Journal of Machine Learning Research 25 (2024) 1-40 Submitted 10/22; Revised 8/23; Published 8/24

On the Hyperparameters in Stochastic Gradient Descent with Momentum

Bin Shi SHIBIN@LSEC.CC.AC.CN
Academy of Mathematics and Systems Science
Chinese Academy of Sciences
Beijing, 100190, China

School of Mathematical Science
University of Chinese Academy of Sciences
Beijing 100049, China

Editor: Animashree Anandkumar

Abstract

Following the same routine as Shi et al. (2023), we continue to present the theoretical analysis for *stochastic gradient descent with momentum* (SGD with momentum) in this paper. Differently, for SGD with momentum, we demonstrate that the two hyperparameters together, the learning rate and the momentum coefficient, play a significant role in the linear convergence rate in non-convex optimizations. Our analysis is based on using a *hyperparameters-dependent stochastic differential equation* (hp-dependent SDE) that serves as a continuous surrogate for SGD with momentum. Similarly, we establish the linear convergence for the continuous-time formulation of SGD with momentum and obtain an explicit expression for the optimal linear rate by analyzing the spectrum of the Kramers-Fokker-Planck operator. By comparison, we demonstrate how the optimal linear rate of convergence and the final gap for SGD only about the learning rate varies with the momentum coefficient increasing from zero to one when the momentum is introduced. Then, we propose a mathematical interpretation of why, in practice, SGD with momentum converges faster and is more robust in the learning rate than standard stochastic gradient descent (SGD). Finally, we show the Nesterov momentum under the presence of noise has no essential difference from the traditional momentum.

Keywords: nonconvex optimization, stochastic gradient descent with momentum, hp-dependent SDE, hp-dependent kinetic Fokker-Planck equation, Kramers operator, Kramers-Fokker-Planck operator, Nesterov momentum

1. Introduction

For a long time, it has been considered a core and fundamental topic to build theories for optimization algorithms, which leads, in practice, to design new algorithms for accelerating and improving performance. Recently, with the blossoming of machine learning, people have spotlighted gradient-based algorithms. However, the mechanism behind them is still mysterious and undiscovered. Significantly, the non-convex structure brings about new and urgent challenges for modern theoreticians.

Recall the minimization problem of a non-convex function f is defined in terms of an expectation:

$$f(x) = \mathbb{E}_{\zeta}[f(x;\zeta)],$$

where the expectation is over the randomness embodied in ζ. Empirical risk minimization, averaged over n data points, is a simple and special case, shown as

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where x denotes a parameter and the datapoint-specific loss, $f_i(x)$, is indexed by i. When n is large, it is prohibitively expensive to compute the full gradient of the objective function. Hence, the algorithms for gradients with incomplete information (noisy gradient) are adopted widely in practice.

Recall standard *stochastic gradient descent*, shortened to SGD,

$$x_{k+1} = x_k - s\nabla f(x_k) + s\xi_k,$$

with any initial $x_0 \in \mathbb{R}^d$, where ξ_k denotes the noise at the k^{th} iteration. In this paper, we consider the most popular variant of SGD — SGD *with momentum*

$$k+1 = x_k - s\nabla f(x_k) + s\xi_k + \alpha(x_k - x_{k-1})$$

with any initial $x_0, x_1 \in \mathbb{R}^d$, where $\alpha(x_k - x_{k-1})$ is named as momentum excluded in SGD. SGD with momentum has been practically proven to be the most effective setting and is widely adopted in deep learning (He et al., 2016). An experimental comparison between SGD and SGD with momentum is shown in Figure 1.

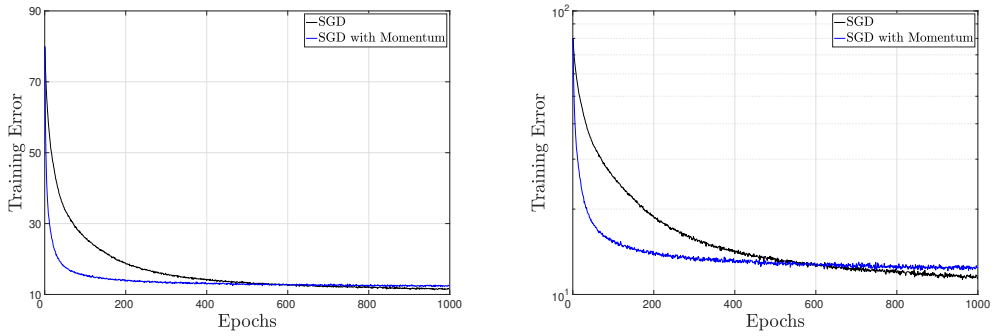

Figure 1: The comparison for the training error between SGD and SGD with momentum. The setting is a 20-layer convolutional neural network on CIFAR-10 (Krizhevsky, 2009) with a mini-batch size of 128. **Learning Rate:** $s = 0.01$. **Momentum Coefficient:** $\alpha = 0.9$. **Left:** Standard Scale; **Right:** Logarithmic Scale. See He et al. (2016) for further investigation of this phenomenon.

In Figure 1, we compare SGD and SGD with momentum using the same learning rate. For the stable (final) training error (the plateau part in the curve of the training error),

SGD with momentum is a little larger than SGD; however, for the least epochs arriving at stabilization, SGD with momentum is far less than SGD. Generally speaking, the momentum coefficient in SGD with momentum is usually set to $\alpha = 0.9$ or beyond in practice. Straightforwardly, people will ask:

Why does SGD with momentum under the setting of momentum coefficient $\alpha = 0.9$ and beyond often perform well?

Actually, it is still a mystery. Now, we can delve into it with two formal and academic questions:

- What happens to the iterative behavior of SGD when the momentum is introduced? Specifically, how does the iterative behavior of SGD with momentum vary with the momentum coefficient α from 0 to 1 when the learning rate s is fixed?
- How does the iterative behavior of SGD with momentum vary with the learning rate s when the momentum coefficient α is fixed?

In other words, do we need to investigate the relationship between SGD and SGD with momentum? Where are the similarities? Where are the differences? In this paper, we will try to answer the questions above from the continuous surrogate. Before discussing the property of SGD with momentum about the hyperparameters, the learning rate s and the momentum coefficient α, we introduce, for convenience, a new mixing parameter as

$$\mu = \frac{(1-\alpha)^2}{(1+\alpha)^2} \cdot \frac{1}{s} \in \left(0,\ \frac{1}{s}\right) \quad \text{suchthat} \quad \alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}} \in (0,1).$$

1.1 Continuous-time approximation

The continuous-time approximation, used widely to study the gradient-based optimization algorithms, is a newly-sprung and vital to assist in the proposal of fundamental new insights and understandings for the performance (Su et al., 2016). Due to its conceptual simplicity in the continuous setting, many properties related to them have been discovered and developed (Shi et al., 2022). Moreover, modern and advanced mathematics and physics methods can propose profound and powerful analyses for them (Shi et al., 2023).

Now, we construct the continuous surrogate for SGD with momentum as follows. Taking a small but nonzero learning rate s, let $t_k = k\sqrt{s}$ $(k = 0, 1, 2, \ldots)$ denote a time step and define $x_k = X_{s,\alpha}(t_k)$ for some sufficiently smooth curve $X_{s,\alpha}(t)$.[1] Applying a Taylor expansion in the power of $\sqrt{s}$, we obtain:

$$\begin{cases} x_{k+1} = X_{s,\alpha}(t_{k+1}) = x_k + \dot{X}_{s,\alpha}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}_{s,\alpha}(t_k)s + \frac{1}{6}\dddot{X}_{s,\alpha}(t_k)(\sqrt{s})^3 + O(s^2) \\ x_{k-1} = X_{s,\alpha}(t_{k-1}) = x_k - \dot{X}_{s,\alpha}(t_k)\sqrt{s} + \frac{1}{2}\ddot{X}_{s,\alpha}(t_k)s - \frac{1}{6}\dddot{X}_{s,\alpha}(t_k)(\sqrt{s})^3 + O(s^2) \end{cases} \tag{1}$$

1. The subscripts outline $X_{s,\alpha}(t)$ is dependent on learning rate s and momentum coefficient α.

Let W be a standard Brownian motion and, for the time being, assume that the noise term ξ_k approximately follows the normal distribution with unit variance. Informally, this leads to

$$\sqrt[4]{s}\xi_k = W(t_{k+1}) - W(t_k) = \sqrt{s}\dot{W}(t_k) + O(s).^2 \tag{2}$$

Taking the straightforward transform for the SGD with momentum, we have

$$\frac{x_{k+1} + x_{k-1} - 2x_k}{s} + \frac{2\sqrt{\mu}}{1+\sqrt{\mu s}} \cdot \frac{x_k - x_{k-1}}{\sqrt{s}} + \nabla f(x_k) = \xi_k.$$

Then, we plug the previous two displays, (1) and (2), into SGD with momentum and get

$$\ddot{X}_{s,\alpha}(t_k) + O(s) + 2\sqrt{\mu}\dot{X}_{s,\alpha}(t_k) + O(\sqrt{s}) + \nabla f(x_k) = \sqrt[4]{s}\dot{W}(t_k) + O(s^{\frac{3}{4}}).$$

Considering the $O(\sqrt[4]{s})$-approximation, that is, retaining both $O(1)$ and $O(\sqrt[4]{s})$ terms but ignoring the smaller terms, we obtain a *hyperparameter-dependent stochastic differential equation* (hp-dependent SDE)

$$\ddot{X}_{s,\alpha} + 2\sqrt{\mu}\dot{X}_{s,\alpha} + \nabla f(X_{s,\alpha}) = \sqrt[4]{s}\dot{W}, \tag{3}$$

where the initial condition is the same value $X_{s,\alpha}(0) = x_0$ and $\dot{X}_{s,\alpha}(0) = (x_1 - x_0)/\sqrt{s}$. Here, we write down the standard form of the hp-dependent SDE (3) as

$$\begin{cases} dX_{s,\alpha} = V_{s,\alpha}, \\ dV_{s,\alpha} = (-2\sqrt{\mu}X_{s,\alpha} - \nabla f(X_{s,\alpha}))\, dt + \sqrt[4]{s}dW. \end{cases} \tag{4}$$

More generally, Li et al. (2019) and Chaudhari and Soatto (2018) consider SDEs with variable-dependent noise covariance as approximating surrogates for SGD with momentum. The hp-dependent SDE (4) is a convenient simple model that allows for a fine-grained analysis, as we will show in this paper.

1.2 An intuitive analysis

In Shi et al. (2023), the authors derive the continuous surrogate for SGD, *lr-dependent* SDE,

$$dX_s = -\nabla f(X_s)dt + \sqrt{s}dW, \tag{5}$$

where the initial is the same value x_0 as its discrete counterpart and W is the standard Brownian motion. By the corresponding lr-dependent Fokker-Planck equation

$$\frac{\partial \rho_s}{\partial t} = \nabla \cdot (\rho_s \nabla f) + \frac{s}{2}\Delta \rho_s, \tag{6}$$

they show the expected excess risk evolves with time as

$$\mathbb{E}[f(X_s(t))] - \min_{x\in\mathbb{R}^d} f(x) = \mathbb{E}[f(X_s(t)) - f(X_s(\infty))] + \underbrace{\mathbb{E}[f(X_s(\infty))] - \min_{x\in\mathbb{R}^d} f(x)}_{\text{final gap}}, \tag{7}$$

2. Although a Brownian motion is not differentiable, the notation $\dot{W}$ has been used in Evans (2012) and Villani (2006).

where the first part $\mathbb{E}[f(X_s(t)) - f(X_s(\infty))]$ converges linearly with the rate as

$$\lambda_s \asymp \exp\left(-\frac{2H_f}{s}\right)$$

and H_f is the height difference (See the detail in Shi et al. (2023, Section 6.2)). Also, the second part is the final gap, which can be bounded by the learning rate s.

Actually, the lr-dependent SDE (5) is the overdamped limit of the hp-dependent SDE (3). Assume the learning rate $s \ll 1$ or $s \to 0$, then we can obtain the friction parameter

$$2\sqrt{\mu} = \frac{1-\alpha}{1+\alpha} \cdot \frac{2}{\sqrt{s}} \gg 1 \qquad \text{or} \qquad 2\sqrt{\mu} = \frac{1-\alpha}{1+\alpha} \cdot \frac{2}{\sqrt{s}} \to \infty.$$

Let us introduce two new variables τ and β and substitute them as

$$\tau = \frac{t}{2\sqrt{\mu}} = \frac{k\sqrt{s}}{2\sqrt{\mu}} = k \cdot \frac{s(1+\alpha)}{2(1-\alpha)} = k\beta(s,\alpha). \tag{8}$$

With the basic calculations

$$\frac{d^2X}{dt^2} = \frac{1}{4\mu}\frac{d^2X}{d\tau^2}, \qquad \frac{dX}{dt} = \frac{1}{2\sqrt{\mu}}\frac{dX}{d\tau}, \qquad \frac{dW(t)}{dt} = \frac{dW(2\sqrt{\mu}\tau)}{d(2\sqrt{\mu}\tau)} = \sqrt{\frac{1}{2\sqrt{\mu}}}\frac{dW(\tau)}{d\tau},$$

we can obtain the equivalent form for the hp-dependent SDE (4) for the variable τ as

$$\frac{1}{4\mu}\ddot{X} + \dot{X} + \nabla f(X) = \sqrt{\beta(s,\alpha)}\dot{W}. \tag{9}$$

Then, we discuss the case of the overdamped friction in (9). In other words, the friction coefficient μ is oversized. Rigorously speaking, when the friction coefficient approximates to infinity, $\mu \to \infty$, the hp-dependent SDE (9) approximates to the following equation as

$$dX(\tau) = -\nabla f(X(\tau))d\tau + \sqrt{\beta(s,\alpha)}dW(\tau), \tag{10}$$

which is similar to the lr-dependent SDE (5) and the only difference is that the coefficient of noise $\sqrt{s}$ is replaced by $\sqrt{\beta(s,\alpha)}$. Recall Shi et al. (2023), we can write down the corresponding Fokker-Planck-Smoluchowski equation for the SDE (10) to govern the evolution of probability density as

$$\frac{\partial \rho_{s,\alpha}}{\partial t} = \nabla \cdot (\rho_{s,\alpha}\nabla f) + \frac{\beta(s,\alpha)}{2}\Delta\rho_{s,\alpha} \tag{11}$$

Hence, we obtain the equilibrium distribution, that is, $\partial\rho_{s,\alpha}/\partial t = 0$, as

$$\mu_{s,\alpha} = \frac{1}{Z_{s,\alpha}}\exp\left(-\frac{2f}{\beta(s,\alpha)}\right) = \frac{1}{Z_{s,\alpha}}\exp\left(-\frac{2f}{s}\cdot\frac{2(1-\alpha)}{1+\alpha}\right) \tag{12}$$

and the rate of linear convergence is

$$\lambda_{s,\alpha} \asymp \exp\left(-\frac{2H_f}{\beta(s,\alpha)}\right) = \exp\left(-\frac{2H_f}{s}\cdot\frac{2(1-\alpha)}{1+\alpha}\right),$$

where f is the potential, and H_f is the height difference. Furthermore, we have

$$\lambda_{s,\alpha}\tau \asymp \exp\left(-\frac{2H_f}{\beta(s,\alpha)}\right)\cdot k\beta(s,\alpha) = \underbrace{\exp\left(-\frac{2H_f}{s}\cdot\frac{2(1-\alpha)}{1+\alpha}\right)}_{\textbf{exponential part}}\cdot\underbrace{\frac{1+\alpha}{2(1-\alpha)}s}_{\textbf{linear part}}\cdot k. \tag{13}$$

Recall (8), the parameter $\beta(s,\alpha)$ is indeed explicitly expressed by the learning rate s and the momentum coefficient α as

$$\beta(s,\alpha) = \frac{1+\alpha}{2(1-\alpha)}\cdot s,$$

where we can find that the parameter $\beta(s,\alpha)$ is linearly dependent on the learning rate s. From (12) and (13), we can also find that the learning rate s here plays the same role as that in the lr-dependent SDE (5). Furthermore, the momentum coefficient α with the mathematical expression $\frac{1+\alpha}{2(1-\alpha)}$ takes the effect on the learning rate s. Hence, we can derive the following conclusions in detail as:

- When $\alpha = 1/3$, then both the equilibrium distribution $\mu_{s,\alpha}$ and the convergence rate $\lambda_{s,\alpha}\tau$ reduces to that in the lr-dependent SDE (5).
- When $\alpha < 1/3$, then the equilibrium distribution $\mu_{s,\alpha}$ concentrates and the convergence rate $\lambda_{s,\alpha}\tau$ decreases sharply. Practically, we rarely adopt this setting.
- When $\alpha > 1/3$, then the equilibrium distribution $\mu_{s,\alpha}$ diverges, but the convergence rate $\lambda_{s,\alpha}\tau$ increases sharply. This is the setting we often adopt in practice. We demonstrate that the parameter $\frac{1+\alpha}{2(1-\alpha)}$ vary with the momentum coefficient α in Table 1 and Figure 2, where we can find that the parameter $\frac{1+\alpha}{2(1-\alpha)}$ does not increas sharply

α	0.5	0.6	0.7	0.8	0.9	0.99	0.999
$\frac{1+\alpha}{2(1-\alpha)}$	1.5	2	2.83	4.5	9.5	99.5	999.5

Table 1: The parameter $\frac{1+\alpha}{2(1-\alpha)}$ varies with the momentum coefficient α.

until the momentum coefficient α is close to 0.9. Then, we can find from (13) that the change of the linear rate is dominated by the exponential part for $\alpha < 0.9$ and by the linear part for $\alpha > 0.9$ since the exponential part is close to 1 when $\alpha > 0.9$. In summary, the linear convergence rate always changes dominantly with the momentum coefficient α.

