A Hessian-Aware Stochastic Differential Equation for Modelling SGD

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

Continuous-time approximation of Stochastic Gradient Descent (SGD) is a crucial tool to study its escaping behaviors from stationary points. However, existing stochastic differential equation (SDE) models fail to fully capture these behaviors, even for simple quadratic objectives. Built on a novel stochastic backward error analysis framework, we derive the Hessian-Aware Stochastic Modified Equation (`HA-SME`), an SDE that incorporates Hessian information of the objective function into both its drift and diffusion terms. Our analysis shows that `HA-SME` matches the order-best approximation error guarantee among existing SDE models in the literature, while achieving a significantly reduced dependence on the smoothness parameter of the objective. Further, for quadratic objectives, under mild conditions, `HA-SME` is proved to be the first SDE model that recovers exactly the SGD dynamics in the distributional sense. Consequently, when the local landscape near a stationary point can be approximated by quadratics, `HA-SME` is expected to accurately predict the local escaping behaviors of SGD.

1. Introduction

We consider unconstrained stochastic minimization problems of the form

$$\min_x f(x) := \mathbb{E}_\xi \left[F(x;\xi) \right], \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a smooth function and ξ is a random vector. In this work, we study Stochastic Gradient Descent (SGD) [52], the most popular approach for solving such problems, which updates iteratively as follows:

$$x_{k+1} = x_k - \eta_k \nabla F(x_k;\xi_k), \tag{2}$$

where $\eta_k > 0$ is the stepsize and $\nabla F(x_k;\xi_k)$ is an unbiased stochastic gradient at x_k. We focus on the constant stepsize regime where $\eta_k \equiv \eta$ for all $k = 0, 1, \ldots$, and try to explore the dynamics of SGD through the lens of Stochastic Differential Equations (SDEs). The primary aim of this paper is to derive an accurate SDE representation of SGD.

Our contributions. We develop a new Stochastic Backward Error Analysis (SBEA) to approximate the dynamics of constant-stepsize SGD. Our analysis yields an SDE, termed "Hessian-Aware Stochastic Modified Equation" (`HA-SME`), that integrates Hessian information into both its drift and diffusion terms. By incorporating local curvature information, `HA-SME` effectively tracks the escape phenomena of SGD. The well-posedness of `HA-SME` is assured under mild regularity assumptions. Moreover, under similar assumptions, the distribution evolution of `HA-SME` *exactly* matches that of SGD on quadratic functions. Our contributions are summarized as follows:

- **Failure modes of existing SDE proxies.** In this study, we identify several shortcomings of current SDE approximations in capturing the escaping behaviors of SGD. We present concrete examples of optimization problems, where SGD exhibits escaping behavior, while existing SDE approximations tend to stabilize around the stationary points. We further pinpoint that this failure mode stems from the negligence of higher-order terms involving Hessians in their derivations.

- **A new Hessian-aware SDE proxy.** We propose the SBEA framework, which enables us to design `HA-SME` capable of encompassing more error terms involving gradients and Hessians than existing SDEs in the error analysis. Specifically, the drift term and diffusion coefficient of `HA-SME` are defined by two power series, respectively. Utilizing a generating function approach, we demonstrate that these power series converge when the stepsize is below a certain constant threshold, and their limits possess analytical expressions. We provide sufficient conditions to ensure `HA-SME` is well-posed.

- **Approximation guarantees on general functions.** We establish that `HA-SME` adheres to an order-2 weak approximation error guarantee, aligning with the best-known results among existing SDE approximations. To further differentiate `HA-SME` from existing Stochastic Modified Equations (summarized in Table 1), we conduct a fine-grained analysis of the weak approximation error, taking into account of the smoothness parameter of the objective, which we denote as λ. For convex functions, our findings indicate that `HA-SME` admits a uniform-in-time approximation error of order $\mathcal{O}\left(\eta^2\right)$. In stark contrast, the error for `SME`-1 model is $\mathcal{O}\left(\eta\lambda\right)$, and for `SME`-2, it is $\mathcal{O}\left(\eta^2\lambda^2\right)$. Our results mark a notable improvement over the dependence on the smoothness parameter compared to existing SDE approximations.

- **Exact recovery on quadratic functions.** We then examine the quadratic functions, a common simplification used to approximate the local landscapes of non-linear objectives. We show that accurately mirroring the SGD dynamics (linear in this setting) with a linear SDE (OU process) is fundamentally impossible, due to a potential mismatch between the ranks of their covariance matrices. Nonetheless, we establish that under certain mild conditions, i.e., when eigenvectors of the Hessian and the noise covariance matrices align or a small stepsize is used, the extension of our proposed `HA-SME` into the complex domain is well-founded, and this extension *exactly* replicates the constant-stepsize SGD dynamics in a distributional sense. This marks the first work, to the best of our knowledge, in achieving exact recovery of SGD using continuous-time models.

We defer a detailed discussion of our motivation and related work to Appendix A. In Appendix B, we introduce the preliminaries, and then in Appendix C, we present our proposed SBEA framework, which serves as the key technique for deriving `HA-SME`.

2. Hessian-Aware Stochastic Modified Equation (`HA-SME`)

In this section, we introduce `HA-SME`, our proposed SDE proxy for SGD. The detailed construction of `HA-SME` is deferred to Appendix D, where we extend the concept of classical Backward Error Analysis to a noisy setting. Additionally, we adapt the term selection approach from Principle Flow [53], an ODE designed to approximate GD.

Referred Name	Stochastic Differential Equations	Weak Approximation Error
Stochastic Modified Equation (`SME`-1) [32, 36, 37]	$dX_t = -\nabla f(X_t)dt + \sqrt{\eta\Sigma(X_t)}dW_t$	$\mathcal{O}(\eta)$
Second-Order SME (`SME`-2) [32]	$dX_t = -\left(\nabla f(X_t) + \frac{\eta}{2}\nabla^2 f(X_t)\nabla f(X_t)\right)dt + \sqrt{\eta\Sigma(X_t)}dW_t$	$\mathcal{O}(\eta^2)$
`HA-SME` (this work)	Definition 1	$\mathcal{O}(\eta^2)$

Table 1: SDE proxies for SGD referred in this paper.

Definition 1. *We denote the following SDE as the Hessian-Aware Stochastic Modified Equation (`HA-SME`):*

$$dX_t = b(X_t)dt + D(X_t)dW_t, \tag{3}$$

$$\text{with } \; b(x) := U(x)\frac{\log(I - \eta\Lambda(x))}{\eta\Lambda(x)}U(x)^\intercal \nabla f(x), \; \text{ and}$$

$$D(x) := \sqrt{U(x)S(x)U(x)^\top}, \text{such that } [S(x)]_{i,j} := \frac{\left[U^\intercal \Sigma U\right]_{i,j}\log(1-\eta\lambda_i)(1-\eta\lambda_j)}{\eta\lambda_i\lambda_j - (\lambda_i + \lambda_j)},$$

where $U(x)$ and $\Lambda(x)$ are defined through the eigen-decomposition $\nabla^2 f(x) = U(x)\Lambda(x)U(x)^\intercal$, and $\Sigma(x)$ is the covariance matrix of the stochastic gradient. The diagonal elements of $\Lambda(x)$ are denoted by $\lambda_i(x)$. For conciseness, we omit the dependence of λ_i, U and Σ on x.

Remark 2 (Comparison with existing `SME`s)**.** *Comparing with `SME`-1 and `SME`-2 (see in Table 1), both the drift term and the diffusion coefficient of `HA-SME` incorporate the Hessian information, hence the name. By integrating Hessian into its diffusion coefficient, `HA-SME` provides a nuanced correction to the SDE noise. This adjustment is crucial for capturing the true dynamics of SGD, even when considering additive, state-independent noise models, such as $\nabla F(x,\xi) = \nabla f(x) + \xi$, where $\xi \sim \mathcal{N}(0,\Sigma)$. While discrete-time models treat noise as state-independent, in continuous time, noise is introduced at infinitesimal time intervals subject to immediate transformation by the drift term, which includes Hessian information. The necessity of incorporating the Hessian matrix into the diffusion term for corrections is also underscored by the failure cases of existing SDE models discussed in Appendix F.1.*

In Lemma 14 and theorem 18 in the appendix, we provide sufficient conditions for the existence of the matrix $D(x)$ and the well-posedness of `HA-SME`. These conditions are either that the Hessian matrix $\nabla^2 f(x)$ and the covariance $\Sigma(x)$ commute with a stepsize $\eta < 1/\|\nabla^2 f(x)\|$, or if they do not commute, that a smaller stepsize is used.

3. Approximation Error Analysis for `HA-SME`

In this section, we discuss theoretical approximation error guarantees for `HA-SME`. First, we demonstrate that for general convex functions, `HA-SME` achieves an order-2 weak approximation error, improving upon existing SDEs by reducing dependence on the smoothness parameter of the objective function. Next, we examine quadratic objectives with additive Gaussian noise. In this scenario, we show that under mild conditions, `HA−SME` exactly recovers the dynamics of SGD.

3.1. Weak Approximation Error Analysis

Existing literature has provided weak approximation error guarantees (Definition 8) for `SME`-1 [32] and `SME`-2 [12, 22, 31, 32]. They demonstrated that the error is of order η for `SME`-1 and η^2 for `SME`-2. In this work, we show in Theorem 20 in the appendix that `HA−SME` also has an order η^2 weak approximation guarantee, achieving the best possible order, the same as `SME`-2. While prior work solely emphasizes the dependence on the stepsize, we further provide a more fine-grained analysis that explicitly accounts for problem-dependent parameters.

We define the Lipschitz parameter as $s := \sup_{x\in\mathbb{R}^d} \|\nabla f(x)\|$ and the smoothness parameter as $\lambda := \sup_{x\in\mathbb{R}^d} \|\nabla^2 f(x)\|$. `HA−SME` enjoys the following approximation error guarantee.

Theorem 3. *Fixing $T > 0$, assume that the test function satisfies $u \in C_b^8\left(\mathbb{R}^d\right)$, $F(\cdot;\xi) \in C_b^9\left(\mathbb{R}^d\right)$ is convex for any ξ, and for any $0 \le i, j \le d$, $[\Sigma(\cdot)]_{i,j} \in C_b^6\left(\mathbb{R}^d\right)$. Let $X(t)$ be the stochastic process described by `HA−SME` and $\{x_k\}$ be the sequence generated by SGD. There exists a constant $\eta_0 > 0$ such that for any $\eta < \eta_0$, it holds that for all $x \in \mathbb{R}^d$,*

$$\sup_{k=1,\ldots,\lfloor T/\eta\rfloor} \left|\mathbb{E}[u\left(x_k\right)|x_0 = x] - \mathbb{E}[u\left(X(k\eta)\right)|X(0) = x]\right| \le \mathcal{O}\left(\eta^2 s^3\right).$$

Remark 4. *Using a similar proof technique, we can establish that the upper bound for `SME`-2 is $\mathcal{O}\left(\eta^2\left(s^3 + s\lambda^2\right)\right)$. It is evident that `HA−SME` improves the dependency on the smoothness parameter λ. Specifically, there is no λ in the leading order term η^2.*

Remark 5. *Since the above theorem addresses the global point-wise approximation, the dependence on the global gradient norm of f appears unavoidable and cannot be neglected. However, to study the escaping behavior of SGD near a critical point, we believe that the global gradient norm s can be relaxed to its local version. This would be an interesting result because, in the vicinity of a critical point of a smooth function, the local gradient norm can be regarded as a negligible constant. Consequently, the leading term in Theorem 3 would be independent of both s and λ (whereas errors of existing SDE proxies always depend on λ).*

3.2. Exact Recovery of SGD on Quadratics

The quadratic objective, despite its simplicity, is significant for studying the behavior of SGD. However, existing SDEs may fail to capture the behavior of SGD even for this basic model. We demonstrate such failures of `SME`-1 and `SME`-2 through both experiments and theoretical analysis in Appendix F.1.

In contrast, under mild conditions, `HA−SME` can exactly match SGD dynamics in this case. Specifically, an SDE is said to *match* the iterates of the SGD if the distributions of X_t matches those of the SGD iterates at all corresponding time stamps, i.e. $t = k\eta$ where k is the iteration index of SGD.

Theorem 6. *Assuming the objective function f is defined as $f(x) = \frac{1}{2}x^\intercal Ax$, with $A \in \mathbb{R}^{d\times d}$, and the stochastic gradients satisfies $\nabla F(x,\xi) = \nabla f(x) + \xi$, where $\xi \sim \mathcal{N}(0,\Sigma)$ with $\Sigma \in \mathbb{R}^{d\times d}$*

independent of x. The solution of `HA-SME` *exactly matches the iterates of SGD if either of the following two conditions holds:*

1. *A and Σ commute, and $\eta\lambda_i \neq 1$ for all eigenvalues λ_i of A.*
2. *Σ is positive definite, and $\eta \le \frac{1}{\|A\|} \min\left\{1 - \sqrt{1 - \frac{\lambda_{\min}(\Sigma)}{\sqrt{d}\lambda_{\max}(\Sigma)}}, 1 - \frac{\sqrt{2}}{2}\right\}$,*

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a matrix.

To the best of our knowledge, this marks the first instance of an SDE that precisely mirrors the distribution of SGD, albeit restricted to quadratic functions.

In the theorem above, we require either that A and Σ commute or that Σ is positive definite. Indeed, we can construct a counterexample where A and Σ do not commute and Σ is only positive semi-definite, demonstrating that no OU process can exactly approximate SGD under these conditions. We restrict our discussion to OU process because the corresponding transition kernel is Gaussian, which matches the transition kernel of SGD (when viewed as a Markov chain) when the objective is quadratic and the gradient noise is additive state-independent Gaussian. This negative result holds even if we extend our discussion to the complex plane.

Theorem 7. *Under the same setting as Theorem 6, consider complex OU process defined as follows:*

$$dX_t = BX_t dt + DdW_t,$$

where $B \in \mathbb{C}^{d\times d}$, $D \in \mathbb{C}^{d\times m}$ and W_t represents an m-dimensional standard Wiener process. One can construct real matrices A and Σ such that, for any given stepsize $\eta > 0$, there exists no complex OU process that matches the iterates of SGD.

