And denote by $\boldsymbol{X} = (\boldsymbol{x},\boldsymbol{p}) \in \mathbb{R}^{2d}$. We consider the following SDE.

$$d\boldsymbol{X}_t = -\mathbf{Q}\nabla H(\boldsymbol{X}_t)dt + \sqrt{2\,\mathrm{sym}(\mathbf{Q})}d\boldsymbol{B}_t\,, \tag{2.15}$$

where $\mathbf{Q} \in \mathbb{R}^{2d\times 2d}$ is of the form

$$\mathbf{Q} = \begin{pmatrix} a\boldsymbol{C} & -\boldsymbol{C} \\ \mathbf{I} & \gamma\mathbf{I} \end{pmatrix}, \tag{2.16}$$

for some constant $a,\gamma \in \mathbb{R}$, and symmetric positive definite $\boldsymbol{C} \in \mathbb{R}^{d\times d}$. $\nabla H(\boldsymbol{X}_t) = (\nabla_x H(\boldsymbol{X}_t), \nabla_p H(\boldsymbol{X}_t))^T$. And $\mathrm{sym}(\mathbf{Q}) = \frac{1}{2}(\mathbf{Q} + \mathbf{Q}^T)$ is the symmetrization of $\mathbf{Q}$. We assume that $\mathrm{sym}(\mathbf{Q})$ is positive semidefinite.

Throughout this paper, we will limit our discussion to $a,\gamma \geq 0$. $\boldsymbol{B}_t$ is a $2d$-dimensional standard Brownian motion. Observe that when $a = 0$, (2.15) reduces to underdamped Langevin dynamics (2.10). When $a > 0$, (2.15) has an additional gradient term $a\boldsymbol{C}\nabla f(\boldsymbol{x}_t)$ in the $d\boldsymbol{x}_t$ equation. Thus, we call (2.15) gradient-adjusted underdamped Langevin dynamics. Let us examine the probability density function $\rho(\boldsymbol{X}, t)$ of the diffusion governed by (2.15). This is described by the following Fokker-Planck equation:

$$\frac{\partial\rho}{\partial t} = \nabla\cdot(\mathbf{Q}\nabla H\rho) + \sum_{i,j=1}^{2d} \frac{\partial^2}{\partial X_i \partial X_j}(Q_{ij}\rho)\,. \tag{2.17}$$

We assume that f is differentiable and ∇f is a smooth Lipschitz vector field. This ensures that the Fokker-Planck equation (2.17) has a smooth solution when $t > 0$ for a given initial condition, such that $\rho(\boldsymbol{X}, 0) \geq 0$ and $\int_{\mathbb{R}^{2d}} \rho(\boldsymbol{X}, 0)d\boldsymbol{X} = 1$.

Denote by $\Pi(\boldsymbol{X}) = \frac{1}{Z}e^{-H(\boldsymbol{X})}$, where Z is a normalization constant such that $\Pi(\boldsymbol{X})$ integrates to one on $\mathbb{R}^{2d}$. We show that $\Pi(\boldsymbol{X})$ is the stationary distribution of (2.17). First, we have the following decomposition for (2.17).

Proposition 2.1 ([28] Proposition 1). *The Fokker-Planck equation* (2.17) *can be decomposed as*

$$\frac{\partial\rho}{\partial t} = \nabla\cdot\left(\rho\,\mathrm{sym}(\mathbf{Q})\nabla\log\frac{\rho}{\Pi}\right) + \nabla\cdot(\rho\Gamma)\,, \tag{2.18}$$

where

$$\begin{aligned}\Gamma(\boldsymbol{X}) :=& \mathrm{sym}(\mathbf{Q})\nabla\log(\Pi(\boldsymbol{X})) + \mathbf{Q}\nabla H(\boldsymbol{X}) \\ =& \frac{1}{2}(\mathbf{Q} - \mathbf{Q}^T)\nabla H(\boldsymbol{X})\,.\end{aligned} \tag{2.19}$$

In particular, the following equality holds:

$$\nabla\cdot(\Pi(\boldsymbol{X})\Gamma(\boldsymbol{X})) = 0.$$

The proof is presented in Appendix C. Observe that the first term on the right-hand side of (2.18) is a Kullback–Leibler (KL) divergence functional that appears in a Fokker-Planck equation associated with the overdamped Langevin dynamics (2.5). The second term is due to the fact that the drift term $-\mathbf{Q}\nabla H$ in (2.15) is a non-gradient vector field.

PROPOSITION 2.2. $\Pi(\boldsymbol{X})$ *is a stationary distribution for* (2.17)*.*

The proof is based on a straightforward calculation: When $\rho = \Pi$, we have $\nabla\cdot(\rho\Gamma) = 0$, and therefore $\frac{\partial \rho}{\partial t} = 0$. For completeness, we have included this calculation in Appendix C. This shows that $\Pi(\boldsymbol{X})$ is indeed the stationary distribution of (2.17). Like the underdamped Langevin dynamics, the $\boldsymbol{x}$-marginal of the stationary distribution is $\exp(-f(\boldsymbol{x}))$ up to some normalization constant. Therefore, (2.15) can be used for sampling $\frac{1}{Z_1}\exp(-f(\boldsymbol{x}))$ by first jointly sampling $\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{p})$ and then taking out the $\boldsymbol{x}$-marginal.

Remark 2.3. GAUL can also be viewed as a preconditioned overdamped Langevin dynamics on the space of $(\boldsymbol{x}, \boldsymbol{p}) \in \mathbb{R}^{2d}$. Designing optimal preconditioning matrix and optimal diffusion matrix have been studied in literature; see [11, 6, 32, 39, 33, 14, 41, 40]. In particular, [41] considered the necessary condition on the optimal diffusion coefficient by studying the spectral gap of the generator assosiated with the SDE, which requires the solution to an optimization subproblem. While the problem considered by [41] is more general, our diffusion matrix (2.16) is much simpler and does not require solving an optimization problem. Another closely related work is [40], which considered preconditioning of the form $\mathbf{Q} = \mathbf{I} + \boldsymbol{J}$. Here $\mathbf{I}$ is the identity matrix and $\boldsymbol{J}$ is skew-symmetric, i.e. $\boldsymbol{J} = -\boldsymbol{J}^T$. [40] studied the optimal $\boldsymbol{J}$ when the potential f is a quadratic function, which is also the focus of this work.

Remark 2.4. In [42], the authors also studied (1.3) with $\boldsymbol{C} = \mathbf{I}$ which they called Hessian-Free High-Resolution (HFHR) dynamics. They considered potential functions f that are L-smooth and m-strongly convex. They proved a convergence rate of $\frac{\sqrt{m}}{2\sqrt{\kappa}}$ in continuous time in terms of Wasserstein-2 distance between the target and sample measure. [42] used the randomized midpoint method [55] combined with as their discretization and showed an interation complexity of $\widetilde{\mathcal{O}}(\sqrt{d}/\varepsilon)$. Specifically, [42] showed that for a two-dimensional Gaussian target measure, under the optimal choice of parameter (damping parameter γ and step size h) for underdamped Langevin dynamics with Euler-Maruyama discretization, the convergence rate is $\mathcal{O}\big((1-\kappa^{-1})^k\big)$. This rate is recovered in Corollary 3.17. On the other hand, [42] showed that under their choice of parameter for HRHF, the convergence rate is $\mathcal{O}\big((1-2\kappa^{-1})^k\big)$, which is a slight improvement compared with underdamped Langevin dynamics. In this work, we performed a detailed eigenvalue analysis of GAUL on Gaussian target measure. We showed that under our choice of parameters (γ, a, h), the convergence rate towards the biased target measure is $\mathcal{O}\big((1-c\sqrt{\kappa})^k\big)$ for some constant c.

3. Analysis of GAUL on quadratic potential functions.

In this section, we establish the convergence rate for the proposed SDE (2.17) towards the target distribution following a Gaussian distribution.

3.1. Problem set-up.

In this subsection, we present the main problem addressed in this paper. We are interested in sampling from a distribution whose probability density function is proportional to $\exp(-f(\boldsymbol{x}))$ for $f : \mathbb{R}^d \to \mathbb{R}$. In this paper, we focus on a concrete example in which the potential function f is quadratic, and thus the target distribution is a Gaussian distribution. Let

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\Sigma_*^{-1}\boldsymbol{x}, \tag{3.1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $\Sigma_* \succ 0$ is a symmetric positive definite matrix in $\mathbb{R}^{d\times d}$. Define

$$\widetilde{\Sigma} = \begin{pmatrix} \Sigma_* & 0 \\ 0 & \mathbf{I} \end{pmatrix}. \tag{3.2}$$

As in the previous section, denote by $\boldsymbol{X} = (\boldsymbol{x}, \boldsymbol{p}) \in \mathbb{R}^{2d}$. And $H(\boldsymbol{X}) = f(\boldsymbol{x}) + \|\boldsymbol{p}\|^2/2$. Then, we can write

$$H(\boldsymbol{X}) = \frac{1}{2}\mathbf{X}^T \begin{pmatrix} \Sigma_*^{-1} & 0 \\ 0 & \mathbf{I} \end{pmatrix} \mathbf{X} := \frac{1}{2}\mathbf{X}^T\widetilde{\Sigma}^{-1}\mathbf{X}. \tag{3.3}$$

Define the target density $\pi : \mathbb{R}^{2d} \to \mathbb{R}$ to be

$$\Pi(\mathbf{X}) = \frac{1}{Z}\exp(-H(\mathbf{X})), \tag{3.4}$$

where $H(\mathbf{X})$ is given by (3.3) and $Z = \int_{\mathbb{R}^{2d}} \exp(-H(\mathbf{X}))d\mathbf{X}$ is a normalization constant such that $\Pi(\mathbf{X})$ integrates to one on $\mathbb{R}^{2d}$. We also define the $\boldsymbol{x}$-marginal target density to be

$$\pi(\boldsymbol{x}) = \frac{1}{Z_1}\exp(-f(\boldsymbol{x})), \tag{3.5}$$

where $f(\boldsymbol{x})$ is given by (3.1) and $Z_1 = \int_{\mathbb{R}^d} \exp(-f(\boldsymbol{x}))d\boldsymbol{x}$ is a normalization constant.

Remark 3.1. Note that for any symmetric positive definite Σ_*, we have that $\Sigma_*^{-1} = \mathbf{P}\Lambda\mathbf{P}^T$ for some orthogonal matrix $\mathbf{P}$ and diagonal matrix $\Lambda = \mathrm{diag}(s_1, \ldots, s_d)$ with $s_1 \geq \cdots s_d > 0$. By a change of variable $\boldsymbol{y} = \mathbf{P}^T\boldsymbol{x}$, one can rewrite $f(\boldsymbol{x})$ in terms of $\boldsymbol{y}$, such that

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\Sigma_*^{-1}\boldsymbol{x} = \frac{1}{2}\boldsymbol{x}^T\mathbf{P}\Lambda\mathbf{P}^T\boldsymbol{x} = \frac{1}{2}\boldsymbol{y}^T\Lambda\boldsymbol{y}.$$

For simplicity of notation, we assume that $\mathbf{P} = \mathbf{I}$ and $\Sigma_*^{-1} = \Lambda$ is a diagonal matrix. We denote by $\kappa = s_1/s_d$ the condition number of f. We will also assume that $s_1 > 1 > s_d$ throughout this paper. Furthermore, to simplify our analysis, we consider $\boldsymbol{C} = \mathrm{diag}(c_1, \ldots, c_d)$.

3.2. Continuous time analysis.

In this subsection, we study the convergence of GAUL. In particular, we analyze the convergence of the Fokker-Planck equation (2.17) to the target density (3.4), (3.5) by directly studying an ODE system of the covariance of the distribution.

PROPOSITION 3.2. *Let $\boldsymbol{X}_t$ be the solution of* (2.15) *where $H(\boldsymbol{X})$ is given by* (3.3), *and $\boldsymbol{X}_0 \sim \mathcal{N}(0, \boldsymbol{I}_{2d\times 2d})$. Then $\boldsymbol{X}_t \sim \mathcal{N}(0, \Sigma(t))$ where the covariance $\Sigma(t)$ satisfies the following matrix ODE:*

$$\dot{\Sigma}(t) = 2\,\mathrm{sym}(\mathbf{Q}(\boldsymbol{I} - \widetilde{\Sigma}^{-1}\Sigma(t))). \tag{3.6}$$

Moreover, equation (3.6) *is well-defined, and has a solution for all $t \geq 0$, such that $\Sigma(t)$ is symmetric semi-positive definite.*

The proof is postponed in Appendix C. We denote by $\Sigma_{ij}(t) \in \mathbb{R}^{d\times d}$ the block components of $\Sigma(t) \in \mathbb{R}^{2d\times 2d}$:

$$\Sigma(t) = \begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{12}^T(t) & \Sigma_{22}(t) \end{pmatrix} .$$

Then we can write (3.6) in terms of the block components.

COROLLARY 3.3. *The componentwise covariance matrix $\Sigma_{ij}(t)$ satisfies the following ODE system*

$$\dot{\Sigma}_{11} = -2a(\mathrm{sym}(\boldsymbol{C}\Sigma_*^{-1}\Sigma_{11}) - \boldsymbol{C}) + 2\,\mathrm{sym}(\boldsymbol{C}\Sigma_{12})\,, \tag{3.7a}$$

$$\dot{\Sigma}_{22} = -2\,\mathrm{sym}(\Sigma_*^{-1}\Sigma_{12}) - 2\gamma(\Sigma_{22} - \boldsymbol{I})\,, \tag{3.7b}$$

$$\dot{\Sigma}_{12} = -a\boldsymbol{C}\Sigma_*^{-1}\Sigma_{12} - (\boldsymbol{C} - \boldsymbol{C}\Sigma_{22}) + (\boldsymbol{I} - \Sigma_{11}\Sigma_*^{-1}) - \gamma\Sigma_{12}\,, \tag{3.7c}$$

Moreover, with initial conditions $\Sigma_{11}(0) = \Sigma_{22}(0) = \boldsymbol{I}$ and $\Sigma_{12}(0) = 0$, the stationary states of $\Sigma_{11}(t)$, $\Sigma_{22}(t)$ and $\Sigma_{12}(t)$ are given by Σ_*, $\boldsymbol{I}$ and 0 respectively.

From now on, we consider $\boldsymbol{C} = \mathbf{I}$ in our analysis. We address our results for $\boldsymbol{C} \neq \mathbf{I}$ in Remark 3.10 and Remark 3.19. Note that when $\boldsymbol{C} = \mathbf{I}$, we have $\mathbf{Q} = \mathrm{sym}(\mathbf{Q})$ is always positive semidefinite for $a, \gamma \geq 0$. Our next theorem makes sure that the stationary state of equation (3.6) is actually unique and characterizes the convergence speed of the convariance matrix towards its stationary state.