Back to Figure 1, the convergence behavior of the training error of SGD with momentum is qualitatively similar to that of SGD in Shi et al. (2023). Concretely, the training error decreases to some stable value instead of incessant convergence. Actually, this final gap is caused by the equilibrium distribution $\mu_{s,\alpha}$. Recall the evolution behavior of expected excess risk described in (7), we know that the final gap is characterized as

$$\mathbb{E}[f(X_s(\infty))] - \min_{x\in\mathbb{R}^d} f(x) = \int_{x\in\mathbb{R}^d} f(x)\mu_{s,\alpha}dx - \min_{x\in\mathbb{R}^d} f(x).$$

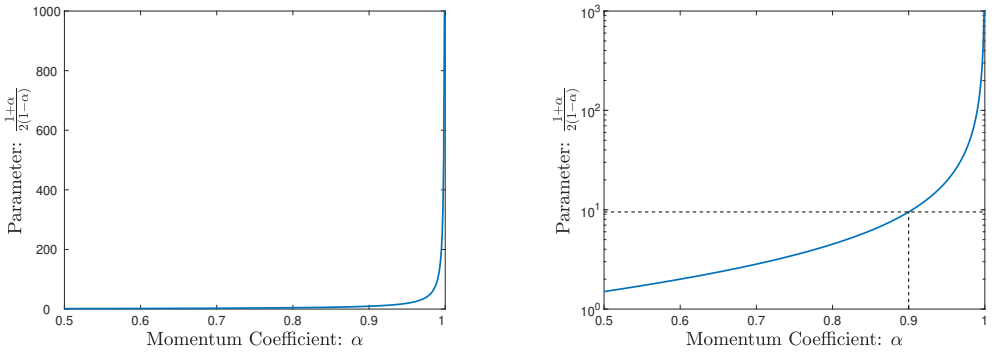

Figure 2: The parameter $\frac{1+\alpha}{2(1-\alpha)}$ varies with the momentum coefficient α. Left: Linear scale; Right: Logarithmic scale.

Recall Shi et al. (2023, Proposition 3.2), we can bound the final gap with the parameter $\beta(s,\alpha)$ instead of the learning rate s as

$$\mathbb{E}[f(X_s(\infty))] - \min_{x\in\mathbb{R}^d} f(x) \leq O(\beta(s,\alpha)) = O\left(\frac{1+\alpha}{2(1-\alpha)}\cdot s\right). \tag{14}$$

Here, we can find that the final gap depends linearly on the learning rate s and the parameter $\frac{1+\alpha}{2(1-\alpha)}$. In the previous analysis from Table 1 and Figure 2, we know that the parameter $\frac{1+\alpha}{2(1-\alpha)}$ changes sharply with the momentum coefficient $\alpha > 0.9$. In other words, when the momentum coefficient α is larger than 0.9, the final gap grows strikingly by increasing the momentum coefficient. Hence, in order to avoid the final gap excessive, the momentum coefficient α set around 0.9 is a reasonable choice.

Based on this intuitive analysis, we know the coefficient of noise is $\sqrt{\beta(s,\alpha)}$ instead of $\sqrt{s}$ in the hp-dependent SDE (4) . Hence, the parameters $\frac{1+\alpha}{2(1-\alpha)}$ for the momentum coefficient α is set as the coefficient of learning rate s, which as a whole, conversely influences the linear rate of convergence and the equilibrium distribution. In Figure 1, we also find that the momentum coefficient $\alpha = 0.9$ balances for the two reverse directions. This paper proposes rigorous proof corresponding to this intuition using modern mathematical techniques: hypocoercity and semi-classical analysis.

1.3 Related work

Currently, the study of deep learning is a fashionable topic. Finding ways to tune the parameters is preoccupying the industry. The seminal work (Bengio, 2012) discusses the significance of the hyperparameters in practice, which not only points out that the learning rate plays the single most crucial role but also suggests that the added momentum will lead to faster convergence in some cases. Moreover, in practice, as proposed in He et al. (2016), the classical Residual Networks adopt SGD with momentum, not SGD, to obtain a sound performance for image recognition.

Recently, in the field of nonlinear optimization, there has been an emerging method called continuous-time approximation for discrete algorithms. By taking the approximating

ODEs, we can simplify the discrete algorithms and use a modern form of analysis to obtain new characteristics, and the formation has not yet been discovered. This method starts to investigate the acceleration phenomenon generated by Nesterov's accelerated gradient methods (Su et al., 2016; Jordan, 2018). Finally, it is solved in Shi et al. (2022) by introducing the high-resolution approximated differential equations based on the dimensional analysis from physics.

In the stochastic setting, this approach has been recently pursued by various authors (Chaudhari et al., 2018; Chaudhari and Soatto, 2018; Mandt et al., 2016; Lee et al., 2016; Caluya and Halder, 2019; Li et al., 2017) to establish the various properties of stochastic optimization. As a notable advantage, the continuous-time perspective allows us to work without assumptions on the boundedness of the domain and gradients, as opposed to the older analyses of SGD (see, for example, Hazan et al. (2008)). Our work is partly motivated by the recent progress on Langevin dynamics, particularly for nonconvex settings (Villani, 2009; Pavliotis, 2014; Helffer et al., 2004; Bovier et al., 2005). In Langevin dynamics, the learning rate s in the hp-dependent SDE can be thought of as the temperature parameter and $2\sqrt{\mu}$ as a function of the learning rate s and the momentum coefficient α, can be thought of as the friction coefficient. In addition, the difference between the strongly convex function and the nonconvex function for the SGD with momentum is also found in Liu et al. (2020). However, the theoretical result for the nonconvex function obtained in Liu et al. (2020) is little related to the observation in practice, since the convergence derived in Liu et al. (2020) is in terms of the iterative average of the expectation of the stochastic gradient's norm square. The convergence of SGD with momentum is also discussed theoretically in Jin and He (2020). However, there are two strong assumptions in Jin and He (2020), one is that the noisy gradient is bounded by the real gradient in the sense of expectation and the other is that the sum of the square of the learning rate is bounded.

1.4 Organization

The remainder of the paper is structured as follows. In Section 2, we introduce the basic concepts and assumptions employed throughout this paper. Next, Section 3 develops our main theorems and some comparisons analytically and numerically with SGD with momentum. Section 4 formally proves the linear convergence, and Section 5 further quantifies the linear rate of convergence. Technical details of the proofs are deferred to the appendices. We conclude the paper in Section 7 with a few directions for future research.

2. Preliminaries

Throughout this paper, we assume that the objective function f is infinitely differentiable in $\mathbb{R}^d$; that is, $f \in C^\infty(\mathbb{R}^d)$. We use $\|\cdot\|$ to denote the standard Euclidean norm. Recall the confining condition for the objective f in Shi et al. (2023, Definition 2.1), also see Markowich and Villani (1999) and Pavliotis (2014): $\lim_{\|x\|\to+\infty} f(x) = +\infty$ and $\exp(-2f/s)$ is integrable for all $s > 0$. This condition is quite mild and requires that the function grows sufficiently rapidly when x is far from the origin. For convenience, we need to define some Hilbert spaces. For any $k = -1, 0, 1$, let $\langle\cdot,\cdot\rangle_{L^2(\mu^k_{s,\alpha})}$ be the inner product in the Hilbert

space $L^2(\mu_{s,\alpha}^k)$, defined as

$$\langle g_1, g_2 \rangle_{L^2(\mu_{s,\alpha}^k)} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} g_1 g_2 \mu_{s,\alpha}^k \mathrm{d}v\mathrm{d}x$$

with any $g_1, g_2 \in L^2(\mu_{s,\alpha}^k)$. For any $g \in L^2(\mu_{s,\alpha}^k)$, the norm induced by the inner product is

$$\|g\|_{L^2(\mu_{s,\alpha}^k)} = \sqrt{\langle g, g \rangle_{L^2(\mu_{s,\alpha}^k)}}.$$

Recall the hp-dependent SDE (4). Similar as Shi et al. (2023), the probability density $\rho_{s,\alpha}(t,\cdot,\cdot)$ of $(X_{s,\alpha}(t), V_{s,\alpha}(t))$ evolves according to the *hp-dependent kinetic Fokker-Planck equation*

$$\frac{\partial \rho_{s,\alpha}}{\partial t} = -v \cdot \nabla_x \rho_{s,\alpha} + \nabla_x f \cdot \nabla_v \rho_{s,\alpha} + 2\sqrt{\mu} \nabla_v \cdot (v \rho_{s,\alpha}) + \frac{\sqrt{s}}{2} \Delta_v \rho_{s,\alpha}, \tag{15}$$

with the initial condition $\rho_{s,\alpha}(0,\cdot,\cdot)$. Here, $\Delta_v \equiv \nabla_v \cdot \nabla_v$ is the Laplacian. For completeness, we derive the hp-dependent kinetic Fokker-Planck equation in Appendix A.1 from the hp-dependent SDE (4) by Itô's formula and Chapman-Kolmogorov equation. If the objective f satisfies the confining condition, then the hp-dependent kinetic Fokker-Planck equation (15) admits a unique invariant Gibbs distribution that takes the form as

$$\mu_{s,\alpha} = \frac{1}{Z_{s,\alpha}} \exp\left(-\frac{2f + \|v\|^2}{\beta(s,\alpha)} \right), \tag{16}$$

where the normalization factor is

$$Z_{s,\alpha} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \exp\left(-\frac{2f + \|v\|^2}{\beta(s,\alpha)} \right) dvdx.$$

The proof of existence and uniqueness is shown in Appendix A.2.

Villani conditions To show the solution's existence and uniqueness to the hp-dependent kinetic Fokker-Planck equation (15), we still need to introduce the Villani conditions first.

Definition 1 (Villani Conditions (Villani, 2009)) *A confining condition f is said to satisfy the Villani conditions if*

(I) *When $\|x\| \to \infty$, we have*

$$\frac{\|\nabla f\|^2}{s} - \Delta f \to \infty;$$

(II) *For any $x \in \mathbb{R}^d$, we have*

$$\|\nabla^2 f\|_2 \le C(1 + \|\nabla f\|),$$

where the matrix 2-norm is $\|\nabla^2 f\|_2 = |\lambda_{\max}(\nabla^2 f)|$.

The Villani condition-(I) has been proposed in Shi et al. (2023, Definition 2.3), which says that the gradient has a sufficiently large squared norm compared with the Laplacian of the function. Generally speaking, the polynomials of degrees no less than two satisfy

the Villani condition-(I). For the hp-dependent kinetic Fokker-Planck equation (15), the existence and uniqueness of the solution are still the Villani condition-(II), which is called the relative bound in Villani (2009). Actually, we will find the Fokker-Planck-Kramers operator $\mathscr{K}_{s,\alpha}$ has a compact resolvent under the two Villani conditions together (See Helffer and Nier (2005, Theorem 5.8, Remark 5.13(a)) and Li (2012)) in Section 5. Hence, taking any initial probability density $\rho_{s,\alpha}(0,\cdot,\cdot) \in L^2(\mu_{s,\alpha}^{-1})$, we have the following guarantee:

Lemma 2 (Existence and uniqueness of the weak solution) *For any confining function f satisfying the Villani Conditions (Definition 1) and any initial $\rho_{s,\alpha}(0,\cdot,\cdot) \in L^2(\mu_{s,\alpha}^{-1})$, the hp-dependent SDE* (4) *admits a weak solution whose probability density in $C^1\left([0,+\infty), L^2(\mu_{s,\alpha}^{-1})\right)$ is the unique solution to the hp-dependent kinetic Fokker-Planck equation* (15)*.*

Basics of Morse theory Similar as Shi et al. (2023), we also need to assume the objective function is a Morse function. Here, we will describe the basic concepts briefly. A point x is called a *critical point* if the gradient $\nabla f(x) = 0$. A function f is said to be a *Morse* function if for any critical point x, the Hessian $\nabla^2 f(x)$ at x is non-degenerate. Note also a *local minimum* $x^\bullet \in \mathcal{X}^\bullet$ is a critical point with all the eigenvalues of the Hessian at $x^\bullet$ positive and an *index-1 saddle point* $x^\circ \in \mathcal{X}^\circ$ is a critical point where the Hessian at x° has exactly one negative eigenvalue, that is, $\eta_1(x^\circ) \geq \cdots \geq \eta_{d-1}(x^\circ) > 0,\ \eta_d(x^\circ) < 0$. Let $\mathcal{K}_{f(x^\circ)} := \left\{x \in \mathbb{R}^d : f(x) < f(x^\circ)\right\}$ denote the sublevel set at level $f(x^\circ)$. If the radius r is sufficiently small, the set $\mathcal{K}_{f(x^\circ)} \cap \{x : \|x - x^\circ\| < r\}$ can be partitioned into two connected components, say $C_1(x^\circ, r)$ and $C_2(x^\circ, r)$. Therefore, we can introduce the most important concept — *index*-1 *separating saddle* as follows.

Definition 3 (Index-1 Separating Saddle) *Let x° be an index-1 saddle point and $r > 0$ be sufficiently small. If $C_1(x^\circ, r)$ and $C_2(x^\circ, r)$ are contained in two different (maximal) connected components of the sublevel set $\mathcal{K}_{f(x^\circ)}$, we call x° an index-1 separating saddle point.*

Intuitively speaking, the index-1 separating saddle point x° is the bottleneck of any path connecting the two local minima. More precisely, along a path connecting $x_1^\bullet$ and $x_2^\bullet$, by definition the function f must attain a value that is at least as large as $f(x^\circ)$. For the detail about the basics of Morse theory, please readers refer to Shi et al. (2023, Section 6.2).

3. Main Results

In this section, we state our main results. Briefly, for SGD with momentum in its continuous formulation, the hp-dependent SDE, we show that the expected excess risk converges linearly to stationarity and estimate the final excess risk by the hyperparameters in Section 3.1. Furthermore, we derive a quantitative expression of the rate of linear convergence in Section 3.2. Finally, we carry the continuous-time convergence guarantees to the discrete case in Section 3.3.

3.1 Linear convergence

In this subsection, we are concerned with the expected excess risk, $\mathbb{E}[f(X_{s,\alpha}(t))] - f^\star$. Recall that $f^\star = \inf_{x\in\mathbb{R}^d} f(x)$.

Theorem 1 *Let f be confined and satisfy the Villani conditions. Then, there exists $\lambda_{s,\alpha} > 0$ for any learning rate $s > 0$ and any momentum coefficient $\alpha \in (0,1)$ such that the expected excess risk satisfies*

$$\mathbb{E}[f(X_{s,\alpha}(t))] - f^\star \leq \epsilon(s,\alpha) + D(s,\alpha)e^{-\lambda_{s,\alpha}t}, \tag{17}$$

for all $t \geq 0$. Here $\epsilon(s,\alpha) = \epsilon(s,\alpha;f) \geq 0$ increases strictly for the mixing parameter $\beta(s,\alpha)$ and depends only on the objective function f, and $D(s,\alpha) = D(s,\alpha;f,\rho) \geq 0$ depends only on s,α,f, and the initial distribution $\rho_{s,\alpha}(0,\cdot,\cdot)$.