The distributional mismatch of the SDE and SGD in this case can be intuitively understood through the following argument: Starting from a deterministic initialization, after one SGD step, the covariance of the iterates is $\eta^2\Sigma$, which is rank-deficient; In contrast, in continuous-time dynamics, the noise injected is rotated by the misaligned linear transformations from the drift term and the covariance matrix quickly becomes full-rank.

4. Conclusion and Future Work

In this work, we present `HA-SME`, an advancement over existing SDEs for approximating the dynamics of SGD. Specifically, `HA-SME` offers improved theoretical approximation guarantees for general smooth objectives over current SDEs in terms of the dependence on the smoothness parameter of the objective. For quadratic objectives, `HA-SME` exactly recovers the dynamics of SGD under mild conditions. The primary innovation lies in integrating Hessian information into both the drift and diffusion terms of the SDE, achieved by extending backward error analysis to the stochastic setting. This integration preserves the interplay of noise and local curvature in approximating SGD, allowing to better capture its escaping behaviors.

In the future, it would be interesting to further analyze the escaping behaviors of `HA-SME`. We anticipate that this will provide a more accurate characterization of SGD's exit time of a local region with respect to the dependence on the Hessian. Second, applying our SBEA to other optimization algorithms, such as momentum and adaptive gradient methods would be intriguing directions for further research.

References

[1] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[2] Hedy Attouch, Zaki Chbani, Jalal Fadili, and Hassan Riahi. First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*, pages 1–43, 2022.

[3] David GT Barrett and Benoit Dherin. Implicit gradient regularization. *arXiv preprint arXiv:2009.11162*, 2020.

[4] Anthony Bloch. *Hamiltonian and gradient flows, algorithms and control*, volume 3. American Mathematical Soc., 1994.

[5] Odile Brandière and Marie Duflo. Les algorithmes stochastiques contournent-ils les pièges ? *Annales de l'I.H.P. Probabilités et statistiques*, 32(3):395–427, 1996.

[6] Andrew A Brown and Michael C Bartholomew-Biggs. Some effective methods for unconstrained optimization based on the solution of systems of ordinary differential equations. *Journal of Optimization Theory and Applications*, 62:211–224, 1989.

[7] Augustin Louis Baron Cauchy. *Analyse algébrique*, volume 1. Debure, 1821.

[8] Arnaud Debussche and Erwan Faou. Weak backward error analysis for SDEs. *SIAM Journal on Numerical Analysis*, 50(3):1735–1752, 2012.

[9] Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020. doi: 10.1214/19-AOS1850.

[10] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

[11] Cong Fang, Zhouchen Lin, and Tong Zhang. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, 2019.

[12] Yuanyuan Feng, Lei Li, and Jian-Guo Liu. Semi-groups of stochastic gradient descent and online principal component analysis: properties and diffusion approximations. *arXiv preprint arXiv:1712.06509*, 2017.

[13] Yuanyuan Feng, Tingran Gao, Lei Li, Jian-Guo Liu, and Yulong Lu. Uniform-in-time weak error analysis for stochastic gradient descent algorithms via diffusion approximation. *arXiv preprint arXiv:1902.00635*, 2019.

[14] Simone Fiori and Yoshua Bengio. Quasi-geodesic neural learning algorithms over the orthogonal group: A tutorial. *Journal of Machine Learning Research*, 6(5), 2005.

[15] Xavier Fontaine, Valentin De Bortoli, and Alain Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. In *Conference on Learning Theory*, 2021.

[16] Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, 2015.

[17] Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[18] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[19] Ernst Hairer, Marlis Hochbruck, Arieh Iserles, and Christian Lubich. Geometric numerical integration. *Oberwolfach Reports*, 3(1):805–882, 2006.

[20] Uwe Helmke and John B Moore. *Optimization and dynamical systems*. Springer Science & Business Media, 2012.

[21] Einar Hille and Ralph Saul Phillips. *Functional analysis and semi-groups*, volume 31. American Mathematical Soc., 1996.

[22] Wenqing Hu, Chris Junchi Li, Lei Li, and Jian-Guo Liu. On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*, 4 (1):3–32, 2019.

[23] Hikaru Ibayashi and Masaaki Imaizumi. Why does SGD prefer flat minima?: Through the lens of dynamical systems. In *When Machine Learning meets Dynamical Systems: Theory and Applications*, 2023.

[24] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[25] Stanisław Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in SGD. *arXiv preprint arXiv:1711.04623*, 2017.

[26] Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, 2017.

[27] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[28] Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

[29] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2011.

[30] Walid Krichene, Alexandre Bayen, and Peter L Bartlett. Accelerated mirror descent in continuous and discrete time. *Advances in Neural Information Processing Systems*, 2015.

[31] Lei Li and Yuliang Wang. On uniform-in-time diffusion approximation for stochastic gradient descent. *arXiv preprint arXiv:2207.04922*, 2022.

[32] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, 2017.

[33] Xiaoyu Li and Francesco Orabona. A high probability analysis of adaptive sgd with momentum. *arXiv preprint arXiv:2007.14294*, 2020.

[34] Tsung-Yu Lin and Subhransu Maji. Improved bilinear pooling with CNNs. *arXiv preprint arXiv:1707.06772*, 2017.

[35] Jun Liu and Ye Yuan. Almost sure saddle avoidance of stochastic gradient methods without the bounded gradient assumption. *arXiv preprint arXiv:2302.07862*, 2023.

[36] Stephan Mandt, Matthew D Hoffman, David M Blei, et al. Continuous-time limit of stochastic gradient descent revisited. In *OPT workshop, NIPS*, 2015.

[37] Stephan Mandt, Matthew Hoffman, and David Blei. A variational analysis of stochastic gradient algorithms. In *International Conference on Machine Learning*, 2016.

[38] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis, and Volkan Cevher. On the almost sure convergence of stochastic gradient descent in non-convex problems. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[39] Grigorii Noikhovich Milstein. *Numerical integration of stochastic differential equations*, volume 313. Springer Science & Business Media, 2013.

[40] Takashi Mori, Liu Ziyin, Kangqiao Liu, and Masahito Ueda. Power-law escape rate of SGD. In *International Conference on Machine Learning*, 2022.

[41] Michael Muehlebach and Michael I Jordan. A dynamical systems perspective on Nesterov acceleration. In *International Conference on Machine Learning*, 2019.

[42] Michael Muehlebach and Michael I Jordan. Optimization with momentum: Dynamical, control-theoretic, and symplectic perspectives. *Journal of Machine Learning Research*, 22(73):1–50, 2021.

[43] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[44] Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018.

[45] Thanh Huy Nguyen, Umut Simsekli, Mert Gurbuzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[46] Antonio Orvieto and Aurelien Lucchi. Continuous-time models for stochastic optimization algorithms. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[47] Rui Pan, Haishan Ye, and Tong Zhang. Eigencurve: Optimal learning rate schedule for SGD on quadratic objectives with skewed hessian spectrums. *arXiv preprint arXiv:2110.14109*, 2021.

[48] Courtney Paquette and Elliot Paquette. Dynamics of stochastic momentum methods on large-scale, quadratic models. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[49] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *arXiv preprint arXiv:2205.07069*, 2022.

[50] Robin Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

[51] Christian Perret. *The stability of numerical simulations of complex stochastic differential equations*. PhD thesis, ETH Zurich, 2010.

[52] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586.

[53] Mihaela Rosca, Yan Wu, Chongli Qin, and Benoit Dherin. On a continuous time model of gradient descent dynamics and instability in deep learning. *Transactions on Machine Learning Research*, 2022.

[54] Peter Ross. Generalized hockey stick identities and n-dimensional blockwalking. *The College Mathematics Journal*, 28(4):325, 1997.

[55] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

[56] Johannes Schropp and I Singer. A dynamical systems approach to constrained minimization. *Numerical functional analysis and optimization*, 21(3-4):537–551, 2000.

[57] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, 2021.

[58] Boris V Shabat. Introduction to complex analysis. Part II, translations of mathematical monographs, vol. 110. *American Mathematical Society, Providence, RI*, 1992.

[59] Tony Shardlow. Modified equations for stochastic differential equations. *BIT Numerical Mathematics*, 46:111–125, 2006.

[60] Bin Shi, Simon S Du, Michael I Jordan, and Weijie J Su. Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, pages 1–70, 2021.

[61] Willi-Hans Steeb and Yorick Hardy. *Matrix calculus and Kronecker product: a practical approach to linear and multilinear algebra*. World Scientific Publishing Company, 2011.

[62] Weijie Su, Stephen Boyd, and Emmanuel J Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

[63] Flemming Topsøe. Some bounds for the logarithmic function. *Inequality theory and applications*, 4:137, 2007.

[64] Mingze Wang and Lei Wu. A theoretical analysis of noise geometry in stochastic gradient descent. *arXiv preprint arXiv:2310.00692*, 2024.

[65] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*, 2020.

[66] Zeke Xie, Xinrui Wang, Huishuai Zhang, Issei Sato, and Masashi Sugiyama. Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum. In *International Conference on Machine Learning*, 2022.

[67] Junchi Yang, Xiang Li, Ilyas Fatkhullin, and Niao He. Two sides of one coin: the limits of untuned SGD and the power of adaptive methods. *Advances in Neural Information Processing Systems*, 36, 2024.

[68] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021. doi: 10.1145/3446776.

[69] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In *International Conference on Machine Learning*, 2019.

Contents

Appendix A. Motivation and Related Work

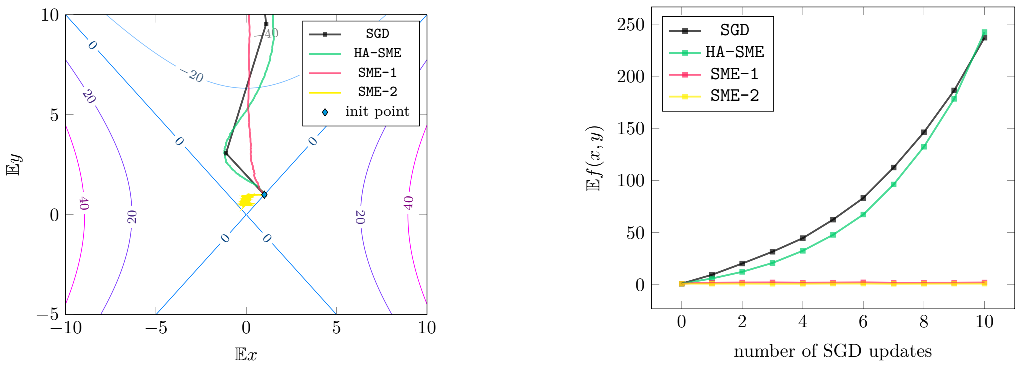
We study the behaviors of SGD in the context of nonconvex optimization. Existing non-convex optimization theory have established the non-asymptotic convergence of SGD towards a *stationary point* using a *diminishing* stepsize; see e.g., Drori and Shamir [10], Fang et al. [11], Ghadimi and Lan [18], Khaled and Richtárik [28], Li and Orabona [33], Sebbouh et al. [57] and Yang et al. [67], to list just a few. Intuitively, the diminishing stepsize reduces the fluctuation caused by SGD's inherent randomness, ensuring the objective descends in expectation. On the other hand, in practice a *constant* stepsize, i.e. $\eta_k \equiv \eta$, is often preferred and yields better performance. This presents a challenge to classical theory, as it fails to capture the non-diminishing role of noise in SGD dynamics. Additionally, it falls short in differentiating between saddle points and local minima, as well as in identifying favorable stationary points — those that not only are local minima but also exhibit good generalization.

The above limitations underscore the need to understand a critical aspect of SGD in nonconvex optimization: *the transitions between different stationary points*, beyond the traditional focus on convergence towards a stationary point. The selection of stepsize plays a crucial role in SGD's transition behaviors:

- **Escaping from stationary points.** At a stationary point, either a saddle point or a local minimum, increasing the stepsize raises the likelihood of SGD to escape from it.
- **Minima selection.** Given a constant stepsize, SGD can rapidly exit local minima with a large Hessian norm, also known as sharp minima, but dwell longer at flat minima where the norm is small.

The latter relates to the notable generalization capabilities of SGD, which in prevailing belief, tends to favor flat minima over sharp ones [27, 43, 44]. These phenomena have been empirically verified [68] and theoretically investigated [5, 35, 38, 50]. Indeed, analyzing the link between stepsize and the escaping dynamics of SGD has emerged as an important topic in non-convex optimization. In particular, there is growing interest in using continuous-time dynamics, specifically Stochastic Differential Equations (SDEs), to understand the mechanisms behind escaping from saddle points and minima selection [22, 66]. This involves quantifying the interplay between escaping speed, stepsizes, and local sharpness [22, 23, 45, 65, 69]. A more detailed discussion of these works is provided in Appendix A.1.