THEOREM 3.4. *Let $\boldsymbol{X}_t$ be the solution of* (2.15) *where $H(\boldsymbol{X})$ is given by* (3.3), *and $\boldsymbol{X}_0 \sim \mathcal{N}(0, \boldsymbol{I}_{2d\times 2d})$. Then $\Sigma(t)$ converges to the unique stationary state $\widetilde{\Sigma}$ given in* (3.2). *The optimal choice of γ is given by $\gamma^* = as_d + 2\sqrt{s_d}$ under which we have $\|\Sigma_{11}(t) - \Sigma_*\|_{\mathrm{F}} = \mathcal{O}(te^{-(2as_d+2\sqrt{s_d})t})$ and $\|\Sigma_{22}(t) - \boldsymbol{I}\|_{\mathrm{F}} = \mathcal{O}(te^{-(2as_d+2\sqrt{s_d})t})$ for $t \geq 1$.*

Proof. As mentioned in Remark 3.1, we consider $\Sigma_*^{-1} = \Lambda$. By our assumption on $\mathbf{X}_0$, (3.7) implies that $\Sigma_{11}(t)$, $\Sigma_{22}(t)$ and $\Sigma_{12}(t)$ will be diagonal matrices for all $t > 0$. This simplifies the ODE system (3.7). After some manipulation, we obtain

$$\begin{pmatrix} \dot{\Sigma}_{11} \\ \dot{\Sigma}_{22} \\ \ddot{\Sigma}_{22} \end{pmatrix} = \underbrace{\begin{pmatrix} -2a\boldsymbol{C}\Sigma_*^{-1} & -2\gamma\boldsymbol{C}\Sigma_* & -\boldsymbol{C}\Sigma_* \\ 0 & 0 & \mathbf{I} \\ 2\Sigma_*^{-2} & 2(-1-a\gamma)\boldsymbol{C}\Sigma_*^{-1} - 2\gamma^2\mathbf{I} & -3\gamma\mathbf{I} - a\boldsymbol{C}\Sigma_*^{-1} \end{pmatrix}}_{\mathcal{D}} \begin{pmatrix} \Sigma_{11} \\ \Sigma_{22} \\ \dot{\Sigma}_{22} \end{pmatrix} + \mathbf{T}\,, \tag{3.8}$$

where

$$\mathbf{T} = \begin{pmatrix} 2a\boldsymbol{C} + 2\gamma\boldsymbol{C}\Sigma_* \\ 0 \\ 2a\gamma\Sigma_*^{-1}\boldsymbol{C} + 2\gamma^2\mathbf{I} + 2\Sigma_*^{-1}\boldsymbol{C} - 2\Sigma_*^{-1} \end{pmatrix} ,$$

And $\boldsymbol{C} = \mathbf{I}$. We have already seen in Corollary 3.3 that the stationary state of $\Sigma(t)$ is $\widetilde{\Sigma}$ given in (3.2). To show uniqueness, we compute the eigenvalues of $\mathcal{D}$:

$$\lambda_0^{(i)} = -as_i - \gamma\,,$$

$$\lambda_1^{(i)} = -as_i - \gamma - \sqrt{\gamma^2 - 2a\gamma s_i + s_i(-4 + a^2 s_i)}\,,$$

$$\lambda_2^{(i)} = -as_i - \gamma + \sqrt{\gamma^2 - 2a\gamma s_i + s_i(-4 + a^2 s_i)}\,,$$

where s_i's are the diagonal elements of Λ for $i = 1, \ldots, d$. It is clear that 0 is not an eigenvalue of $\mathcal{D}$. Therefore, $\widetilde{\Sigma}$ is the unique stationary state for $\Sigma(t)$. The convergence speed of (3.8) is essentially controlled by the largest real part of the eigenvalues of $\mathcal{D}$. Note that for all i,

$$\Re(\lambda_2^{(i)}) \geq \Re(\lambda_0^{(i)}) \geq \Re(\lambda_1^{(i)}) ,$$

where $\Re(z)$ denotes the real part of $z \in \mathbb{C}$. Therefore, to characterize the convergence speed of (3.8), it suffices to control $\max_i \Re(\lambda_2^{(i)})$. By Lemma B.7, we know that for any given $a \geq 0$, the optimal choice of γ is

$$\gamma^* = \arg\min_{\gamma>0} \max_i \Re(\lambda_2^{(i)}) = as_d + 2\sqrt{s_d} .$$

With $\gamma = \gamma^*$, we get that

$$\max_{i,j} \Re(\lambda_j^{(i)}) \leq \max_i \Re(\lambda_2^{(i)}) \leq -2as_d - 2\sqrt{s_d} .$$

This leads to

$$\left\| \begin{pmatrix} \Sigma_{11}(t) - \Sigma_* \\ \Sigma_{22}(t) - \mathbf{I} \\ \dot{\Sigma}_{22}(t) \end{pmatrix} \right\|_{\mathrm{F}} \leq C_1 t e^{-(2as_d + 2\sqrt{s_d})t} , \tag{3.9}$$

□

which is valid for $t \geq 1$. The constant C_1 depends on d, s_1, s_d^{-1} at most polynomially according to Lemma B.8. Note that the extra t dependence comes from the repeated eigenvalue $\lambda_0^{(d)} = \lambda_1^{(d)} = \lambda_2^{(d)}$ when $\gamma = \gamma^*$. By a triangle inequality, we get

$$\|\Sigma_{11} - \Sigma_*\|_{\mathrm{F}} \leq \left\| \begin{pmatrix} \Sigma_{11}(t) - \Sigma_* \\ \Sigma_{22}(t) - \mathbf{I} \\ \dot{\Sigma}_{22}(t) \end{pmatrix} \right\|_{\mathrm{F}} \leq C_1 t e^{-(2as_d + 2\sqrt{s_d})t} .$$

And similarly,

$$\|\Sigma_{22} - \mathbf{I}\|_{\mathrm{F}} \leq C_1 t e^{-(2as_d + 2\sqrt{s_d})t} .$$

Remark 3.5. The choice $a = 0$ corresponds to underdamped Langevin dynamics (UL). Taking $a > 0$ gives an extra factor of $e^{-2as_d t}$ in terms of convergence.

DEFINITION 3.6 (Mixing time). *The total variation between two probability measures $\mathcal{P}$ and $\mathcal{Q}$ over a measurable space $(\mathbb{R}^d, \mathcal{F})$ is*

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) = \sup_{A \in \mathcal{F}} |\mathcal{P}(A) - \mathcal{Q}(A)| .$$

Let $\mathcal{T}_p$ be an operator on the space of probability distributions. Assume that $\mathcal{T}_p^k(\nu_0) \to \nu$ as $k \to \infty$ for some initial distribution ν_0 and stationary distribution ν. The discrete δ-mixing time ($\delta \in (0,1)$) is given by

$$t_{\mathrm{mix}}^{\mathrm{dis}}(\delta; \nu_0, \nu) = \min\{k \,|\, \mathrm{TV}(\mathcal{T}_p^k(\nu_0), \nu) \leq \delta\} .$$

Similarly, if $\mathcal{T}_p(t; \cdot)$ is an operator for each $t \geq 0$ with $\mathcal{T}_p(0; \cdot) = \mathrm{id}(\cdot)$ and assume that $\mathcal{T}_p(t; \nu_0) \to \nu$ as $t \to \infty$. The continuous δ-mixing time ($\delta \in (0,1)$) is given by

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\delta; \nu_0, \nu) = \min\{t \,|\, \mathrm{TV}(\mathcal{T}_p(t; \nu_0), \nu) \leq \delta\} .$$

THEOREM 3.7 ([24]). *Let $\mu \in \mathbb{R}^d$, Σ_1, Σ_2 be two positive definite covariance matrices, and $\lambda_1, \dots, \lambda_d$ denote the eigenvalues of $\Sigma_1^{-1}\Sigma_2 - \boldsymbol{I}$. Then the total variation satisfies*

$$\mathrm{TV}(\mathcal{N}(\mu,\Sigma_1),\mathcal{N}(\mu,\Sigma_2)) \le \frac{3}{2}\min\left\{1, \sqrt{\sum_{i=1}^{d}\lambda_i^2}\right\}.$$

A straightforward corollary follows from Schur decomposition theorem.

COROLLARY 3.8. *We have*

$$\mathrm{TV}(\mathcal{N}(\mu,\Sigma_1),\mathcal{N}(\mu,\Sigma_2)) \le \frac{3}{2}\min\left\{1, \|\Sigma_1^{-1}\Sigma_2 - \boldsymbol{I}\|_{\mathrm{F}}\right\}.$$

Using Theorem 3.4 and Corollary 3.8, we obtain the following mixing time theorem when the potential function f is quadratic.

THEOREM 3.9 (Continuous mixing time). *Consider the same setting as in Theorem* 3.4. *Consider $0 < \delta \ll 1$. Then*

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\delta;\nu_0,\pi) \le \frac{\mathcal{O}(\log(d)+\log(\kappa)) + \log(1/\delta)}{as_d + 2\sqrt{s_d}}.$$

Here ν_0 is the distribution of $\boldsymbol{x}$, which is $\mathcal{N}(0, \boldsymbol{I}_{d\times d})$. π is the target density in the $\boldsymbol{x}$ variable given in (3.5).

Proof. We shall use Corollary 3.8 with

$$\Sigma_1 = \Sigma_*, \qquad \Sigma_2 = \Sigma_{11}(t).$$

We have

$$\begin{aligned}\|\Sigma_1^{-1}\Sigma_2 - \mathbf{I}\|_{\mathrm{F}} &= \left\|\Sigma_*^{-1}(\Sigma_{11}(t) - \Sigma_*)\right\|_{\mathrm{F}} \\ &\le C_1 t e^{-(2as_d+2\sqrt{s_d})t} s_1.\end{aligned}$$

By a direct computation, we get

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\delta;\nu_0,\pi) \le \frac{\log(\tilde{C}_1/\delta)}{as_d + 2\sqrt{s_d}}, \qquad \square$$

where $\tilde{C}_1 = \frac{3}{2}C_1 s_1$. By Lemma B.8, we have that

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\delta;\nu_0,\pi) \le \frac{\mathcal{O}(\log(d\kappa)) + \log(1/\delta)}{as_d + 2\sqrt{s_d}}.$$

Remark 3.10. When $\boldsymbol{C} = \mathrm{diag}(c_1, \dots, c_d)$ and $\mathrm{sym}(\mathbf{Q}) \succeq 0$ in (2.16), our proof can be easily adapted to show similar results in Theorem 3.9:

$$t_{\mathrm{mix}}^{\mathrm{cont}}(\delta;\nu_0,\pi) \le \frac{\mathcal{O}(\log(d)+\log(\hat{\kappa})) + \log(1/\delta)}{a\hat{s}_d + 2\sqrt{\hat{s}_d}},$$

where $\hat{s}_i$ is the i-th largest eigenvalue of matrix $\boldsymbol{C}\Sigma_*^{-1}$. And $\hat{\kappa} = \hat{s}_1/\hat{s}_d$. In other words, the matrix $\boldsymbol{C}$ can be viewed as a preconditioner for the target covariance matrix in the sampling problem.

3.3. Discrete time analysis. To implement (2.15), we need to consider its time discretization. As discretization is not the focus of this paper, we will only analyze the simplest discretization using the Euler-Maruyama method in Appendix A.

Let us first make a few observations regarding the discretization in Appendix A. After a straightforward computation, we obtain the following update rule.

PROPOSITION 3.11. *The Euler-Maruyama discretization of* (2.15) *given in Appendix* A *with step size h can be written in the following form*

$$\begin{pmatrix}\boldsymbol{x}_{n+1}\\ \boldsymbol{p}_{n+1}\end{pmatrix} = \boldsymbol{A}\begin{pmatrix}\boldsymbol{x}_{n}\\ \boldsymbol{p}_{n}\end{pmatrix} + \boldsymbol{L}\boldsymbol{z}\,, \tag{3.10}$$

where

$$\boldsymbol{A} = \boldsymbol{I}_{2d\times 2d} - \underbrace{h\begin{pmatrix} a\Lambda & -\boldsymbol{I}_{d\times d}\\ \Lambda & \gamma\boldsymbol{I}_{d\times d}\end{pmatrix}}_{\boldsymbol{G}}, \qquad \boldsymbol{L} = \begin{pmatrix}\sqrt{2ah}\boldsymbol{I} & 0\\ 0 & \sqrt{2\gamma h}\boldsymbol{I}\end{pmatrix}. \tag{3.11}$$

And $\boldsymbol{z}$ is a $2d$-dimensional Brownian motion, i.e., $\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}_{2d\times 2d})$.

Using (3.10), we can derive the evolution of the mean and covariance at each time step. As before, let us denote by $\boldsymbol{X}_n = (\boldsymbol{x}_n,\boldsymbol{p}_n)$.

COROLLARY 3.12. *Suppose that $\mathbb{E}(\boldsymbol{x}_0)=\mathbb{E}(\boldsymbol{p}_0)=0$. Then*

$$\mathrm{cov}(\boldsymbol{X}_{n+1},\boldsymbol{X}_{n+1}) = \boldsymbol{A}\mathrm{cov}(\boldsymbol{X}_n,\boldsymbol{X}_n)\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{L}^T\,.$$

Proof. From (3.10), it is clear that $\mathbb{E}(\boldsymbol{x}_n)=\mathbb{E}(\boldsymbol{p}_n)=0$ for all $n\geq 0$. We calculate

$$\begin{aligned}\mathrm{cov}(\boldsymbol{X}_{n+1},\boldsymbol{X}_{n+1}) &= \mathbb{E}\big(\boldsymbol{A}\boldsymbol{X}_n\boldsymbol{X}_n^T\boldsymbol{A}^T + \boldsymbol{A}\boldsymbol{X}_n\boldsymbol{z}^T\boldsymbol{L}^T + \boldsymbol{L}\boldsymbol{z}\boldsymbol{X}_n^T\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{L}^T\big)\\ &= \boldsymbol{A}\mathrm{cov}(\boldsymbol{X}_n,\boldsymbol{X}_n)\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{L}^T\,. \qquad\square\end{aligned}$$

COROLLARY 3.13. *Denote by $\boldsymbol{Y}^*$ a solution to the fixed point equation $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{Y}\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{L}^T$. And let $\boldsymbol{Y}_n = \mathrm{cov}(\boldsymbol{X}_n,\boldsymbol{X}_n) - \boldsymbol{Y}^*$. Then*

$$\boldsymbol{Y}_{n+1} = \boldsymbol{A}\boldsymbol{Y}_n\boldsymbol{A}^T\,.$$

THEOREM 3.14. *Suppose $a \geq \frac{2}{\sqrt{s_1}-\sqrt{s_d}}$ and the step size h satisfies $0<h<1/(as_1+\gamma)$ and $\gamma=\gamma^*=as_d+2\sqrt{s_d}$. Then there exists a unique $\boldsymbol{Y}^*$ satisfying*

$$\boldsymbol{Y}^* = \boldsymbol{A}\boldsymbol{Y}^*\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{L}^T\,.$$

Moreover, the iteration $\boldsymbol{Y}_{k+1} = \boldsymbol{A}\boldsymbol{Y}_k\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{L}^T$ converges to $\boldsymbol{Y}^*$ linearly: $\|\boldsymbol{Y}_k - \boldsymbol{Y}^*\|_{\mathrm{F}} \leq \widetilde{C}h^2k^2(1-\frac{h}{2}(as_d+\sqrt{s_d}))^{2k-2}$, where the constant $\widetilde{C} = d^2\cdot\mathcal{O}(\mathrm{poly}(\kappa))$.

Proof. Existence: we directly compute this stationary point in Lemma B.17. Uniqueness: by Lemma B.14 and Corollary B.10 we see that $\boldsymbol{Y}^*$ is unique. The convergence rate is proved in Lemma B.14 and Theorem B.16.