Similar to as Shi et al. (2023), the proof of this theorem is based on the following decomposition of the expected excess risk:

$$\mathbb{E}[f(X_{s,\alpha}(t))] - f^\star = \mathbb{E}[f(X_{s,\alpha}(t))] - \mathbb{E}[f(X_{s,\alpha}(\infty))] + \mathbb{E}[f(X_{s,\alpha}(\infty))] - f^\star, \tag{18}$$

where $\mathbb{E}[f(X_{s,\alpha}(\infty))]$ denotes $\mathbb{E}_{x\sim\mu_{s,\alpha}}[f(x)]$ in light of the fact that $X_{s,\alpha}(t)$ converges weakly to $\mu_{s,\alpha}$ as $t \to +\infty$ (see Lemma 11). The question is thus separated into quantifying how fast $\mathbb{E}[f(X_{s,\alpha}(t))]$ converges weakly to $\mathbb{E}[f(X_{s,\alpha}(\infty))]$ as $t \to +\infty$ and how the expected excess risk at stationarity $\mathbb{E}[f(X_{s,\alpha}(\infty))] - f^\star$ depends on the hyperparameters. The following two propositions address these two questions. Recall that $\rho_{s,\alpha}(0,\cdot,\cdot) \in L^2(\mu_{s,\alpha}^{-1})$ is the probability density of the initial iterate in SGD with momentum.

Proposition 4 *Under the assumptions of Theorem 1, there exists $\lambda_{s,\alpha} > 0$ for any learning rate s and any momentum coefficient $\alpha \in (0,1)$ such that*

$$|\mathbb{E}[f(X_{s,\alpha}(t))] - \mathbb{E}[f(X_{s,\alpha}(\infty))]| \leq C(s,\alpha)\,\|\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha}\|_{L^2(\mu_{s,\alpha}^{-1})}\, e^{-\lambda_{s,\alpha}t},$$

for any $t \geq 0$, where the constant $C(s,\alpha)$ depends only s, α and f, and where

$$\|\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha}\|_{L^2(\mu_{s,\alpha}^{-1})} = \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha})^2\, \mu_{s,\alpha}^{-1} dv dx\right)^{\frac{1}{2}}$$

measures the gap between the initialization and the Gibbs invariant distribution.

Loosely speaking, it takes $O(1/\lambda_{s,\alpha})$ to converge to stationarity from the beginning. In Theorem 1, $D(s,\alpha)$ can be set to $C(s,\alpha)\|\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha}\|_{L^2(\mu_{s,\alpha}^{-1})}$. Notably, the proof of Proposition 4 shall reveal that $C(s,\alpha)$ increases as the learning rate s increases. Turning to the analysis of the second term, $\mathbb{E}[f(X_{s,\alpha}(\infty))] - f^\star$, we write henceforth $\epsilon(s,\alpha) := \mathbb{E}[f(X_{s,\alpha}(\infty))] - f^\star$.

Proposition 5 *Under the assumptions of Theorem 1, the expected excess risk at stationarity, $\epsilon(s,\alpha)$, is a strictly increasing function $\beta(s,\alpha)$. Moreover, for any $S > 0$, there exists a constant A that depends only on S and f and satisfies*

$$\epsilon(s,\alpha) \equiv \mathbb{E}\left[X_{s,\alpha}(\infty)\right] - f^\star \leq A\beta(s,\alpha) = \frac{A(1+\alpha)}{2(1-\alpha)} \cdot s,$$

for any learning rate $0 < s \leq S$.

Proposition 5 verifies rigorously our intuitive estimate (14) in Section 1.2. The two propositions are proved in Section 4. The proof of Theorem 1 is a direct consequence of Proposition 4 and Proposition 5. More precisely, the two propositions taken together give

$$\mathbb{E}f(X_{s,\alpha}(t)) - f^\star \leq O\left(\frac{1+\alpha}{2(1-\alpha)} \cdot s\right) + C(s,\alpha)e^{-\lambda_{s,\alpha}t}, \tag{19}$$

for a bounded learning rate s related to the momentum coefficient α. Furthermore, when $\alpha > 0.9$, the parameter $\frac{1+\alpha}{2(1-\alpha)}$ increases strikingly with the increase of the momentum α, which leads to an oversize training error. The following result gives the iteration complexity of SGD with momentum in its continuous-time formulation.

Corollary 6 *Under the assumptions of Theorem 1, for any $\epsilon > 0$, if the learning rate $s \leq \min\{\epsilon/A \cdot (1-\alpha)/(1+\alpha), S\}$ and $t \geq \frac{1}{\lambda_{s,\alpha}} \log \frac{2C(s,\alpha)\|\rho_{s,\alpha}(0,\cdot,\cdot)-\mu_{s,\alpha}\|_{L^2(\mu_{s,\alpha}^{-1})}}{\epsilon}$, then*

$$\mathbb{E}\left[f(X_{s,\alpha}(t))\right] - f^\star \leq \epsilon.$$

3.2 The rate of linear convergence

Now, we turn to the key issue of understanding how the linear rate $\lambda_{s,\alpha}$ depends on the hyperparameters, the learning rate s, and the momentum coefficient α. Here, we propose an explicit expression for the linear rate $\lambda_{s,\alpha}$ to interpret this.

Theorem 2 *In addition to the assumptions of Theorem 1, assume that the objective f is a Morse function and has at least two local minima.Then the constant $\lambda_{s,\alpha}$ in* (17) *satisfies*

$$\lambda_{s,\alpha} = \frac{v(\gamma + o(s))}{\sqrt{\mu} + \sqrt{\mu + v}} e^{-\frac{2H_f}{\beta(s,\alpha)}} = \frac{v(\gamma + o(s))}{\sqrt{\mu} + \sqrt{\mu + v}} e^{-\frac{2H_f}{s} \cdot \frac{2(1-\alpha)}{1+\alpha}}, \tag{20}$$

for $0 < s \leq s_0$, where $s_0 > 0, \alpha > 0$, $H_f > 0$ are constants completely depending only on f and $-v$ is the unique negative eigenvalue of the Hessian at the highest index-1 separating saddle.

It should here be noted that our assumption of having at least two local minima is important. Otherwise, there may probably be no saddle points, much less the index-1 separating saddle. Recall Shi et al. (2023, Theorem 2), for the lr-dependent SDE (5), the exponential decay constant is obtained as

$$\lambda_s = v(\gamma + o(s))e^{-\frac{2H_f}{s}}. \tag{21}$$

Here, we first come to analyze the ratio of two exponential decays, $\lambda_{s,\alpha}$ in (20) and λ_s in (21), as

$$\frac{\lambda_{s,\alpha}}{\lambda_s} = \frac{1}{\sqrt{\mu} + \sqrt{\mu + v}} e^{-\frac{2H_f}{s} \cdot \frac{1-3\alpha}{1+\alpha}}.$$

Then, we compute the ratio of the exponential decay constants in (20) with different learning rates, s_1 and s_2, as

$$\frac{\lambda_{s_1,\alpha}}{\lambda_{s_2,\alpha}} = \sqrt{\frac{s_1}{s_2}} \left(\frac{\lambda_{s_1}}{\lambda_{s_2}}\right)^{\frac{2(1-\alpha)}{1+\alpha}}.$$

For any $\alpha > 1/3$, when $s, s_1, s_2 \to 0$, we can obtain

$$\frac{\lambda_{s,\alpha}}{\lambda_s} \asymp \frac{1+\alpha}{1-\alpha} \cdot 2\sqrt{s} e^{\frac{2H_f}{s} \cdot \frac{3\alpha-1}{1+\alpha}} \to +\infty, \tag{22}$$

and there exist two constants C_1 and C_2 in $(0, 1)$ such that

$$\left(\frac{\lambda_{s_1}}{\lambda_{s_2}}\right)^{C_1} \leq \frac{\lambda_{s_1,\alpha}}{\lambda_{s_2,\alpha}} \leq \left(\frac{\lambda_{s_1}}{\lambda_{s_2}}\right)^{C_2}. \tag{23}$$

From (22), we can find when the learning rate s is sufficiently small, $\lambda_{s,\alpha}$ is larger than λ_s. In other words, the SGD with momentum converges faster than SGD. Furthermore, from (23), the ratio $\lambda_{s_1,\alpha}/\lambda_{s_2,\alpha}$ is weaker than $\lambda_{s_1}/\lambda_{s_2}$, that is, $\lambda_{s_1,\alpha}/\lambda_{s_2,\alpha}$ is closer to 1 than $\lambda_{s_1}/\lambda_{s_2}$. In Shi et al. (2023), we show the exponential decay constant λ_s changes sharply with the learning rate s. Here, we can find the exponential decay constant $\lambda_{s,\alpha}$ in SGD with momentum is more robust with the learning rate s than λ_s.

Numerical Demonstration We demonstrate a numerical comparison based on the analysis and description above. Recall (Shi et al., 2023, Figure 7), we compare the iteration number of lr-dependent SDE for the learning rates $s = 0.001$ and $s = 0.1$ as

$$k_{0.001} \approx 2.5 \times 10^{47}, \quad k_{0.1} \approx 2.5 \times 10^{2}, \quad \text{and} \quad k_{0.001}/k_{0.1} \approx 10^{45}.$$

From (18), we know that the expected risk is separated into two parts. Theorem 1 characterizes the linear convergence of $\mathbb{E}[f(X_{s,\alpha}(t))] - \mathbb{E}[f(X_{s,\alpha}(\infty))]$ and estimates the final gap $\mathbb{E}[f(X_{s,\alpha}(\infty))] - f^\star = 100$. Theorem 2 show the explicit form of the linear convergence rate. Let $\mathbb{E}[f(X_{s,\alpha}(0))] - f^\star = 100$ be an idealized case. Then, the idealized risk function for hp-dependent SDE (4) can be given with the following form as

$$R(t) = (100 - \beta(s,\alpha))\, e^{-\frac{1}{2\sqrt{\mu}} \cdot e^{-\frac{0.1}{\beta(s,\alpha)}} t} + \beta(s,\alpha),$$

shown in Figure 3. With the basic calculation, we can obtain that

$$k_{0.001,0.9} \approx 1.5 \times 10^{9}, \quad k_{0.1,0.9} \approx 2.5 \times 10^{1}, \quad \text{and} \quad k_{0.001,0.9}/k_{0.1,0.9} \approx 6 \times 10^{7}.$$

By comparison of the two cases, without momentum and within momentum, we can obtain the following results as

(1) The iterative number satisfies $k_{0.001,0.9} < k_{0.001}$ and $k_{0.1,0.9} < k_{0.1}$, which verifies that the iteration number for SGD with momentum is far less than that for SGD.

(2) Moreover, the ratio for the iterative number $k_{0.001,0.9}/k_{0.1,0.9}$ is also far less than $k_{0.001}/k_{0.1}$, which manifests the iteration number in SGD with momentum is more robust (**or stable**) than SGD.

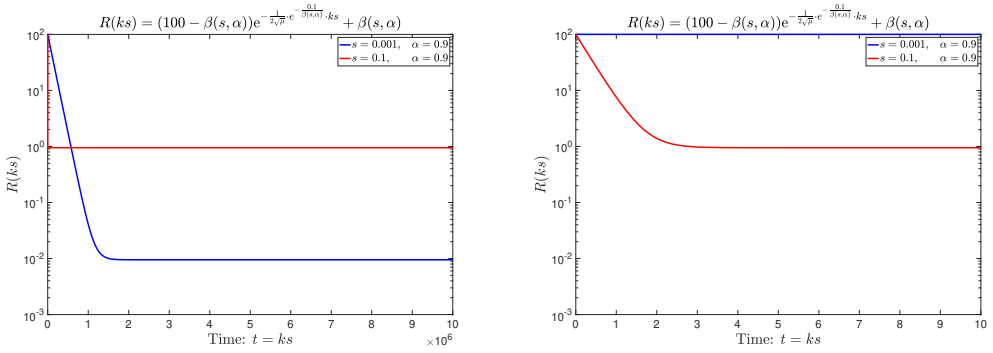

Figure 3: Idealized risk function of the form $R(t) = (100 - \beta(s,\alpha))\, e^{-\frac{1}{2\sqrt{\mu}}\cdot e^{-\frac{0.1}{\beta(s,\alpha)}}\, t} + \beta(s,\alpha)$ with the identification $t = ks$, which is adapted from (17). Similar as (Shi et al., 2023, Figure 7), the learning rate is selected as $s = 0.1$ and 0.001. The right plot is a locally enlarged image of the left.

3.3 Discretization

This subsection presents the results developed from the continuous perspective of the discrete regime. For the discrete algorithms, we still need to assume f to be L-smooth, that is, the gradient of f is L-Lipschitz continuous in the sense that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$. Therefore, we can restrict the learning rate s to be no larger than $1/L$. The following proposition is the key tool for translation to the discrete regime.

Proposition 7 *For any L-smooth objective f and any initialization $X_{s,\alpha}(0)$ drawn from a probability density $\rho_{s,\alpha}(0,\cdot,\cdot) \in L^2(\mu_{s,\alpha}^{-1})$, the hp-dependent SDE (4) has a unique global solution $X_{s,\alpha}$ in expectation; that is, $\mathbb{E}[X_{s,\alpha}(t)]$ as a function of t in $C^1([0,+\infty);\mathbb{R}^d)$ is unique. Moreover, there exists $B(T) > 0$ such that the SGD with momentum iterates x_k satisfy*

$$\max_{0\leq k\leq T/s} |\mathbb{E}f(x_k) - \mathbb{E}f(X_s(ks))| \leq B(T)s,$$

for any fixed $T > 0$.

We note a sharp bound on $B(T)$ in Bally and Talay (1996). For completeness, we also remark that the convergence can be strengthened to the strong sense:

$$\max_{0\leq k\leq T/s} \mathbb{E}\,\|x_k - X_s(ks)\| \leq B'(T)s.$$

This result has appeared in Mil'shtein (1975); Talay (1982); Pardoux and Talay (1985); Talay (1984); Kloeden and Platen (1992) and we provide a self-contained proof in Appendix B.1. We now state the main result of this subsection.

Theorem 3 *In addition to the assumptions of Theorem 1, assume that f is L-smooth. Then, the following two conclusions hold:*

(a) *For any $T > 0$ and any learning rate $0 < s \leq 1/L$, the iterates of SGD with momentum satisfy*

$$\mathbb{E}f(x_k) - f^\star \leq \left(\frac{A(1+\alpha)}{2(1-\alpha)} + B(T) \right) s + C \left\| \rho - \mu_{s,\alpha} \right\|_{L^2(\mu_{s,\alpha}^{-1})} \mathrm{e}^{-s\lambda_{s,\alpha} k}, \tag{24}$$

for all $k \leq T/s$, where $\lambda_{s,\alpha}$ is the exponential decay constant in (17)*, A as in Proposition 5 depends only on $1/L$ and f, $C = C_{1/L}$ is as in Proposition 4, and $B(T)$ depends only on the time horizon T and the Lipschitz constant L.*

(b) *If f is a Morse function with at least two local minima, with $\lambda_{s,\alpha}$ appearing in* (24) *being given by* (20)*.*

Theorem 3 follows as a direct consequence of Theorem 1 and Proposition 7. Note that the second part of Theorem 3 is simply a restatement of Theorem 2. Moreover, we also mention that the dimension parameter d is not essential for characterizing the linear rate of convergence.

4. Proof of the Linear Convergence

In this section, we prove Proposition 3.1 and Proposition 5, which lead to a complete proof of Theorem 1.

4.1 Linear operators and convergence

Before proving Proposition 3.1, we first take a simple analysis for the linear operators derived from the hp-dependent kinetic Fokker-Planck equation (15). Then, we point out that only the Villani condition-(I) and the Poincaré inequality cannot work here. To obtain the estimate for the hp-dependent kinetic Fokker-Planck equation (15), we still need to introduce the Villani condition-(II) and demonstrate a vital inequality, which is named the relative bound (Villani, 2009).