Existing continuous-time approximations of SGD and their failure modes. Previous study on the escaping behavior of constant-stepsize SGD through its continuous-time approximation typically entails two steps: (1) Proposing a proxy SDE model that approximately tracks the distributional



(a) $f\left([x,y]^{\intercal}\right) = \frac{1}{2}\left(x^2 - y^2\right)$ (b) $f\left([x,y]^{\intercal}\right) = \frac{1}{2}\left(x^2 + y^2\right)$

Figure 1: Escaping behaviors of SGD and SDEs on quadratic functions. The left sub-figure demonstrates the trajectories around a saddle point $(0,0)$, while the right sub-figure represents the evolution of the function value around a minimum. We set $\eta = 2.1$, the initial point to $(1,1)$, and the covariance of the additive noise to I for both cases. We conducted the experiments 1000 times to estimate the expectation. For the right sub-figure, we evaluated the SDEs at time stamps corresponding to η times the number of SGD updates. In these two cases, the proposed `HA−SME` behaves similarly as SGD. However, `SME`-2 fails to escape the saddle point in case (a), and both `SME`-1 and `SME`-2 fail to escape the minimum in case (b). We further provide in Appendix F.1 mathematical proofs for the failure cases of these two `SME`s in approximating SGD.

evolution of the SGD dynamics; (2) Analyzing the escaping behavior of the proposed SDE model to predict or interpret the behavior of SGD. Evidently, whether the behavior of SGD can be precisely captured hinges on the accuracy of the proxy models. Among existing SDE models, the second-order stochastic modified equation (`SME`-2) [32] provides the order-best approximation guarantee:

$$dX_t = -\nabla\left(f(X_t) + \frac{\eta}{4}\|\nabla f(X_t)\|^2\right)dt + \sqrt{\eta\Sigma(X_t)}dW_t, \tag{4}$$

where $\Sigma(X_t)$ denotes the covariance matrix of the stochastic gradient at X_t, W_t represents the standard Wiener process, and η denotes the constant stepsize of SGD. However, we note that even for simple quadratic objectives, `SME`-2 and SGD show different behaviors in escaping from saddle points and minima: `SME`-2 exhibits locally stabilizing behavior while SGD with constant stepsize quickly escapes from the same local region, as illustrated in Figure 1. Our study also identifies similar failure modes for other existing SDE models, detailed in Appendix F.1. This disparity in behavior suggests the inadequacy of existing SDEs in modeling the escaping dynamics of SGD, even with high-order approximation such as `SME`-2. Consequently, this brings us to a central question:

> *How to accurately model SGD in continuous time while preserving its escaping behaviors?*

A.1. Related Work

Continuous-time approximation of GD and SGD. The connection between differential equations and gradient-based methods has been extensively studied previously, such as in [2, 4, 6, 14, 20, 30, 41, 42, 56, 60, 62]. For GD, the most straightforward continuous-time approximation is the gradient flow

$$dX_t = -\nabla f(X_t)dt. \tag{5}$$

While this flow approximates GD accurately as the stepsize approaches zero, the approximation error becomes critical when the stepsize is large. Barrett and Dherin [3] took a step further by employing backward error analysis (detailed in Appendix B.2) to derive a second-order continuous-time approximation for GD, as given by:

$$dX_t = -\nabla \left(f(X_t) + \frac{\eta}{4} \|\nabla f(X_t)\|^2 \right) dt. \tag{6}$$

However, this second-order model still has limitations, particularly in capturing certain critical behaviors of GD in discrete-time, such as overshooting, i.e., the iterates oscillate around the minimum or even diverge because of large stepsize. To tackle this issue, Rosca et al. [53] introduced the Principle Flow (PF), a continuous-time approximation for GD that operates in complex space and can exactly match the discrete-time dynamics of GD on quadratics. This work is closely related to ours, and we will dive deeper in Appendix B.3.

When taking noises into account, Mandt et al. [36, 37] introduced the following SDE dynamics for approximating SGD (`SME`-1):

$$dX_t = -\nabla f(X_t)dt + \sqrt{\eta \Sigma(X_t)} dW_t. \tag{7}$$

Later, Li et al. [32] introduced `SME`-2 (described in Equation (4)) and established rigorous weak approximation for these SDEs. These SDEs are named according to the order of their weak approximation accuracy (see precise definition in Definition 8). Additional contributions by Orvieto and Lucchi [46] and Fontaine et al. [15] examined SDEs where the drift and diffusion terms are scaled by varying stepsizes. Their SDEs reduce to `SME`s when adopting constant stepsizes. Another line of work has considered Homogenized SGD, which replaces the diffusion term in `SME`-1 with $\sqrt{\eta f(X_t) \nabla^2 f(X_t)/B}$ [48, 49] or $\sqrt{\eta f(X_t) \nabla^2 f(X^*)/B}$ [40], where B denotes the batch size and X^* the local minimum. However, their analyses are limited to the least square loss, lacking approximation guarantees for general non-convex objectives.

Escaping Behaviors Analysis Back in the 90s, Pemantle [50] and Brandière and Duflo [5] showed that SGD with a small stepsize can avoid hyperbolic saddle points, i.e., $\lambda_{\min}\left(\nabla^2 f(x)\right) < 0$ and $\det\left(\nabla^2 f(x)\right) \neq 0$. Recent studies [35, 38] extended this understanding by showing SGD can escape strict saddle points ($\lambda_{\min}\left(\nabla^2 f(x)\right) < 0$). Additionally, with slight modifications to the algorithm, such as adding artificial noises when certain conditions are met, Ge et al. [16] and Jin et al. [26] established convergence of noisy SGD to approximate local minima.

In terms of minima escaping, leveraging continuous-time frameworks, Jastrzebski et al. [25] highlighted three crucial factors that influence the minima found by SGD: step-size, batchsize, and gradient covariance. Zhu et al. [69] focused on the role of anisotropic noise and concluded that such noise enables SGD to evade sharp minima more effectively. Quantitatively, the time required for escaping has been theoretically shown to exponentially depend on the inverse of the stepsize, using theory for random perturbations of dynamical systems [22]. Subsequent work expanded on these ideas, considering heavy-tail noise and utilizing Lévy-driven SDEs to examine SGD's stability around minima [45]. While the aforementioned research largely focused on parameter-independent gradient noise, Xie et al. [65] analyzed parameter-dependent anisotropic noise. Using the analysis for the classic Kramers' escape problem, they showed that compared to Stochastic Gradient Langevin Dynamics, SGD can quickly escape sharp minima, with escape speed exponentially depending on the determinant of the Hessian. Their analysis assumes that the system first forms a stationary distribution around a basin before escaping. However, SGD typically selects minima dynamically. New insights into the exponential escape time have been developed by employing Large Deviation Theory in non-stationary settings [23].

Appendix B. Preliminaries

In this section, we first introduce the notations used throughout our analysis, and then review the Backward Error Analysis (BEA), a technique instrumental to the development of our stochastic backward error analysis. Additionally, we discuss the Principle Flow (PF) introduced by Rosca et al. [53], which inspires the derivation of our HA-SME model.

B.1. Notations

Uppercase variables, such as X, denote continuous-time variables, while lowercase variables like x represent discrete-time variables. The function $X(x_0, t)$ indicates the solution of a continuous-time process starting from the initial point x_0 after a time duration t. For matrix notation, we use $[A]_{i,j}$ to specify the element at the (i, j)-th position in matrix A. For two matrices, A and B, of the same size, we denote $A\!:\!B = \mathrm{tr}\big(A^\top B\big)$. Applying the logarithmic function to a matrix and dividing between two matrices are by default element-wise. We use the notation I to represent the identity matrix and $\odot$ to denote the Hadamard product between matrices. The square root of a matrix, i.e., $\sqrt{A}$, is the matrix B such that $BB^\intercal = A$. In the context of functions, $\|f\|_{C^m}$ is defined as $\sum_{|\alpha|\leq m} |D^\alpha f|_\infty$, and we define the set $C_b^m\left(\mathbb{R}^d\right) = \{f \in C^m\left(\mathbb{R}^d\right) \mid \|f\|_{C^m} < \infty\}$, where $C^m\left(\mathbb{R}^d\right)$ is the set of all m-th differentiable functions on $\mathbb{R}^d$. For complex number x, $\mathrm{Re}(x)$ and $\mathrm{Im}(x)$ extract the real and imaginary parts, respectively, and $\overline{x}$ denotes the complex conjugate of x. For random variables x and y, the covariance is represented as $\mathrm{Cov}\,[x, y] = \mathbb{E}\left[(x - \mathbb{E}(x))\,(y - \mathbb{E}(y))^\intercal\right]$. Lastly, we use W_t to represent the standard Wiener process.

B.2. Backward Error Analysis (BEA)

Consider the GD dynamics[1]

$$x_{k+1} = x_k - \eta\nabla f(x_k), \tag{8}$$

with constant stepsize η. Let $\widetilde{X} : [0, \infty) \to \mathbb{R}^d$ be a continuous trajectory. The idea of BEA is to identify a series of functions, denoted as $g_i : \mathbb{R}^d \to \mathbb{R}^d$ for $i \in \{0, 1, \ldots\}$, which constitute a modified equation as:

$$\frac{d}{dt}\widetilde{X} = \widetilde{G}_\eta(\widetilde{X}) := \sum_{p=0}^{\infty} \eta^p g_p(\widetilde{X}), \tag{9}$$

1. Similar arguments can be made for other discrete-time dynamics.

such that $\widetilde{X}((k+1)\eta)$ closely approximates x_{k+1}, given $\widetilde{X}(k\eta) = x_k$. To achieve this goal, we consider the Taylor expansion of the above continuous-time process at time $k\eta$,

$$\widetilde{X}((k+1)\eta) = \sum_{j=0}^{\infty} \frac{\eta^j}{j!} \left. \frac{d^j \widetilde{X}}{(dt)^j} \right|_{\widetilde{X}=x_k} = x_k + \eta \sum_{p=0}^{\infty} \eta^p g_p(x_k) + \frac{\eta^2}{2} \sum_{p=0}^{\infty} \eta^p \nabla g_p(x_k) \sum_{p=0}^{\infty} \eta^p g_p(x_k) + \cdots . \tag{10}$$

By matching the powers of η in Equation (8) and Equation (10) up to the $(p+1)$-order, we solve for g_i for $i = 0, ..., p$. This procedure ensures that the leading error term between the continuous-time variable $\widetilde{X}((k+1)\eta)$ and the discrete-time variable x_{k+1}, given $\widetilde{X}(k\eta) = x_k$, is of the order $\mathcal{O}\left(\eta^{p+2}\right)$. For a more comprehensive exploration of BEA, please refer to Section IX of Hairer et al. [19].

When $p = 0$, the system reduces to gradient flow (Equation (5)) and then $p = 1$, we recover Equation (6). Increasing the order p in the above BEA procedure improves the accuracy of the approximation. The higher order modified flows generated by BEA involve higher-order gradients of the objective function f. For example, for $p = 2$, according to BEA, we obtain the third-order Gradient Flow [53]:

$$dX_t/dt = -\nabla f(X_t) - \frac{\eta}{2}\nabla^2 f(X_t)\nabla f(X_t) - \eta^2 \left(\frac{1}{3}\nabla^2 f(X_t)\nabla f(X_t) + \frac{1}{12}\nabla f(X_t)^\intercal \nabla^3 f(X_t)\nabla f(X_t)\right).$$

B.3. Principle Flow

While theoretically, one could solve for all component function g_i's to achieve an arbitrary target error order, the resulting modified flow quickly becomes cumbersome due to the involvement of higher-order gradients. Moreover, the resulting power series (w.r.t. the exponent of η) may not even be convergent [19]. In light of this limitation of BEA, Rosca et al. [53] suggested the following term selection strategy:

Principle 1 (Term selection of PF, BEA)**.** *In BEA, when solving g_i's in Equation (9), derivatives of $f(x)$ up to the second order, i.e., terms containing $\nabla^p f(x)$ for $p \geq 3$, are omitted.*

With this term selection scheme, the resulting power series not only converges (with convergence radius $\eta\|\nabla^2 f(x)\| = 1$) but admits a concise form. The corresponding modified equation is:

$$dX_t = U(X_t)\frac{\log(I - \eta\Lambda(X_t))}{\eta\Lambda(X_t)}U(X_t)^\intercal \nabla f(X_t)dt, \tag{11}$$

where for any fixed x, we denote $\nabla^2 f(x) = U(x)\Lambda(x)U(x)^\intercal$ the eigen-decomposition of $\nabla^2 f(x)$. When the stepsize is large, the logarithmic function could result in complex numbers, as elements of $I - \eta\Lambda$ might be negative. Operating in the complex space is beneficial for modeling the divergent and oscillatory behaviors of gradient descent Rosca et al. [53]. As only higher-order gradients are omitted in Principle 1, PF provides an *exact* characterization of the discrete GD dynamics on quadratic objectives.

Appendix C. Stochastic Backward Error Analysis

We extend the BEA approach to the stochastic domain, termed Stochastic Backward Error Analysis (SBEA), providing a systematic approach for constructing SDEs that approximate stochastic discrete-time processes. While we focus on the analysis for SGD, similar argument can be generalized to other dynamics.

C.1. Differences between BEA and SBEA

Compared with BEA, there are two major modifications in SBEA. First, SBEA includes an additional Brownian motion term in its ansatz, compared to the one in Equation (9), with the diffusion coefficients to be determined; Second, to evaluate the quality of the approximation, SBEA adopts the weak approximation error (see Definition 8) as a metric in the distributional sense.

(1) Ansatz with Brownian motion. In BEA, we define the modified equation with a series of functions $\{g_i\}_{i\in\mathbb{N}}$. To further model the randomness in SGD, we introduce another series $\{h_i\}_{i\in\mathbb{N}_{>0}}$, $h_i : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, to represent the diffusion coefficients of the Brownian motion. Specifically, we consider a hypothesis continuous-time dynamics as follows:

$$dX = \left(g_0(X) + \eta g_1(X) + \eta^2 g_2(X) + \cdots\right) dt + \sqrt{\eta h_1(X) + \eta^2 h_2(X) + \cdots} dW_t. \tag{12}$$

The rationale stems from the observation that the noise introduced in the ansatz undergoes continuous modification by the drift term. Consequently, it is natural to employ a similar power series representation for the diffusion coefficient[2]. While one can truncate the series $\{h_i\}_{i\in\mathbb{N}}$ to simplify the analysis, as we illustrate in Figure 2 and theoretically justify in Appendix F.1, such truncation can compromise the approximation quality and lead to misaligned escaping behaviors.

(2) Weak approximation error as quality metric. Our goal is to solve for the functions g_i's and h_i's so that the ansatz in Equation (12) approximates the SGD dynamics with high quality. Instead of seeking a path-wise alignment like BEA, we focus on aligning the continuous-time and discrete-time dynamics in a distributional sense, characterized by the following weak approximation error.

Definition 8 (Weak Approximation Error)**.** *Fixing $T > 0$, we say a continuous-time process $X(t)_{t\in[0,T]}$ is an order p weak approximation to the sequence $\{x_k\}$ if for any $u \in C_b^{2(p+1)}\left(\mathbb{R}^d\right)$, there exists $C > 0$ and $\eta_0 > 0$ that are independent of η (but may depend on T, u and its derivatives) such that for all $x \in \mathbb{R}^d$,*

$$\Big|\mathbb{E}[u\left(x_k\right)|x_0 = x] - \mathbb{E}[u\left(X(k\eta)\right)|X(0) = x]\Big| \leq C\eta^p, \tag{13}$$

for all $k = 1, \ldots, \lfloor T/\eta \rfloor$ and all $0 < \eta < \eta_0$.

This metric is commonly used for analyzing discretization errors of SDE, as referenced in textbooks [29, 39] and various studies on modeling SGD behavior with SDEs [13, 22, 31, 32].