THEOREM 3.15 (Discrete mixing time). *Suppose $\sqrt{s_1}-\sqrt{s_d}\geq 2$. We take $a=1$, $\gamma=\gamma^*=s_d+2\sqrt{s_d}$, $h=1/5s_1$. If we use the Euler-Maruyama scheme for* (2.15)*, then for $0<\delta\ll 1$,*

$$t_{\mathrm{mix}}^{\mathrm{dis}}(\delta;\nu_0,\tilde{\pi}) = \mathcal{O}\left(\frac{\log(\kappa)+\log(1/\delta)+\log(d)}{\frac{1}{\kappa}+\frac{1}{\sqrt{\kappa s_1}}}\right). \tag{3.12}$$

Here ν_0 is the distribution of $\boldsymbol{x}$, which is $\mathcal{N}(0,\boldsymbol{I}_{d\times d})$. $\tilde{\pi}$ is the target density in the $\boldsymbol{x}$ variable which is a zero mean Gaussian distribution with a variance given by (B.24)*.*

Proof. Note that from our previous notation, we have that

$$\operatorname{cov}(\boldsymbol{x}_k, \boldsymbol{x}_k) = \begin{pmatrix}\mathbf{I}_{d\times d} & 0\end{pmatrix} \operatorname{cov}(\boldsymbol{X}_k, \boldsymbol{X}_k) \begin{pmatrix}\mathbf{I}_{d\times d} \\ 0\end{pmatrix} =: \widetilde{\boldsymbol{Y}}_k\,.$$

Moreover, let us define

$$\widetilde{\boldsymbol{Y}}^* = \begin{pmatrix}\mathbf{I}_{d\times d} & 0\end{pmatrix} \boldsymbol{Y}^* \begin{pmatrix}\mathbf{I}_{d\times d} \\ 0\end{pmatrix}$$

to be the limiting covariance in the $\boldsymbol{x}$ variable for the discretization ($\boldsymbol{Y}^*$ is defined in Theorem 3.14). Clearly, we have that

$$\|\widetilde{\boldsymbol{Y}}_k - \widetilde{\boldsymbol{Y}}^*\|_{\mathrm{F}} \le \|\boldsymbol{Y}_k - \boldsymbol{Y}^*\|_{\mathrm{F}} \le \widetilde{C}h^2k^2(1 - \frac{h}{2}(as_d + \sqrt{s_d}))^{2k-2}\,. \tag{3.13}$$

Using Corollary 3.8, we compute

$$\begin{aligned}\|(\widetilde{\boldsymbol{Y}}^*)^{-1}\widetilde{\boldsymbol{Y}}_k - \mathbf{I}\|_{\mathrm{F}} &= \|(\widetilde{\boldsymbol{Y}}^*)^{-1}(\widetilde{\boldsymbol{Y}}_k - \widetilde{\boldsymbol{Y}}^*)\|_{\mathrm{F}} \\ &\le \|(\widetilde{\boldsymbol{Y}}^*)^{-1}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{Y}}_k - \widetilde{\boldsymbol{Y}}^*\|_{\mathrm{F}}\,.\end{aligned}$$

By Lemma B.17, $\widetilde{\boldsymbol{Y}}^*$ is a diagonal matrix. Therefore $(\widetilde{\boldsymbol{Y}}^*)^{-1}$ is also a diagonal matrix. Moreover, from (B.24), we see that $\|(\widetilde{\boldsymbol{Y}}^*)^{-1}\|_{\mathrm{F}} \le \sqrt{d}\mathcal{O}(\mathrm{poly}(\kappa))$. Therefore, we obtain

$$\begin{aligned}\|(\widetilde{\boldsymbol{Y}}^*)^{-1}\widetilde{\boldsymbol{Y}}_k - \mathbf{I}\|_{\mathrm{F}} &\le d^{5/2}\cdot\mathcal{O}(\mathrm{poly}(\kappa))h^2k^2(1 - \frac{h}{2}(s_d + \sqrt{s_d}))^{2k-2} \\ &\le d^{5/2}\cdot\mathcal{O}(\mathrm{poly}(\kappa))h^2k^2e^{-(k-1)h(s_d+\sqrt{s_d})}\,,\end{aligned}$$ ∎

where we used $1 - x \le e^{-x}$ for $x \in \mathbb{R}$ to get the second inequality. Letting $h = 1/5s_1$ and taking logarithm on both hand sides, we conclude that

$$t_{\mathrm{mix}}^{\mathrm{dis}}(\delta; \nu_0, \tilde{\pi}) \le \frac{\mathcal{O}(\log(d)) + \mathcal{O}(\log(\kappa)) + \log(1/\delta)}{\frac{1}{10}(\frac{1}{\kappa} + \frac{1}{\sqrt{\kappa s_1}})}\,.$$

THEOREM 3.16 (A better choice of a). *The denominator of the mixing time given in Theorem 3.15 can be improved to $\kappa^{-1/2}$ by choosing $a = \frac{2}{\sqrt{s_1}-\sqrt{s_d}}$, $\gamma = as_d + 2\sqrt{s_d}$ and $h = \frac{1}{2(as_1+\gamma)}$. To be more precise, we have*

$$t_{\mathrm{mix}}^{\mathrm{dis}}(\delta; \nu_0, \tilde{\pi}) = \mathcal{O}\left(\frac{\log(\kappa) + \log(1/\delta) + \log(d)}{\frac{1}{\sqrt{\kappa}}}\right)\,. \tag{3.14}$$

Proof. The proof will be very similar to that of Theorem 3.15. We start with (3.13). And we can explicitly calculate

$$\begin{aligned}1 - \frac{h}{2}(as_d + \sqrt{s_d}) &= 1 - \frac{as_d + \sqrt{s_d}}{4(as_1 + as_d + 2\sqrt{s_d})} \\ &= 1 - \frac{2s_d + \sqrt{s_d}(\sqrt{s_1} - \sqrt{s_d})}{8(s_1 + s_d + \sqrt{s_d}(\sqrt{s_1} - \sqrt{s_d}))} \\ &= 1 - \frac{\sqrt{s_1 s_d} + s_d}{8(s_1 + \sqrt{s_1 s_d})} \\ &\le 1 - \frac{1}{16\sqrt{\kappa}}\,.\end{aligned}$$

The rest of the proof is the same as the proof of Theorem 3.15 and we will suppress it for brevity. ∎

The following corollary follows from Lemma B.15 and the proof of Theorem 3.15.

COROLLARY 3.17 (Underdamped Langevin mixing time). *Suppose* $a = 0$, $\gamma = 2\sqrt{s_d}, h = \sqrt{s_d}/s_1$. *If we use the Euler-Maruyama scheme for* (2.15)*, then for* $0 < \delta \ll 1$,

$$t_{\text{mix}}^{\text{dis}}(\delta; \nu_0, \tilde{\pi}) = \mathcal{O}\left(\frac{\log(\kappa) + \log(1/\delta) + \log(d)}{\frac{1}{\kappa}}\right), \tag{3.15}$$

ν_0 *is the distribution of* $\boldsymbol{x}$*, which is* $\mathcal{N}(0, \boldsymbol{I}_{d\times d})$*.* $\tilde{\pi}$ *is the target density in the* $\boldsymbol{x}$ *variable which is a zero mean Gaussian with variance given by* (B.24) *with* $a = 0$*.*

Remark 3.18. $a = 0$ in (2.15) corresponds to the underdamped Langevin dynamics. In this case, we show in Lemma B.15 that to guarantee convergence (to a biased target) the step size restriction on h is more strict than when $a = 1$. In particular, when $a = 0$ it follows from Lemma B.15 that the choice $h = 1/5s_1$ does not guarantee convergence if $s_d < 10^{-2}$. Comparing (3.14) and (3.15), we see that the mixing time for GAUL beats that of underdamped Langevin dynamics under the Euler-Maruyama discretization. We are aware that this does not imply the same result will hold when comparing the mixing time towards the true target distribution $\pi(\boldsymbol{x})$ given in (3.5), due to the presence of bias in the Euler-Maruyama scheme. Designing better discretization and reducing the bias in the stationary distribution is left as future works.

Remark 3.19. When $\boldsymbol{C} = \text{diag}(c_1, \dots, c_d)$ and $\text{sym}(\mathbf{Q}) \succeq 0$ in (2.16), we also have a similar mixing time described in Theorem 3.16, which is $\mathcal{O}\big(\sqrt{\hat{\kappa}}(\log(\hat{\kappa}) + \log(1/\delta) + \log(d))\big)$ when $a = \frac{2}{\sqrt{\hat{s}_1} - \sqrt{\hat{s}_d}}$, $\gamma = a\hat{s}_d + 2\sqrt{\hat{s}_d}$ and $h = \frac{1}{2(a\hat{s}_1 + \gamma)}$. The notation $\hat{s}_i$ and $\hat{\kappa}$ are defined in Remark 3.10.

Remark 3.20. When the target potential f is not a quadratic function, it is more technical in proving the convergence speed. A common technique to prove convergence in the Wasserstein-2 distance is by a coupling argument (see [16, 22]). [9] proved L_2 convergence under a Poincarè-type inequality using Bochner's formula. In the L_1 distance and KL divergence, [28] design convergence analysis towards these problems. We leave the convergence analysis of general f with optimal choices of preconditioned matrices $\mathbf{Q}$ in future works.

4. Numerical experiment.

In this section, we implement several numerical examples to compare the proposed SDE with the overdamped (labeled 'ol') and underdamped (labeled 'ul') Langevin dynamics. We use the same step size for all three algorithms. Recall that 'ol' corresponds to the choice $a = 1, \gamma = 0$ and 'ul' corresponds to $a = 0$ in (2.15). We set $\boldsymbol{C} = \mathbf{I}$.

4.1. Gaussian examples.

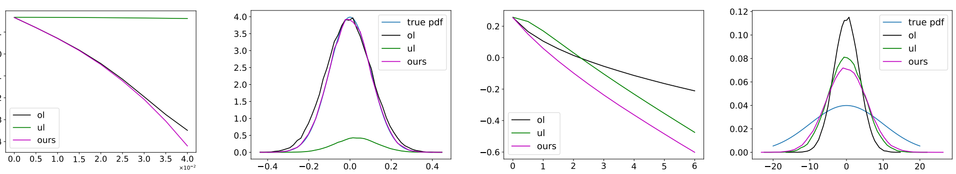
4.1.1. One dimension.

We begin with a simple example, a one dimensional Gaussian distribution with zero mean. In Figure 1, we consider two cases where the variances are given by 0.01 and 100 respectively. We first sample $M = 10^5$ particles from $\mathcal{N}(0, \mathbf{I}_{2\times 2})$ (although our experiment is in one dimension, we need both $\boldsymbol{x}$ and $\boldsymbol{p}$ variables). When measuring the convergence speed, we use KL divergence in Gaussian distributions to measure the change of covariances. Note that we will only measure the KL divergence in the x variable, since we are primarily interested in sampling distribution of the form $\frac{1}{Z}e^{-f(x)}$. In this experiment, we can make use of the fact that the sample distribution and the target distribution are both Gaussians. And the

KL divergence between two centered Gaussians has a closed form expression:

$$\mathrm{D}_{\mathrm{KL}}(\Sigma(t), \widetilde{\Sigma}) = \frac{1}{2}\left(\operatorname{tr}(\Sigma(t)\widetilde{\Sigma}^{-1}) - \log\det(\Sigma(t)\widetilde{\Sigma}^{-1}) - d\right). \tag{4.1}$$

In this one dimensional example, we study two cases where $\widetilde{\Sigma} = 0.01$ or 100. $\Sigma(t)$ can be approximated by the unbiased sample variance. For $\widetilde{\Sigma} = 0.01$, we choose time step size $h = 10^{-4}$, total number of steps $N = 400$, $\gamma_{ul} = 2\widetilde{\Sigma}^{-1/2} = 20$, $\gamma_{pdd} = 2\widetilde{\Sigma}^{-1/2} + \widetilde{\Sigma}^{-1} = 120$. For $\widetilde{\Sigma} = 100$, we choose the time step size $h = 10^{-2}$, total number of steps $N = 600$, $\gamma_{ul} = 2\widetilde{\Sigma}^{-1/2} = 0.2$, $\gamma_{pdd} = 2\widetilde{\Sigma}^{-1/2} + \widetilde{\Sigma}^{-1} = 0.21$. In Figure 1, we observe that our proposed method outperforms both overdamped and underdamped Langevin dynamics in both cases.



(a) KL decay (b) Density (c) KL decay (d) Density

Fig. 1: Convergence and density comparisons of three methods. (a) and (c): KL divergence between the sample and the target distribution, which is a one-dimensional Gaussian with zero mean and variance 0.01 (a), 100 (c). 'ol' represents overdamped Langevin dynamics; 'ul' represents underdamped Langevin dynamics. x-axis represents time and y-axis is in $\log_{10}$ scale. (b) and (d): density comparison at the end of the experiment between the three methods and the true density.

4.1.2. 20 dimensions. Let the target distribution be a 20-dimensional Gaussian with zero mean and covariance given by a diagonal matrix with entries $0.05 + 5i$ for $i = 0, \ldots, 19$. The last dimension has the largest variance, which is $\sigma_{\max}^2 = 95.05$. Therefore, we choose $a = \frac{2}{\sigma_{\min}^{-1/2} - \sigma_{\max}^{-1/2}}$, $\gamma_{ul} = 2\sigma_{\max}^{-1}$ and $\gamma_{pdd} = 2\sigma_{\max}^{-1} + a\sigma_{\max}^{-2}$. In this experiment, we use (1) time step size $h = 5 \times 10^{-3}$ and run for 4000 steps; (2) time step size $h = 5 \times 10^{-2}$ and run for 400 steps. The KL divergence can still be computed using (4.1). To visualize the final distribution in a two-dimensional plane, we plot the scatter plot of the samples in the first and the last dimensions. All results are presented in Figure 2.

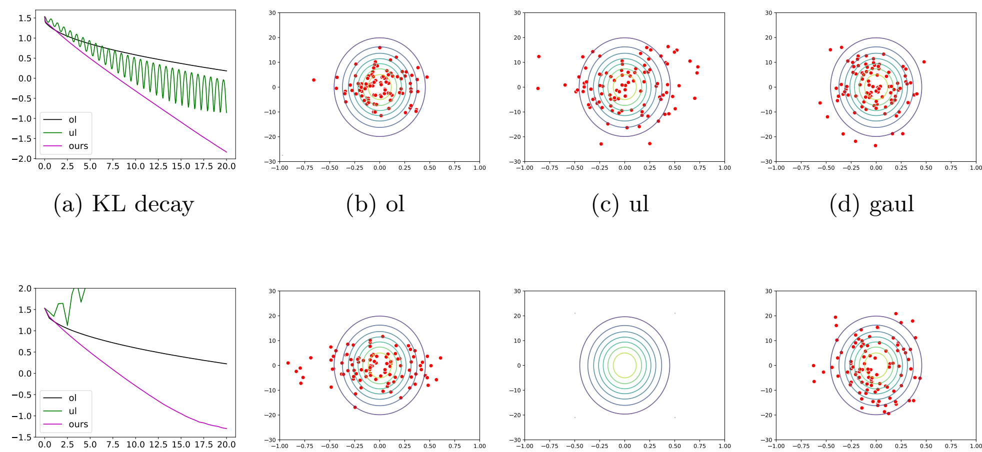
4.2. Mixture of Gaussian.

4.2.1. Strongly log-concave. Consider the problem of sampling from a mixture of Gaussian distributions $\mathcal{N}(\alpha, \mathbf{I})$ and $\mathcal{N}(-\alpha, \mathbf{I})$, whose density satisfies:

$$p(\boldsymbol{x}) = \frac{1}{2(2\pi)^{d/2}}\left(e^{-\|\boldsymbol{x}-\alpha\|_2^2/2} + e^{-\|\boldsymbol{x}+\alpha\|_2^2/2}\right).$$

The corresponding potential is given as

$$f(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \alpha\|_2^2 - \log\left(1 + e^{-2\boldsymbol{x}^\top \alpha}\right), \tag{4.2}$$



(a) KL decay (b) ol (c) ul (d) gaul

(e) KL decay (f) ol (g) ul (h) gaul

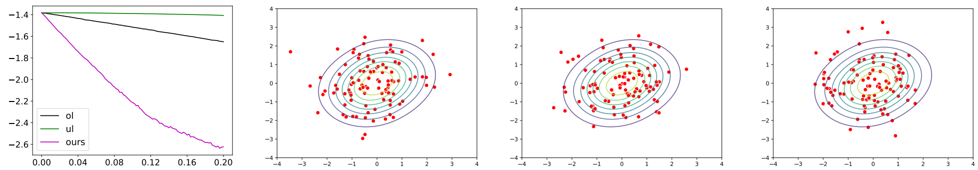
Fig. 2: Convergence and scatter plots. (a)–(d): $h = 0.005$. (e)–(h): $h = 0.05$. (a) and (e): KL divergence between the sample and target distribution. The x-axis represents time and the y-axis is in $\log_{10}$ scale. Rest panels: scatter plot of the three methods at the end of the experiment for different step sizes. Contours of the true density are also provided for comparisons. In (g) there are no scatter points shown as 'ul' does not converge for this choice of h.

$$\nabla f(\boldsymbol{x}) = \boldsymbol{x} - \alpha + 2\alpha(1 + e^{2\boldsymbol{x}^\top \alpha})^{-1}. \tag{4.3}$$

Following [27, 20], we set $\alpha = (1/2, 1/2)$ and $d = 2$. This choice of parameters yields strong convexity parameter $m = 1/2$ and Lipschitz constant $L = 1$. We choose $a = \frac{2}{\sqrt{L}-\sqrt{m}}$, $\gamma_{ul} = 2m^{1/2}$ and $\gamma_{pdd} = 2m^{1/2} + am$. Initially particles are sampled from $\mathcal{N}(0, \mathbf{I})$. We use time step $h = 2 \times 10^{-4}$ and run for 2000 steps. We use 5×10^5 particles and $n^2 = 2500$ bins to approximate the KL divergence between the sample points and the target distribution (see Remark 4.1). The results are shown in Figure 3.