Basic Property of Linear Operators Similar to the transformation used in (Shi et al., 2023, Section 5), we obtain the time-evolving probability density over the Gibbs invariant measure given as

$$h_{s,\alpha}(t,\cdot,\cdot) = \rho_{s,\alpha}(t,\cdot,\cdot)\mu_{s,\alpha}^{-1} \in C^1\left([0,+\infty), L^2(\mu_{s,\alpha})\right),$$

which allows us to work in the space $L^2(\mu_{s,\alpha})$ in place of $L^2(\mu_{s,\alpha}^{-1})$. It is not hard to show that $h_{s,\alpha}$ satisfies the following differential equation

$$\frac{\partial h_{s,\alpha}}{\partial t} = -\mathscr{L}_{s,\alpha} h_{s,\alpha}, \tag{25}$$

with the initial $h_{s,\alpha}(0,\cdot,\cdot) = \rho_{s,\alpha}(0,\cdot,\cdot)\mu_{s,\alpha}^{-1} \in L^2(\mu_{s,\alpha})$, where the linear operator is defined as

$$\mathscr{L}_{s,\alpha} = v \cdot \nabla_x - \nabla_x f \cdot \nabla_v + 2\sqrt{\mu} v \cdot \nabla_v - \frac{\sqrt{s}}{2}\Delta_v. \tag{26}$$

Simple observation tells us that the linear operator (26) can be separated into two parts

$$\mathscr{L}_{s,\alpha} = \mathscr{T} + \mathscr{D}_{s,\alpha} \tag{27}$$

where the first part $\mathscr{T} = v \cdot \nabla_x - \nabla_x f \cdot \nabla_v$ is named as *transport operator* and the second part $\mathscr{S}_{s,\alpha} = 2\sqrt{\mu} v \cdot \nabla_v - \frac{\sqrt{s}}{2}\Delta_v$ is named as *diffusion operator*.

Let $[\cdot, \cdot]$ be the commutator operation and the two new operators

$$\mathscr{A}_{s,\alpha} = \sqrt[4]{\frac{s}{4}}\nabla_v \quad \text{and} \quad \mathscr{C}_{s,\alpha} = \sqrt[4]{\frac{s}{4}}\nabla_x, \tag{28}$$

then we can obtain the basic facts described in the following lemma.

Lemma 8 *In the Hilbert space $L^2(\mu_{s,\alpha})$, with the notation of the linear operators above, we have*

(i) The conjugate operators of the linear operators, $\mathscr{A}_{s,\alpha}$ and $\mathscr{C}_{s,\alpha}$, are

$$\mathscr{A}^\star_{s,\alpha} = \sqrt[4]{\frac{s}{4}}\left(-\nabla_v + \frac{2v}{\beta}\right) \quad \text{and} \quad \mathscr{C}^\star_{s,\alpha} = \sqrt[4]{\frac{s}{4}}\left(-\nabla_x + \frac{2\nabla_x f}{\beta}\right). \tag{29}$$

Moreover, the diffusion operator is $\mathscr{D}_{s,\alpha} = \mathscr{A}^\star_{s,\alpha}\mathscr{A}_{s,\alpha}$.

(ii) The linear operators $\mathscr{A}_{s,\alpha}$ and $\mathscr{A}^\star_{s,\alpha}$ commutes with $\mathscr{C}_{s,\alpha}$. Also, the commutator between $\mathscr{A}_{s,\alpha}$ and $\mathscr{A}^\star_{s,\alpha}$ is

$$[\mathscr{A}_{s,\alpha}, \mathscr{A}^\star_{s,\alpha}] = 2\sqrt{\mu}\mathbf{I}_{d\times d}. \tag{30}$$

(iii) The transport operator $\mathscr{T}$ is anti-symmetric. The commutator between $\mathscr{A}_{s,\alpha}$ and $\mathscr{T}$ is

$$[\mathscr{A}_{s,\alpha}, \mathscr{T}] = \mathscr{C}_{s,\alpha} \tag{31}$$

and that between $\mathscr{C}_{s,\alpha}$ and $\mathscr{T}$ is

$$[\mathscr{C}_{s,\alpha}, \mathscr{T}] = -\nabla_x^2 f \cdot \mathscr{A}_{s,\alpha}. \tag{32}$$

The proof is only based on the basic operations and integration by parts, which is shown in Appendix C.1.

Convergence to Gibbs invariant measure Similarly as Shi et al. (2023), we need to claim the function in $L^2(\mu_{s,\alpha}^{-1})$ is integrable, that is, $L^2(\mu_{s,\alpha}^{-1})$ is a subset of $L^1(\mathbb{R}^d)$.

Lemma 9 *Let f satisfy the confining condition. Then, $L^2(\mu_{s,\alpha}^{-1}) \subset L^1(\mathbb{R}^d)$.*

The proof of Lemma 9 is simple and shown in Appendix C.2. With Lemma 8, we claim the basic fact as the following lemma.

Lemma 10 *The linear operator $\mathscr{L}_{s,\alpha}$ is nonpositive in $L^2(\mu_{s,\alpha})$. Explicitly, for any $g \in L^2(\mu_{s,\alpha})$, the linear operator $\mathscr{L}_{s,\alpha}$ obeys*

$$\langle \mathscr{L}_{s,\alpha} g, g\rangle_{L^2(\mu_{s,\alpha})} = \langle \mathscr{D}_{s,\alpha} g, g\rangle_{L^2(\mu_{s,\alpha})} = \|\mathscr{A}_{s,\alpha} g\|^2_{L^2(\mu_{s,\alpha})} = \frac{\sqrt{s}}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|\nabla_v g\|^2 \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x.$$

With Lemma 10, we can show the solution to the hp-dependent SDE (15) converges to the Gibbs invariant measure in terms of the dynamics of its probability densities over time.

Lemma 11 *Let f satisfy the confining condition and denote the initial distribution as $\rho_{s,\alpha}(0,\cdot,\cdot) \in L^2(\mu_{s,\alpha}^{-1})$. Then, the unique solution $\rho_{s,\alpha}(t,\cdot,\cdot) \in C^1\left([0,+\infty), L^2(\mu_{s,\alpha}^{-1})\right)$ to the hp-dependent Fokker-Planck equation* (15) *converges in $L^2(\mu_{s,\alpha}^{-1})$ to the Gibbs invariant measure $\mu_{s,\alpha}$, which is specified by* (16).

Proof [Proof of Lemma 11] Taking a derivative for the L^2 distance between the time involving probability density $\rho_{s,\alpha}(t,\cdot,\cdot)$ and the Gibbs invariant measure $\mu_{s,\alpha}$, we have

$$\begin{aligned}\frac{d}{dt}\|\rho_{s,\alpha}(t,\cdot,\cdot)-\mu_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha}^{-1})} &= \frac{d}{dt}\|h_{s,\alpha}(t,\cdot,\cdot)-1\|^2_{L^2(\mu_{s,\alpha})}\\ &= \frac{d}{dt}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(h_{s,\alpha}(t,x,v)-1\right)^2\mu_{s,\alpha}\mathrm{d}v\mathrm{d}x\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}-\mathscr{L}_{s,\alpha}\left(h_{s,\alpha}(t,x,v)-1\right)\left(h_{s,\alpha}(t,x,v)-1\right)\mu_{s,\alpha}\mathrm{d}v\mathrm{d}x,\end{aligned}$$

The last equality is due to the equation (25). Next, by use of Lemma 10, we can proceed to

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\rho_{s,\alpha}(t,\cdot,\cdot)-\mu_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha}^{-1})} = -\frac{\sqrt{s}}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_v h_{s,\alpha}(t,x,v)\|^2\mu_{s,\alpha}\mathrm{d}v\mathrm{d}x \le 0. \tag{33}$$

Thus, $\|\rho_{s,\alpha}(t,\cdot,\cdot)-\mu_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha}^{-1})}$ is decreasing strictly and asymptotically towards the equilibrium state

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_v h_{s,\alpha}(t,x,v)\|^2 dvdx = 0.$$

This equality holds, however, only if $h_{s,\alpha}(t,\cdot,\cdot)$ is constant about v. Back to the differential equation (25), we have

$$\frac{\partial h_{s,\alpha}}{\partial t} = -v\cdot\nabla_x h_{s,\alpha},$$

of which the solution is $h_{s,\alpha} = h_{s,\alpha}(0, x+tv)$. Furthermore, we can obtain $\nabla_v h_{s,\alpha} = t\nabla_x h_{s,\alpha}$. With the time t's arbitrarity, we know $\nabla_x h_{s,\alpha} = 0$. Therefore, we obtain that $h_{s,\alpha}(t,\cdot,\cdot)$ is constant. The fact that both $\rho_{s,\alpha}(t,\cdot,\cdot)$ and $\mu_{s,\alpha}$ are probability densities implies that $h_{s,\alpha}(t,\cdot)\equiv 1$; that is, $\rho_{s,\alpha}(t,\cdot,\cdot)\equiv\mu_{s,\alpha}$. ■

Failure of Poincaré inequality Recall the Villani Condition-(I) in Definition 1

$$\frac{\|\nabla f\|^2}{s} - \Delta f \to +\infty, \qquad \text{with} \quad \|x\|\to+\infty.$$

In Shi et al. (2023, Lemma 5.4), we obtain the Poincaré inequality to derive the linear convergence, which is based on the technique in Villani (2009, Theorem A.1). Here, for the differential equation (25), what we consider is not only the potential $f(x)$ itself but is the Hamiltonian.

$$H(x,v) = f(x) + \frac{1}{2}\|v\|^2.$$

With the new norm $\|\nabla H\|^2 = \|\nabla_x H\|^2 + \|\nabla_v H\|^2$, the Villani condition-(I) directly leads to

$$\frac{\|\nabla H\|^2}{\beta(s,\alpha)} - \Delta H = \frac{\|\nabla_x f\|^2 + \|v\|^2}{\beta(s,\alpha)} - \Delta_x f - d \to +\infty \qquad \text{with} \quad \|x\|^2+\|v\|^2\to+\infty.$$

Based on the same technique in Villani (2009, Theorem A.1), we can obtain the following Poincaré inequality

$$\|h_{s,\alpha}-1\|^2_{L^2(\mu_{s,\alpha})} \leq \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left(\|\nabla_x h_{s,\alpha}\|^2 + \|\nabla_v h_{s,\alpha}\|^2\right) \mu_{s,\alpha} dxdv. \tag{34}$$

Different from Shi et al. (2023), this inequality above cannot connect with the equation (33), because there is only the derivative of $h_{s,\alpha}$ about the variable v on the right-hand side of the equality (33).

4.2 Proof of Proposition 3.1

To better appreciate the linear convergence of the hp-dependent SDE (3), as established in Proposition 3.1, we need to show the linear convergence for H^1-norm instead of L^2-norm, which is the key difference with the lr-dependent SDE (5) in Shi et al. (2023). The technique is named hypocoercivity, introduced by Villani (2009).

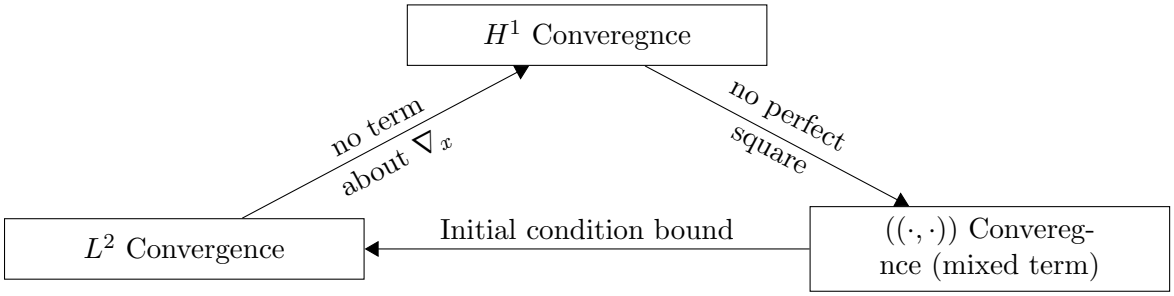

Figure 4: The Diagram of the proof framework for L^2-convergence.

4.2.1 H^1-Norm

Due to the failure of Poincaré inequality, we need to introduce a new Hilbert space $H^1(\mu_{s,\alpha})$, where the inner product is

$$\langle g_1, g_2\rangle_{H^1(\mu_{s,\alpha})} = \langle g_1, g_2\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{A}_{s,\alpha} g_1, \mathscr{A}_{s,\alpha} g_2\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{C}_{s,\alpha} g_1, \mathscr{C}_{s,\alpha} g_2\rangle_{L^2(\mu_{s,\alpha})}$$

for any $g_1, g_2 \in H^1(\mu_{s,\alpha})$. Then, the induced H^1-norm for $h_{s,\alpha}-1$ is

$$\|h_{s,\alpha}-1\|^2_{H^1(\mu_{s,\alpha})} = \|h_{s,\alpha}-1\|^2_{L^2(\mu_{s,\alpha})} + \|\mathscr{A}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + \|\mathscr{C}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}.$$

For convenience, we use $h_{s,\alpha}$ instead of $h_{s,\alpha}-1$. Then, we can obtain the derivative as

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\langle h_{s,\alpha}, h_{s,\alpha}\rangle_{H^1(\mu_{s,\alpha})} =& (1+2\sqrt{\mu})\|\mathscr{A}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + \|\mathscr{A}^2_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\ &+ \langle \mathscr{C}_{s,\alpha} h_{s,\alpha}, \mathscr{A}_{s,\alpha} h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\ &- \left\langle \nabla^2_x f \cdot \mathscr{A}_{s,\alpha} h_{s,\alpha}, \mathscr{C}_{s,\alpha} h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})},\end{aligned} \tag{35}$$

where the detailed calculation is shown in Appendix C.3. From (35), we can find only the last term on the right-hand side is negative. The simple Cauchy-Schwartz inequality tells us that there are two terms needed to be bounded as

$$-\left\langle \nabla_x^2 f \cdot \mathscr{A}_{s,\alpha} h_{s,\alpha}, \mathscr{C}_{s,\alpha} h_{s,\alpha} \right\rangle_{L^2(\mu_{s,\alpha})} \geq -\frac{1}{2}\|\nabla_x^2 f \cdot \mathscr{A}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} - \frac{1}{2}\|\mathscr{C}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \quad (36)$$

The first part needs a bound for $\nabla_x^2 f$, which requires us to consider the Villani condition-(II), shown in Section 4.2.2. The second part requires us to consider a mixed term, shown in Section 4.2.3.

4.2.2 Relative bound

With the introduction of Villani condition-(II), we can obtain a relative bound as

Lemma 12 (Lemma A.24 in Villani (2009)) *Let $f \in C^2(\mathbb{R}^d)$ and satisfy the Villani condition-(II)*

$$\left\|\nabla^2 f\right\|_2 \leq C\left(1 + \|\nabla f\|\right),$$

Then, for all $g \in H^1(\mu_{s,\alpha})$, we have

(i) $\|(\nabla_x f)g\|^2_{L^2(\mu_{s,\alpha})} \leq \kappa_1(s,\alpha)\left(\|g\|^2_{L^2(\mu_{s,\alpha})} + \|\nabla_x g\|^2_{L^2(\mu_{s,\alpha})}\right)$;

(ii) $\|\|\nabla_x^2 f\|_2 g\|^2_{L^2(\mu_{s,\alpha})} \leq \kappa_2(s,\alpha)\left(\|g\|^2_{L^2(\mu_{s,\alpha})} + \|\nabla_x g\|^2_{L^2(\mu_{s,\alpha})}\right)$.

The proof is shown in Appendix C.4. Then, with Lemma 12, we can bound the first term on the right-hand side of (36) as

$$\begin{aligned}
\left\|\nabla_x^2 f \cdot \mathscr{A}_{s,\alpha} h_{s,\alpha}\right\|^2_{L^2(\mu_{s,\alpha})} &= \left\| \left[\mathscr{T}, \mathscr{C}_{s,\alpha}\right] h_{s,\alpha}\right\|^2_{L^2(\mu_{s,\alpha})} \\
&= \frac{\sqrt{s}}{2}\|\nabla_x^2 f \cdot \nabla_v h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\
&\leq \frac{\kappa_2(s,\alpha)\sqrt{s}}{2}\left(\|\nabla_x \nabla_v h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + \|\nabla_v h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}\right) \\
&\leq \frac{2\kappa_2(s,\alpha)}{\sqrt{s}}\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + \kappa_2(s,\alpha)\|\mathscr{A}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\
&\leq \kappa_3(s,\alpha)\left(\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + \|\mathscr{A}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}\right),
\end{aligned}$$

where $\kappa_3(s,\alpha) = \max\{2\kappa_2(s,\alpha)/\sqrt{s}, \kappa_2(s,\alpha)\}$.