Remark 9. *We highlight the difference between SBEA and the weak backward error analysis introduced by Debussche and Faou [8]. The latter provides a framework for constructing a sequence of modified Kolmogorov generators with increasing orders to approximate discrete-time processes.*

2. We do not include h_0 in the ansatz as SGD degenerates to gradient flow when $\eta \to 0$, which implies $h_0 \equiv 0$.

However, since the high-order Kolmogorov generators involve high-order gradients, it does not *yield an equivalent SDE representation (recall that the Kolmogorov generator corresponding to an Itô SDE contains only second-order gradients). Consequently, their framework is fundamentally different from our approach in terms of the final outcome.*

C.2. Procedures in SBEA

We determine the components $\{h_i\}_{i\in\mathbb{N}}$ and $\{g_i\}_{i\in\mathbb{N}}$ by first calculating the semi-group expansions of the two conditional expectations involved in Equation (13) after one step, i.e. $k = 1$, and then match the resulting terms according to the power of η. Such a one-step approximation analysis can be translated to the weak approximation error guarantee via a martingale argument, as we will elaborate in Appendix E.

Semi-group expansions. We expand the conditional expectation of the test function u after one SGD step as:

$$\mathbb{E}[u\left(x - \eta\nabla F(x;\xi)\right) \mid x] = u(x) + \eta\sum_{p=0}^{\infty}\eta^p\Phi_p(x), \tag{14}$$

where we use Φ_p to denote the aggregation of terms with η^{p+1} in the expansion. Simple calculation shows

$$\Phi_0(x) = -\nabla f(x)^\intercal\nabla u(x),\ \Phi_1(x) = \frac{1}{2}\nabla^2 u(x) : \left(\nabla f(x)\nabla f(x)^\intercal + \Sigma(x)\right),\ \cdots \tag{15}$$

where $\Sigma(x)$ denotes the covariance matrix of the stochastic gradient at x, defined as

$$\Sigma(x) = \mathbb{E}\left[\left(\nabla F(x;\xi) - \mathbb{E}[\nabla F(x;\xi)]\right)\left(\nabla F(x;\xi) - \mathbb{E}[\nabla F(x;\xi)]\right)^\intercal\right]. \tag{16}$$

For the continuous-time system, by the semi-group expansion [21], the conditional expectation of a test function u under the SDE model yields:

$$\mathbb{E}[u\left(X(\eta)\right) \mid X(0) = x] = \left(e^{\eta\mathcal{L}}u\right)(x) = u(x) + \eta\mathcal{L}u(x) + \frac{1}{2}\eta^2\mathcal{L}^2u(x) + \frac{1}{6}\eta^3\mathcal{L}^3u(x) + \cdots, \tag{17}$$

where $\mathcal{L}$ is corresponding the infinitesimal generator [55]. For the SDE described in Equation (12), the generator is given explicitly by

$$\mathcal{L} = (g_0 + \eta g_1 + \cdots)\cdot\nabla + \frac{1}{2}(\eta h_1 + \cdots) : \nabla^2 = g_0\cdot\nabla + \sum_{i=1}^{+\infty}\eta^i\left(g_i\cdot\nabla + \frac{1}{2}h_i : \nabla^2\right) = \sum_{i=0}^{+\infty}\eta^i\mathcal{L}_i,$$

where for the ease of notation, we denote

$$\mathcal{L}_0 \coloneqq g_0\cdot\nabla \text{ and } \mathcal{L}_i \coloneqq g_i\cdot\nabla + \frac{1}{2}h_i : \nabla^2 \text{ for } i \geq 1.$$

Gathering terms with power η^p in Equation (17) leads to the following lemma:

Lemma 10. *Define the function*

$$\Psi_p(x) = \sum_{n=1}^{p+1} \frac{1}{n!} \sum_{\{l_i^n \geq 0\}_{i=1}^n : \sum_{i=1}^n l_i^n = p+1-n} \mathcal{L}_{l_1^n} \mathcal{L}_{l_2^n} \cdots \mathcal{L}_{l_n^n} u(x), \tag{18}$$

where $\mathcal{L}_{l_j^n} \mathcal{L}_{l_{j+1}^n}$ denotes the composition of the operators $\mathcal{L}_{l_j^n}$ and $\mathcal{L}_{l_{j+1}^n}$. We have that

$$\left(e^{\eta \mathcal{L}} u\right)(x) = u(x) + \eta \sum_{p=0}^{\infty} \eta^p \Psi_p(x). \tag{19}$$

In the following, we determine the components g_p's and h_p's in an iterative manner.

Identify components by iteratively matching terms. Our aim is to determine g_i's and h_i's by aligning the terms from the discrete-time expansion (Equation (14)) with those from continuous time (Equation (19)), according to the order of η. Note that every application of the generator $\mathcal{L}$ on a function generates terms with all orders of η. Consequently, p-th order terms in the continuous-time expansion become complicated as p grows, as evident in Lemma 10. Fortunately, we can identify g_p's and h_p's in an iterative manner. In the following discussion, g_i and h_i for $i = 0, 1, \ldots, p-1$ are treated as *known* from our iterative construction, and we solve the *unknowns* g_p and h_p by matching Φ_p and Ψ_p. We make the following two observations:

[$\Psi_p(x)$ (Equation (18)) is linear on g_p and h_p.] Recall that g_p and h_p are only contained linearly in $\mathcal{L}_p u$, which appears only once in $\Psi_p(x)$ by choosing $n = 1$ and $l_1^1 = p$.

[Terms in $\Psi_p(x)$ with $n > 1$ contains only known terms.] For $n > 1$, the terms in Equation (18) will only include g_i and h_i for $i \leq p-1$, since for any choice of the indices $\{l_j^{n+1}\}_{j=1}^n$, we must have $l_j^n \leq p-1$ if $n > 1$. These terms are already determined by our iterative strategy.

Following Rosca et al. [53], we also explore various term selection strategies, starting with the exact matching case.

Principle 2 (Term selection of exact-matching, SBEA)**.** *When determining the component functions g_p and h_p, all terms in Φ_p from Equation* (14) *must match the ones in Ψ_p from Equation* (19)*. In particular, given the arbitrariness of the test function, the coefficients of $\nabla^p u$ ($p \geq 1$) should all be matched.*

With the above principle, we can identify the unknowns, g_p and h_p, for $p \leq 1$.

Remark 11 (Recovering `SME`-2 by SBEA)**.** *By matching Ψ_0 and Φ_0, we deduce that $g_0(x) = -\nabla f(x)$. By matching Ψ_1 with Φ_1, we solve that $g_1(x) = -\frac{1}{2}\nabla^2 f(x) \nabla f(x)$ and $h_1(x) = \Sigma(x)$. This leads to the formulation of the `SME`-2 (Equation* (4)*).*

C.3. Exact Matching Fails for Order η^3 in SBEA

We have shown that for $p = 0, 1$, it is possible to exactly solve g_p and h_p. However, for $p \geq 2$, adhering to Principle 2 can be infeasible. Specifically, following Principle 2 leads to an over-determined system when determining the unknown component functions g_p and h_p for $p \geq 2$. Recall that the test function u is chosen arbitrarily and let us consider all the possible occurrence of $\nabla^q u(x)$, $q = 1, 2, \ldots$, in Φ_p and Ψ_p.

- For the continous-time dynamics, the term Ψ_p in Equation (18) can contain terms with factors $\nabla^q u(x)$ for q ranging from 1 to $p+1$. For example, the term $\nabla^{p+1}u(x)$ can be generated from Equation (18) with the choice of $n = p+1$ and $l_j^n = 0$ for all $j \in \{1, \ldots, n\}$; the term $\nabla u(x)$ can be generated with the choice $n = 1$ and $l_1^1 = 0$.
- For discrete-time expansion as per Equation (14), the term Φ_p involves only $\nabla^p u(x)$.

For exact matching, the coefficient of $\nabla^p u$ in Ψ_p must match that in Φ_p; besides, for any $q \neq p$, the coefficient of $\nabla^q u$ in Ψ_p must be zero. This leads to an over-determined system comprising $p+1$ conditions for 2 free variables (g_p and h_p) when $p \geq 2$. In other words, the above iterative construction of the components g_p and h_p fails for $p \geq 2$. The infeasibility result highlights a drastic difference between SBEA and BEA. Shardlow [59] also noted such infeasibility when attempting to find stochastic modified flows to achieve higher orders of weak approximation guarantees.

Inspired by PF, we introduce a relaxed term selection scheme, compared to Principle 2, so that only a subset of terms in Ψ_p (the continuous-time dynamics) and Φ_p (the discrete-time dynamics) match. Importantly, this term selection scheme should ensure that the resulting continuous-time dynamics still preserve the favorable escaping behaviors of the discrete-time dynamics of interest. This will be the focus of the following section.

Appendix D. Derivation of Hessian-Aware Stochastic Modified Equation

In this section, we first derive the proposed SDE, `HA-SME`, to model SGD using the idea of SBEA. Then we discuss sufficient conditions for its well-posedness.

Due to the lack of degree of freedom, exactly matching all terms in Φ_p and Ψ_p for $p \geq 2$ is in general infeasible. Instead, we adopt the following term selection principle.

Principle 3 (Term selection in `HA-SME`, SBEA)**.** *When determining the components g_p and h_p in SBEA, ignore the terms that involve $\nabla^{(r)} f(x)$, $\nabla^{(s)} u(x)$ and $\nabla^{(m)} \Sigma(x)$ for $r, s \geq 3$ and $m \geq 1$.*

Under this principle of term selection, we obtain the following result.

Lemma 12. *Under Principle 3, when applying SBEA on SGD, we have the following results:*

1. *The component functions g_p and h_p in Equation (12) are uniquely determined;*
2. *The component functions g_p and h_p in Equation (12) admit the form*

$$g_p(x) = c_p \cdot \left(\nabla^2 f(x)\right)^p \nabla f(x), \tag{20}$$

$$h_p(x) = \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot \left(\nabla^2 f(x)\right)^k \Sigma(x) \left(\nabla^2 f(x)\right)^{p-1-k}; \tag{21}$$

3. *The coefficients $\{a_{s,m}\}$ and $\{c_s\}$ in $\{g_p\}$ and $\{h_p\}$ satisfy*

$$\frac{\log(1-x)}{x} = \sum_{s=0}^{+\infty} c_s x^s \text{ and } \frac{\log(1-x)(1-y)}{xy - (x+y)} = \sum_{s,m \geq 0}^{+\infty} a_{s,m} x^s y^m.$$

We now derive the limit of the power series defined in Equation (12).

Theorem 13. *Let the components $\{g_p\}$ and $\{h_p\}$ be determined by the SBEA framework applied on SGD under Principle 3. Denote the limits*

$$b(x) = \sum_{p\geq 0} \eta^p g_p(x) \text{ and } \mathcal{D}(x) = \sum_{p\geq 1} \eta^p h_p(x). \tag{22}$$

Under the assumption that $\eta < 1/\left\|\nabla^2 f(x)\right\|$, we have that the limits $b(x)$ and $\mathcal{D}(x)$ exist with

$$b(x) = U(x)\frac{\log(I-\eta\Lambda(x))}{\eta\Lambda(x)}U(x)^\intercal \nabla f(x), \tag{23}$$

$$\mathcal{D}(x) = U(x)S(x)U(x)^\top, \text{such that } [S(x)]_{i,j} = \frac{\left[U^\intercal \Sigma U\right]_{i,j}\log(1-\eta\lambda_i)(1-\eta\lambda_j)}{\eta\lambda_i\lambda_j - (\lambda_i+\lambda_j)}, \tag{24}$$

where $U(x)$ and $\Lambda(x)$ are defined through the eigen-decomposition $\nabla^2 f(x) = U(x)\Lambda(x)U(x)^\intercal$. The diagonal elements of $\Lambda(x)$ are denoted by $\lambda_i(x)$. For conciseness, we omit the dependence of λ_i, U and Σ on x.

The derivation of the above theorem is highly nontrivial, and represents one of the major contributions of our paper. Its proof includes several critical steps: (1) Identifying the structure of the components g_p and h_p in the SBEA ansatz; (2) Determining the coefficient of the components g_p and h_p; (3) Computing the limit of the resulting power series, which we defer to Appendix G.1.

The following lemma provides sufficient conditions for the existence of the diffusion coefficient $\sqrt{\mathcal{D}(x)}$.

Lemma 14. *$\mathcal{D}(x)$ is positive semi-definite at point x if either of the following two conditions holds:*

1. *The Hessian $\nabla^2 f(x)$ and covariance matrix $\Sigma(x)$ commute, and $\eta < 1/\left\|\nabla^2 f(x)\right\|$.*
2. *The covariance matrix $\Sigma(x)$ is positive definite, and η satisfies*

$$\eta \leq \frac{1}{\|\nabla^2 f(x)\|}\min\left\{1-\sqrt{1-\frac{\lambda_{\min}(\Sigma(x))}{\sqrt{d}\lambda_{\max}(\Sigma(x))}}, 1-\frac{\sqrt{2}}{2}\right\}, \tag{25}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a matrix.

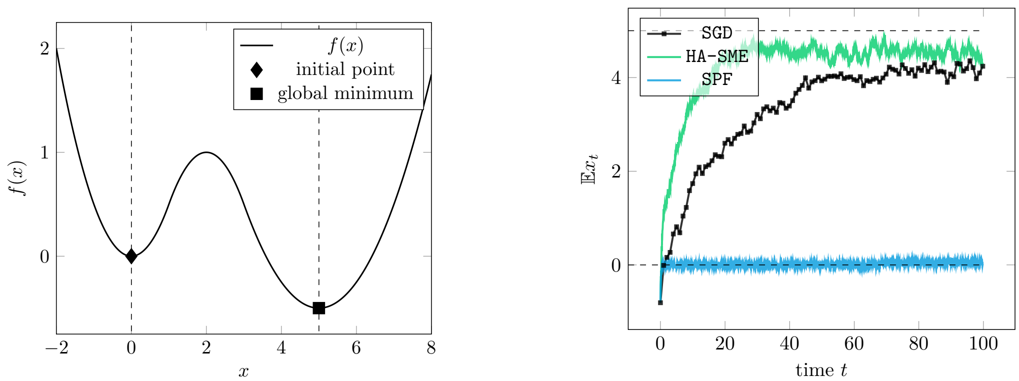
The validity of the above two conditions is discussed at the end of this subsection. With Theorem 13 and Lemma 14, we now present the proposed SDE model:

Definition 15. *Let the assumptions in Theorem 13 and Lemma 14 hold for any x, then the drift term $b(x)$ and diffusion coefficient $D(x) := \sqrt{\mathcal{D}(x)}$ in Equation (22) are well-defined. We denote the SDE as the Hessian-Aware Stochastic Modified Equation (`HA-SME`):*

$$dX_t = b(X_t)dt + D(X_t)dW_t. \tag{26}$$

Comparing with `SME`-1 in Equation (7) and `SME`-2 in Equation (4), both the drift term and the diffusion coefficient of `HA-SME` incorporate the Hessian information, hence the name.