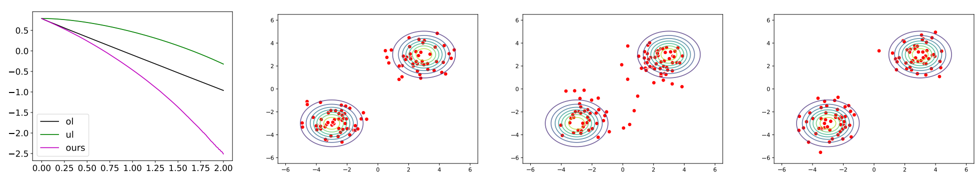
Remark 4.1. To compute the KL divergence between sample points and a non-Gaussian target distribution in two dimension, we first get the 2d histogram of the samples points using n^2 bins (n in each dimension). We then use this 2d histogram as an approximation of the empirical distribution of the samples. Similarly, we can get a discretized target distribution by evaluating the target distribution at the center of each bins. Finally, we can compute the discrete KL divergence using n^2 values from the histogram and the discretized target distribution.

4.2.2. Non log-concave . We also consider the same example as in Subsection 4.2.1 with $\alpha = (3, 3)$. As the distance between the two Gaussians increases, the target density is no longer log-concave. We use time step size $h = 10^{-3}$ and run for 2000 steps. We use $a = 1$, $\gamma_{ul} = \sqrt{2}$, and $\gamma_{pdd} = \sqrt{2} + 1/2$. We use 5×10^5 particles and $n^2 = 2500$ bins to evaluate the KL divergence. The results are demonstrated in Figure 4.



(a) KL decay (b) ol (c) ul (d) gaul

Fig. 3: Convergence and scatter plots. (a): KL divergence between the sample and target distribution, which is a mixture of two unit variance Gaussians located at $(1/2, 1/2)$ and $(-1/2, -1/2)$. x-axis represents time and y-axis is in $\log_{10}$ scale. (b)–(d): scatter plot of the three methods a the end of the experiment. Contour of the true density is also provided for comparison.



(a) KL decay (b) ol (c) ul (d) gaul

Fig. 4: Convergence and scatter plots for mixture of Guassians centered at $(3, 3)$ and $(-3, -3)$.

4.3. Quadratic cosine. Consider a potential function given by a quadratic function and a cosine term:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T B^{-1}\boldsymbol{x} - \cos(\boldsymbol{c}^T\boldsymbol{x})$$

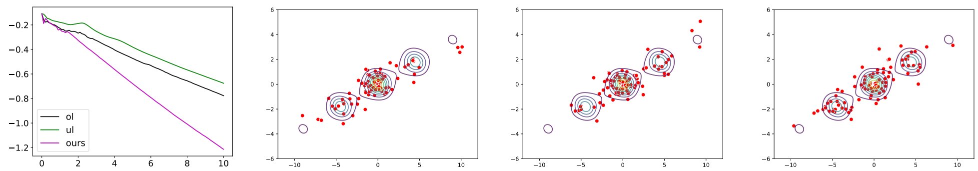
where $B = \boldsymbol{P}\,\mathrm{diag}(1, 25)\,\boldsymbol{P}^T$ for an orthogonal matrix $\boldsymbol{P}$ and $\boldsymbol{c} = \sqrt{0.95}\,(1, 1)^T$. Here $\boldsymbol{P}$ is generated by using torch.linalg.qr(torch.randn(d)) in Pytorch, where $d = 2$ is the dimension. We set $a = 1$, $\gamma_{ul} = 2m^{1/2}$ and $\gamma_{pdd} = 2m^{1/2} + m$ where we choose $m = 1/25$. We use time step size $h = 10^{-2}$ and run for 1000 steps. We use 5×10^5 particles and $n^2 = 2500$ bins to evaluate the KL divergence. The results are demonstrated in Figure 5.

4.4. Bimodal. We consider a two-dimensional bimodal distribution studied in [64] whose target density has the following form:

$$p(\boldsymbol{x}) \propto \exp\big(-2(\|\boldsymbol{x}\| - 3)^2\big)\Big[\exp\big(-2(x_1 - 3)^2\big) + \exp\big(-2(x_1 + 3)^2\big)\Big].$$

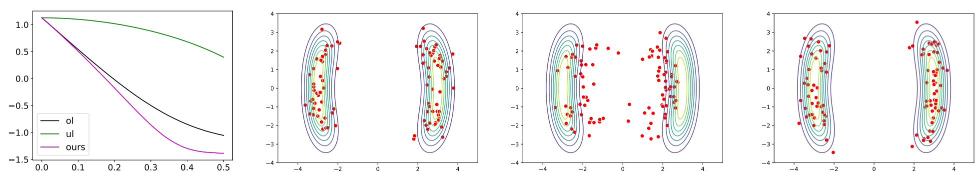
The corresponding potential function is given by

$$f(\boldsymbol{x}) = 2(\|\boldsymbol{x}\| - 3)^2 - 2\log\Big[\exp\big(-2(x_1 - 3)^2\big) + \exp\big(-2(x_1 + 3)^2\big)\Big].$$



(a) KL decay (b) ol (c) ul (d) gaul

Fig. 5: Convergence and scatter plots for the quadratic cosine example.



(a) KL decay (b) ol (c) ul (d) gaul

Fig. 6: Convergence and scatter plots for the bimodal example.

The gradient is

$$\nabla f(\boldsymbol{x}) = \frac{4(x_1-3)\exp\big(-2(x_1-3)^2\big)+4(x_1+3)\exp\big(-2(x_1+3)^2\big)}{\exp\big(-2(x_1-3)^2\big)+\exp\big(-2(x_1+3)^2\big)}\boldsymbol{e}_1$$
$$+4\frac{(\|\boldsymbol{x}\|-3)\boldsymbol{x}}{\|\boldsymbol{x}\|},$$

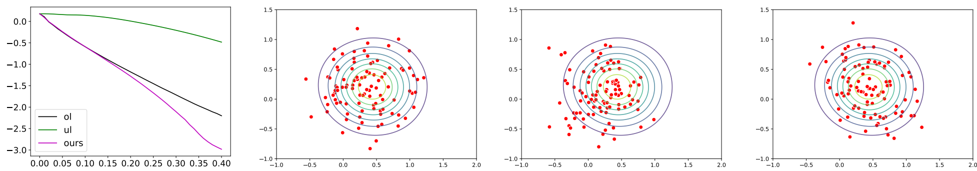
where $\boldsymbol{e}_1 = (1,0)^T$ is the first standard coordinate vector. We set $\gamma_{ul} = 2m^{1/2}$ and $\gamma_{pdd} = 2m^{1/2} + m$ where we choose $m = 1/2$. We use time step size $h = 10^{-3}$ and run for 500 iterations. We use 10^6 particles and $n^2 = 2500$ bins to evaluate the KL divergence. The results are shown in Figure 6.

4.5. Bayesian logistic regression. We consider the Bayesian logistic regression problem studied in [27, 20, 60]. We give a brief description of the problem. Suppose we are given a feature matrix $X \in \mathbb{R}^{n\times d}$ with rows $x_i \in \mathbb{R}^d$. Correspondingly we are given $Y \in \{0,1\}^n$ the binary response vector for each of the covariates in our feature matrix. The logistic model for the probability of $y_i = 1$ given $x_i \in \mathbb{R}^d$ and a parameter $\theta \in \mathbb{R}^d$ is

$$\mathbb{P}(y_i = 1|x_i,\theta) = \frac{\exp(\theta^T x_i)}{1+\exp(\theta^T x_i)}\,. \tag{4.4}$$

Suppose we impose a prior distribution on the parameter $\theta \sim \mathcal{N}(0,\Sigma_X)$, where $\Sigma_X = \frac{1}{n}X^TX$ is the sample covariance of X. Then the posterior distribution for θ can be calculated by

$$p(\theta|X,Y) \propto \exp\left[Y^TX\theta - \sum_{i=1}^n \log\big(1+\exp(\theta^T x_i)\big) - \frac{\alpha}{2}\theta^T\Sigma_X\theta\right],$$



(a) KL decay (b) ol (c) ul (d) gaul

Fig. 7: Convergence and scatter plots for Bayesian logistic regression.

where $\alpha > 0$ is a regularization parameter. The potential function is

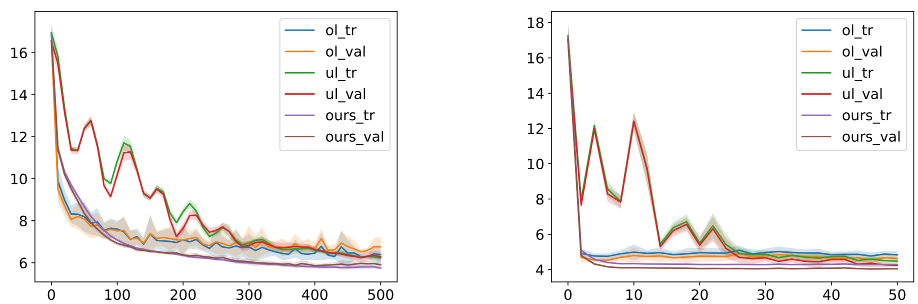
$$f(\theta) = -Y^T X\theta + \sum_{i=1}^{n} \log\left(1 + \exp(\theta^T x_i)\right) + \frac{\alpha}{2}\theta^T \Sigma_X \theta\,.$$

Its gradient is

$$\nabla f(\theta) = -X^T Y + \sum_{i=1}^{n} \frac{x_i}{1 + \exp(-\theta^T x_i)} + \alpha \Sigma_X \theta\,.$$

As shown in [27], the Hessian of f is upper bounded by $L = (0.25n + \alpha)\lambda_{\max}(\Sigma_X)$ and lower bounded by $m = \alpha\lambda_{\min}$. To generate X and Y, we set $x_{i,j}$ to be independent Rademacher random variables for each i and j. And each y_i is generated according to (4.4) with $\theta = \theta^* = (1, 1)^T$. We set $\alpha = 0.5$, $d = 2$, $n = 50$, $\gamma_{ul} = 2m^{1/2}$ and $\gamma_{pdd} = 2m^{1/2} + m$. To sample the posterior distribution, we use time step size $h = 10^{-3}$ and run for 400 iterations. The initial distribution of particles is $\mathcal{N}(0, L^{-1}\mathbf{I})$. As for evaluation metric, we directly evaluate the KL divergence between the sampled posterior and the true posterior. We use 10^6 particles and $n^2 = 2500$ bins to evaluate the KL divergence as before. This is different from the choice by [27] and [60], where [27] compared the samples with θ^*. [60] compared samples with the true minimizer of $f(\theta)$, i.e. the maximum a posteriori (MAP) estimate in the Bayesian optimization literature. We believe that directly measuring the KL divergence gives a better understanding of how 'close' our samples are to the true posterior distribution. The results are presented in Figure 7.

4.6. Bayesian neural network. In this section, we compare GAUL with overdamped ('ol') and underdamped Langevin ('ul') dynamics in training Bayesian neural network. We test a one-hidden-layer fully connected neural network with 50 hidden neurons and ReLU activation function on the UCI concrete dataset. We use $h = 10^{-3}$, $a = 0.1$, $\gamma = 0.5$. For each method, we sample $M = 20$ particles (each particle corresponds to a neural network) and take the average output as the final output. Figure 8a and Table 1 show the rMSE averaged over 10 experiments. We see that 'ul' can achieve smaller training and validation error than 'ol'. However, 'ul' also exhibits a slow start and an oscillatory behavior at the beginning of training as is commonly seen in acceleration methods in optimization. GAUL can get rid of the oscillation and achieve a even smaller training and validation error as is demonstrated in Table 1. We have also tested out the three methods using the Combined Cycle Power Plant (CCPP) dataset. We choose the same parameter as the concrete experiment. The results are presented in Figure 8b and Table 1.



(a) Concrete dataset

(b) CCPP dataset

Fig. 8: Convergence comparison. x-axis represents number of epochs. y-axis represents rMSE averaged over 10 experiments.

	ol	ul	gaul
concrete tr err	6.39 ± 0.44	6.23 ± 0.15	$\mathbf{5.74 \pm 0.06}$
concrete val err	6.76 ± 0.49	6.28 ± 0.24	$\mathbf{5.90 \pm 0.14}$
ccpp tr err	4.84 ± 0.22	4.48 ± 0.11	$\mathbf{4.28 \pm 0.03}$
ccpp val err	4.63 ± 0.25	4.25 ± 0.11	$\mathbf{4.04 \pm 0.04}$

Table 1: Training and validation rmse.

5. Conclusions. In this work, we introduced gradient-adjusted underdamped Langevin dynamics (GAUL) inspired by primal-dual damping dynamics and Hessian-driven damping dynamics. We demonstrated that GAUL admitted the correct stationary target distribution $\pi \propto \exp(-f)$ under appropriate conditions and achieves exponential convergence for quadratic functions, outperforming both the overdamped and underdamped Langevin dynamics in terms of convergence speed. Our numerical experiments further illustrate the practical advantages of GAUL, showcasing faster convergence and more efficient sampling compared to classical methods, such as overdamped and underdamped Langevin dynamics.

We also note a connection between the primal-dual damping dynamics and GAUL. A key challenge in the primal-dual damping algorithm is the design of preconditioner matrices, which can accelerate the algorithm's convergence compared to the gradient descent method. In the context of solving a linear problem where f is a quadratic function and the diffusion constant is zero, [67] demonstrates that the convergence rate depends on the square root of the smallest eigenvalue. In this paper, we extend the study from a sampling perspective, where f is also a quadratic function but the diffusion is non-zero. Towards a Gaussian target distribution, GAUL converges to a biased target distribution with the mixing time depending on $\sqrt{\kappa}$. This is in contrast with overdamped and underdamped Langevin sampling algorithms.

Several possible future directions are worth exploring. First, can we show that GAUL converges faster than overdamped and underdamped Langevin dynamics for more general potentials, which is beyond the current study of Gaussian distributions?

One common assumption is that the potential f is strongly log-concave [8, 17, 18, 19, 25, 27, 34, 38, 42]. Recently, [9] proved that for a class of distributions that satisfy a Poincaré-type inequality, underdamped Langevin dynamics converges in L_2 with rate $\exp(-\sqrt{m}t)$ where m is the Poincaré constant. Then it is interesting to study for the same class of distributions, whether GAUL could converge at an even faster rate. Another direction is to study the convergence of GAUL under different metrics. From a more practical perspective, designing new time discretization schemes [55, 16, 50, 60, 42] for implementing GAUL is also an important direction. We proved that using the Euler-Maruyama discretization, GAUL will converge to a biased target distribution, which is not surprising since ULA is also biased. Therefore, another promising direction could be to combine GAUL with MCMC methods [7, 27], such as Metropolis-Hastings algorithms, to design a hybrid method with accept/reject options so that the algorithm converges to the correct target distribution in the discrete-time update. Finally, choosing the preconditioner $\boldsymbol{C}$ to accelerate convergence is an important topic. The difficulty of picking $\boldsymbol{C}$ arises from the positive semidefinite constraint on $\mathrm{sym}(\mathbf{Q})$ in (2.16), which we should explore in future work.