4.2.3 Mixed term

For the second term on the right-hand side of (36), $-\|\mathscr{C}_{s,\alpha} h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}$, in order to make it a perfect sum of squares, we need to consider the mixed term

$$\left\langle \mathscr{A}_{s,\alpha} h_{s,\alpha}, \mathscr{C}_{s,\alpha} h_{s,\alpha} \right\rangle_{L^2(\mu_{s,\alpha})},$$

of which the derivative is

$$\begin{aligned}
&\frac{d}{dt}\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&\qquad \geq -2\|\mathscr{A}_{s,\alpha}^2 h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} \\
&\qquad\quad -2\sqrt{\mu}\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} + \|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\
&\qquad\quad -\sqrt{\kappa_3(s,\alpha)}\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\left(\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} + \|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\right).
\end{aligned} \tag{37}$$

The detailed calculation is shown in Appendix C.5.

4.2.4 New inner product

From (35) and (37), we have obtained the four following squared terms.

$$\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}, \quad \|\mathscr{A}_{s,\alpha}^2 h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}, \quad \|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}, \quad \|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}.$$

Intuitively, we can construct a new inner product satisfying

$$-\frac{1}{2}\frac{d}{dt}\left((h_{s,\alpha}, h_{s,\alpha})\right) \geq \lambda(s,\alpha)\left((h_{s,\alpha}, h_{s,\alpha})\right),$$

where $\lambda(s,\alpha)$ is some positive constant which depends on the parameters s and α. If $b^2 \leq ac$, we can define the new inner product as

$$\begin{aligned}
\left((h_{s,\alpha}, h_{s,\alpha})\right) = \langle h_{s,\alpha}, h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} &+ a\left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\
&+ 2b\left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + c\left\langle \mathscr{C}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}
\end{aligned} \tag{38}$$

Apparently, the norm $\|\cdot\|_{((\cdot,\cdot))}$ induced by the new inner product $((\cdot,\cdot))$ is equivalent to H^1-norm, that is, there exists two positive reals $\mathcal{C}_1$ and $\mathcal{C}_2$ such that

$$\mathcal{C}_1\|h_{s,\alpha}\|^2_{((\cdot,\cdot))} \leq \|h_{s,\alpha}\|^2_{H^1} \leq \mathcal{C}_2\|h_{s,\alpha}\|^2_{((\cdot,\cdot))}. \tag{39}$$

Then, from the basic calculations in Section 4.2.1 and Section 4.2.3, we can obtain the following expression by combining like terms as

$$\begin{aligned}
-\frac{1}{2}\frac{d}{dt}\left((h_{s,\alpha}, h_{s,\alpha})\right) &= \left((\mathscr{L}_{s,\alpha}h_{s,\alpha}, h_{s,\alpha})\right) \\
&\geq \left(1 + 2a\sqrt{\mu} - 2b\sqrt{\kappa_3(s,\alpha)}\right)\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + a\|\mathscr{A}_{s,\alpha}^2 h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\
&\quad + c\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} + 2b\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} \\
&\quad - 2b\sqrt{\kappa_3(s,\alpha)}\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} \\
&\quad - \left(a + c\sqrt{\kappa_3(s,\alpha)} + 4b\sqrt{\mu}\right)\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} \\
&\quad - 4b\|\mathscr{A}_{s,\alpha}^2 h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} \\
&\quad - c\sqrt{\kappa_3(s,\alpha)}\,\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\,\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}
\end{aligned} \tag{40}$$

From (40), we can find this problem is transformed to guarantee the following matrix K_1 positive definite as

$$K_1 = \begin{pmatrix} 1+2a\sqrt{\mu}-2b\sqrt{\kappa_3(s,\alpha)} & 0 & -b\sqrt{\kappa_3(s,\alpha)} & -\frac{1}{2}\left(a+c\sqrt{\kappa_3(s,\alpha)}+4b\sqrt{\mu}\right) \\ 0 & a & -2b & 0 \\ -b\sqrt{\kappa_3(s,\alpha)} & -2b & c & -\frac{1}{2}c\sqrt{\kappa_3(s,\alpha)} \\ -\frac{1}{2}\left(a+c\sqrt{\kappa_3(s,\alpha)}+4b\sqrt{\mu}\right) & 0 & -\frac{1}{2}c\sqrt{\kappa_3(s,\alpha)} & 2b \end{pmatrix}.$$

Let $M := \max\{1, \sqrt{\mu}, \sqrt{\kappa_3(s,\alpha)}\}$ and fix $1 \geq a \geq b \geq 2c$, in order to make the matrix K positive definite, we can make the following matrix K_2 positive definite as

$$K_2 = \begin{pmatrix} 1-2Ma & 0 & -Mb & -3Ma \\ 0 & a & -2Mb & 0 \\ -Mb & -2Mb & c & -\frac{1}{2}Mc \\ -3Ma & 0 & -\frac{1}{2}Mc & 2b \end{pmatrix}.$$

Furthermore, if we assume that $M \leq 1/4a$, we here use the matrix L instead of K_2 as

$$L = [l_{ij}] \equiv \begin{pmatrix} \frac{1}{2} & 0 & -Mb & -3Ma \\ 0 & a & -2Mb & 0 \\ -Mb & -2Mb & c & -\frac{1}{2}Mc \\ -3Ma & 0 & -\frac{1}{2}Mc & 2b \end{pmatrix}.$$

Recall the fact that if the element l_{ij} of the matrix L satisfies that

$$|l_{ij}| = |l_{ji}| \leq \frac{\sqrt{l_{ii}l_{jj}}}{4} \leq \frac{l_{ii}+l_{jj}}{8}, \tag{41}$$

the following quadratic inequality will be obtained as

$$\sum_{i=1}^{4}\sum_{j=1}^{4} l_{ij}x_ix_j \geq \frac{1}{4}\sum_{i=1}^{d} l_{ii}x_i^2.$$

To ensure (41), we require the coefficient M satisfies that

$$Mb \leq \frac{\sqrt{c/2}}{4}, \qquad 2Mb \leq \frac{\sqrt{ac}}{4}, \qquad 3Ma \leq \frac{\sqrt{b}}{4}, \qquad \frac{1}{2}Mc \leq \frac{\sqrt{2bc}}{4},$$

that is,

$$M = \sqrt{\min\left\{1,\ \frac{1}{4a},\ \frac{c}{32b^2},\ \frac{ac}{64b^2},\ \frac{b}{144a^2},\ \frac{b}{2c}\right\}}.$$

Then, we can obtain the following estimate for (40) as

$$\begin{aligned}
-\frac{1}{2}\frac{d}{dt}\left((h_{s,\alpha},h_{s,\alpha})\right) &= \left((\mathscr{L}_{s,\alpha}h_{s,\alpha},h_{s,\alpha})\right)\\
&\geq \frac{1}{4}\min\left\{\frac{1}{2},2b\right\}\left(\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}+\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}\right)\\
&\geq \frac{1}{8}\min\left\{\frac{1}{2},2b\right\}\left(\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}+\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}\right)\\
&\quad+\frac{1}{8}\min\left\{\frac{1}{2},2b\right\}\chi_{s,\alpha}\|h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})}\\
&\geq \min\left\{\frac{1}{8}\min\left\{\frac{1}{2},2b\right\},\frac{1}{8}\min\left\{\frac{1}{2},2b\right\}\chi_{s,\alpha}\right\}\|h_{s,\alpha}\|^2_{H^1(\mu_{s,\alpha})}
\end{aligned}$$

where the last but one inequality follows Poincaré inequality (34). With the norm equivalence (39) and taking

$$\lambda_{s,\alpha}=\mathcal{C}_1\cdot\min\left\{\frac{1}{8}\min\left\{\frac{1}{2},2b\right\},\frac{1}{8}\min\left\{\frac{1}{2},2b\right\}\chi_{s,\alpha}\right\},$$

we can furthermore obtain the following inequality as

$$-\frac{1}{2}\frac{d}{dt}\left((h_{s,\alpha},h_{s,\alpha})\right)\geq\lambda_{s,\alpha}\left((h_{s,\alpha},h_{s,\alpha})\right).$$

Continuing with the norm equivalence (39), we have

$$\|h_{s,\alpha}(t,\cdot,\cdot)\|^2_{L^2(\mu_{s,\alpha})}\leq e^{-2\lambda_{s,\alpha}t}\left((h_{s,\alpha}(0),h_{s,\alpha}(0))\right)\leq\mathcal{C}_2e^{-2\lambda_{s,\alpha}t}\|h_{s,\alpha}(0,\cdot,\cdot)\|^2_{H^1(\mu_{s,\alpha})}.$$

Finally, from (Villani, 2009, Theorem A.8), we know that the H^1-norm $\|h_{s,\alpha}(t,\cdot,\cdot)\|_{H^1(\mu_{s,\alpha})}$ at $0<t\leq 1$ can be bounded by L^2-norm $\|h_{s,\alpha}(0,\cdot,\cdot)\|_{L^2(\mu_{s,\alpha})}$ at $t=0$. Some basic operations tell us that there exists $\mathcal{C}_3>0$ such that

$$\|\rho_{s,\alpha}(t,\cdot,\cdot)-\mu_{s,\alpha}\|^2_{L^2(\mu^{-1}_{s,\alpha})}\leq\mathcal{C}_3e^{-2\lambda_{s,\alpha}t}\|\rho_{s,\alpha}(0,\cdot,\cdot)-\mu_{s,\alpha}\|^2_{L^2(\mu^{-1}_{s,\alpha})}.$$

4.3 Proof of Proposition 5

Since the potential, $f(x)-f^\star$, is only for x in the proof of Proposition 5, we can integrate the variable v. Hence, except for the mixing parameter $\beta(s,\alpha)$ instead of the learning rate s, the technique here is the same as that used in Shi et al. (2023).

5. Estimate of Exponential Decay Constant

In this section, we complete the proof of Theorem 2 to quantify the linear rate of linear convergence, the exponential decay constant $\lambda_{s,\alpha}$. This is crucial for us to understand the dynamics of SGD with momentum, especially its dependence on the hyperparameters, the learning rate s and the momentum coefficient α.

5.1 Connection to a Kramers-Fokker-Planck operators

Similar as Shi et al. (2023), we start to derive the relationship between the hp-dependent SDE (4) and the Kramers operator. Recall that the probability density $\rho_{s,\alpha}(t,\cdot,\cdot)$ of the solution to the hp-dependent SDE (4) is assumed in $L^2(\mu_{s,\alpha}^{-1})$. Here, we consider the transformation

$$\psi_{s,\alpha}(t,\cdot,\cdot) = \frac{\rho_{s,\alpha}(t,\cdot,\cdot)}{\sqrt{\mu_{s,\alpha}}} \in L^2(\mathbb{R}^d,\mathbb{R}^d).$$

With this transformation, we can equivalently rewrite the kinetic Fokker-Planck equation (15) as

$$\frac{\partial \psi_{s,\alpha}}{\partial t} = -\left[v\cdot\nabla_x - \nabla_x f\cdot\nabla_v - \frac{\sqrt{s}}{2}\left(\Delta_v + \frac{d}{\beta(s,\alpha)} - \frac{\|v\|^2}{\beta^2(s,\alpha)}\right)\right] \tag{42}$$

with the initial $\psi_{s,\alpha}(0,\cdot,\cdot) = \rho_{s,\alpha}(0,\cdot,\cdot)/\sqrt{\mu_{s,\alpha}} \in L^2(\mathbb{R}^d,\mathbb{R}^d)$. Here, the operator on the right-hand side of the kinetic Fokker-Planck equation (42) is named as Kramers operator:

$$\mathscr{K}_{s,\alpha} = v\cdot\nabla_x - \nabla_x f\cdot\nabla_v - \frac{\sqrt{s}}{2}\left(\Delta_v + \frac{d}{\beta(s,\alpha)} - \frac{\|v\|^2}{\beta^2(s,\alpha)}\right). \tag{43}$$

Similarly, we collect some essential facts concerning the spectrum of the Kramers operator $\mathscr{K}_{s,\alpha}$. Here, we first need to show the spectrum of the Kramers operator $\mathscr{K}_{s,\alpha}$ is discrete, that is, the unbounded Kramers operator $\mathscr{K}_{s,\alpha}$ has a compact resolvent. Based on the Villani condition-(I), the unbounded Kramers operator $\mathscr{K}_{s,\alpha}$ has a compact resolvent, first shown in Helffer and Nier (2005, Corollary 5.10, Remark 5.13). However, it still needs a polynomial-controlled condition. Until Li (2012), the polynomial-controlled condition has not been removed. Here, we state the well-known result in spectral theory — that the unbounded Kramers operator $\mathscr{K}_{s,\alpha}$ has a compact resolvent (Li, 2012).

Lemma 13 (Corollary 1.4 in Li (2012)) *Let the potential $f(x)$ satisfy the Villani conditions (Definition 1). The Kramers operator $\mathscr{K}_{s,\alpha}$ has a compact resolvent.*

Recall the existence and uniqueness of Gibbs distribution $\mu_{s,\alpha}$ in Appendix A.2, it is not hard to show that $\sqrt{\mu_{s,\alpha}}$ is the unique eigenfunction of $\mathscr{K}_{s,\alpha}$ corresponding to the eigenvalue zero. Taking (33), we can obtain

$$\begin{aligned}
\langle \mathscr{K}_{s,\alpha}\psi_{s,\alpha}(t,\cdot,\cdot), \psi_{s,\alpha}(t,\cdot,\cdot)\rangle_{L^2(\mathbb{R}^d,\mathbb{R}^d)} &= \left\langle \mathscr{K}_{s,\alpha}(\psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}}), \psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}}\right\rangle_{L^2(\mathbb{R}^d,\mathbb{R}^d)} \\
&= -\frac{1}{2}\frac{d}{dt}\left\langle \psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}}, \psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}}\right\rangle_{L^2(\mathbb{R}^d,\mathbb{R}^d)} \\
&= -\frac{1}{2}\frac{d}{dt}\left\|\rho_{s,\alpha}(t,\cdot,\cdot) - \mu_{s,\alpha}\right\|^2_{L^2(\mu_{s,\alpha})} \geq 0.
\end{aligned}$$

With Lemma 13, this verifies the unbounded Kramers operator $\mathscr{K}_{s,\alpha}$ is positive semidefinite. Hence, we can order the eigenvalues of $\mathscr{K}_{s,\alpha}$ in $L^2(\mathbb{R}^d,\mathbb{R}^d)$ as

$$0 = \zeta_{s,\alpha;0} \leq \zeta_{s,\alpha;1} \leq \cdots \leq \zeta_{s,\alpha;\ell} \leq \cdots \leq +\infty.$$

Recall Theorem 1, the exponential decay constant in (17) is set to

$$\lambda_{s,\alpha} = \zeta_{s,\alpha;1}.$$

To see this, note that $\psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}}$ also satisfies (42) and is orthogonal to the null eigenfunction $\sqrt{\mu_{s,\alpha}}$. Therefore, we can obtain

$$\begin{aligned}\left\langle \mathscr{K}_{s,\alpha}(\psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}}), \psi_{s,\alpha}(t,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}} \right\rangle_{L^2(\mathbb{R}^d,\mathbb{R}^d)} &\geq \zeta_{s,\alpha;1} \|\psi_{s,\alpha}(0,\cdot,\cdot) - \sqrt{\mu_{s,\alpha}})\|^2_{L^2(\mathbb{R}^d,\mathbb{R}^d)} \\ &= \zeta_{s,\alpha;1} \|\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha}^{-1})}.\end{aligned}$$

Furthermore, we can obtain the exponential estimate for $\rho_{s,\alpha}(t,\cdot,\cdot) - \mu_{s,\alpha}$ as

$$\|\rho_{s,\alpha}(t,\cdot,\cdot) - \mu_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha}^{-1})} \leq e^{-\zeta_{s,\alpha;1} t} \|\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha}^{-1})}.$$

5.2 The spectrum of Kramers-Fokker-Planck operators

Similar in Shi et al. (2023), the Schrödinger operator is equivalent to the Witten-Laplacian, here the unbounded Kramers operator $\mathscr{K}_{s,\alpha}$ is equivalent to the Kramers-Fokker-Planck operator as

$$\mathscr{P}_{s,\alpha} := \beta(s,\alpha)\mathscr{K}_{s,\alpha} = v \cdot \beta(s,\alpha)\nabla_x - \nabla_x f \cdot \beta(s,\alpha)\nabla_v + \sqrt{\mu}(v - \beta(s,\alpha)\nabla_v)(v + \beta(s,\alpha)\nabla_v), \tag{44}$$

where $\mathscr{P}_{s,\alpha}$ is only a simple rescaling of $\mathscr{K}_{s,\alpha}$. Denote the eigenvalues of the Kramers-Fokker-Planck operator as $0 = \delta_{s,\alpha;0} \leq \delta_{s,\alpha;1} \leq \cdots \leq \delta_{s,\alpha;\ell} \leq \cdots \leq +\infty$, we get the simple relationship as

$$\delta_{s,\alpha;\ell} = \beta(s,\alpha)\zeta_{s,\alpha;\ell}$$

for all $\ell \in \mathbb{N}$.