Notice that the drift term $b(x)$ of `HA-SME` exactly corresponds to PF defined in Equation (11). A naive idea would be to combine PF with the conventional diffusion coefficient $\sqrt{\eta\Sigma}$ as in `SME`-1 and `SME`-2. The following remark comments on the drawback of such a construction.



(a) function $f(x)$ (b) expected value of x_t through time

Figure 2: The illustration depicts the failure case for SPF, a naive stochastic extension of PF. The objective function is defined as $f(x) = \frac{1}{2}x^2$ for $x < 1$; $-\frac{1}{2}(x-2)^2 + 1$ for $1 \leq x < 3$; and $\frac{1}{4}(x-5)^2 - \frac{1}{2}$ for $x \geq 3$. We use dashed lines to indicate the two minima. The initial point is set to 0, and we set $\eta = 0.999$ and the noise variance to 1 for all methods. We run the simulation 100 times to estimate the expectation. SGD and HA-SME escape the initial minimum and arrive near the global minimum, while SPF stays around the initial one. A theoretical justification for the failure of SPF is presented in Appendix F.1.

Remark 16 (A naive SDE derived from PF)**.** *Adding the diffusion coefficient $\sqrt{\eta\Sigma}$ to the PF in Equation (11) gives rise to the following SDE:*

$$dX_t = U(X_t)\frac{\log(I - \eta\Lambda(X_t))}{\eta\Lambda(X_t)}U(X_t)^\intercal \nabla f(X_t)dt + \sqrt{\eta\Sigma(X_t)}dW_t, \tag{27}$$

which we refer to as Stochastic Principle Flow (SPF). Such a straightforward combination could lead to escaping behaviors different from SGD. For example, when considering a multi-mode function, as illustrated in Figure 2, SGD and our proposed HA-SME can easily escape the local minimum and find the global one, whereas SPF remains trapped in the valley of the initial local minimum. A theoretical comparison of HA-SME and SPF on general objective functions is provided in Appendix E.2.

Remark 17 (Discussion on conditions of Lemma 14)**.** *The first condition states that the eigenvectors of $\nabla^2 f(x)$ and $\Sigma(x)$ are aligned. In the vicinity of minima, this condition finds support from both theoretical analysis and empirical evidence, even in deep learning. Theoretically, Jastrzebski et al. [25] showed that $\Sigma(x^*) \approx \nabla^2 f(x^*)$ when the model fits all the data at x^*, which is further empirically verified by Xie et al. [65]. For mean-square loss, Mori et al. [40] and Paquette et al. [49] derived $\Sigma(x) \approx \frac{2f(x)}{B}\nabla^2 f(x^*)$ near local minima x^*, where B is the mini-batch size. Wang and Wu [64] further theoretically justified the approximation for nonlinear networks. In addition, the approximation $\Sigma(x) \approx \nabla^2 f(x^*)$ is commonly used in local escaping analysis of SGD [23, 65, 66, 69]. Our requirement here is more relaxed, as we only need the eigenvectors to be the same, regardless of eigenvalues.*

To guarantee the existence and uniqueness of a solution for an SDE, it is often sufficient to impose regularity conditions on both the drift and diffusion terms. In this context, we show that, when Σ is positive definite, conditions in Lemma 14 lead to the well-posedness of the proposed HA-SME model.

Theorem 18. *Assume $f \in C_b^3\left(\mathbb{R}^d\right)$, Σ being positive definite, and that at least one of the two conditions in Lemma 14 is satisfied everywhere.* `HA-SME` *has a unique strong solution.*

Remark 19. *While for general smooth objectives, Lemma 14 requires a small stepsize, in the particular case of quadratic functions with a constant noise covariance, the requirement of small stepsize is not necessary, as will be discussed in Appendix F. In this case, the diffusion coefficient is constant, and the drift term depends linearly on X_t, naturally fulfilling the Lipschitz criterion. This is sufficient for proving the well-posedness of SDEs, independent of the stepsize.*

Appendix E. Weak Approximation Error Analysis of `HA-SME`

In this section, we analyze the approximation errors of `HA-SME` in approximating SGD on general smooth functions. Our first result establishes the order 2 weak approximation for `HA-SME`, as defined in Definition 8, matching the order-best guarantee in the literature. Subsequently, we conduct a more fine-grained analysis, elucidating the explicit dependence of the approximation error on the smoothness parameter of the objective function. We observe a significant improvement of `HA-SME` over the existing `SME` models. In particular, for convex objectives, the leading error term of `HA-SME` is independent of the smoothness parameter.

E.1. Weak Approximation Error Guarantee

Below, we show a weak approximation error guarantee on sufficiently smooth functions. The assumptions we impose are common in deriving weak approximation errors for SDEs that approximate SGD [12, 22, 32].

Theorem 20. *Assuming for any ξ, $F(\cdot;\xi) \in C_b^7\left(\mathbb{R}^d\right)$,* `HA-SME` *is an order 2 weak approximation of SGD.*

An order 2 weak approximation is the order-best approximation guarantee known for SGD in existing literature. Typically, achieving weak approximation results requires bounding the (high-order) derivatives of the drift and diffusion terms of the SDE. While this condition is readily met for `SME`s, ensuring boundedness for `HA-SME` presents a challenge, as the drift and diffusion terms in `HA-SME` are defined as limits of power series and involve logarithmic components. Our proof builds upon the following lemma.

Lemma 21 (Regularity of the drift term and diffusion coefficient)**.** *Consider a fixed $n \geq 0$. Assume for that any ξ, we have $F(\cdot;\xi) \in C_b^{n+2}(\mathbb{R}^d)$. Then, there exists a constant $\eta_0 > 0$ such that, for any $\eta < \eta_0$, $\max_{0\leq i\leq d} \|[b(x)]_i\|_{C^n} < \infty$ and $\max_{0\leq i,j\leq d} \left\|\left[D(x)D(x)^\intercal\right]_{i,j}\right\|_{C^n} < \infty$.*

Remark 22. *A similar proof shows that* `SPF` *also admits the order 2 weak approximation error. Although* `SME`*-2,* `SPF`*, and* `HA-SME` *share the same order of weak approximation error, as demonstrated in Appendix F, these models can have drastically different behaviors near critical points. This discrepancy between the practice and the above theory arises because the classical analysis solely emphasizes the dependence on the stepsize, while neglecting other crucial factors, such as the norm of the Hessian matrix. To better differentiate these models, a more fine-grained analysis is necessary, where the dependence on the problem-dependent parameters is explicitly accounted for.*

E.2. Fine-Grained Error Analysis with Hessian Dependence

Our subsequent analysis explicitly examines the dependence of the error on stepsize η, Lipschitz parameter $s := \sup_{x\in\mathbb{R}^d} \|\nabla f(x)\|$, and smoothness parameter $\lambda := \sup_{x\in\mathbb{R}^d} \left\|\nabla^2 f(x)\right\|$. Through this detailed analysis, we differentiate the approximation guarantees of `HA-SME` from `SME`-2 and `SPF`, highlighting its advantage in modeling SGD.

Theorem 23. *Assume for any ξ, we have $F(\cdot;\xi) \in C_b^8\left(\mathbb{R}^d\right)$, and for any $0 \le i,j \le d$, we have $[\Sigma(\cdot)]_{i,j} \in C_b^6\left(\mathbb{R}^d\right)$. Let $X(t)$ be the stochastic process described by `HA-SME` and $\{x_k\}$ be the sequence generated by SGD. There exists $\eta_0 > 0$ such that for any $\eta < \eta_0$ and $T > 0$, it holds that for all $x \in \mathbb{R}^d$,*

$$\sup_{k=1,\ldots,\lfloor T/\eta\rfloor} \left|\mathbb{E}[u(x_k)\,|x_0 = x] - \mathbb{E}[u(X(k\eta))\,|X(0)=x]\right| \le \mathcal{O}\left(\left(\eta^2 s^3 + \eta^3 s^4 \lambda^3\right) M(T)\right), \quad (28)$$

where for any $p \ge 1$, $M(T) := \eta \sum_{k=0}^{\lfloor T/\eta\rfloor -1} \sum_{1\le|J|\le 8} \left|D^J u^k(x)\right|_\infty$ with $u^k(x) = \mathbb{E}\left[u(x_k) \mid x_0 = x\right]$.

Note that $M(T)$ characterizes how the regularity of the test function u deteriorates along the SGD trajectory. By definition, this quantity is solely determined by the SGD dynamics, thus independent of the SDE models.

Similarly, we derive the following results for `SME`-2 and `SPF`.

Theorem 24. *Under the same settings as Theorem 23,*

1. *when $X(t)$ is described by `SME`-2, it holds that for all $x \in \mathbb{R}^d$,*

$$\sup_{k=1,\ldots,\lfloor T/\eta\rfloor} \left|\mathbb{E}[u(x_k)\,|x_0 = x] - \mathbb{E}[u(X(k\eta))\,|X(0)=x]\right| \le \mathcal{O}\left(\left(\eta^2\left(s^3 + s\lambda^2\right) + \eta^3 s^4 \lambda^3\right) M(T)\right),$$

2. *when $X(t)$ is described by `SPF`, it holds that for all $x \in \mathbb{R}^d$,*

$$\sup_{k=1,\ldots,\lfloor T/\eta\rfloor} \left|\mathbb{E}[u(x_k)\,|x_0 = x] - \mathbb{E}[u(X(k\eta))\,|X(0)=x]\right| \le \mathcal{O}\left(\left(\eta^2\left(s^3 + \lambda\right) + \eta^3 s^4 \lambda^3\right) M(T)\right).$$

The above theorems implies that `HA-SME` demonstrates a clear improvement in terms of dependency on λ. In particular, it eliminates the dependence on λ in the leading error term involving η^2.

As is common in the finite-time approximation error analysis of SDE [12, 22, 32], we point out that $M(T)$ could exhibit exponential growth in terms of T and λ in general. However, in cases when the function is convex, the following result shows that these constants remain independent of λ.

Lemma 25. *Assuming that for any ξ, $F(\cdot;\xi) \in C_b^9\left(\mathbb{R}^d\right)$ is convex, i.e., $F(y;\xi) - F(x;\xi) \ge \nabla F(x;\xi)^\intercal (y-x)$ for any $x, y \in \mathbb{R}^d$. There exists a constant η_0 such that, for any $\eta < \eta_0$, it holds that*

$$M(T) \le \mathcal{O}\left(\|u\|_{C^8}\right),$$

where $M(T)$ is defined in Theorem 23, and $\mathcal{O}(\cdot)$ hides terms that do not depend on η, λ or s.

The above lemma allows us to develop the following approximation guarantee.

Theorem 26. *Fixing $T > 0$, assume that the test function satisfies $u \in C_b^8\left(\mathbb{R}^d\right)$, $F(\cdot;\xi) \in C_b^9\left(\mathbb{R}^d\right)$ is convex for any ξ, and for any $0 \le i, j \le d$, $[\Sigma(\cdot)]_{i,j} \in C_b^6\left(\mathbb{R}^d\right)$. Let $X(t)$ be the stochastic process described by* `HA-SME` *and $\{x_k\}$ be the sequence generated by SGD. There exists a constant $\eta_0 > 0$ such that for any $\eta < \eta_0$, it holds that for all $x \in \mathbb{R}^d$,*

$$\sup_{k=1,\dots,\lfloor T/\eta \rfloor} \left|\mathbb{E}[u\left(x_k\right)|x_0 = x] - \mathbb{E}[u\left(X(k\eta)\right)|X(0) = x]\right| \le \mathcal{O}\left(\eta^2 s^3\right).$$

Following a similar proof together with Theorem 24, we can establish that the upper bounds for `SME`-2 and `SPF` are $\mathcal{O}\left(\eta^2\left(s^3 + s\lambda^2\right)\right)$ and $\mathcal{O}\left(\eta^2(s^3 + \lambda)\right)$, respectively. This difference partially explains the limitations of existing SDE models in capturing the escape dynamics as illustrated in Figures 1 and 2.

Intuitively, the advantage of `HA-SME` is inherent in its construction by SBEA. Even under Principle 3 of term selection, `HA-SME` incorporates infinite sequences of terms associated with λ, $\left\{\left(\nabla^2 f(x)\right)^p\right\}_{p=0}^{\infty}$, in both its drift and diffusion terms. This is in stark contrast to `SME`-1, `SME`-2, and `SPF`, all of which truncate the error series at a certain point. All three of these models can be recovered by SBEA with a proper truncated series of $\{g_p\}$ and $\{h_p\}$: `SME`-1 includes g_0, h_0, and h_1; `SME`-2 includes g_0, g_1, h_0, and h_1; and `SPF` includes $\{g_i\}_{i=0}^{\infty}$, h_0, and h_1. We conjecture that the λ-dependence in the error analysis for `SME`-2 and `SPF` cannot be improved. This limitation arises because error terms that exhibit such λ-dependence are inherently truncated during their construction within the SBEA framework.