Appendix A. Euler-Maruyama Discretization. The Euler-Maruyama scheme of (2.15) with step size h and $\boldsymbol{C} = \mathbf{I}$ reads

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - a\nabla f(\boldsymbol{x}_t)h + \boldsymbol{p}_t h + \sqrt{2ah}\boldsymbol{z}^{(1)}\,, \tag{A.1a}$$

$$\boldsymbol{p}_{t+1} = \boldsymbol{p}_t - \nabla f(\boldsymbol{x}_t)h - \gamma \boldsymbol{p}_t h + \sqrt{2\gamma h}\boldsymbol{z}^{(2)}\,. \tag{A.1b}$$

$\boldsymbol{z}^{(i)}$ is a standard Gaussian random variable for $i = 1, 2$.

Appendix B. A matrix lemma. Let $a \geq 0$, $s > 0$, $\gamma > 0$, and consider the 3×3 matrix

$$\mathbf{D} = \begin{pmatrix} -2as & -2\gamma s^{-1} & -s^{-1} \\ 0 & 0 & 1 \\ 2s^2 & 2(-1-a\gamma)s^{-1} - 2\gamma^2 & -3\gamma - as \end{pmatrix}. \tag{B.1}$$

A direct calculation shows that the eigenvalues are given by

$$\lambda_0(a,\gamma,s) = -as - \gamma\,, \tag{B.2a}$$

$$\lambda_-(a,\gamma,s) = -as - \gamma - \sqrt{\gamma^2 - 2a\gamma s + s(-4 + a^2 s)}\,, \tag{B.2b}$$

$$\lambda_+(a,\gamma,s) = -as - \gamma + \sqrt{\gamma^2 - 2a\gamma s + s(-4 + a^2 s)}\,. \tag{B.2c}$$

We have the following lemmas regarding the eigenvalues given by (B.2).

LEMMA B.1. *Let* $\mathbf{D}$ *be as* (B.1)*. If* $a = 0$*, then*

$$\arg\min_{\gamma > 0} \Re\big(\lambda_+(0,\gamma,s)\big) = 2\sqrt{s}\,.$$

Proof. We have that $\lambda_+(0,\gamma,s) = \frac{1}{2}\left(-\gamma + \sqrt{\gamma^2 - 4s}\right)$. If $\gamma \leq 2\sqrt{s}$, then $\Re\big(\lambda_+(0,\gamma,s)\big) \geq -\sqrt{s}$. When $\gamma \geq 2\sqrt{s}$, we have that $\Re\big(\lambda_+(0,\gamma,s)\big) = \lambda_+(0,\gamma,s)$. And $\frac{\partial}{\partial\gamma}\lambda_+(0,\gamma,s) \geq 0$. Therefore, the minimum of $\Re\big(\lambda_+(0,\gamma,s)\big)$ takes place at $\gamma = 2\sqrt{s}$. ∎

LEMMA B.2. *Let* $\mathbf{D}$ *be as* (B.1)*. Let* $\gamma > 0$ *be fixed. Then*

$$\arg\min_{a > 0} \Re\big(\lambda_+(a,\gamma,s)\big) = \frac{\gamma}{s} + \frac{2}{\sqrt{s}}\,. \tag{B.3}$$

Proof. Let us define $\Delta(a) = \gamma^2 - 2a\gamma s + s(a^2 s - 4)$. It can be seen that Δ is a quadratic function of a. The two roots of Δ are given by

$$a_\pm = \frac{\gamma}{s} \pm \frac{2}{\sqrt{s}}\,.$$

When $a \in [a_-, a_+]$, $\Delta(a) \le 0$ and

$$\Re\big(\lambda_+(a,\gamma,s)\big) = \frac{1}{2}(-\gamma - as) \ge \frac{1}{2}(-\gamma - a_+ s) = \Re\big(\lambda_+(a_+,\gamma,s)\big) = -\gamma - \sqrt{s}\,.$$

When $a < a_-$, we can calculate that

$$\frac{\partial}{\partial a}\lambda_+(a,\gamma,s) = -s + \frac{-\gamma s + as^2}{\sqrt{\Delta}} < 0\,.$$

This implies that $\lambda_+(a_- - \varepsilon, \gamma, s) > \lambda_+(a_-,\gamma,s)$ for any $\varepsilon > 0$. Similarly, when $a > a_+$, we have that $\frac{\partial}{\partial a}\lambda_+(a,\gamma,s) > 0$. Thus, $\lambda_+(a_+ + \varepsilon,\gamma,s) > \lambda_+(a_-,\gamma,s)$ for any $\varepsilon > 0$. Combining the above results, we conclude our proof. ∎

LEMMA B.3. *Let* $\mathbf{D}$ *be as* (B.1)*. Let* $a > 0$ *be fixed. Then*

$$\arg\min_{\gamma>0} \Re\big(\lambda_+(a,\gamma,s)\big) = as + 2\sqrt{s}\,. \tag{B.4}$$

Proof. The proof will be similar to that of Lemma B.2. This time we define $\Delta(\gamma) = \gamma^2 - 2a\gamma s + s(a^2 s - 4)$. It can be seen that $\Delta(\gamma)$ is a quadratic function of γ. The two roots of $\Delta(\gamma)$ are given by

$$\gamma_\pm = as \pm 2\sqrt{s}\,.$$

When $\gamma \in [\gamma_-, \gamma_+], \Delta(\gamma) < 0$ and

$$\Re\big(\lambda_+(a,\gamma,s)\big) = \frac{1}{2}(-\gamma - as) \ge \frac{1}{2}(-\gamma_+ - as) = \Re\big(\lambda_+(a,\gamma_+,s)\big) = -as - \sqrt{s}\,.$$

When $\gamma < \gamma_-$, we have

$$\begin{aligned}\frac{\partial}{\partial \gamma}\lambda_+(a,\gamma,s) &= -1 + \frac{\gamma - as}{\sqrt{(\gamma - as)^2 - 4s}}\\ &\le -1 < 0\,,\end{aligned}$$

since $\gamma - as < 0$. When $\gamma > \gamma_+$, we have

$$\begin{aligned}\frac{\partial}{\partial \gamma}\lambda_+(a,\gamma,s) &= -1 + \frac{\gamma - as}{\sqrt{(\gamma - as)^2 - 4s}}\\ &\ge -1 + 1 = 0\,.\end{aligned}$$

Combining the above arguments, we conclude that the optimal γ is γ_+. ∎

We now turn to a more general setting. Let $a \ge 0$, $\gamma > 0$ and define

$$\mathcal{D} = \begin{pmatrix} -2a\boldsymbol{S} & -2\gamma \boldsymbol{S}^{-1} & -\boldsymbol{S}^{-1} \\ 0 & 0 & \mathbf{I} \\ 2\boldsymbol{S}^2 & 2(-1 - a\gamma)\boldsymbol{S}^{-1} - 2\gamma^2 \mathbf{I} & -3\gamma \mathbf{I} - a\boldsymbol{S} \end{pmatrix}, \tag{B.5}$$

where now $\boldsymbol{S}$ is a diagonal matrix whose diagonal is given by $s_1 \ge s_2 \ge \ldots \ge s_d > 0$. And $\mathbf{I}$ is the identity matrix. Just like Lemma B.1, Lemma B.2, and Lemma B.7 we want to characterize the eigenvalues of $\mathcal{D}$. In particular, we would like to characterize the largest real part of the eigenvalue of $\mathcal{D}$ in terms of a and γ.

Proposition B.4. *The eigenvalues for $\mathcal{D}$ are given by*

$$\lambda_0^{(i)}(a,\gamma,\boldsymbol{S}) = -as_i - \gamma\,, \tag{B.6a}$$

$$\lambda_-^{(i)}(a,\gamma,\boldsymbol{S}) = -as_i - \gamma - \sqrt{\gamma^2 - 2a\gamma s_i + s_i(-4+a^2 s_i)}\,, \tag{B.6b}$$

$$\lambda_+^{(i)}(a,\gamma,\boldsymbol{S}) = -as_i - \gamma + \sqrt{\gamma^2 - 2a\gamma s_i + s_i(-4+a^2 s_i)}\,, \tag{B.6c}$$

for $i = 1,\dots,d$. The corresponding eigenvectors are sparse and take the following form. (Here we only write out the non-zero part of the eigenvectors)

$$v_{0,i}^{(i)} = \frac{-1}{s_i(\gamma + as_i)}\,, \tag{B.7a}$$

$$v_{0,i+d}^{(i)} = \frac{-1}{\gamma + as_i}\,, \tag{B.7b}$$

$$v_{0,i+2d}^{(i)} = 1\,, \tag{B.7c}$$

$$v_{-,i}^{(i)} = \frac{2\gamma - \sqrt{\gamma^2 - 2a\gamma s_i + s_i(a^2 s_i - 4)} - \frac{2(\gamma^2 + s_i + a\gamma s_i)}{\gamma + as_i + \sqrt{\gamma^2 - 2a\gamma s_i + s_i(a^2 s_i - 4)}}}{2s_i^2}\,, \tag{B.8a}$$

$$v_{-,i+d}^{(i)} = \frac{-1}{\gamma + as_i + \sqrt{\gamma^2 - 2a\gamma s_i + s_i(a^2 s_i - 4)}}\,, \tag{B.8b}$$

$$v_{-,i+2d}^{(i)} = 1\,, \tag{B.8c}$$

$$v_{+,i}^{(i)} = \frac{2\gamma + \sqrt{\gamma^2 - 2a\gamma s_i + s_i(a^2 s_i - 4)} - \frac{2(\gamma^2 + s_i + a\gamma s_i)}{\gamma + as_i - \sqrt{\gamma^2 - 2a\gamma s_i + s_i(a^2 s_i - 4)}}}{2s_i^2}\,, \tag{B.9a}$$

$$v_{+,i+d}^{(i)} = \frac{-1}{\gamma + as_i - \sqrt{\gamma^2 - 2a\gamma s_i + s_i(a^2 s_i - 4)}}\,, \tag{B.9b}$$

$$v_{+,i+2d}^{(i)} = 1\,. \tag{B.9c}$$

In the above, $v_{*,j}^{(i)}$ represents the j-th component of the eigenvector corresponding to the eigenvalue $\lambda_*^{(i)}$, where $* \in \{0,+,-\}$.

Moreover, when γ is chosen according to Lemma B.7*, we have a defective eigenvalue $\lambda_0^{(d)} = \lambda_\pm^{(d)} = -as_d - \gamma$, which is accompanied with two generalized eigenvectors η, ξ that satisfy $(\mathcal{D} - \lambda_0^{(d)})\eta = v_0^{(d)}$, $(\mathcal{D} - \lambda_0^{(d)})\xi = v_0^{(d)}$. In details, the nonzero components of $v_0^{(d)}$, η and ξ are given by*

$$v_{0,d}^{(d)} = \frac{-1}{s_d(\gamma + as_d)}\,, \tag{B.10a}$$

$$v_{0,2d}^{(d)} = \frac{-1}{\gamma + as_d}\,, \tag{B.10b}$$

$$v_{0,3d}^{(d)} = 1\,, \tag{B.10c}$$

$$\eta_d = \frac{\gamma - as}{2s_d^2}\,, \tag{B.11a}$$

$$\eta_{3d} = 1\,, \tag{B.11b}$$

$$\xi_d = \frac{\gamma^2 - (1 + a\gamma)s_d}{s_d^2}\,, \tag{B.12a}$$

$$\xi_{2d} = 1\,. \tag{B.12b}$$

Proof. One can directly verify that the above computation gives an eigensystem for $\mathcal{D}$. ∎

From the sparsity structure of $v_\pm^{(j)}$ and $v_0^{(j)}$, we immediately have the following corollary.

COROLLARY B.5. *$v_*^{(j)}$ is orthogonal to $v_\star^{(k)}$ for $*, \star \in \{0, +, -\}$ if $j \neq k$.*

LEMMA B.6. *Let $\mathcal{D}$ be as* (B.5)*. If $a = 0$, then*

$$\arg\min_{\gamma > 0} \max_j \Re(\lambda_+^{(j)}(0, \gamma, \boldsymbol{S})) = 2\sqrt{s_d}\,. \tag{B.13}$$

Proof. Plugging $a = 0$ into (B.6) we have

$$\lambda_+^{(j)}(0, \gamma, \boldsymbol{S}) = \frac{1}{2}\left(-\gamma + \sqrt{\gamma^2 - 4s_j}\right)\,.$$

We first note that since $s_d \le s_{d-1} \le \ldots \le s_1$, if $\gamma \le 2\sqrt{s_d}$ then $\Re(\lambda_+^{(j)}(0, \gamma, \boldsymbol{S})) = -\gamma/2$ for all $1 \le j \le d$. In particular, this implies that

$$\arg\min_{0 < \gamma \le 2\sqrt{s_d}} \max_j \Re(\lambda_+^{(j)}(0, \gamma, \boldsymbol{S})) = 2\sqrt{s_d}\,.$$

We then need to show that if $\gamma > 2\sqrt{s_d}$, $\max_j \Re(\lambda_+^{(j)}(0, \gamma, \boldsymbol{S})) > -\sqrt{s_d}$. This will be very similar to the argument in the proof of Lemma B.1. Now consider $\gamma > 2\sqrt{s_d}$. We showed in the proof of Lemma B.1 that $\Re\big(\lambda_+^{(n)}(0, \gamma, \boldsymbol{S})\big) = \lambda_+^{(n)}(0, \gamma, \boldsymbol{S})$. And $\frac{\partial}{\partial\gamma}\lambda_+^{(n)}(0, \gamma, \boldsymbol{S}) \ge 0$. Hence, we have

$$\max_j \Re(\lambda_+^{(j)}(0, \gamma, \boldsymbol{S})) > \Re(\lambda_+^{(n)}(0, \gamma, \boldsymbol{S})) = \lambda_+^{(n)}(0, \gamma, \boldsymbol{S}) \ge \lambda_+^{(n)}(0, 2\sqrt{s_d}, \boldsymbol{S}) = -\sqrt{s_d}\,.$$

This concludes our proof. ∎

LEMMA B.7. *Let $\mathcal{D}$ be as* (B.5)*. Let $a > 0$. Then*

$$\arg\min_{\gamma > 0} \max_j \Re(\lambda_+^{(j)}(a, \gamma, \boldsymbol{S})) = as_d + 2\sqrt{s_d}\,. \tag{B.14}$$

Proof. Let us define $\Delta(\gamma, s) = \gamma^2 - 2a\gamma s + s(a^2 s - 4)$. A straightforward calculation shows that the two roots of $\Delta(\gamma, s_j)$ (when viewing Δ as a function of γ) are given by

$$\gamma_\pm^{(j)} = as_j \pm 2\sqrt{s_j}\,.$$

We have shown in Lemma B.3 that

$$\arg\min_{\gamma > 0} \Re(\lambda_+^{(d)}(a, \gamma, \boldsymbol{S})) = as_d + 2\sqrt{s_d}\,.$$

Denote by $\gamma^*(a) = as_d + 2\sqrt{s_d}$. Let us consider $\tilde{s} > s_d$. If $\Delta(\gamma^*(a), \tilde{s}) \leq 0$, then we have

$$\begin{aligned} \Re\left(-\gamma^*(a) - a\tilde{s} + \sqrt{\gamma^*(a)^2 - 2a\gamma^*(a)\tilde{s} + \tilde{s}(a^2\tilde{s} - 4)}\right) &= -\gamma^*(a) - a\tilde{s} \\ &\leq -\gamma^*(a) - as_d \\ &= \Re(\lambda_+^{(d)}(a, \gamma^*(a), \boldsymbol{S}))\,, \end{aligned} \tag{B.15}$$

where the last line follows from $\Delta(\gamma^*(a), s_d) = 0$ by definition of $\gamma^*(a)$. If $\Delta(\gamma^*(a), \tilde{s}) > 0$, we compute

$$\begin{aligned} &\frac{\partial}{\partial s}\left(-\gamma^*(a) - as + \sqrt{\gamma^*(a)^2 - 2a\gamma^*(a)s + s(a^2 s - 4)}\right)|_{s=\tilde{s}} \\ &= -a + \frac{-a\gamma^*(a) + a^2\tilde{s} - 2}{\sqrt{\gamma^*(a)^2 - 2a\gamma^*(a)\tilde{s} + \tilde{s}(a^2\tilde{s} - 4)}} > 0\,. \end{aligned} \tag{B.16}$$

We now verify that the above derivative is indeed positive. First observe that given $\tilde{s} > s_d$, the two roots for $\Delta(\gamma, \tilde{s})$ are

$$\tilde{\gamma}_\pm = a\tilde{s} \pm 2\sqrt{\tilde{s}}\,.$$

Clearly, $\tilde{\gamma}_+ > \gamma^*(a)$. Hence, $\Delta(\gamma^*(a), \tilde{s}) > 0$ implies that $\gamma^*(a) < \tilde{\gamma}_-$, or equivalently $\tilde{s} > s_d + (2\sqrt{s_d} + 2\sqrt{\tilde{s}})/a$. This further implies $\sqrt{\tilde{s}} > 2/a$. Therefore,

$$\begin{aligned} -a\gamma^*(a) + a^2\tilde{s} - 2 &> a^2(s_d + (2\sqrt{s_d} + 2\sqrt{\tilde{s}})/a) - a\gamma^*(a) - 2 \\ &= 2a\sqrt{\tilde{s}} - 2 \\ &> 2a\frac{2}{a} - 2 > 0\,. \end{aligned}$$

Knowing that the numerator in the second term of (B.16) is positive, we know that (B.16) is positive if and only if

$$(-a\gamma^*(a) + a^2\tilde{s} - 2)^2 > a^2(\gamma^*(a)^2 - 2a\gamma^*(a)\tilde{s} + \tilde{s}(a^2\tilde{s} - 4))\,,$$

which can be verified by expanding the square on the left hand side and comparing with the right hand side directly.