The spectrum of the Kramers-Fokker-Planck operator $\mathscr{P}_{s,\alpha}$ has been investigated in Helffer and Nier (2005); Hérau et al. (2011). Here, we exploit this literature to derive the closed-form expression for the first positive eigenvalue of the Kramers-Fokker-Planck operator, thereby obtaining the dependence of the exponential decay constant on the hyperparameters, the learning rate s and the momentum coefficient α, for a certain class of nonconvex objective functions.

Recall in Section 2, and we show the basic concepts of the Morse function. For detail about the ideas of the Morse function, please refer to Shi et al. (2023, Section 6.2). Notably, the most crucial concept, the index-1 separating saddle (See Shi et al. (2023, Figure 9)), is introduced. Here, we briefly describe the general assumption for a Morse function.

Assumption 14 (Generic case in Hérau et al. (2011), also see Assumption 6.4 in Shi et al. (2023
For every critical component E_j^i selected in the labeling process above, where $i = 0, 1, \ldots, I$, we assume that

- *The minimum $x_{i,j}^{\bullet}$ of f in any critical component E_j^i is unique.*
- *If $E_j^i \cap \mathcal{X}^\circ \neq \varnothing$, there exists a unique $x_{i,j}^{\circ} \in E_j^i \cap \mathcal{X}^\circ$ such that $f(x_{i,j}^{\circ}) = \max_{x \in E_j^i \cap \mathcal{X}^\circ} f(x)$. In particular, $E_j^i \cap \mathcal{K}_{f(x_{i,j}^{\circ})}$ is the union of two distinct critical components.*

With the generic assumption (Assumption 14), the labeling process for the index-1 separating saddle and local minima (See Shi et al. (2023, Figure 10)) is introduced, which

reveals a remarkable result: there exists a bijection between the set of local minima and the set of index-1 separating saddle points (including the fictive one) $\mathcal{X}^\circ \cup \{\infty\}$. Interestingly, this shows that the number of local minima is always larger than the number of index-1 separating saddle points by one; that is, $n^\circ = n^\bullet - 1$. Here, we relabel the index-1 separating saddle points x_ℓ° for $\ell = 0, 1, \ldots, n^\circ$ with $x_0^\circ = \infty$, and the local minima $x_\ell^\bullet$ for $\ell = 0, 1, \ldots, n^\bullet - 1$ with $x_0^\bullet = x^\star$, such that

$$f(x_0^\circ) - f(x_0^\bullet) > f(x_1^\circ) - f(x_1^\bullet) \geq \ldots \geq f(x_{n^\bullet-1}^\circ) - f(x_{n^\bullet-1}^\bullet), \tag{45}$$

where $f(x_0^\circ) - f(x_0^\bullet) = f(\infty) - f(x^\star) = +\infty$. A detailed description of this bijection is given in (Hérau et al., 2011, Proposition 5.2). With the pairs $(x_\ell^\circ, x_\ell^\bullet)$ in place, we state the fundamental result concerning the first $n^\bullet - 1$ smallest positive eigenvalues of the Fokker-Planck-Kramers operator as

$$\mathscr{P}_{\beta(s,\alpha)} = v \cdot \beta(s,\alpha)\nabla_x - \nabla_x f \cdot \beta(s,\alpha)\nabla_v + \frac{1}{2}\gamma(v - \beta(s,\alpha)\nabla_v)(v + \beta(s,\alpha)\nabla_v). \tag{46}$$

Proposition 15 (Theorem 1.2 in Hérau et al. (2011)) *Under Assumption 14 and the assumptions of Theorem 2, there exists $\beta_0 > 0$ such that for any $\beta \in (0, \beta_0]$, the first $n^\bullet - 1$ smallest positive eigenvalues of the Kramers-Fokker-Planck operator $\mathscr{P}_{s,\alpha}$ associated with f satisfy*

$$\delta_{\beta(s,\alpha),\ell} = \beta(s,\alpha)|\eta_d(x_\ell^\circ)|\left(\gamma_\ell + o(\beta(s,\alpha))\right) e^{-\frac{2(f(x_\ell^\circ) - f(x_\ell^\bullet))}{\beta(s,\alpha)}} \tag{47}$$

for $\ell = 1, 1, \ldots, n^\bullet - 1$, where

$$\gamma_\ell = \frac{1}{\pi}\left(\frac{\det(\nabla^2 f(x_\ell^\bullet))}{-\det(\nabla^2 f(x_\ell^\circ))}\right)^{\frac{1}{2}},$$

and $\eta_d(x_\ell^\circ)$ is the unique negative eigenvalue of the block matrix

$$\begin{pmatrix} 0 & \mathbf{I}_{d\times d} \\ -\nabla^2 f(x_\ell^\circ) & \gamma \mathbf{I}_{d\times d} \end{pmatrix}.$$

Recall Theorem 2, we assume the negative eigenvalue of $\nabla^2 f(x_1^\circ)$ is $-v$. Apply Proposition 15 into the Kramers operator $\mathscr{K}_{s,\alpha}$ in (44), we can obtain there exists $H_f > 0$ completely depending only on f such that

$$\zeta_{s,\alpha;1} = |\eta_d(x_1^\circ)|\left(\gamma_1 + o(\beta(s,\alpha))\right) e^{-\frac{2H_f}{\beta(s,\alpha)}} = |\eta_d(x_1^\circ)|\left(\gamma_1 + o(s)\right) e^{-\frac{2H_f}{s}\cdot\frac{2(1-\alpha)}{1+\alpha}},$$

where $\eta_d(x_1^\circ) = \sqrt{\mu} - \sqrt{\mu + v}$. With some basic substitution of notations, we complete the proof of Theorem 2.

6. On Nesterov's Momentum with Noise

This section briefly discusses Nesterov's accelerated gradient descent method (NAG) under the gradient with incomplete information. Recall the NAG, and the algorithms are set as

$$\begin{cases} y_{k+1} = x_k - s\nabla f(x_k) + s\xi_k \\ x_{k+1} = y_{k+1} + \alpha(y_{k+1} - y_k). \end{cases} \tag{48}$$

Recall the classical analysis for NAGs in Shi et al. (2022), and there are two kinds of NAGs, NAG-`SC` and NAG-`C`.

6.1 NAG-SC

For NAG-SC, Nesterov's accelerated gradient descent method for the μ-strongly convex objective, the momentum coefficient is set as $\alpha = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$. Notably, though the name, NAG-SC, comes from the μ-strongly convex case, here, the algorithm works on the non-convex objective. Similarly, plugging the two high-resolution Taylor expansions, (1) and (2), into the NAG-SC, we then get the high-resolution SDE for NAG-SC as

$$\ddot{X} + \left(2\sqrt{\mu} + \sqrt{s}\nabla^2 f(X)\right)\dot{X} + (1+\sqrt{\mu s})\nabla f(X) = \sqrt[4]{s}\dot{W}. \tag{49}$$

Here, when the momentum coefficient is set $\alpha = 0.9$ and the learning rate s is small enough, a simple transformation tells us that the main part for the logarithm of exponential decay constant in Theorem 2 for the NAG-SC high-resolution SDE (49) here is

$$-\frac{2H_{(1+\sqrt{\mu s})f}}{\sqrt{s}/\left(2\sqrt{\mu} + \sqrt{s}\nabla^2 f\right)} = -\frac{2}{1+\alpha}\cdot\left(1+\beta\nabla^2 f(x)\right)\cdot\frac{2H_f}{\beta(s,\alpha)} \approx -\frac{2H_f}{\beta(s,\alpha)}.$$

Meanwhile, taking a simple calculation, when the learning rate s is very small, we can obtain the asymptotic estimate for $|\eta_d(x_1^\circ)|$ in (47) as

$$|\eta_d(x_1^\circ)| = \sqrt{\mu + \frac{sv^2}{4} + v} - \sqrt{\mu} - \frac{\sqrt{s}v}{2} \approx \sqrt{\mu+v} - \sqrt{\mu}.$$

Therefore, we find when the noise exists, the quantitive estimate of the exponential decay constant, $\lambda_{s,\alpha}$, in (17) is almost the same between the high-resolution SDE (49) and the low-resolution SDE (4) (the hp-dependent SDE) for NAG-SC.

6.2 NAG-C

Similarly, we consider NAG-C, Nesterov's accelerated gradient descent method for the general convex objective, which works on the non-convex objective function. Plugging the two high-resolution Taylor expansions, (1) and (2), into the NAG-C, we then get the high-resolution SDE for NAG-C as

$$\ddot{X} + \left(\frac{3}{t} + \sqrt{s}\nabla^2 f(x)\right)\dot{X} + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(X) = \sqrt[4]{s}\dot{W}. \tag{50}$$

Consider that the learning rate s is small enough. When the time t is small enough, that is, $t \to 0$, the central part for the logarithm of exponential decay constant in Theorem 2 is

$$-\frac{2H_{\left(1+\frac{3\sqrt{s}}{2t}\right)f}}{\sqrt{s}/\left(\frac{3}{t} + \sqrt{s}\nabla^2 f(X)\right)} \approx -\frac{1}{k}\left(1+\frac{3}{2k}\right)\cdot\frac{6H_f}{s}.$$

Here, we can find the convergence speed is faster at the beginning. However, the change is very sharp, thus, the convergence rate decays very fast. Here, we can also find the high-resolution SDE (50) for NAG-C is not faster than the hp-dependent SDE (4) under the existence of noise.

7. Conclusion

In (Shi et al., 2023), the authors show the linear convergence of SGD in non-convex optimization and quantify the linear rate as the exponential of the negatively inverse learning rate by taking its continuous surrogate. This paper presents the theoretical perspective on the linear convergence of SGD with momentum. Different from SGD, the two hyperparameters together, the learning rate s and the momentum coefficient α, play a significant role in the linear convergence rate in the non-convex optimization. Taking the hp-dependent SDE, we proceed further to use modern mathematical tools such as hypocoercivity and semiclassical analysis to analyze the dynamics of SGD with momentum in a continuous-time model. We demonstrate how the linear rate of convergence and the final gap for SGD only for the learning rate s varies with the momentum coefficient α when the momentum is added. Finally, we also briefly analyze the Nesterov momentum under the existence of noise, which has no essential difference from the standard momentum.

Similarly, for understanding theoretically the stochastic optimization via SDEs, the pressing question is to characterize better the gap between the stationary distribution of the hp-dependent SDE and that of the discrete SGD with momentum (Pavliotis, 2014). Moreover, Theorem 3 is a bound for the algorithms in finite time T, which is based on the numerical analysis and not an algorithmic bound. More generally, it would be of interest to extend our results to SDEs with variable-dependence noise variance (Li et al., 2019; Chaudhari and Soatto, 2018). A related question is whether Theorem 3 can be improved to an algorithmic bound

$$\mathbb{E}f(x_k) - f^\star \leq O(\beta(s,\alpha) + (1 - \lambda_{s,\alpha}\beta(s,\alpha))^k),$$

or more possible output

$$\mathbb{E}f(x_k) - f^\star \leq O(\beta(s,\alpha) + (1 + \lambda_{s,\alpha}\beta(s,\alpha))^{-k}).$$

A straightforward direction is to extend our SDE-based analysis to various learning rate schedules used in training deep neural networks, such as the diminishing learning rate, cyclical learning rates, RMSProp, and Adam (Bottou et al., 2018; Smith, 2017; Tieleman and Hinton, 2012; Kingma and Ba, 2014). From Theorem 2, we know that the linear convergence rate has the explicit form as

$$\lambda_{s,\alpha} \asymp \exp\left(-\frac{2H_f}{s} \cdot \frac{2(1-\alpha)}{1+\alpha}\right).$$

Practically, we speed up the convergence by adjusting the two parameters, learning rate s and momentum coefficient α. If we consider from the view taking $-\frac{2H_f}{s} \cdot \frac{2(1-\alpha)}{1+\alpha}$ as a unity, the adjustment of the two parameters with the Morse barrier fixed is indeed equivalent to the change of the Morse barrier with the two parameters fixed. Based on the description above, the Morse barrier is also a factor to influence the convergence rate. Hence, perhaps these results may be used to choose the neural network architecture and the loss function to get a small value of the Morse saddle barrier H_f. Similarly, the hp-dependent SDE might give insights into the generalization properties of neural networks, such as implicit regularization (Zhang et al., 2016; Gunasekar et al., 2018).

Acknowledgments

We would like to thank Yu Sun help us practically run the deep learning in the University of California, Berkeley. We thank the action editor and two referees for their constructive comments that helped improve the presentation of this work. This work was supported by grant no.YSBR-034 of CAS and grant no. 12241105 of NSFC.

Appendix A. Technical Details for Section 2

Here, we rewrite the hp-dependent SDE (4) in the form of vector field as

$$d\begin{pmatrix} X \\ V \end{pmatrix} = \begin{pmatrix} V \\ -2\sqrt{\mu}V - \nabla f(X) \end{pmatrix} dt + \begin{pmatrix} 0 \\ \sqrt[4]{s}I \end{pmatrix} dW.$$

A.1 Derivation of the hp-dependent kinetic Fokker-Planck equation

First, we derive the corresponding Itô's formula for the hp-dependent SDE (4) as

Lemma 16 (Itô's lemma) *For any $f \in C^\infty(\mathbb{R}^d)$ and $u \in C^\infty\left([0,+\infty)\times\mathbb{R}^d\times\mathbb{R}^d\right)$, let (X,V) be the solution to the hp-dependent SDE (4). Then, we have*

$$\begin{aligned} du(t,X,V) = &\left(\frac{\partial u}{\partial t} + V\cdot\nabla_x u - \nabla_x f\cdot\nabla_v u - 2\sqrt{\mu}V\cdot\nabla_v u + \frac{\sqrt{s}}{2}\Delta_v u\right)dt \\ &+\sqrt[4]{s}\left(\sum_{i=1}^{d}\frac{\partial u}{\partial v_i}\right)dW. \end{aligned} \tag{51}$$

Let $g \in C^\infty\left(\mathbb{R}^d\times\mathbb{R}^d\right)$. Then, for any $\tau < t$, we assume

$$u(\tau,x,v) = \mathbb{E}[g(X(t),V(t))|X(\tau)=x, V(\tau)=v]. \tag{52}$$

Directly, from (52), we can obtain $u(t,x,v) = g(x,v)$. Hence, we have

$$\mathbb{E}\left[u(t,X(t),V(t)) - u(\tau,X(\tau),V(\tau))\big|X(\tau)=x,V(\tau)=v\right] = 0. \tag{53}$$

According to Ito's integral $\mathbb{E}\left[\int_\tau^t h dW\right] = 0$, then we get the backward Kolmogorov equation for $u(\tau,x,v)$ defined in (52) as

$$\begin{cases} \dfrac{\partial u}{\partial \tau} = -v\cdot\nabla_x u + \nabla_x f\cdot\nabla_v u + 2\sqrt{\mu}v\cdot\nabla_v u - \dfrac{\sqrt{s}}{2}\Delta_v u \\ u(t,x,v) = g(x,v). \end{cases} \tag{54}$$

Now, let $\rho_{s,\alpha}(t,x,v) = \rho_{s,\alpha}(X(t)=x, V(t)=v)$ be the evolving probability density with the initial as $\rho_{s,\alpha}(0,x,v)$. Taking $\tau = 0$, then (53) can be written as

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(t,y,w)p(t,y,w|0,x,v)dydw - u(0,x,v) = 0.$$

According to Chapman-Kolmogorov equation, we can obtain the expectation $\mathbb{E}[u(t,x,v)]$ is a constant, that is,

$$\begin{aligned}
\mathbb{E}[u(t,x,v)] &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(t,x,v)\rho_{s,\alpha}(t,x,v)dxdv \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(t,y,w)\rho_{s,\alpha}(t,y,w)dydw \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(t,y,w)\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} p(t,y,w|0,x,v)\rho_{s,\alpha}(0,x,v)dxdv\right)dydw \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(t,y,w)p(t,y,w|0,x,v)dydw\right)\rho_{s,\alpha}(0,x,v)dxdv \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u(0,x,v)\rho_{s,\alpha}(0,x,v)dxdv \\
&= \mathbb{E}[u(0,x,v)].
\end{aligned}$$

Hence, by differentiating the expectation $\mathbb{E}[u(t,x,v)]$, we can obtain

$$\begin{aligned}
\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u\frac{\partial\rho_{s,\alpha}}{\partial t}dxdv &= -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{\partial u}{\partial t}\rho_{s,\alpha}dxdv \\
&= -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}(-v\cdot\nabla_x u + \nabla_x f\cdot\nabla_v u + 2\sqrt{\mu}v\cdot\nabla_v u - \frac{\sqrt{s}}{2}\Delta_v u)\rho_{s,\alpha}dxdv \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u\left[-v\cdot\nabla_x\rho_{s,\alpha} + \nabla_x f\cdot\nabla_v\rho_{s,\alpha} + 2\sqrt{\mu}\nabla_v\cdot(v\rho_{s,\alpha}) + \frac{\sqrt{s}}{2}\Delta_v\rho_{s,\alpha}\right]dxdv.
\end{aligned}$$

Since u is an arbitrary function, we complete the derivation of the hp-dependent kinetic Fokker-Planck equation.