Remark 27. *Since Theorem 26 addresses the global point-wise approximation, the dependence on the global gradient norm of f appears unavoidable and cannot be neglected. However, to study the escaping behavior of SGD near a critical point, we believe that the global gradient norm s can be relaxed to its local version. This would be an interesting result because, in the vicinity of a critical point of a smooth function, the local gradient norm can be regarded as a negligible constant. Consequently, the leading term in Theorem 26 would be independent of both s and λ (whereas errors of existing SDE proxies always depend on λ).*

Remark 28. *If we further assume that $f(\cdot, \xi)$ is strongly-convex for any ξ, then $M(T)$ can be shown to be uniformly bounded w.r.t. T (the proof is provided in* **??** *G.2.1), leading to a uniform-in-time approximation guarantee[3]. A similar guarantee for* `SME`-2 *was established in the literature under the strong-convexity assumption [31], but without considering the explicit dependence on the problem parameters s and λ.*

Appendix F. Exact Recovery of SGD by `HA-SME` on Quadratics

The quadratic objective, despite its simplicity, holds significance in studying the behaviors of SGD. Its relevance comes from not only its application in linear regression but also its use in modeling the local curvature of complex models through second-order Taylor expansions. Moreover, insights from Neural Tangent Kernel [1, 24] suggests that in regression tasks, the objectives of sufficiently wide neural networks closely resemble quadratic functions. The convergence of SGD on quadratic models has been extensively studied, for example, for constant stepsizes [9] and stepsize schedulers [17, 47].

3. Our result assumes uniformly bounded gradients, which may not hold for strongly-convex functions. However, this can be alleviated by restricting our discussion to a compact set, and modifying `HA-SME` to prevent it from escaping this set, akin to the approach by Li and Wang [31].

Additionally, the exploration of locally escaping behaviors of SGD often relies on a local quadratic assumption [22, 23, 65, 66, 69].

In this section, we aim to compare different SDEs in terms of their ability to approximate SGD on quadratic functions. In this context, We first discuss the failure cases of existing SDE proxies (including `SPF` in Equation (27)) and then illustrate how `HA-SME` can accurately match SGD under certain conditions.

F.1. Failure Cases for Existing SDEs

We now offer a theoretical justification for the phenomena observed in Figures 1 and 2 — specifically, why `SME`-1, `SME`-2, and `SPF` fail in modeling SGD on quadratic functions. We restrict the noise in this analysis to be additive and state-independent.

Assumption 1 (quadratic objective[4])**.** *The objective function f is defined as $f(x) = \frac{1}{2}x^\intercal Ax$, where $A \in \mathbb{R}^{d\times d}$ is a symmetric real matrix. We denote its eigen-decomposition by $A = U\Lambda U^\intercal$, where $U \in \mathbb{R}^{d\times d}$ satisfies $U^\top U = I_d$ and $\Lambda \in \mathbb{R}^{d\times d}$ is diagonal.*

Assumption 2 (state-independent noise)**.** *The stochastic gradient satisfies*

$$\nabla F(x, \xi) = \nabla f(x) + \xi, \quad \xi \sim \mathcal{N}(0, \Sigma),$$

where Σ is a constant positive semi-definite matrix.

With state-independent noise, the aforementioned SDEs applied to quadratics can be represented as Ornstein–Uhlenbeck (OU) processes. For simplicity, this subsection considers isotropic noise, i.e. $\Sigma = \sigma^2 I$ for some scalar σ. We show that even within this highly simplified noise setting, existing SDEs may struggle to accurately capture the escaping behaviors exhibited by SGD. The following proposition computes the distributions of the iterates of SGD and the SDE proxies.

Proposition 29. *Under Assumptions 1 and 2 with $\Sigma = \sigma^2 I$ for some scalar σ, it holds that*

1. *The iterates of SGD (Equation (2)) with stepsize $\eta_k \equiv \eta$ satisfy*

$$x_k \sim \mathcal{N}\left(U\left(I - \eta\Lambda\right)^k U^\intercal x_0,\ \eta^2\sigma^2 \sum_{m=0}^{k-1} U\left(I - \eta\Lambda\right)^{2m} U^\intercal\right).$$

2. *The solution of `SME`-1 (Equation (7)) satisfies*

$$X(x_0, t) \sim \mathcal{N}\left(\exp(-At)x_0,\ \eta\sigma^2 U \frac{I - \exp(-2\Lambda t)}{2\Lambda} U^\intercal\right).$$

3. *The solution of `SME`-2 (Equation (4)) satisfies*

$$X(x_0, t) \sim \mathcal{N}\left(\exp\left(-\left(A + \frac{\eta}{2}A^2\right)t\right)x_0,\ \eta\sigma^2 U \frac{I - \exp\left(-\left(2\Lambda + \eta\Lambda^2\right)t\right)}{2\Lambda + \eta\Lambda^2} U^\intercal\right).$$

4. We focus on the simply quadratic $\frac{1}{2}x^\intercal Ax$, while noting that all results in this section can easily generalize to any quadratic function of the form $\frac{1}{2}x^\intercal Ax + bx + c$.

4. *The solution of* SPF *(Equation (27)), if $\eta < 1/\|A\|$, satisfies*

$$X(x_0,t) \sim \mathcal{N}\left(U\left(1-\eta\Lambda\right)^{t/\eta}U^\intercal x_0,\ \eta^2\sigma^2 U\frac{1-\left(1-\eta\Lambda\right)^{2t/\eta}}{-2\log\left(1-\eta\Lambda\right)}U^\intercal\right).$$

Now we delve into the analysis of different escaping scenarios.

Escaping Saddle Points in Figure 1(*a*) Consider the function defined in Figure 1(*a*), i.e., $f\left([x,y]^\intercal\right) = \frac{1}{2}\left(x^2-y^2\right)$, where the saddle point is at $(0,0)$. According to Proposition 29, along the direction of y, the iterate x_k generated by SGD has a mean of $(1+\eta\lambda_2)^k y_0$ and a variance of $\eta^2\sigma^2\sum_{m=0}^{k-1}(1+\eta\lambda_2)^{2m}$, which drifts away from 0 for any $\eta > 0$. Thus, SGD can successfully escape for any stepsize $\eta > 0$. However, for SME-2, when $\eta \geq 2$, as $t \to \infty$, its stationary distribution is a Gaussian distribution with mean 0 and covariance $\eta\sigma^2\left(2\Lambda+\eta\Lambda^2\right)^{-1}$. In such a regime of constant stepsize, the dynamics of SME-2 converges to a stationary distribution around 0, while SGD escapes exponentially fast. This observation confirms the numerical simulation presented in Figure 1(*a*).

Escaping Minimum in Figure 1(*b*) Consider the objective $f\left([x,y]^\intercal\right) = \frac{1}{2}\left(x^2+y^2\right)$ where all eigenvalues of the Hessian are positive. From Proposition 29, SGD will escape from $(0,0)$ if $\eta > 2$ — both the mean and covariance of the iterates grow exponentially. However, for SME-1 and SME-2, their stationary distributions are zero-mean Gaussians with covariance $\eta\sigma^2 U(2\Lambda)^{-1}U^\intercal$ and $\eta\sigma^2\left(2\Lambda+\eta\Lambda^2\right)^{-1}$, respectively. This elucidates why, in Figure 1(*b*), their dynamics oscillate around the minimum.

Escaping Behavior on Bimodal Function in Figure 2 Consider the one-dimensional piece-wise quadratic function depicted in Figure 2(*a*). When we initialize the point within the range $[-1,1]$, the local function is quadratic. At least for the first step, the behaviors of SGD and SPF should be accurately predicted by Proposition 29. We consider small stepsize close to $1/\|\nabla^2 f(x)\|$ (which is 1 in this case). Due to the additive noise, the iterates of SGD always have a variance no smaller than $\eta^2\sigma^2$. Such a variance allows the iterates to escape from the local basin and reach the basin around the global minimum at $x = 5$. However, after the first step, SPF would have a mean around 0 and variance also near 0 after time η. It can be shown by substituting $t = \eta$, setting $\eta\Lambda$ close to 1 in Proposition 29 and by noting that $\lim_{z\to 1}\frac{1-(1-z)^2}{\log(1-z)} = 0$. This explains why SPF tends to stay closely around the initial minimum in our simulation, as shown in Figure 2. This phenomenon underscores the need for a more nuanced correction of the diffusion term to accurately model the escaping behaviors of SGD in such settings.

F.2. Hardness of Approximating SGD with OU Process on Quadratics

Ideally, we would like to approximate SGD on quadratics exactly using SDEs, i.e., the iterates of SGD share the same distribution as its continuous-time counterpart at time stamps $k\eta$ for any natural number k. However, this is not possible for general quadratic objective with arbitrary gradient noises, even if they are additive and state-independent Gaussian noises. We provide a hard instance, demonstrating that no OU process (even extended to complex domain as elaborated below) can achieve this. We focus on the class of OU process because the corresponding transition kernel is Gaussian, which matches the transition kernel of SGD (when viewed as a Markov chain) under Assumptions 1 and 2.

Complex OU Process We note that the iterates of SGD in this context is always *real-valued* and follows a Gaussian distribution, as outlined in Proposition 29[5]. However, to allow the continuous-time proxy to have the maximum capability of approximating SGD, we allow it to be *complex-valued*. The benefit of extending to complex space has been observed by Rosca et al. [53] when matching the dynamics of GD by PF.

A random vector, denoted by $z = x + iy$, is a complex Gaussian random vector if x and y are two (possibly correlated) real Gaussian random vectors. In the real-valued case, a Gaussian random vector can be characterized by its mean and covariance, while in the complex-valued case, one additional parameter called *pseudo-covariance* is involved. Formally, the covariance is defined as $\Gamma(z) = \mathbb{E}\left[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^H\right]$ and pseudo-covariance is defined as $C(z) = \mathbb{E}\left[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^\intercal\right]$. We refer to **??** G.3.1 for a detailed discussion of properties of complex Gaussian random vectors.

We now state the desideratum when matching the distribution of a real normal variable, denoted as $\widetilde{z}$, and a complex normal variable z.

Desideratum 1. *The distribution of a complex Gaussian random vector z is said to* match *the distribution of real normal variable $\widetilde{z}$ if the following conditions hold:*

$$\mathbb{E}[\mathrm{Re}(z)] = \mathbb{E}[\widetilde{z}], \qquad \mathbb{E}[\mathrm{Im}(z)] = 0, \qquad \Gamma(z) = C(z) = \mathrm{Cov}[\widetilde{z}, \widetilde{z}].$$

The above result is equivalent to the statement: $\widetilde{z}$ and $\mathrm{Re}(z)$ have the same distribution, and $\mathrm{Im}(z)$ have 0 mean and 0 variance. Moreover, consider the setting described by Assumptions 1 and 2 so that the SGD iterates are all real Gaussian random vectors. A complex OU process

$$dX_t = BX_t dt + DdW_t,$$

is said to match *the iterates of the SGD if the distributions of X_t matches those of the SGD iterates at all corresponding time stamps, i.e. $t = k\eta$ where k is the iteration index of SGD. Here $B \in \mathbb{C}^{d\times d}$, $D \in \mathbb{C}^{d\times m}$ and W_t represents an m-dimensional standard Wiener process.*

Proposition 30. *Consider the setting described by Assumptions 1 and 2. One can construct real matrices A and Σ such that, for any given step-size $\eta > 0$, there exists no complex OU process[6] that matches the iterates of SGD, in the sense of Desideratum 1.*

The hard instance is constructed by choosing a *full-rank* matrix A and a *degenerate* covariance matrix Σ such that the eigenvectors of Σ are not aligned with those of A. The distributional mismatch of the SDE and SGD can be intuitively understood through the following argument: Starting from a deterministic initialization, after one SGD step, the covariance of the iterates is $\eta^2\Sigma$, which is rank-deficient; In contrast, in continuous-time dynamics, the noise injected is rotated by the misaligned linear transformations from the drift term and the covariance matrix quickly becomes full-rank. Please find the rigorous proof in **??** G.3.5.

5. This proposition, while initially framed for isotropic noises, can be easily generalized to anisotropic noises.
6. In our formulation of complex OU process, we use real Brownian motion W_t. It should be noted, however, that the framework can also include complex Brownian motion [51]. Consider a complex Brownian motion defined by $\widetilde{W}_t = W_t^r + iW_t^i$, where W_t^r and W_t^i are two independent real Brownian motions. Consequently, any SDE of the form $dX_t = BX_t dt + Dd\widetilde{W}_t$ can be equivalently expressed as $dX_t = BX_t dt + D\begin{bmatrix} I & iI \end{bmatrix}\begin{bmatrix} dW_t^r \ dW_t^i \end{bmatrix}^\intercal$.

F.3. Exact Approximation From HA-SME

While in general, complex OU process cannot exactly model SGD on quadratic functions, HA-SME when extended to complex-valued, achieves an exact match when the matrices A and Σ commute or when the stepsize is small enough. These conditions are slightly weaker than the sufficient conditions for the existence of our HA-SME (Lemma 14).

The following lemma outline conditions for a complex OU process to match SGD on quadratics.

Lemma 31. *Consider the same setting as Proposition 30, and denote the diagonal elements of Λ as $\lambda_1, \lambda_2, \ldots, \lambda_d$. After time $k\eta$ for $k \geq 0$, the mean of $X(k\eta)$ equals the mean of x_k, i.e., $\mathbb{E}[X(k\eta)] = \mathbb{E}[x_k]$, if and only if $B = U\frac{\log(1-\eta\Lambda)}{\eta}U^\intercal$. In addition,*

1. *the covariance of $X(k\eta)$ equals the covariance of x_k if and only if for all $0 \leq i, j \leq d$,*

$$\left[\left(U^\intercal D\right)\left(U^\intercal D\right)^H\right]_{i,j} = \frac{\left[U^\intercal \Sigma U\right]_{i,j}\left(\log(1-\eta\lambda_i) + \overline{\log(1-\eta\lambda_j)}\right)}{\eta\lambda_i\lambda_j - (\lambda_i + \lambda_j)}. \tag{29}$$

2. *pseudo-covariance of $X(k\eta)$ equals the covariance of x_k if and only if for all $0 \leq i, j \leq d$,*

$$\left[\left(U^\intercal D\right)\left(U^\intercal D\right)^\intercal\right]_{i,j} = \frac{\left[U^\intercal \Sigma U\right]_{i,j}\left(\log(1-\eta\lambda_i) + \log(1-\eta\lambda_j)\right)}{\eta\lambda_i\lambda_j - (\lambda_i + \lambda_j)}. \tag{30}$$

Satisfying any of the existence conditions for HA-SME in Lemma 14 ensures the existence of matrix D that adheres to the two required conditions in Lemma 31. Under these existence conditions, the matrices, whose elements are defined in the RHS of Equation (29) and Equation (30), become the same real positive semi-definite matrix. By taking the square roots of this matrix's eigenvalues, we can construct D that satisfies both Equations (29) and (30), thereby achieving an exact match for SGD on quadratic functions.