Since the derivative in (B.16) is positive, let us examine the limit

$$\begin{aligned} &\lim_{s\to\infty} -\gamma^*(a) - as + \sqrt{\gamma^*(a)^2 - 2a\gamma^*(a)s + s(a^2 s - 4)} \\ &= \lim_{s\to\infty} -\gamma^*(a) - as + s\sqrt{\gamma^*(a)^2 s^{-2} - 2a\gamma^*(a)s^{-1} + a^2 - 4s^{-1}} \\ &= \lim_{s\to\infty} -\gamma^*(a) - as + as - (\gamma^*(a) + \frac{2}{a}) + \mathcal{O}(s^{-1}) \\ &= -2\gamma^*(a) - \frac{2}{a} \\ &= -2(as_d + 2\sqrt{s_d}) - \frac{2}{a} < \Re(\lambda_+^{(d)}(a, \gamma^*(a), \boldsymbol{S}))\,. \end{aligned} \tag{B.17}$$

□

Combining (B.15), (B.16) and (B.17), we obtain that for $1 \leq j \leq d$

$$\Re(\lambda_+^{(j)}(a, \gamma^*(a), \boldsymbol{S})) \leq \lambda_+^{(d)}(a, \gamma^*(a), \boldsymbol{S}) = \Re(\lambda_+^{(d)}(a, \gamma^*(a), \boldsymbol{S}))\,,$$

which implies

$$\min_{\gamma>0}\max_{j}\Re(\lambda_+^{(j)}(a,\gamma,\boldsymbol{S}))\leq\max_{j}\Re(\lambda_+^{(j)}(a,\gamma^*(a),\boldsymbol{S}))=\Re(\lambda_+^{(d)}(a,\gamma^*(a),\boldsymbol{S}))\,.$$

Finally, by Lemma B.3 again, we have

$$\min_{\gamma>0}\max_{j}\Re(\lambda_+^{(j)}(a,\gamma,\boldsymbol{S}))\geq\min_{\gamma>0}\Re(\lambda_+^{(d)}(a,\gamma,\boldsymbol{S}))=\Re(\lambda_+^{(d)}(a,\gamma^*(a),\boldsymbol{S}))\,.$$

We now conclude that

$$\arg\min_{\gamma>0}\max_{j}\Re(\lambda_+^{(j)}(a,\gamma,\boldsymbol{S}))=\gamma^*(a)\,.$$

LEMMA B.8. *The constant C_1 in Equation* (3.9) *depends at most polynomially on d, s_1, $1/s_d$, i.e. $C_1=\mathrm{poly}(d,s_1,s_d^{-1})\leq\mathrm{poly}(d,\kappa)$.*

Proof. First, we show that C_1 depends linearly on the dimension d. Let us recall the following fact from linear ODE: if $\dot{x}=Ax$ for some constant matrix $A\in\mathbb{R}^{d\times d}$, with eigenvalues $\lambda_1,\ldots,\lambda_d$ and eigenvectors $v_1,\ldots,v_d$, then the solution is of the form $x(t)=\sum_i a_ie^{\lambda_it}v_i$. In case there are repeated eigenvalues (e.g. λ_i) and generalized eigenvectors, the corresponding term in the sum will be replaced with some $\sum_j b_jt^{k-j}e^{\lambda_it}v_i$ where the sum is over $j=1,\ldots,k$ and k is the dimension of the generalized eigenspace associated with λ_i. Let $\mathcal{D}$ and $\mathbf{T}$ be as defined in (3.8). By our choice of γ, we know that eigenvalues of $\mathcal{D}$ are nonzero. Therefore, $\mathcal{D}$ is invertible. Denote by

$$\mathbf{Y}(t)=\begin{pmatrix}\Sigma_{11}(t)\\ \Sigma_{22}(t)\\ \dot{\Sigma}_{22}(t)\end{pmatrix}+\mathcal{D}^{-1}\mathbf{T}\,.$$

Then (3.8) reads

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{Y}=\mathcal{D}\mathbf{Y}\,. \tag{B.18}$$

We follow the notation in Proposition B.4 and use $(\lambda_*^{(i)},v_*^{(i)})$ to represent an eigenvalue eigenvector pair of $\mathcal{D}$, for $i=1,\ldots,d$, and $*\in\{0,+,-\}$. Note that for our choice of $\gamma=as_d+2\sqrt{s_d}$, we have $\lambda_0^{(d)}=\lambda_\pm^{(d)}$. Correspondingly, there will be generalized eigenvectors. Following the notation in Proposition B.4, we use $v_0^{(d)}$ to represent the eigenvector associated with $\lambda_0^{(d)}$; and we use η and ξ to represent the generalized eigenvectors associated with $\lambda_0^{(d)}$. We have already shown in Proposition B.4 that both η and ξ are generalized eigenvector of rank 2. Therefore, the solution to (B.18) takes the form

$$\begin{aligned}\mathbf{Y}(t)=&\left(\sum_{i=1}^{d-1}\sum_{*\in\{0,+,-\}}\alpha_*^{(i)}e^{\lambda_*^{(i)}t}v_*^{(i)}\right)+\alpha_0^{(d)}e^{\lambda_0^{(d)}t}v_0^{(d)}+\alpha_-^{(d)}e^{\lambda_0^{(d)}t}(tv_0^{(d)}+\eta)\\ &+\alpha_+^{(d)}e^{\lambda_0^{(d)}t}(tv_0^{(d)}+\xi)\,,\end{aligned} \tag{B.19}$$

where the constants $\alpha_*^{(i)}$ are to be determined by $\mathbf{Y}(0)$. By Lemma B.7 and our choice of γ, we have that

$$\max_{i\leq d}\max_{*\in\{0,+,-\}}\Re(\lambda_*^{(i)})=\lambda_0^{(d)}=-2as_d-2\sqrt{s_d}\,.$$

Without loss of generality, consider $t \geq 1$. We have

$$
\begin{aligned}
\|\mathbf{Y}(t)\|^2 &= \left\| \left(\sum_{i=1}^{d-1} \sum_{* \in \{0,+,-\}} \alpha_*^{(i)} e^{\lambda_*^{(i)} t} v_*^{(i)} \right) + \alpha_0^{(d)} e^{\lambda_0^{(d)} t} v_0^{(d)} + \alpha_-^{(d)} e^{\lambda_0^{(d)} t} (t v_0^{(d)} + \eta) \right. \\
&\qquad \left. + \alpha_+^{(d)} e^{\lambda_0^{(d)} t} (t v_0^{(d)} + \xi) \right\|^2 \\
&= \sum_{i=1}^{d-1} \left\| \sum_{* \in \{0,+,-\}} \alpha_*^{(i)} e^{\lambda_*^{(i)} t} v_*^{(i)} \right\|^2 + \left\| \alpha_0^{(d)} e^{\lambda_0^{(d)} t} v_0^{(d)} + \alpha_-^{(d)} e^{\lambda_0^{(d)} t} (t v_0^{(d)} + \eta) \right. \\
&\qquad \left. + \alpha_+^{(d)} e^{\lambda_0^{(d)} t} (t v_0^{(d)} + \xi) \right\|^2 \\
&\leq \sum_{i=1}^{d-1} \sum_{* \in \{0,+,-\}} 3 \left\| \alpha_*^{(i)} e^{\lambda_*^{(i)} t} v_*^{(i)} \right\|^2 + 3 \left\| \alpha_0^{(d)} e^{\lambda_0^{(d)} t} v_0^{(d)} \right\|^2 + 3 \left\| \alpha_-^{(d)} e^{\lambda_0^{(d)} t} (t v_0^{(d)} + \eta) \right\|^2 \\
&\qquad + 3 \left\| \alpha_+^{(d)} e^{\lambda_0^{(d)} t} (t v_0^{(d)} + \xi) \right\|^2 \\
&\leq 3 t^2 e^{2\lambda_0^{(d)} t} \left[\left(\sum_{i=1}^{d-1} \sum_{* \in \{0,+,-\}} \left\| \alpha_*^{(i)} v_*^{(i)} \right\|^2 \right) + \left\| v_0^{(d)} \right\|^2 \left((\alpha_0^{(d)})^2 + 2(\alpha_-^{(d)})^2 + 2(\alpha_+^{(d)})^2 \right) \right. \\
&\qquad \left. + 2 \|\eta\|^2 (\alpha_-^{(d)})^2 + 2 \|\xi\|^2 (\alpha_+^{(d)})^2 \right] \\
&\leq 6 t^2 e^{2\lambda_0^{(d)} t} \left[\left(\sum_{i=1}^{d-1} \sum_{* \in \{0,+,-\}} \left\| \alpha_*^{(i)} v_*^{(i)} \right\|^2 \right) + \left\| v_0^{(d)} \right\|^2 \left((\alpha_0^{(d)})^2 + (\alpha_-^{(d)})^2 + (\alpha_+^{(d)})^2 \right) \right. \\
&\qquad \left. + \|\eta\|^2 (\alpha_-^{(d)})^2 + \|\xi\|^2 (\alpha_+^{(d)})^2 \right] \\
&\leq 6 t^2 e^{2\lambda_0^{(d)} t} \left[\left(\sum_{i=1}^{d-1} \sum_{* \in \{0,+,-\}} \left\| \alpha_*^{(i)} v_*^{(i)} \right\|^2 \right) + \left(\left\| \alpha_+^{(d)} v_0^{(d)} \right\|^2 + \left\| \alpha_-^{(d)} \eta \right\|^2 + \left\| \alpha_+^{(d)} \xi \right\|^2 \right) \right. \\
&\qquad \left. \left(1 + \frac{\left\| v_0^{(d)} \right\|^2}{\|\xi\|^2} + \frac{\left\| v_0^{(d)} \right\|^2}{\|\eta\|^2} \right) \right] \\
&\leq 6 t^2 e^{2\lambda_0^{(d)} t} \left(1 + \frac{\left\| v_0^{(d)} \right\|^2}{\|\xi\|^2} + \frac{\left\| v_0^{(d)} \right\|^2}{\|\eta\|^2} \right) \left[\sum_{i=1}^{d-1} \sum_{* \in \{0,+,-\}} \left\| \alpha_*^{(i)} v_*^{(i)} \right\|^2 \right. \\
&\qquad \left. + \left\| \alpha_+^{(d)} v_0^{(d)} \right\|^2 + \left\| \alpha_-^{(d)} \eta \right\|^2 + \left\| \alpha_+^{(d)} \xi \right\|^2 \right].
\end{aligned} \tag{B.20}
$$

$$
\|\mathbf{Y}(0)\|^2 = \sum_{i=1}^{d-1} \left\| \sum_{* \in \{0,+,-\}} \alpha_*^{(i)} v_*^{(i)} \right\|^2 + \left\| \alpha_0^{(d)} v_0^{(d)} + \alpha_-^{(d)} \eta + \alpha_+^{(d)} \xi \right\|^2 .
$$

Denote by $\mathbf{Y}(0)^{(i)}$ the projection of $\mathbf{Y}(0)$ onto the subspace $\Phi_i = \mathrm{Span}(\{v_0^{(i)}, v_+^{(i)}, v_-^{(i)}\})$. And accordingly, $\Phi_d = \mathrm{Span}(\{v_0^{(d)}, \eta, \xi\})$. By Corollary B.5, we know that Φ_i is orthogonal to Φ_j for $i \neq j$. Therefore, $|\alpha_*^{(i)}|$ depends on the inverse of the Gram matrix of $\{v_0^{(i)}, v_+^{(i)}, v_-^{(i)}\}$ as well as $\|\mathbf{Y}(0)^{(i)}\|$. This inverse Gram matrix can be computed analytically since it is a 3 by 3 matrix for each $1 \le i \le d$. However, the exact computation does not add more insights to the proof and we will not include the computation. Since each eigenvector and generalized eigenvector depends on $\{s_1, \dots s_d, s_1^{-1}, \dots s_d^{-1}\}$ polynomially, we know that the inverse of the Gram matrix also also depends on $\{s_1, \dots s_d, s_1^{-1}, \dots s_d^{-1}\}$ polynomially. From (B.20), we conclude that

$$\|\mathbf{Y}(t)\|^2 = \mathcal{O}\big(t^2 e^{2\lambda_0^{(d)} t} d^2 \cdot \mathrm{poly}(s_1, s_d^{-1})\big) = \mathcal{O}\big(t^2 e^{2\lambda_0^{(d)} t} d^2 \cdot \mathrm{poly}(\kappa)\big)\,.$$

LEMMA B.9. *Suppose $X \in \mathbb{S}^n$ satisfies $X = AXA^T$ for some $A \in \mathbb{R}^n$. If all eigenvalues of A has absolute value less than 1, then X is the zero matrix.*

Proof. Let us first assume that A^T is diagonalizable: $A^T = QDQ^{-1}$, where D is a diagonal matrix of eigenvalues $d_1, \dots, d_n$, and the columns of Q contains the eigenvectors $q_1, \dots, q_n$. Then it follows that

$$|q_i^T X q_j| = |d_i d_j| |q_i^T X q_j|\,.$$

This implies $|q_i^T X q_j| = 0$ for all $1 \le i, j \le n$, since $|d_i d_j| < 1$ by assumption. Now suppose that A has some generalized eigenvalues. Without loss of generality, assume that d_{n-1} is a generalized eigenvalue such that $A^T q_{n-1} = d_{n-1} q_{n-1}$ and $A^T q_n = d_{n-1} q_n + q_{n-1}$. Let q_i be an eigenvector. Then we still have $q_i^T X q_{n-1} = 0$ as before. And

$$|q_i^T X q_n| = |d_i d_{n-1} q_i^T X q_n + d_i q_i^T X q_{n-1}| = |d_i d_{n-1} q_i^T X q_n| = |d_i d_{n-1}|\, |q_i^T X q_n|\,.$$

Again this implies $|q_i^T X q_n| = 0$. The case where d_{n-1} has algebraic multiplicity greater than 2 or q_i is a generalized eigenvector can be proved in a similar fashion. Therefore, we have shown that if A has Jordan decomposition $A = PJP^{-1}$, then $q_i^T X q_j = 0$ where q_i and q_j are the i-th and j-th column of P. Equivalently, we have $P^T X P = 0$. This proves that $X = 0$. ◻

COROLLARY B.10. *Suppose $X, Y \in \mathbb{S}^n$ satisfy $X = AXA^T + B$, $Y = AYA^T + B$ for some $B \in \mathbb{S}^n$. If all eigenvalues of A have absolute value less than 1, then $X = Y$.*

Proof. The proof follows by Lemma B.9 and that $X - Y = A(X - Y)A^T$. ◻

Taking inspiration from system of linear ODE, we have the following lemma regarding the solution to the iteration $X_{k+1} = AX_kA^T$.