A.2 The existence and uniqueness of Gibbs invariant distribution

Recall the equivalent form of the hp-dependent kinetic Fokker-Planck equation (25)

$$\frac{\partial h_{s,\alpha}}{\partial t} = -\mathscr{L}_{s,\alpha}h_{s,\alpha},$$

where the linear operator $\mathscr{L}_{s,\alpha}$ is expressed as

$$\mathscr{L}_{s,\alpha}h_{s,\alpha} = \mathscr{T} + \mathscr{D}_{s,\alpha} = v\cdot\nabla_x - \nabla_x f\cdot\nabla_v + 2\sqrt{\mu}v\cdot\nabla_v - \frac{\sqrt{s}}{2}\Delta_v.$$

Now, we start to prove the existence and uniqueness of the Gibbs invariant distribution. Apparently, the Gibbs invariant distribution $\mu_{s,\alpha}$ is a steady solution. Assume there exists another steady solution $\vartheta_{s,\alpha}$, then $h_{s,\alpha} = \vartheta_{s,\alpha}\mu_{s,\alpha}^{-1}$ should satisfy

$$\mathscr{L}_{s,\alpha}h_{s,\alpha} = 0.$$

Taking integration by parts, for the transport operator $\mathscr{T}$, we have

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{s,\alpha}\mu_{s,\alpha}(\mathscr{T}h_{s,\alpha})dxdv = 0;$$

while for the diffusion operator $\mathscr{D}_{s,\alpha}$, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{s,\alpha}\mu_{s,\alpha}(\mathscr{D}_{s,\alpha}h_{s,\alpha})dxdv &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{s,\alpha}\mu_{s,\alpha}\left(-2\sqrt{\mu}v\cdot\nabla_v + \frac{\sqrt{s}}{2}\Delta_v\right)h_{s,\alpha}dxdv\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} -2\sqrt{\mu}h_{s,\alpha}\mu_{s,\alpha}v\cdot\nabla_v h_{s,\alpha} - \frac{\sqrt{s}}{2}\nabla_v h_{s,\alpha}\cdot\nabla_v(h_{s,\alpha}\mu_{s,\alpha})dxdv\\ &= -\frac{\sqrt{s}}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_v h_{s,\alpha}\|^2\mu_{s,\alpha}dxdv + \left(-2\sqrt{\mu}+\frac{\sqrt{s}}{\beta}\right)...\\ &= \frac{\sqrt{s}}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_v h_{s,\alpha}\|^2\mu_{s,\alpha}dxdv.\end{aligned}$$

Then, some basic calculations tell us that the linear operator $\mathscr{L}_{s,\alpha}$ satisfies

$$\begin{aligned}0 &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{s,\alpha}\mathscr{L}_{s,\alpha}h_{s,\alpha}\mu_{s,\alpha}dxdv\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{s,\alpha}\mathscr{T}(h_{s,\alpha}\mu_{s,\alpha})dxdv + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} h_{s,\alpha}\mathscr{D}_{s,\alpha}(h_{s,\alpha}\mu_{s,\alpha}dxdv\\ &= -\frac{\sqrt{s}}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_v h_{s,\alpha}\|^2\mu_{s,\alpha}dxdv \le 0.\end{aligned}$$

Hence, we obtain $h_{s,\alpha}(x,v) = g_{s,\alpha}(x)$. Furthermore, according to $\mathscr{L}_{s,\alpha}h_{s,\alpha} = 0$, we have $-v\cdot\nabla_x g_{s,\alpha} = 0$. Because v is arbitrary, we know $g_{s,\alpha} = C$. With both $\vartheta_{s,\alpha}$ and $\mu_{s,\alpha}$ being probability densities, $C = 1$. Hence, the proof of the existence and uniqueness of the Gibbs invariant distribution is complete.

A.3 Proof of Lemma 2

Recall that Section 5.1 shows that the transition probability density $\rho_{s,\alpha}(t,\cdot,\cdot) \in C^1([0,+\infty), L^2(\mu_{s,\alpha}^{-1}))$ governed by the hp-dependent kinetic Fokker-Planck equation (15) is equivalent to the function $\psi_s(t,\cdot,\cdot)$ in $C^1([0,+\infty), L^2(\mathbb{R}^d,\mathbb{R}^d))$ governed by the differential equation (42). Moreover, in Section 5.1, we have shown that the spectrum of the Kramers operator $\mathscr{K}_{s,\alpha}$ satisfies

$$0 = \zeta_{s,\alpha;0} < \zeta_{s,\alpha;1} \le \cdots \le \zeta_{s,\alpha;\ell} \le \cdots < +\infty.$$

Since $L^2(\mathbb{R}^d,\mathbb{R}^d)$ is a Hilbert space, there exists a standard orthogonal basis corresponding to the spectrum of the Kramers operator $\mathscr{K}_{s,\alpha}$:

$$\mu_{s,\alpha} = \phi_{s,\alpha;0},\ \phi_{s,\alpha;1},\ \ldots,\ \phi_{s,\alpha;\ell},\ \ldots \in L^2(\mathbb{R}^d,\mathbb{R}^d).$$

Then, for any initialization $\psi_s(0,\cdot,\cdot) \in L^2(\mathbb{R}^d,\mathbb{R}^d)$, there exist a family of constants c_ℓ $(\ell = 1,2,\ldots)$ such that

$$\psi_{s,\alpha}(0,\cdot,\cdot) = \sqrt{\mu_{s,\alpha}} + \sum_{\ell=1}^{+\infty} c_\ell\phi_{s,\alpha;\ell}.$$

Thus, the solution to the partial differential equation (42) is

$$\psi_{s,\alpha}(t,\cdot,\cdot) = \sqrt{\mu_{s,\alpha}} + \sum_{\ell=1}^{+\infty} c_\ell e^{-\zeta_{s,\alpha;\ell}t}\phi_{s,\alpha;\ell}.$$

Recognizing the transformation $\psi_{s,\alpha}(t,\cdot,\cdot) = \rho_{s,\alpha}(t,\cdot,\cdot)/\sqrt{\mu_{s,\alpha}}$, we recover

$$\rho_{s,\alpha}(t,\cdot,\cdot) = \mu_{s,\alpha} + \sum_{\ell=1}^{+\infty} c_\ell e^{-\zeta_{s,\alpha;\ell}t}\phi_{s,\alpha;\ell}\sqrt{\mu_{s,\alpha}}.$$

Note that $\zeta_{s,\alpha;\ell}$ is positive for $\ell \geq 1$. Thus, the proof is complete.

Appendix B. Technical Details for Section 3

B.1 Proof of Proposition 7

By Lemma 2, let $\rho_{s,\alpha}(t,\cdot,\cdot) \in C^1([0,+\infty), L^2(\mu_{s,\alpha}^{-1}))$ denote the unique transition probability density of the solution to the hp-dependent SDE (4). Taking an expectation, we get

$$\mathbb{E}[X_{s,\alpha}(t)] = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} x\rho_{s,\alpha}(t,x,v)dvdx.$$

Hence, the uniqueness has been proved. Using the Cauchy–Schwarz inequality, we obtain:

$$\begin{aligned}
\|\mathbb{E}[X_{s,\alpha}(t)]\| &\leq \left\|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} x(\rho_s(t,\cdot,\cdot) - \mu_{s,\alpha})dvdx\right\| + \left\|\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} x\mu_{s,\alpha}dvdx\right\| \\
&\leq \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x\|^2\mu_{s,\alpha}dvdx\right)^{\frac{1}{2}} \left(\|\rho_{s,\alpha}(t,\cdot,\cdot) - \mu_{s,\alpha}\|_{L^2(\mu_s^{-1})} + 1\right) \\
&\leq \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x\|^2\mu_{s,\alpha}dvdx\right)^{\frac{1}{2}} \left(e^{-\lambda_{s,\alpha}t}\|\rho_{s,\alpha}(0,\cdot,\cdot) - \mu_{s,\alpha}\|_{L^2(\mu_s^{-1})} + 1\right) \\
&< +\infty,
\end{aligned}$$

where the integrability $\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x\|^2\mu_{s,\alpha}dvdx$ is due to the fact that the objective f satisfies the Villani conditions. The existence of a global solution to the hp-dependent SDE (4) is thus esta- -blished.

For the strong convergence, the hp-dependent SDE (4) corresponds to the Milstein scheme in numerical methods. The original result is obtained by Mil'shtein (1975) and Talay (1982); Pardoux and Talay (1985), independently. We refer the readers to Kloeden and Platen (1992, Theorem 10.3.5 and Theorem 10.6.3), which studies numerical schemes for stochastic differential equation. For the weak convergence, we can obtain numerical errors by using both the Euler-Maruyama scheme and Milstein scheme. The original result is obtained by Mil'shtein (1986) and Pardoux and Talay (1985); Talay (1984) independently and Kloeden and Platen (1992, Theorem 14.5.2) is also a well-known reference. Furthermore, there exists a more accurate estimate of $B(T)$ shown in Bally and Talay (1996). The original proofs in the aforementioned references only assume finite smoothness such as $C^6(\mathbb{R}^d)$ for the objective function.

Appendix C. Technical Details for Section 4

C.1 Proof of Lemma 8

Here, we show the basic calculations in detail for the readers of all fields. Since the calculations are basic operations, if the readers are familiar with them, you can ignore them.

(i) For any $g_1, g_2 \in L^2(\mu_{s,\alpha})$, with the definition of conjugate linear operators, we have

$$\begin{aligned}\langle \mathscr{A}^\star_{s,\alpha} g_1, g_2\rangle_1 = \langle g_1, \mathscr{A}_{s,\alpha} g_2\rangle_1 &= \sqrt[4]{\frac{s}{4}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (g_1 \cdot \nabla_v g_2)\, \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x \\ &= -\sqrt[4]{\frac{s}{4}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left[\nabla_v \cdot (g_1 \mu_{s,\alpha})\right] g_2 \mathrm{d}v\mathrm{d}x \\ &= -\sqrt[4]{\frac{s}{4}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(\nabla_v \cdot g_1 - \frac{2v}{\beta} g_1\right) g_2 \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x \\ &= \left\langle \sqrt[4]{\frac{s}{4}} \left(-\nabla_v + \frac{2v}{\beta}\right) g_1, g_2 \right\rangle_1\end{aligned}$$

and

$$\begin{aligned}\langle \mathscr{C}^\star_{s,\alpha} g_1, g_2\rangle_1 = \langle g_1, \mathscr{C} g_2\rangle_1 &= \sqrt[4]{\frac{s}{4}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (g_1 \cdot \nabla_x g_2)\, \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x \\ &= -\sqrt[4]{\frac{s}{4}} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \left[\nabla_x \cdot (g_1 \mu_{s,\alpha})\right] g_2 \mathrm{d}v\mathrm{d}x \\ &= -\sqrt[4]{\frac{s}{4}} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left(\nabla_x \cdot g_1 - \frac{2\nabla_x f}{\beta} g_1\right) g_2 \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x \\ &= \left\langle \sqrt[4]{\frac{s}{4}} \left(-\nabla_x + \frac{2\nabla_x f}{\beta}\right) g_1, g_2 \right\rangle_1 .\end{aligned}$$

Hence, we obtain the diffusion operator as $\mathscr{D}_{s,\alpha} = \mathscr{A}^\star_{s,\alpha}\mathscr{A}_{s,\alpha}$.

(ii) Since both $\mathscr{A}_{s,\alpha}$ and $\mathscr{A}^\star_{s,\alpha}$ are only the linear operators about v and the linear operator $\mathscr{C}_{s,\alpha}$ about x, both $\mathscr{A}_{s,\alpha}$ and $\mathscr{A}^\star_{s,\alpha}$ commutes with $\mathscr{C}_{s,\alpha}$, that is,

$$[\mathscr{A}_{s,\alpha}, \mathscr{C}_{s,\alpha}] = [\mathscr{A}^\star_{s,\alpha}, \mathscr{C}_{s,\alpha}] = 0.$$

For the commutator between $\mathscr{A}$ and $\mathscr{A}^\star$, we have

$$\begin{aligned}[\mathscr{A}, \mathscr{A}^\star] &= \frac{\sqrt{s}}{2}\left[\nabla_v, -\nabla_v + \frac{2v}{\beta}\right] \\ &= \frac{\sqrt{s}}{2}\left[\nabla_v\left(-\nabla_v + \frac{2v}{\beta}\right) - \left(-\nabla_v + \frac{2v}{\beta}\right)\nabla_v\right] \\ &= 2\sqrt{\mu}\mathbf{I}_{d\times d}.\end{aligned}$$

(iii) Taking the simple equation $(v\cdot\nabla_x - \nabla_x f\cdot\nabla_v)\,\mu_{s,\alpha} = 0$, for any $g_1, g_2 \in L^2(\mu_{s,\alpha})$,we have

$$\begin{aligned}\langle \mathscr{T} g_1, g_2\rangle_1 &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (v\cdot\nabla_x - \nabla_x f\cdot\nabla_v)\, g_1 g_2 \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x \\ &= -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (v\cdot\nabla_x - \nabla_x f\cdot\nabla_v)\, (g_2\mu_{s,\alpha}) g_1 \mathrm{d}v\mathrm{d}x \\ &= -\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (v\cdot\nabla_x - \nabla_x f\cdot\nabla_v)\, g_2 g_1 \mu_{s,\alpha} \mathrm{d}v\mathrm{d}x \\ &= -\langle g_1, \mathscr{T} g_2\rangle_1 .\end{aligned}$$

The commutator between $\mathscr{A}_{s,\alpha}$ and $\mathscr{T}$ is

$$\begin{aligned}[\mathscr{A}_{s,\alpha}, \mathscr{T}] &= \mathscr{A}_{s,\alpha}\mathscr{T} - \mathscr{T}\mathscr{A}_{s,\alpha} \\ &= \sqrt[4]{\frac{s}{4}}\left[\nabla_v\left(v\cdot\nabla_x - \nabla_x f\cdot\nabla_v\right) - \left(v\cdot\nabla_x - \nabla_x f\cdot\nabla_v\right)\nabla_v\right] \\ &= \sqrt[4]{\frac{s}{4}}\nabla_x = \mathscr{C}_{s,\alpha}.\end{aligned}$$

Similarly, the commutator between $\mathscr{C}_{s,\alpha}$ and $\mathscr{T}$ is

$$\begin{aligned}[\mathscr{C}_{s,\alpha}, \mathscr{T}] &= \mathscr{C}_{s,\alpha}\mathscr{T} - \mathscr{T}\mathscr{C}_{s,\alpha} \\ &= \sqrt[4]{\frac{s}{4}}\left[\nabla_x\left(v\cdot\nabla_x - \nabla_x f\cdot\nabla_v\right) - \left(v\cdot\nabla_x - \nabla_x f\cdot\nabla_v\right)\nabla_x\right] \\ &= -\sqrt[4]{\frac{s}{4}}\nabla_x^2 f\cdot\nabla_v.\end{aligned}$$

C.2 Proof of Lemma 9

For any $g \in L^2(\mu_{s,\alpha}^{-1})$, we can write down it as

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |g(v,x)|^2 \mu_{s,\alpha}^{-1} dvdx < +\infty.$$

Then, we can take the following simple calculation as

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |g(x,v)|dvdx &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |g(x,v)|\mu_{s,\alpha}^{-\frac{1}{2}}\mu_{s,\alpha}^{\frac{1}{2}}dvdx \\ &\le \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |g(x,v)|^2\mu_{s,\alpha}^{-1}dvdx\right)^{\frac{1}{2}}\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\mu_{s,\alpha}dvdx\right)^{\frac{1}{2}} \\ &= \left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |g(x,v)|^2\mu_{s,\alpha}^{-1}dvdx\right)^{\frac{1}{2}} < +\infty.\end{aligned}$$

This completes the proof.