Theorem 32. *Under Assumptions 1 and 2, the solution of HA-SME exactly matches the iterates of SGD if either of the following two conditions holds:*

1. *A and Σ commute, and $\eta\lambda_i \neq 1$ for all eigenvalues λ_i of A.*
2. *Σ is positive definite, and $\eta \leq \frac{1}{\|A\|}\min\left\{1 - \sqrt{1 - \frac{\lambda_{\min}(\Sigma)}{\sqrt{d}\lambda_{\max}(\Sigma)}}, 1 - \frac{\sqrt{2}}{2}\right\}$.*

To the best of our knowledge, this marks the first instance of an SDE that precisely mirrors the distribution of SGD, albeit restricted to quadratic functions.

Remark 33. *The first condition in this theorem relaxes the constraints from Lemma 14 to allow larger stepsizes. As noted in Rosca et al. [53], complex flow is helpful for capturing the instabilities caused by large stepsizes in GD on quadratics. This is also the case for HA-SME. When $\eta < 1/\|A\|$, HA-SME operates in real space. However, when $\eta > 1/\|A\|$, imaginary components emerge due to the logarithmic function.*

Appendix G. Detailed Proofs

G.1. Construction of HA-SME

G.1.1. PROOF SKETCH OF LEMMA 12

We start with a sketch of the proof for Lemma 12, where we first determine the functional structure of the components g_i's and h_i's in the SBEA ansatz in Equation (12) and then determine their exact expression using a generating function based approach. We prove via induction. The complexity in our analysis emerges from identifying all possible terms generated by the expansion of $\mathcal{L}_{l_1}\mathcal{L}_{l_2}\cdots\mathcal{L}_{l_n}$ in Equation (18). The collective expansion of these operators unfolds as a multinomial series. This sets our work apart from the PF, as in the latter, in the absence of diffusion terms, the expansion generates only a monomial. The rigorous proofs of the statements made in the induction are provided afterwards.

Lemma 10 is the building block for our analysis, it can be seen from the following.

Proof for Lemma 10. In the semi-group expansion of Equation (17), consider the term $\frac{1}{n!}\eta^n\mathcal{L}^n u(x)$ for $1 \leq n \leq p+1$. Since we already have η^n, to get η^{p+1} , $\mathcal{L}^n u(x)$ must contribute η^{p+1-n}. As $\mathcal{L}_i$ contains η^i, we obtain the conclusion. □

Before providing the inductive proof, statement (1) in Lemma 12, i.e. the uniqueness of g_p and h_p, can be easily obtained from the following argument.

Proof of point (1) in Lemma 12. We first give examples for solving the first few g_i and h_i, then we use an argument of induction to finish the proof. For the terms associated with η^0, Equation (14) and Equation (17) already match, i.e., $u(x)$. For terms associated with η, we can solve that $g_0(x) = -\nabla f(x)$ and $h_0(x) = 0$. Considering terms with η^2, we have for discrete-time (Equation (14))

$$\eta^2 \cdot \frac{1}{2}\nabla^2 u(x) : \left(\nabla f(x)\nabla f(x)^\intercal + \Sigma(x)\right),$$

and for continuous-time(Equation (17))

$$\eta^2 \cdot \left(\underbrace{g_1(x)^\intercal \nabla u(x) + \frac{1}{2}h_1(x) : \nabla^2 u(x)}_{\text{from } \eta\mathcal{L}u(x)} + \underbrace{\frac{1}{2}(-\nabla f \cdot \nabla(-\nabla f \cdot \nabla u))(x)}_{\text{from } \frac{1}{2}\eta^2\mathcal{L}^2 u(x)} \right).$$

Note that this holds for any proper u, so we group terms associated with ∇u and $\nabla^2 u$ and match the terms in discrete-time and continuous-time. In this way, we obtain 2 equations, which gives us the following g_1 and h_1:

$$g_1(x) = -\frac{1}{2}\nabla^2 f(x)\nabla f(x) \quad \text{and} \quad h_1(x) = \Sigma(x). \tag{31}$$

Now assume we already know $\{g_i\}_{1\leq i<p}$ and $\{h_i\}_{1\leq i<p}$. We will proceed to solve g_p and h_p. This is done by solving terms with η of order $p+1$. The terms associated with η^{p+1} are shown in Lemma 10. When $n = 1$, we will have linear terms with g_p and h_p, respectively. When $n \geq 2$, since $l_1 + l_2 + \cdots + l_n$ could only sum to $p-1$, we only have g_i and h_i for $i < p$. Given that we assume $\{g_i\}_{1\leq i<p}$ and $\{h_i\}_{1\leq i<p}$ are already solved, by grouping terms with ∇u and $\nabla^2 u$, we obtain two

linear equations and can solve for g_p and h_p. Note that the coefficients for g_p and h_p in these two linear equations are non-zero, therefore, the solution exists and is unique. □

We now give a sketch of the inductive proof of the second statement in Lemma 12, on the functional structure of the components h_p's and g_p's. Recall point (2) in Lemma 12: The components g_p and h_p admit the form

$$g_p(x) = c_p \cdot \left(\nabla^2 f(x)\right)^p \nabla f(x), \tag{32}$$

$$h_p(x) = \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot \left(\nabla^2 f(x)\right)^k \Sigma(x) \left(\nabla^2 f(x)\right)^{p-1-k}, \tag{33}$$

where $\{c_k\}$ and $\{a_{k,p-1-k}\}$ are constants to be determined. Let q be the induction index and let Equations (20) and (21) be the induction hypothesis for $q = p \geq 1$. Such a hypothesis clearly holds for $q = 1$ (recall the calculation in Remark 11). The following lemma shows that the nice functional structures in Equations (20) and (21) are preserved for $q = p + 1$.

Lemma 34. *Under Principle 3, suppose that Equations* (20) *and* (21) *hold in step $q = p$. They also hold for $q = p + 1$.*

Proof sketch. To establish the induction, we need to enumerate all possible terms generated by the expansion in Equation (18). Viewing the application of each $\mathcal{L}_{l_i}$ as a layer, this process entails a dual-stage selection mechanism:

1. Within each layer we consider the terms generated from either $\eta^{l_i} g_{l_i} \cdot \nabla$ or $\frac{\eta^{l_i}}{2} h_{l_i} : \nabla^2$ in $\mathcal{L}_{l_i}$.

2. For a fixed layer, we consider how the operators ∇ and ∇^2 are applied on the subsequent layer. For example, for the operator ∇, it could be acting on $g_{l_{i+1}}$ to yield a concrete function $\nabla g_{l_{i+1}}$, or it can engage with the operator ∇^2 from the subsequent layer to form the operator ∇^3.

Finally, for every possible `sequence-of-selections` made in the above mechanism, any remaining differential operators ∇^j will be applied on the test function u. Hereafter, we use the typewriter font to emphasize that the `sequence-of-selections` is with respect to the above selection mechanism.

Unraveling all potential selection in Equation (18) is inherently difficult. However, Principle 3 allows us to ensures the particular structures of g_p and h_p. For the following discussion, we introduce the concept of `free-nabla-`j: During unraveling of the operator compositions in Equation (18), a `free-nabla-`j is generated if we obtain an operator ∇^j during any stage of the above selection mechanism.

We have the following result.

Lemma 35. *Once `free-nabla-`j, for $j \geq 3$, is generated, one can terminate the subsequent operator expansion as all the resulting terms will be excluded by Principle 3.*

See an elaborated discussion in **??** G.1.2. From this observation, the expansion of Equation (18) can be significantly simplified and one can show that if the induction in Equations (20) and (21) holds for $q = p$, g_{p+1} and h_{p+1} admit the following form:

$$g_{p+1}(x) = c_{p+1} \cdot \left(\nabla^2 f(x)\right)^{p+1} \nabla f(x) \tag{34}$$

$$h_{p+1}(x) = \sum_{k=0}^{p} (a_{k,p-k} + b_{k,p-k} \cdot \|\nabla f(x)\|^2) \cdot \left(\nabla^2 f(x)\right)^k \Sigma(x) \left(\nabla^2 f(x)\right)^{p-k}. \tag{35}$$

Further, a detailed examination of the combinations reveals the following result.

Lemma 36. $b_{k,p-k} \equiv 0$.

The above results together establish the induction step of $q = p + 1$. □

Once we have determine the functional structure of g_p and h_p, the next step is to determine the coefficients $a_{i,j}$ and c_p. As we obtain the structures of g_p and h_p by induction, the coefficients admit recursion forms. To resolve these coefficients, we utilize an argument based on the generating functions. As a result, we obtain point (3) in Lemma 12.

G.1.2. Proof of Lemma 34

To establish the inductive step in Lemma 34, we prove the following points in this section.

(a) We first prove Lemma 35, which allows us to significantly simplify the expansion of the compositions of the generators $\mathcal{L}_i$'s.

(b) We analyze all possible terms generated without incurring `free-nabla-`j, for $j \geq 3$, and find that there are only three possibilities. This leads to the functional structure described in Equations (34) and (35).

(c) We further provide the recursive definitions of the coefficients $\{c_i\}$, $\{a_{i,j}\}$, and $\{b_{i,j}\}$. This allows us to conclude that $b_{k,p-k} \equiv 0$ (Lemma 36) and also it allows us to calculate the limit of the power series $\{g_p\}$ and $\{h_p\}$, i.e. point (3) in Lemma 12.

G.1.3. Proof of Lemma 35

Let us recall the concept of `free-nabla-`j: During unraveling of the operator compositions in Equation (18), a `free-nabla-`j is generated if we obtain an operator ∇^j during any stage of the selection mechanism mentioned in the proof sketch of Lemma 34.

In the following, we show that once `free-nabla-`j for $j \geq 3$ is generated in a `sequence-of-selections` mentioned in the proof sketch of Lemma 34, the term resulting from this sequence is excluded by Principle 3, i.e. it contains at least one of the terms $\nabla^{(r)} f(x)$, $\nabla^{(s)} u(x)$ and $\nabla^{(m)} \Sigma(x)$ for $r, s \geq 3$ and $m \geq 1$. To establish this result, there are two possibilities once `free-nabla-`j for $j \geq 3$ is generated:

1. The `free-nabla-`j operator is to be directly applied on the test function u. In this case, all the resulting term from this `sequence-of-selections` contains $\nabla^{(j)} u(x)$, $j \geq 3$, which is excluded by Principle 3.

2. The `free-nabla-`j operator is to be applied on a generator $\mathcal{L}_i$, i.e. we encounter the term

$$\nabla^{(j)}(\eta^i g_i \cdot \nabla + \frac{1}{2}\eta^i h_i : \nabla^2). \tag{36}$$

> One can prove that (see the discussion below) all terms in the expansion of the above expression either is excluded by Principle 3 or it contains `free-nabla-`q for $q \geq j$[7]. Consequently, the order j does not decrease after the operator `free-nabla-`j being applied on a generator $\mathcal{L}_i$.

One can repeat the case 2 until the `free-nabla-`j operator is directly applied on the test function u, which reduces to case 1. From the above argument, all terms generated are to be excluded by Principle 3.

G.1.4. PROOF OF THE FUNCTIONAL STRUCTURES IN EQUATIONS (34) AND (35)

According to Lemma 35, we only need to consider the `sequence-of-selections` mentioned in the proof sketch of Lemma 34 such that no `free-nabla-`j operator is generated, for $j \geq 3$. It turns out that there are only three possible `sequence-of-selections`, as depicted in Figure 3.

In this section, we elaborate on these three possible sequences, based on which we then establish the functional structures in Equations (34) and (35). In the meantime, we also derive the recursive definitions of the coefficients $\{c_i\}$, $\{a_{i,j}\}$, and $\{b_{i,j}\}$.

Notation For the ease of the proof, we define for $n \geq 1$, $m \geq 0$ and sequence $\{c_i\}_{i=0}^{+\infty}$,

$$\rho\left(n, m, \{c_i\}_{i=0}^{+\infty}\right) = \sum_{l_1+l_2+\ldots l_n=m} c_{l_1} c_{l_2} \cdots c_{l_n},$$

i.e., sum of all possible combinations of n items from sequence $\{c_i\}_{i=0}^{+\infty}$ such that the sum of indices is m. A way to look at this is through generating functions. Let $c(x) = \sum_{i=0}^{+\infty} c_i x^i$, then $\rho\left(n, m, \{c_i\}_{i=0}^{+\infty}\right)$ is equivalent to the coefficient of x^m in $c(x)^n$.

Solve for g_p We claim that we can solve g_p by matching terms with ∇u. To show the claim, we observe that ∇u terms could only be achieved by choosing g_{l_i} and never selecting h_{l_i} for for all $1 \leq i \leq n$. This is the case of our Sequence Type I illustrated in Figure 3(a). To see why this is the case, we start with $\mathcal{L}_{l_n}$, since we want terms with $\nabla u(x)$, we must select g_{l_n} instead of h_{l_n}. Next, when selecting from $\mathcal{L}_{l_{n-1}}$, we could also only select $g_{l_{n-1}}$, since if $h_{l_{n-1}}$ is chosen, we would have $h_{l_{n-1}} : \nabla^2 \left(g_{l_n} \cdot \nabla u\right)$. The ∇^2 operation must be applied to g_{l_n} otherwise we will have $\nabla^2 u$ instead of ∇u. However, when we apply the gradient operator twice in g_{l_n}, there must be a third-order gradient factor, which we choose to exclude, popping up as we have g_{l_n} of form Equation (20). Following this logic recursively, we have that ∇u terms contain only g_i (with $0 \leq i \leq p-1$). Also note that the gradient operator in $g_{l_i} \cdot \nabla$ must be applied to $g_{l_{i+1}}$ except when $i = n$ where the gradient operator is applied to u, otherwise we will have ∇^2 operator and face the same issue as when we selecting h_{l_i}.