LEMMA B.11. *Let $A \in \mathbb{R}^{n\times n}$ be given by $A = \boldsymbol{I} - h\tilde{G}$ for some $\tilde{G} \in \mathbb{R}^{n\times n}$, $h > 0$. Suppose $\tilde{G}$ has Jordan decomposition $\tilde{G} = PJP^{-1}$. And consider the iteration $X_{k+1} = AX_kA^T$. If q_i is an eigenvector of $\tilde{G}$ with associated eigenvalue d_i and $X_0 = q_i q_i^T$, then $X_k = (1 - hd_i)^{2k} X_0$. Moreover, if q_i is a generalized eigenvector of $\tilde{G}$ of algebraic multiplicity 2, i.e. $\tilde{G} q_i = d_j q_i + q_j$ for some eigenvector q_j and eigenvalue d_j, and $X_0 = q_i q_i^T$, then $X_k = \big((1 - hd_j)^k q_i - kh(1 - hd_j)^{k-1} q_j\big)\big((1 - hd_j)^k q_i - kh(1 - hd_j)^{k-1} q_j\big)^T$*

LEMMA B.12. *The eigenvalues of $\boldsymbol{G}$ in* (3.11) *are given by the following*

$$\tilde{\lambda}_{\pm}^{(i)} = h\,\frac{(as_i + \gamma) \pm \sqrt{(as_i - \gamma)^2 - 4s_i}}{2}\,. \tag{B.21}$$

Proof. The proof follows by a direct computation. ∎

LEMMA B.13. *Consider $\gamma = \gamma^* = as_d + 2\sqrt{s_d}$. Let $s > s_d$. Then $a \le \frac{2}{\sqrt{s}-\sqrt{s_d}}$ if and only if $(as - \gamma^*)^2 - 4s \le 0$.*

Proof. Multiplying by $s - s_d$, we obtain

$$a \le \frac{2}{\sqrt{s} - \sqrt{s_d}} \iff a(s - s_d) \le 2\sqrt{s} + 2\sqrt{s_d} \iff as - 2\sqrt{s} \le \gamma^* \,.$$

And it is straightforward to verify that $2\sqrt{s} > -as + \gamma^*$ always holds. Squaring on both hand sides completes the proof. ∎

LEMMA B.14. *Consider $\tilde{\lambda}_\pm^{(i)}$ given by* (B.21)*. Suppose $a \ge \frac{2}{\sqrt{s_1}-\sqrt{s_d}}$. If the step size h satisfies $0 < h \le 1/(as_1 + \gamma)$, and $\gamma = \gamma^* = as_d + 2\sqrt{s_d}$, then*

$$\max_i |1 - \tilde{\lambda}_\pm^{(i)}| \le 1 - \frac{h}{2}(as_d + \sqrt{s_d}) \,.$$

Proof. Observe that the eigenvalues given in (B.21) is almost the same as the eigenvalues given in (B.6) except for an extra factor of $h/2$. This allows us to use previous lemma regarding the eigenvalues from (B.6). We consider two cases. Define

$$j = \inf \left\{ n : a \le \frac{2}{\sqrt{s_n} - \sqrt{s_d}} \right\} \,.$$

Case 1: Consider $i \le j-1$ (if $j = 1$, we directly consider Case 2). Then $a \ge \frac{2}{\sqrt{s_i}-\sqrt{s_d}}$. By Lemma B.13 and our assumption on a, we have $(as_i - \gamma^*)^2 - 4s_i \ge 0$. Then, one can verify that $0 < h \le \frac{1}{as_1+\gamma^*}$ is a sufficient condition for $1 - \tilde{\lambda}_\pm^{(i)} > 0$. Indeed, we compute

$$\begin{aligned}
\tilde{\lambda}_\pm^{(i)} &\le \frac{1}{as_1 + \gamma^*} \frac{(as_i + \gamma^*) + \sqrt{(as_i - \gamma^*)^2 - 4s_i}}{2} \\
&\le \frac{1}{as_1 + \gamma^*} \frac{(as_i + \gamma^*) + \sqrt{(as_i + \gamma^*)^2}}{2} \\
&= \frac{as_i + \gamma^*}{as_1 + \gamma^*} \\
&\le 1 \,. \qquad \text{(B.22)}
\end{aligned}$$

Moreover, we clearly have $\tilde{\lambda}_\pm^{(i)} > 0$. Therefore, $|1 - \tilde{\lambda}_\pm^{(i)}| \le 1$. On the other hand, by (B.16) and (B.17), we have that

$$\begin{aligned}
\tilde{\lambda}_\pm^{(i)} &\ge \lim_{s\to\infty} \frac{h}{2}\Big((as + \gamma^*) + \sqrt{(as - \gamma^*)^2 - 4s}\Big) \\
&= \frac{h}{2}\Big(2\gamma^* + \frac{2}{a}\Big) \\
&\ge h\gamma^* \,.
\end{aligned}$$

Therefore,

$$\max_{i \le j-1} |1 - \tilde{\lambda}_\pm^{(i)}| \le 1 - h(as_d + 2\sqrt{s_d}) \,.$$

Case 2: Consider $i \geq j$. Note that for a complex number $z = z_1 + iz_2$ and $h > 0$, we have that

$$|1 - hz|^2 = (1 - hz_1)^2 + h^2 z_2^2 \leq 1 - hz_1 \leq (1 - hz_1/2)^2\,,$$

where the first inequality holds if and only if $h \leq z_1/(z_1^2 + z_2^2)$. Therefore, we have

$$|1 - \tilde{\lambda}_{\pm}^{(i)}|^2 \leq \Big(1 - \frac{\Re(\tilde{\lambda}_{\pm}^{(i)})}{2}\Big)^2\,,$$

if

$$h \leq \frac{2(as_i + \gamma^*)}{(as_i + \gamma^*)^2 + 4s_i - (as_i - \gamma^*)^2} = \frac{as_i + \gamma^*}{2as_i\gamma^* + 2s_i}\,.$$

We now verify that $h \leq 1/(as_1 + \gamma^*)$ is a sufficient condition. We have

$$\frac{1}{as_1 + \gamma^*} \leq \frac{as_i + \gamma^*}{2as_i\gamma^* + 2s_i} \iff a^2 s_1 s_i + \gamma^* a s_1 + (\gamma^*)^2 \geq as_i\gamma^* + 2s_i\,.$$

By Lemma B.13, we have that

$$\begin{aligned}
a^2 s_1^2 + (\gamma^*)^2 - 2as_1\gamma^* &\geq 4s_1 \\
a^2 s_1 + \frac{(\gamma^*)^2}{s_1} - 2a\gamma^* &\geq 4 \\
a^2 s_1 &\geq 4 + 2a\gamma^* - \frac{(\gamma^*)^2}{s_1}\,.
\end{aligned}$$

Then

$$\begin{aligned}
a^2 s_1 s_i + \gamma^* a s_1 + (\gamma^*)^2 &\geq s_i\Big(4 + 2a\gamma^* - \frac{(\gamma^*)^2}{s_1}\Big) + \gamma^* a s_1 + (\gamma^*)^2 \\
&= 4s_i + 2a\gamma^* s_i - \frac{s_i(\gamma^*)^2}{s_1} + \gamma^* a s_1 + (\gamma^*)^2 \\
&\geq 4s_i + 2a\gamma^* s_i + \gamma^* a s_1 \\
&> as_i\gamma^* + 2s_i\,.
\end{aligned}$$

This shows that $h \leq 1/(as_1 + \gamma^*)$ is sufficient. By Lemma B.7 and our choice of h, we obtain that

$$\max_{i \geq j} |1 - \tilde{\lambda}_{\pm}^{(i)}| < \max_i 1 - \frac{\Re(\tilde{\lambda}_{\pm}^{(i)})}{2} \leq 1 - \frac{h}{2}(as_d + \sqrt{s_d})\,.$$

Combining the two cases, we complete the proof. ∎

Lemma B.15. *Consider* $\tilde{\lambda}_{\pm}^{(i)}$ *given by* (B.21). *Suppose* $a = 0$ *and* $\gamma = \gamma^* = 2\sqrt{s_d}$. *Then*

$$\max_i |1 - \tilde{\lambda}_{\pm}^{(i)}| \leq 1\,,$$

if and only if $h \leq 2\sqrt{s_d}/s_1$.

Proof. We directly compute

$$\begin{aligned}
|1 - \tilde{\lambda}_{\pm}^{(i)}|^2 \leq 1 &\iff |1 - h\sqrt{s_d} \mp h\sqrt{s_d - s_i}|^2 \leq 1 \\
&\iff 1 - 2h\sqrt{s_d} + h^2 s_i \leq 1 \\
&\iff h \leq 2\sqrt{s_d}/s_1\,.
\end{aligned}$$

∎

THEOREM B.16. *Consider the iteration given in Corollary* 3.13. *Suppose* $a \geq \frac{2}{\sqrt{s_1}-\sqrt{s_d}}$. *We choose* $\gamma = \gamma^* = as_d + 2\sqrt{s_d}$ *and* $0 < h \leq 1/(as_1 + \gamma^*)$. *Then for* $k \geq 1/h$ *we have* $\|Y_k\|_{\mathrm{F}} \leq \widetilde{C}h^2k^2(1-\frac{h}{2}(as_d+\sqrt{s_d})^{2k-2}$, *where the constant* $\widetilde{C} = d^2 \cdot \mathcal{O}(\mathrm{poly}(\kappa))$.

Proof. Let us denote by $A = PJP^{-1}$ the Jordan decomposition of A. Then we know from (B.21) that A has precisely $2d-1$ eigenvectors and one generalized eigenvector of algebraic multiplicity 2. Let $q_{\pm}^{(i)}, \ldots, q_{\pm}^{(d-1)}$ be the eigenvectors with associated eigenvalues $\lambda_{\pm}^{(i)} = 1 - \tilde{\lambda}_{\pm}^{(i)}$, where $\tilde{\lambda}_{\pm}^{(i)}$ are from (B.21). With $\gamma = \gamma^*$, one has that $\tilde{\lambda}_{+}^{(d)} = \tilde{\lambda}_{-}^{(d)}$ is a generalized eigenvalue. Abusing notation, let us use $q_{+}^{(d)}$ to represent the eigenvector and $q_{-}^{(d)}$ to represent the generalized eigenvector of $\lambda_{-}^{(d)} = \lambda_{+}^{(d)}$. This means

$$Aq_{+}^{(d)} = \lambda_{+}^{(d)} q_{+}^{(d)}, \qquad Aq_{-}^{(d)} = \lambda_{-}^{(d)} q_{-}^{(d)} + q_{+}^{(d)}.$$

We can express Y_0 by a basis representation

$$Y_0 = \sum_{\star,\diamond\in\{\pm\}} \sum_{i,j\leq d} \alpha_{\star,\diamond}^{i,j} q_{\star}^{(i)} (q_{\diamond}^{(j)})^T.$$

Then using Lemma B.11, we have that for $k \geq 1/h$,

$$\|Y_k\|_{\mathrm{F}} \leq 4d^2h^2k^2 \max_i |\lambda_{\pm}^{(i)}|^{2k-2} \max_{i,j,\star,\diamond} |\alpha_{\star,\diamond}^{i,j}| \|q_{\star}^{(i)} q_{\diamond}^{(j)}\|_{\mathrm{F}}$$

$$\leq 4d^2h^2k^2 \left(1 - \frac{h}{2}(as_d + \sqrt{s_d})\right)^{2k-2} \max_{i,j,\star,\diamond} |\alpha_{\star,\diamond}^{i,j}| \|q_{\star}^{(i)} q_{\diamond}^{(j)}\|_{\mathrm{F}}. \tag{B.23}$$

The second inequality is due to Lemma B.14. The maximum in the above is over $1 \leq i, j \leq d$ and $\star, \diamond \in \{\pm\}$. It remains to show that $\max_{i,j,\star,\diamond} |\alpha_{\star,\diamond}^{i,j}| \|q_{\star}^{(i)} q_{\diamond}^{(j)}\|_{\mathrm{F}} = \mathcal{O}(\mathrm{poly}(\kappa))$. Note that $\boldsymbol{A}$ in Corollary 3.13 can be written as $\boldsymbol{A} = \mathbf{I} - h\tilde{G}$ where $\tilde{G}$ does not depend on h when taking the first order approximation as in Lemma B.12. The rest of the argument is very similar to the proof of Lemma B.8 which we will not present due to brevity. We conclude that

$$\|Y_k\|_{\mathrm{F}} \leq d^2h^2k^2 \left(1 - \frac{h}{2}(as_d + \sqrt{s_d})\right)^{2k-2} \mathcal{O}(\mathrm{poly}(\kappa))$$

$$= \widetilde{C}h^2k^2 \left(1 - \frac{h}{2}(as_d + \sqrt{s_d})\right)^{2k-2}. \qquad \square$$

LEMMA B.17. *A solution to the fixed point equation* $\boldsymbol{Y}^* = \boldsymbol{A}\boldsymbol{Y}^*\boldsymbol{A}^T + \boldsymbol{L}\boldsymbol{L}^T$ *where* $\boldsymbol{A}$ *and* $\boldsymbol{L}$ *are given in Proposition* 3.11, *is given by*

$$\boldsymbol{Y}^* = \begin{pmatrix} Y_{11}^* & Y_{12}^* \\ Y_{12}^* & Y_{22}^* \end{pmatrix},$$

where $Y_{ij}^* \in \mathbb{R}^d$ are diagonal matrices. And the diagonal elements of Y_{ij}^* are given by

$$Y_{11,i}^* = \frac{1}{s_i}\Big(1 - \frac{hs_i(4+(h+a(h\gamma-2))(hs_i-\gamma+as_i(h\gamma-1)))}{(hs_i-\gamma+as_i(h\gamma-1))(4+h(hs_i-2\gamma+as_i(h\gamma-2)))}\Big)\,, \tag{B.24}$$

$$Y_{12,i}^* = \frac{2h(as_i-\gamma)}{(hs_i-\gamma+as_i(h\gamma-1))(4+h(hs_i-2\gamma+as_i(h\gamma-2)))}\,, \tag{B.25}$$

$$Y_{22,i}^* = \frac{-4\gamma-2as_i(2+h(hs_i-3\gamma+as_i(h\gamma-1)))}{(hs_i-\gamma+as_i(h\gamma-1))(4+h(hs_i-2\gamma+as_i(h\gamma-2)))}\,. \tag{B.26}$$

Appendix C. Postponed proofs.

proof of Proposition 2.1. We directly plug (2.19) into (2.18) and verify that we recover (2.17).