C.3 Technical Details in Section 4.2.1

Here, we compute the derivative in detail as

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\left\langle h_{s,\alpha}, h_{s,\alpha}\right\rangle_{H^1(\mu_{s,\alpha})} = &-\underbrace{\left\langle \mathscr{L}_{s,\alpha}h_{s,\alpha}, h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{I}} - \underbrace{\left\langle \mathscr{A}_{s,\alpha}\mathscr{L}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{II}} \\ &-\underbrace{\left\langle \mathscr{C}_{s,\alpha}\mathscr{L}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{III}}.\end{aligned}$$

(1) For the term **I**, we have

$$\mathbf{I} = \left\langle \mathscr{L}_{s,\alpha}h_{s,\alpha}, h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} = \left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} = \left\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\right\|^2_{L^2(\mu_{s,\alpha})}.$$

(2) For the term **II**, we have

$$\mathbf{II} = \underbrace{\left\langle \mathscr{A}_{s,\alpha}\mathscr{A}_{s,\alpha}^{\star}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{II}_a} + \underbrace{\langle \mathscr{A}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{II}_b}$$

– For the term $\mathbf{II}_a$, we have

$$\begin{aligned}
\mathbf{II}_a &= \left\langle \mathscr{A}_{s,\alpha}\mathscr{A}_{s,\alpha}^{\star}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\
&= \left\langle \mathscr{A}_{s,\alpha}^{\star}\mathscr{A}_{s,\alpha}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + \left\langle \left[\mathscr{A}_{s,\alpha}, \mathscr{A}_{s,\alpha}^{\star}\right]\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\
&= \left\langle \mathscr{A}_{s,\alpha}^{2}h_{s,\alpha}, \mathscr{A}_{s,\alpha}^{2}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + \left\langle \left[\mathscr{A}_{s,\alpha}, \mathscr{A}_{s,\alpha}^{\star}\right]\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\
&= \left\langle \mathscr{A}_{s,\alpha}^{2}h_{s,\alpha}, \mathscr{A}_{s,\alpha}^{2}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + 2\sqrt{\mu}\left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\
&= \|\mathscr{A}_{s,\alpha}^{2}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}^{2} + 2\sqrt{\mu}\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}^{2}
\end{aligned}$$

– For the term $\mathbf{II}_b$, we have

$$\begin{aligned}
\mathbf{II}_b &= \langle \mathscr{A}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle \mathscr{T}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle [\mathscr{A}_{s,\alpha}, \mathscr{T}]h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle \mathscr{C}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})}
\end{aligned}$$

(3) For the term **III**, we have

$$\mathbf{III} = \underbrace{\left\langle \mathscr{C}_{s,\alpha}\mathscr{A}_{s,\alpha}^{\star}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{III}_a} + \underbrace{\langle \mathscr{C}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{III}_b}$$

(1) For the term $\mathbf{III}_a$, we have

$$\begin{aligned}
\mathbf{III}_a &= \left\langle \mathscr{C}_{s,\alpha}\mathscr{A}_{s,\alpha}^{\star}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} = \left\langle \mathscr{A}_{s,\alpha}^{\star}\mathscr{C}_{s,\alpha}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle \mathscr{C}_{s,\alpha}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} = \langle \mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}^{2}
\end{aligned}$$

(2) For the term $\mathbf{III}_b$, we have

$$\begin{aligned}
\mathbf{III}_b &= \langle \mathscr{C}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle \mathscr{T}\mathscr{C}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle [\mathscr{C}_{s,\alpha}, \mathscr{T}]\, h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle [\mathscr{C}_{s,\alpha}, \mathscr{T}]\, h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= -\left\langle \nabla_x^2 f \cdot \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}
\end{aligned}$$

C.4 Proof of Lemma 12

(i) By a density argument, we may assume that g is smooth and decays fast enough at infinity. Then, taking the identity $\nabla_x \mu_{s,\alpha} = -\left(2\nabla f/\beta(s,\alpha)\right)\mu_{s,\alpha}$ and by integration by parts, we have

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2 \mu_{s,\alpha}dxdv &= -\frac{\beta(s,\alpha)}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\nabla_x f\cdot\nabla_x\mu_{s,\alpha}dxdv\\ &= \frac{\beta(s,\alpha)}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\nabla_x\cdot\left(g^2\nabla_x f\right)\mu_{s,\alpha}dxdv\\ &= \frac{\beta(s,\alpha)}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2(\Delta_x f)\mu_{s,\alpha}dxdv + \beta(s,\alpha)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g\left(\nabla_x g\cdot\nabla_x f\right)\mu_{s,\alpha}dxdv\end{aligned}$$

By Cauchy-Schwarz inequality

$$\begin{aligned}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv \le& \frac{\beta(s,\alpha)}{2}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2(\Delta_x f)^2\mu_{s,\alpha}dxdv}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv}\\ &+\beta(s,\alpha)\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x g\|^2\mu_{s,\alpha}dxdv}\end{aligned}$$

With the Villani Condition-(II), $\|\nabla_x^2 f\|_2 \le C(1+\|\nabla f\|)$, we have

$$(\Delta f)^2 \le d^2\|\nabla^2 f\|_2^2 \le d^2C^2\left(1+\|\nabla f\|\right)^2 \le 4d^2C^2\left(1+\|\nabla f\|^2\right).$$

Hence, we can obtain

$$\begin{aligned}&\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv\\ \le& dC\beta(s,\alpha)\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv}\\ &+\beta(s,\alpha)\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x g\|^2\mu_{s,\alpha}dxdv}\\ \le& dC\beta(s,\alpha)\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv + \sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv}\right)\\ &+\beta(s,\alpha)\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv}\sqrt{\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x g\|^2\mu_{s,\alpha}dxdv}\\ \le& dC\beta(s,\alpha)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv + \frac{1}{4}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv + d^2C^2\beta^2(s,\alpha)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv\\ &+\frac{1}{4}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv + \beta^2(s,\alpha)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x g\|^2\mu_{s,\alpha}dxdv\end{aligned}$$

where the last inequality follows Cauchy-Schwartz inequality. Therefore, we have

$$\begin{aligned}\frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2 g^2\mu_{s,\alpha}dxdv \le& \left(dC\beta(s,\alpha)+d^2C^2\beta^2(s,\alpha)\right)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g^2\mu_{s,\alpha}dxdv\\ &+\ \beta^2(s,\alpha)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x g\|^2\mu_{s,\alpha}dxdv.\end{aligned}$$

Multiplied by two, taking $\kappa_1(s,\alpha) = \max\{2(dC\beta(s,\alpha) + d^2C^2\beta^2(s,\alpha)),\ 2\beta^2(s,\alpha)\}$, we have

$$\|(\nabla_x f)g\|^2_{L^2(\mu_{s,\alpha})} \le \kappa_1(s,\alpha)\left(\|g\|^2_{L^2(\mu_{s,\alpha})} + \|\nabla_x g\|^2_{L^2(\mu_{s,\alpha})}\right).$$

(ii) Similarly, with the Villani Condition-(II), $\|\nabla^2 f\|_2^2 \le 2C^2\left(1 + \|\nabla f\|^2\right)$, we have

$$\begin{aligned}\left\|\|\nabla_x^2 f\|_2 g\right\|^2_{L^2(\mu_{s,\alpha})} &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|\nabla_x^2 f\|_2^2 \cdot \|g\|^2 \mu_{s,\alpha} dxdv \\ &\le 2C^2\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|g\|^2\mu_{s,\alpha}dxdv + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x f\|^2\|g\|^2\mu_{s,\alpha}dxdv\right).\end{aligned}$$

With the inequality (i), we can obtain directly

$$\begin{aligned}\left\|\|\nabla_x^2 f\|_2 g\right\|^2_{L^2(\mu_{s,\alpha})} \le\ & 2C^2(1+\kappa_1(s,\alpha))\left(\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|\nabla_x g\|^2\mu_{s,\alpha}dxdv + \int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\|g\|^2\mu_{s,\alpha}dxdv\right) \\ =\ & 2C^2(1+\kappa_1(s,\alpha))\left(\|g\|^2_{L^2(\mu_{s,\alpha})} + \|\nabla_x g\|^2_{L^2(\mu_{s,\alpha})}\right)\end{aligned}$$

Taking $\kappa_2(s,\alpha) = 2C^2(1+\kappa_1(s,\alpha))$, we complete the proof.

C.5 Technical Details in Section 4.2.3

Here, we compute the derivative of the mixed term in detail as

$$\begin{aligned}&\frac{d}{dt}\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\ &\quad= \underbrace{\langle \mathscr{A}_{s,\alpha}\mathscr{L}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}\mathscr{L}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{IV}} \\ &\quad= \underbrace{\langle \mathscr{A}_{s,\alpha}\mathscr{A}^\star\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}\mathscr{A}^\star_{s,\alpha}\mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{IV}_a} \\ &\qquad+ \underbrace{\langle \mathscr{A}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}\mathscr{T}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})}}_{\mathbf{IV}_b}\end{aligned}$$

- For the term $\mathbf{IV}_a$, we have

$$\begin{aligned}\mathbf{IV}_a &= \left\langle \mathscr{A}_{s,\alpha}\mathscr{A}^\star_{s,\alpha}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + \left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}\mathscr{A}^\star_{s,\alpha}\mathscr{A}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\ &= \left\langle \mathscr{A}^\star_{s,\alpha}\mathscr{A}^2_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + \left\langle \left[\mathscr{A}_{s,\alpha}, \mathscr{A}^\star_{s,\alpha}\right]\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\ &\quad+ \left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{A}^\star_{s,\alpha}\mathscr{C}_{s,\alpha}\mathscr{A}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\ &= 2\left\langle \mathscr{A}^2_{s,\alpha}h_{s,\alpha}, \mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} + 2\sqrt{\mu}\left\langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\right\rangle_{L^2(\mu_{s,\alpha})} \\ &\ge -2\|\mathscr{A}^2_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} - 2\sqrt{\mu}\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\end{aligned}$$

- For the term $\mathbf{IV}_b$, we have

$$\begin{aligned}
\mathbf{IV}_b &= \langle \mathscr{A}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}\mathscr{T}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle \mathscr{A}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{T}\mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} + \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, [\mathscr{C}_{s,\alpha}, \mathscr{T}]\, h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle \mathscr{A}_{s,\alpha}\mathscr{T}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} - \langle \mathscr{T}\mathscr{A}_{s,\alpha}h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} - \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, [\mathscr{T}, \mathscr{C}_{s,\alpha}]\, h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&= \langle (\mathscr{A}_{s,\alpha}\mathscr{T} - \mathscr{T}\mathscr{A}_{s,\alpha})h_{s,\alpha}, \mathscr{C}_{s,\alpha}h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} - \langle \mathscr{A}_{s,\alpha}h_{s,\alpha}, [\mathscr{T}, \mathscr{C}_{s,\alpha}]\, h_{s,\alpha}\rangle_{L^2(\mu_{s,\alpha})} \\
&\geq \|\mathscr{C}_{s,\alpha}h_{s,\alpha}\|^2_{L^2(\mu_{s,\alpha})} - \sqrt{\kappa_3(s,\alpha)}\|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} \left(\|\mathscr{A}_{s,\alpha}\mathscr{C}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})} + \|\mathscr{A}_{s,\alpha}h_{s,\alpha}\|_{L^2(\mu_{s,\alpha})}\right)
\end{aligned}$$

I

References

V. Bally and D. Talay. The law of the Euler scheme for stochastic differential equations: Ii. convergence rate of the density. *Monte Carlo Methods and Applications*, 2(2):93–128, 1996.

Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer, 2012.

L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes II: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7(1):69–99, 2005.

K. Caluya and A. Halder. Gradient flow algorithms for density propagation in stochastic systems. *IEEE Transactions on Automatic Control*, 2019.

P. Chaudhari and S. Soatto. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

P. Chaudhari, A. Oberman, S. Osher, S. Soatto, and G. Carlier. Deep relaxation: partial differential equations for optimizing deep neural networks. *Research in the Mathematical Sciences*, 5(3):30, 2018.

L. Evans. *An introduction to stochastic differential equations*, volume 82. American Mathematical Society, 2012.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

E. Hazan, A. Rakhlin, and P. Bartlett. Adaptive online gradient descent. In *Advances in Neural Information Processing Systems*, pages 65–72, 2008.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

B. Helffer and F. Nier. *Hypoelliptic estimates and spectral theory for Fokker-Planck operators and Witten Laplacians*. Springer, 2005.

B. Helffer, M. Klein, and F. Nier. Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach. *Mat. Contemp.*, 26:41–85, 2004.

F. Hérau, M. Hitrik, and J. Sjöstrand. Tunnel effect and symmetries for Kramers–Fokker–Planck type operators. *Journal of the Institute of Mathematics of Jussieu*, 10(3):567–634, 2011.

Ruinan Jin and Xingkang He. Convergence of momentum-based stochastic gradient descent. In *2020 IEEE 16th International Conference on Control & Automation (ICCA)*, pages 779–784. IEEE, 2020.

M. I. Jordan. Dynamical, symplectic and stochastic perspectives on gradient-based optimization. In *Proceedings of the International Congress of Mathematicians, Rio de Janeiro*, volume 1, pages 523–550, 2018.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

P. E. Kloeden and E. Platen. The approximation of multiple stochastic integrals. *Stochastic Analysis and Applications*, 10(4):431–441, 1992.

A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Toronto*, 2009.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Q. Li, C. Tai, and W. E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 2101–2110. JMLR. org, 2017.

Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520, 2019.

W.-X. Li. Global hypoellipticity and compactness of resolvent for fokker-planck operator. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 11(4):789–815, 2012.

Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. *Advances in Neural Information Processing Systems*, 33:18261–18271, 2020.

S. Mandt, M. Hoffman, and D. Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 354–363, 2016.

P. A. Markowich and C. Villani. On the trend to equilibrium for the Fokker-Planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*, pages 1–29, 1999.

G. N. Mil'shtein. Approximate integration of stochastic differential equations. *Theory of Probability & Its Applications*, 19(3):557–562, 1975.

G. N. Mil'shtein. Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766, 1986.

E. Pardoux and D. Talay. Discretization and simulation of stochastic differential equations. *Acta Applicandae Math*, 3:23–47, 1985.

G. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60. Springer, 2014.

Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195 (1-2):79–148, 2022.

Bin Shi, Weijie Su, and Michael I Jordan. On learning rates and schrödinger operators. *Journal of Machine Learning Research*, 24(379):1–53, 2023.

L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

W. Su, S. Boyd, and E. Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17:1–43, 2016.

D. Talay. *Analyse numérique des équations différentielles stochastiques*. PhD thesis, Université Aix-Marseille I, 1982.

D. Talay. Efficient numerical schemes for the approximation of expectations of functionals of the solution of a SDE and applications. In *Filtering and Control of Random Processes*, pages 294–313. Springer, 1984.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

C. Villani. Hypocoercive diffusion operators. In *International Congress of Mathematicians*, volume 3, pages 473–498, 2006.

C Villani. *Hypocoercivity*, volume 202. Memoirs of the American Mathematical Society, 2009.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.