$$\begin{aligned}
&\nabla\cdot\Big(\rho\,\mathrm{sym}(\mathbf{Q})\nabla\log\frac{\rho}{\Pi}\Big)+\nabla\cdot\big(\rho(\mathrm{sym}(\mathbf{Q})\nabla\log(\Pi)+\mathbf{Q}\nabla H)\big)\\
&=\mathrm{sym}(\mathbf{Q}):\nabla^2\rho+\nabla\rho\,\mathrm{sym}(\mathbf{Q})\nabla H+\rho\,\mathrm{sym}(\mathbf{Q}):\nabla^2 H+\nabla\rho\,\mathrm{sym}(\mathbf{Q})\nabla\log(\Pi)\\
&\qquad+\rho\,\mathrm{sym}(\mathbf{Q}):\nabla^2\log(\Pi)+\nabla\cdot\big(\rho\mathbf{Q}\nabla H\big)\\
&=\mathrm{sym}(\mathbf{Q}):\nabla^2\rho+\nabla\cdot\big(\rho\mathbf{Q}\nabla H\big)\\
&=\nabla\cdot(\mathbf{Q}\nabla H\rho)+\sum_{i,j=1}^{2d}\frac{\partial^2}{\partial X_i\partial X_j}(Q_{ij}\rho)\,,
\end{aligned}$$

where we denote by $\mathbf{A}:\mathbf{B}=\sum_{i,j=1}^{2d}A_{ij}B_{ij}$ for $\mathbf{A},\mathbf{B}\in\mathbb{R}^{d\times d}$. We have used $\nabla\log(\Pi)=-\nabla H$ and $\nabla^2\log(\Pi)=-\nabla^2 H$ to get the second equality. ◻

proof of Proposition 2.2. We just need to verify that when $\rho(\boldsymbol{X},t)=\Pi(\boldsymbol{X})$, we have $\frac{\partial\rho}{\partial t}=0$. It is clear that when $\rho(\boldsymbol{X},t)=\Pi(\boldsymbol{X})$, the first term on the right hand side of (2.18) is 0, since $\nabla\log(\frac{\rho}{\Pi})=0$. For the second term, let us use (2.19) to get

$$\begin{aligned}
\nabla\cdot(\Pi\Gamma)&=\nabla\cdot\big(\Pi\mathbf{Q}\nabla H-\Pi\,\mathrm{sym}(\mathbf{Q})\nabla\log(\Pi)\big)\\
&=\nabla\Pi\mathbf{Q}\nabla H+\Pi\mathbf{Q}:\nabla^2H+\nabla\Pi\,\mathrm{sym}(\mathbf{Q})\nabla\log(\Pi)+\Pi\,\mathrm{sym}(\mathbf{Q}):\nabla^2\log(\Pi)\\
&=-\Pi\nabla H\mathbf{Q}\nabla H+\Pi\mathbf{Q}:\nabla^2H+\Pi\nabla H\mathrm{sym}(\mathbf{Q})\nabla H+\Pi\,\mathrm{sym}(\mathbf{Q}):\nabla^2\log(\Pi)\\
&=\Pi\mathbf{Q}:\nabla^2H+\Pi\,\mathrm{sym}(\mathbf{Q}):\nabla^2\log(\Pi)\\
&=\Pi\mathbf{Q}:\nabla^2H-\Pi\,\mathrm{sym}(\mathbf{Q}):\nabla^2H\\
&=0\,,
\end{aligned}$$ ◻

We have used $\nabla\Pi=-\Pi\nabla H$ to get the third equality. And we used $\nabla^2\log(\Pi)=-\nabla^2H$ to get the fifth equality. This proves that when $\rho=\Pi$, we indeed have

$$\frac{\partial\rho}{\partial t}\Big|_{\rho=\Pi}=\nabla\cdot\Big(\Pi\,\mathrm{sym}(\mathbf{Q})\nabla\log\frac{\Pi}{\Pi}\Big)+\nabla\cdot(\Pi\Gamma)=0+0=0\,.$$

proof of Proposition 3.2. With our choice of H, (2.15) is a multidimensional OU process. And since $\mathbf{X}_0$ follows a Gaussian distribution, it shows that $\mathbf{X}_t$ will also be a Gaussian distribution. It is well known that the solution to (2.15) with H given by (3.3) is

$$\boldsymbol{X}_t=e^{-t\mathbf{Q}\widetilde{\Sigma}^{-1}}\boldsymbol{X}_0+\int_0^t e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\sqrt{2\,\mathrm{sym}(\mathbf{Q})}\,d\boldsymbol{B}_\tau\,.$$

The mean of $\mathbf{X}_t$ is given by

$$\mathbb{E}\mathbf{X}_t = e^{-t\mathbf{Q}\widetilde{\Sigma}^{-1}}\mathbb{E}\mathbf{X}_0 = 0.$$

We can compute the covariance $\Sigma(t)$ of $\mathbf{X}_t$. Since $\mathbf{X}_t$ has zero mean, we obtain the following using Ito's isometry

$$\Sigma(t) = \mathbb{E}\mathbf{X}_t\mathbf{X}_t^T = 2\int_0^t e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\mathrm{sym}(\mathbf{Q})\left(e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\right)^T d\tau + \mathbb{E}\mathbf{X}_0\mathbf{X}_0^T\,. \tag{C.1}$$

From the above expression, $\Sigma(t)$ is clearly well-defined, symmetric, positive definite for all $t > 0$. We proceed by differentiating $\Sigma(t)$

$$\begin{aligned}
\dot{\Sigma}(t) &= 2\frac{d}{dt}\int_0^t e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\mathrm{sym}(\mathbf{Q})\left(e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\right)^T d\tau \\
&= 2\,\mathrm{sym}(\mathbf{Q}) + \int_0^t \frac{d}{dt}e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\mathrm{sym}(\mathbf{Q})\left(e^{-(t-\tau)\mathbf{Q}\widetilde{\Sigma}^{-1}}\right)^T d\tau \\
&= 2\,\mathrm{sym}(\mathbf{Q}) - \mathbf{Q}\widetilde{\Sigma}^{-1}\Sigma(t) - \Sigma(t)\widetilde{\Sigma}^{-1}\mathbf{Q}^T \\
&= 2\,\mathrm{sym}(\mathbf{Q}(\mathbf{I} - \widetilde{\Sigma}^{-1}\Sigma))\,.
\end{aligned}$$

This finishes the proof. ∎

Acknowledgments. X. Zuo is partially supported by AFOSR YIP award No. FA9550-23-1-0087. X. Zuo and S. Osher are partially funded by STROBE NSF STC DMR 1548924, AFOSR MURI FA9550-18-502 and ONR N00014-20-1-2787. W. Li is partially supported by the AFOSR YIP award No. FA9550-23-1-0087, NSF DMS-2245097, and NSF RTG: 2038080.

REFERENCES

[1] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*, Springer Science & Business Media, 2008.

[2] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, *An introduction to MCMC for machine learning*, Machine learning, 50 (2003), pp. 5–43.

[3] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi, *First-order optimization algorithms via inertial systems with Hessian driven damping*, Mathematical Programming, (2020), pp. 1–43.

[4] H. Attouch, Z. Chbani, J. Fadili, and H. Riahi, *Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping*, Optimization, (2021), pp. 1–40.

[5] H. Attouch, Z. Chbani, and H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM Journal on Optimization, 29 (2019), pp. 2227–2256.

[6] C. H. Bennett, *Mass tensor molecular dynamics*, Journal of Computational Physics, 19 (1975), pp. 267–279.

[7] J. Besag, *Comments on "Representations of knowledge in complex systems" by U. Grenander and MI Miller*, J. Roy. Statist. Soc. Ser. B, 56 (1994), p. 4.

[8] Y. Cao, J. Lu, and L. Wang, *Complexity of randomized algorithms for underdamped Langevin dynamics*, arXiv preprint arXiv:2003.09906, (2020).

[9] Y. Cao, J. Lu, and L. Wang, *On explicit L_2-convergence rate estimate for underdamped Langevin dynamics*, Archive for Rational Mechanics and Analysis, 247 (2023), p. 90.

[10] J. A. Carrillo, Y.-P. Choi, and O. Tse, *Convergence to Equilibrium in Wasserstein Distance for Damped Euler Equations with Interaction Forces*, Communications in Mathematical Physics, 365 (2019), pp. 329–361.

[11] F. Casas, J. M. Sanz-Serna, and L. Shaw, *Split hamiltonian monte carlo revisited*, Statistics and Computing, 32 (2022), p. 86.

[12] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of mathematical imaging and vision, 40 (2011), pp. 120–145.
[13] S. Chen, Q. Li, O. Tse, and S. J. Wright, *Accelerating optimization over the space of probability measures*, arXiv preprint arXiv:2310.04006, (2023).
[14] Y. Chen, D. Z. Huang, J. Huang, S. Reich, and A. M. Stuart, *Gradient flows for sampling: mean-field models, gaussian approximations and affine invariance*, arXiv preprint arXiv:2302.11024, (2023).
[15] X. Cheng and P. Bartlett, *Convergence of Langevin MCMC in KL-divergence*, in Algorithmic Learning Theory, PMLR, 2018, pp. 186–211.
[16] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, in Conference on learning theory, PMLR, 2018, pp. 300–323.
[17] S. Chewi, P. R. Gerber, C. Lu, T. Le Gouic, and P. Rigollet, *The query complexity of sampling from strongly log-concave distributions in one dimension*, in Conference on Learning Theory, PMLR, 2022, pp. 2041–2059.
[18] S. Chewi, C. Lu, K. Ahn, X. Cheng, T. Le Gouic, and P. Rigollet, *Optimal dimension dependence of the Metropolis-adjusted Langevin algorithm*, in Conference on Learning Theory, PMLR, 2021, pp. 1260–1300.
[19] A. Dalalyan, *Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent*, in Conference on Learning Theory, PMLR, 2017, pp. 678–689.
[20] A. S. Dalalyan, *Theoretical guarantees for approximate sampling from smooth and log-concave densities*, Journal of the Royal Statistical Society Series B: Statistical Methodology, 79 (2017), pp. 651–676.
[21] A. S. Dalalyan and A. Karagulyan, *User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient*, Stochastic Processes and their Applications, 129 (2019), pp. 5278–5311.
[22] A. S. Dalalyan and L. Riou-Durand, *On sampling from a log-concave density using kinetic Langevin diffusions*, Bernoulli, 26 (2020), pp. 1956–1988.
[23] M. Dashti and A. M. Stuart, *The Bayesian approach to inverse problems*, arXiv preprint arXiv:1302.6989, (2013).
[24] L. Devroye, A. Mehrabian, and T. Reddad, *The total variation distance between high-dimensional Gaussians with the same mean*, arXiv preprint arXiv:1810.08693, (2018).
[25] A. Durmus, S. Majewski, and B. Miasojedow, *Analysis of Langevin Monte Carlo via convex optimization*, Journal of Machine Learning Research, 20 (2019), pp. 1–46.
[26] A. Durmus and E. Moulines, *Nonasymptotic convergence analysis for the unadjusted Langevin algorithm*, Annals of Applied Probability, 27 (2017), pp. 1551–1587.
[27] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu, *Log-concave sampling: Metropolis-Hastings algorithms are fast*, Journal of Machine Learning Research, 20 (2019), pp. 1–42.
[28] Q. Feng, X. Zuo, and W. Li, *Fisher information dissipation for time inhomogeneous stochastic differential equations*, arXiv preprint arXiv:2402.01036, (2024).
[29] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart, *Interacting langevin diffusions: Gradient structure and ensemble kalman sampler*, SIAM Journal on Applied Dynamical Systems, 19 (2020), pp. 412–441.
[30] S. B. Gelfand and S. K. Mitter, *Simulated annealing type algorithms for multivariate optimization*, Algorithmica, 6 (1991), pp. 419–436.
[31] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Chapman and Hall/CRC, 1995.
[32] M. Girolami and B. Calderhead, *Riemann manifold langevin and hamiltonian monte carlo methods*, Journal of the Royal Statistical Society Series B: Statistical Methodology, 73 (2011), pp. 123–214.
[33] J. Goodman and J. Weare, *Ensemble samplers with affine invariance*, Communications in applied mathematics and computational science, 5 (2010), pp. 65–80.
[34] Y. He, K. Balasubramanian, and M. A. Erdogdu, *On the ergodicity, bias and asymptotic normality of randomized midpoint sampling method*, Advances in Neural Information Processing Systems, 33 (2020), pp. 7366–7376.
[35] J. Idier, *Bayesian approach to inverse problems*, John Wiley & Sons, 2013.
[36] P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. Wilson, *What are Bayesian neural network posteriors really like?*, in International conference on machine learning, PMLR, 2021, pp. 4629–4640.
[37] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–*

Planck equation, SIAM journal on mathematical analysis, 29 (1998), pp. 1–17.
[38] Y. T. Lee, R. Shen, and K. Tian, *Logsmooth gradient concentration and tighter runtimes for Metropolized Hamiltonian Monte Carlo*, in Conference on learning theory, PMLR, 2020, pp. 2565–2597.
[39] B. Leimkuhler, C. Matthews, and J. Weare, *Ensemble preconditioning for markov chain monte carlo simulation*, Statistics and Computing, 28 (2018), pp. 277–290.
[40] T. Lelievre, F. Nier, and G. A. Pavliotis, *Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion*, Journal of Statistical Physics, 152 (2013), pp. 237–274.
[41] T. Lelièvre, G. A. Pavliotis, G. Robin, R. Santet, and G. Stoltz, *Optimizing the diffusion of overdamped langevin dynamics*, arXiv preprint arXiv:2404.12087, (2024).
[42] R. Li, H. Zha, and M. Tao, *Hessian-free high-resolution nesterov acceleration for sampling*, in International Conference on Machine Learning, PMLR, 2022, pp. 13125–13162.
[43] J. S. Liu, *Monte Carlo strategies in scientific computing*, vol. 10, Springer, 2001.
[44] Y.-A. Ma, N. S. Chatterji, X. Cheng, N. Flammarion, P. L. Bartlett, and M. I. Jordan, *Is there an analog of nesterov acceleration for gradient-based MCMC?*, Bernoulli, 27 (2021), pp. 1942–1992.
[45] D. J. MacKay, *Bayesian neural networks and density networks*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 354 (1995), pp. 73–80.
[46] D. J. MacKay, *Information theory, inference and learning algorithms*, Cambridge university press, 2003.
[47] C. J. Maddison, D. Paulin, Y. W. Teh, B. O'Donoghue, and A. Doucet, *Hamiltonian descent methods*, arXiv preprint arXiv:1809.05042, (2018).
[48] J. C. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDEs and approximations: locally lipschitz vector fields and degenerate noise*, Stochastic processes and their applications, 101 (2002), pp. 185–232.
[49] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer Science & Business Media, 2012.
[50] W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, *High-order Langevin diffusion yields an accelerated MCMC algorithm*, arXiv preprint arXiv:1908.10859, (2019).
[51] R. M. Neal, *Bayesian learning for neural networks*, vol. 118, Springer Science & Business Media, 2012.
[52] Y. E. Nesterov, *A method of solving a convex programming problem with convergence rate* $\mathcal{O}(\frac{1}{k^2})$, in Doklady Akademii Nauk, vol. 269, Russian Academy of Sciences, 1983, pp. 543–547.
[53] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*, vol. 2, Springer, 1999.
[54] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, (1996), pp. 341—-363.
[55] R. Shen and Y. T. Lee, *The randomized midpoint method for log-concave sampling*, Advances in Neural Information Processing Systems, 32 (2019).
[56] A. M. Stuart, *Inverse problems: a Bayesian perspective*, Acta numerica, 19 (2010), pp. 451–559.
[57] W. Su, S. Boyd, and E. J. Candes, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, Journal of Machine Learning Research, 17 (2016), pp. 1–43.
[58] A. Taghvaei and P. Mehta, *Accelerated flow for probability distributions*, in International conference on machine learning, PMLR, 2019, pp. 6076–6085.
[59] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic analysis and applications, 8 (1990), pp. 483–509.
[60] H. Y. Tan, S. Osher, and W. Li, *Noise-free sampling algorithms via regularized Wasserstein proximals*, arXiv preprint arXiv:2308.14945, (2023).
[61] Y. W. Teh, A. Thiéry, and S. J. Vollmer, *Consistency and fluctuations for stochastic gradient Langevin dynamics*, Journal of Machine Learning Research, 17 (2016).
[62] T. Valkonen, *A primal–dual hybrid gradient method for nonlinear operators with applications to mri*, Inverse Problems, 30 (2014), p. 055012.
[63] S. Vempala and A. Wibisono, *Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices*, Advances in neural information processing systems, 32 (2019).
[64] Y. Wang and W. Li, *Accelerated information gradient flow*, Journal of Scientific Computing, 90 (2022), pp. 1–47.
[65] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*, in

Proceedings of the 28th international conference on machine learning (ICML-11), Citeseer, 2011, pp. 681–688.

[66] S. Zhang, S. Chewi, M. Li, K. Balasubramanian, and M. A. Erdogdu, *Improved discretization analysis for underdamped Langevin Monte Carlo*, in The Thirty Sixth Annual Conference on Learning Theory, PMLR, 2023, pp. 36–71.

[67] X. Zuo, S. Osher, and W. Li, *Primal-dual damping algorithms for optimization*, Annals of Mathematical Sciences and Applications, 9 (2024), pp. 467–504.