Since now in the ∇u terms only contains g_p (with non-zero coefficient) without h_p, we can solve for g_p. Again, since the gradient operator from g_{l_i} (except $i = n$) could only be used for the next $g_{l_{i+1}}$, specifically applied on the $\nabla f(x)$ factor of $g_{l_{i+1}}$ otherwise we would have third-order gradient

7. For simplicity, we take $j = 3$ as example. For $j > 3$, the logic is similar. One has for the first term in Equation (36)

$$\nabla^{(3)}\left(g_i \cdot \nabla\right) = \underbrace{g_i \cdot \nabla^{(4)}}_{\texttt{free-nabla-4}} + \underbrace{\nabla g_i \cdot \nabla^{(3)}}_{\texttt{free-nabla-3}} + \underbrace{\nabla^2 g_i \cdot \nabla^{(2)} + \nabla^3 g_i \cdot \nabla}_{\text{every term contains } \nabla^r f,\, r \geq 3}.$$

By the product rule of the differential and based on the induction in Equation (34), one can check that in the expansions of $\nabla^2 g_i$ and $\nabla^3 g_i$, all terms contain $\nabla^r f$, for $r \geq 3$. Hence, the last two terms in the above expression are excluded by Principle 3. The second term in Equation (36) can be excluded with a similar argument.

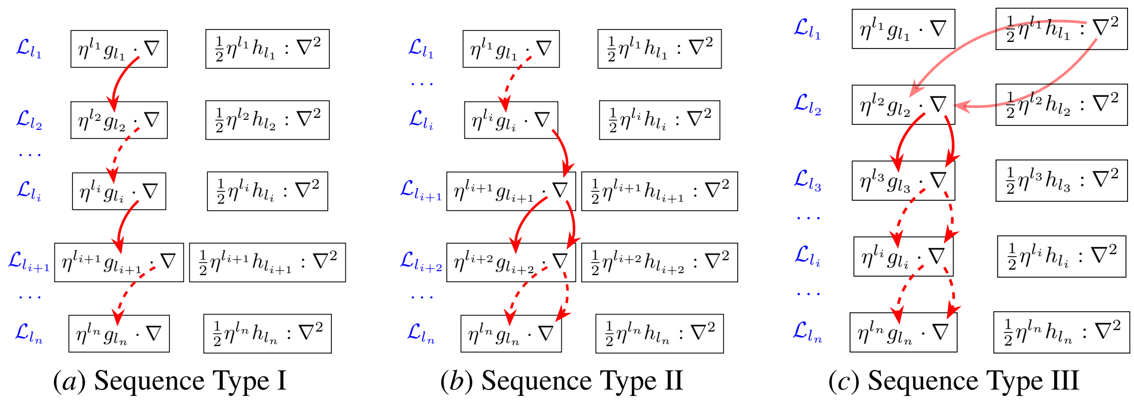


Figure 3: Illustrations of possible combinations in the construction of HA−SME. In constructing HA−SME, we examine the product of $\mathcal{L}_{l_1}\mathcal{L}_{l_2}\cdots\mathcal{L}_{l_n}$, where each $\mathcal{L}_{l_i}$ contains two terms. The red arrows indicate the remaining Sequences after filtering out those with higher-order derivatives of f, u, or Σ, which are excluded in the construction rules of HA−SME. These arrows originate from ∇ or ∇^2 and lead to g_{l_i} or ∇, demonstrating how the ∇ from the previous operator is applied in the subsequent operator. The dashed arrows represent the exclusion of repeated operations.

of $f(x)$, we have $n-1$ factors of $\nabla f(x)$ changed to $\nabla^2 f(x)$ and only one remains. Therefore, g_p is of form Equation (20).

The remaining proof for finding c_p is almost the same as the proof for Theorem A.2 of Rosca et al. [53]. The resulting c_i is the coefficients of the Taylor expansion of $c(x) = \frac{\log(1-x)}{x}$, i.e., $c(x) = \sum_{i=0}^{+\infty} c_i x^i$.

Solve for h_p In the proof g_p, we have shown that the ∇u terms allow us to solve g_p, and now we proceed to solve the form of h_p by considering $\nabla^2 u$ terms. Again, to solve h_p, we need terms with the order of η summing up to $p+1$ and consider Lemma 10. Let us discuss what are the valid possibilities of selections from such an expansion.

Let us call the selection between g_{l_i} and h_{l_i} the i-th step of selection. Assume we have the gradient operator ∇^q after the i-th step. If in the next step, we select $g_{l_{i+1}} \cdot \nabla$, since $g_{l_{i+1}}$ has only one factor of ∇f, which means we can only apply the gradient operator once on this $g_{l_{i+1}}$, the order of the remaining gradient operator passed to the next step is still at least q. If in the $i+1$-th step, we select $h_{l_{i+1}} : \nabla^2$ instead, all the gradient operators should go to the next step (otherwise, if applied on $h_{l_{i+1}}$, factors we exclude will emerge), resulting in a ∇^{q+2}. Note that by our construction, in the end, the operator applied on u should not exceed 2. Therefore no matter what we select from the first step, which produces at least ∇, we should only select g_{l_i} instead h_{l_i} for $i>1$. Because otherwise we would have at least ∇^3 passing to u. According to the above reasoning, we have only two cases left:

1. **If we select g_{l_1} in the first step**, which corresponds to Figure 3(*b*), then in order to get $\nabla^2 u$ at the end, we will need at a step i with $2 \leq i \leq n$, the gradient operator is not applied to g_{l_i} but passed to the next step. For other steps, the operator is applied to g. This is the only way to have $\nabla^2 u$. Plugging in the solution of g_i, we know that the resulting term in Equation (17) is

of form

$$\eta^{p+1} \sum_{s=0}^{p-1} b_{s,p-1-s} \operatorname{Tr}\left\{ \nabla f(x) \left(\nabla^2 f(x)\right)^s \nabla^2 u(x) \left(\nabla^2 f(x)\right)^{p-1-s} \nabla f(x)^\intercal \right\},$$

where $b_{s,p-1-s}$ can be written as

$$\begin{aligned} & b_{s,p-1-s} \\ &= \sum_{n=2}^{p+1} \frac{1}{n!} \sum_{i=2}^{n} \sum_{q=0}^{n-i} \binom{n-i}{q} \rho\left(q+1, p-1-s-q, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-1-q, s+q-n+2, \{c_k\}_{k=0}^{+\infty}\right), \end{aligned} \tag{37}$$

where $c_k = -\frac{1}{1+k}$, the coefficient associated with g_i. Note that the coefficient is for terms with $\nabla^2 f(x)$ to the power of s on the left of $\nabla^2 u$ and $p-1-s$ on the right. To see why this is the case, we have the following selection procedure:

(a) Such terms can be found in $\frac{\eta^n}{n!}\mathcal{L}^n u$ for $n = 2$ to $p+1$.

(b) As mentioned before, we have at step $2 \leq i \leq n$, the gradient operator is not applied to g_{l_i}. Note that in this case the $\nabla^2 f(x)$ coming from g_{l_i} will contribute to the RHS of $\nabla^2 u$. Then we will have for the remaining steps of selection,

$$\nabla^2 \left(g_{l_{i+1}} \cdot \nabla \left(g_{l_{i+2}} \cdot \nabla \left(\cdots g_{l_n} \cdot \nabla u\right)\right)\right).$$

(c) According to Lemma 40, there are $\binom{n-i}{q}$ combinations such that the number of ∇g resulting in the RHS of $\nabla^2 u$ is q.

(d) For the RHS of $\nabla^2 u$, we have g_{l_i} and q of ∇g selected from the last step. Note that g_k would contains $\left(\nabla^2 f(x)\right)^k$ for $k \geq 0$, and for the q of ∇g, we obtain additional $\left(\nabla^2 f(x)\right)^q$ because of the ∇ operator. Therefore, the total number of $\nabla^2 f(x)$ on the RHS of $\nabla^2 u$ comes from g_{l_i} and q of ∇g and an additional q. Therefore, we select from $q+1$ of g, whose indices sum up to $p-1-s-q$ (so that the total order of $\nabla f(x)$ is $p-1-s-q+q = p-1-s$).

(e) For the LHS of $\nabla^2 u$, we have g_1, $\{\nabla g_{l_k}\}_{k=2}^{i-1}$ and $n-i-q$ of ∇g selected according to Step 1c. Following similar logic as the last step, in total we select $1+i-2+n-i-q = n-1-q$ of g, whose indices sum up to $s+q-n+2$.

We note that for any matrix A, $\operatorname{Tr}\{A\} = \operatorname{Tr}\{A^\intercal\}$, so the traces associated with $b_{s,m}$ and $b_{m,s}$ are the same for $s, m \geq 0$. According to Lemma 37, $b_{s,m} + b_{m,s} = 0$ for $m + s \geq 1$. Therefore, the coefficients cancel each other. For the case of $b_{s,s}$ with $s \geq 1$, Lemma 37 also implies $b_{s,s} + b_{s,s} = 0 \implies b_{s,s} = 0$.

Therefore, in summary, in this whole case, there is no term generated.

2. **If we select h_{l_1} in the first step**, which corresponds to Figure 3(*c*), then in order to get $\nabla^2 u$ at the end, we will in the following steps apply the gradient at g_{l_k} for $k \geq 2$ at once and pass to

the next step the other gradient operator. In this way, all terms of Equation (17) with order of η summing to $p+1$ and coming from $\frac{1}{n!}\mathcal{L}^n u$, has the following form:

$$\frac{1}{2}\eta^{p+1}\tilde{a}_{i,j,k}\operatorname{Tr}\Big\{\left(\nabla^2 f(x)\right)^i \Sigma(x)\left(\nabla^2 f(x)\right)^j \nabla^2 u(x)\left(\nabla^2 f(x)\right)^k\Big\}, \tag{38}$$

for some absolute constants $\tilde{a}_{i,j,k}$ with $i+j+k=p-1$. The reason why we have this form is because of our induction assumption and Lemma 40. The reason why $i+j+k=p-1$ is that to have η^{p+1}, we must have $l_1+l_2+\cdots+l_n=p+1-n$. For the first step, we have h_{l_1}. According to the induction assumption, it can offer (l_1-1)-th order of $\nabla^2 f(x)$. For the remaining steps, we have ∇g_{l_s}, which can offer (l_s+1)-th order of $\nabla^2 f(x)$. Therefore, in total we have $\nabla f(x)$ to the power of $l_1-1+l_2+l_3+\cdots+l_n+n-1=p+1-n-1+n-1=p-1$. According to the property of the trace operator, Equation (38) can also be written as

$$\frac{1}{2}\eta^{p+1}\tilde{a}_{i,j,k}\operatorname{Tr}\Big\{\left(\nabla^2 f(x)\right)^{k+i} \Sigma(x)\left(\nabla^2 f(x)\right)^j \nabla^2 u(x)\Big\}.$$

Now we try to match the η^{p+1} terms in Equation (14) and Equation (17). Note that in Equation (14), the term associated with η^{p+1} for $p\geq 2$ is 0, therefore we have the following equality for solving h_p:

$$\frac{1}{2}\eta^{p+1}\left(\operatorname{Tr}\big\{h_p(x)^\intercal\nabla^2 u(x)\big\}+\sum_{n=2}^{p+1}\tilde{a}_{i,j,k}\operatorname{Tr}\Big\{\left(\nabla^2 f(x)\right)^{k+i} \Sigma(x)\left(\nabla^2 f(x)\right)^j \nabla^2 u(x)\Big\}\right)=0,$$

which implies (since this should hold for any u)

$$h_p(x)=-\sum_{n=2}^{p+1}\tilde{a}_{i,j,k}\left(\nabla^2 f(x)\right)^j \Sigma(x)\left(\nabla^2 f(x)\right)^{k+i}.$$

So far, by induction, we have proved that h_p has the form of Equation (20) (regardless of the constants). Next we will try to find these constants. We will stick to the notation in Lemma 12, i.e., using $a_{i,j}$ for the rest of the proof, i.e.,

$$h_p(x)=\sum_{k=0}^{p-1}a_{k,p-1-k}\cdot\left(\nabla^2 f(x)\right)^k \Sigma(x)\left(\nabla^2 f(x)\right)^{p-1-k}.$$

Similar to the proof of Equation (37), we have

$$\begin{aligned}&a_{k,p-1-k}\\&=-\sum_{n=2}^{p}\frac{1}{n!}\sum_{l=0}^{k}\sum_{r=0}^{p-1-k}a_{l,r}\sum_{q=0}^{n-1}\binom{n-1}{q}\rho\left(q,k-q-l,\{c_k\}_{k=0}^{+\infty}\right)\rho\left(n-1-q,p-k-n+q-r,\{c_k\}_{k=0}^{+\infty}\right).\end{aligned} \tag{39}$$

The reason is that we have the following selection process:

(a) Such terms can be found in $\frac{\eta^n}{n!}\mathcal{L}^n u$ for $n=2$ to $p+1$.

(b) We try to find term with

$$\mathrm{Tr}\Big\{\left(\nabla^2 f(x)\right)^{p-1-k}\Sigma\left(\nabla^2 f(x)\right)^{k}\nabla^2 u\Big\},$$

which can be found from Sequence

$$\underbrace{\left(\nabla^2 f(x)\right)^{l}\Sigma\left(\nabla^2 f(x)\right)^{r}}_{h_{l_1}} : \left(\nabla^2 f(x)\right)^{k-l}\nabla^2 u\left(\nabla^2 f(x)\right)^{p-1-k-r}.$$

(c) For $\left(\nabla^2 f(x)\right)^k$, we can have $\left(\nabla^2 f(x)\right)^l$ coming from h_{l_1}. Also h_{l_1} can contribute to $(\nabla f(x))^{p-1-k}$ with $\left(\nabla^2 f(x)\right)^r$. That is why we have summations over l and r with coefficients $a_{l,r}$.

(d) As mentioned before, we have for the selection at step $2 \le k \le n$,

$$\nabla^2\left(g_{l_2}\cdot\nabla\left(g_{l_3}\cdot\nabla\left(\cdots g_{l_n}\cdot\nabla u\right)\right)\right).$$

According to Lemma 40, there are $\binom{n-1}{q}$ combinations such that the number of ∇g resulting in the LHS of $\nabla^2 u$ is q.

(e) For the LHS of $\Sigma(x)$, we have q of ∇g selected from the last step. The indices should sum up to $k-q-l$, since we will generate additional q of $\nabla^2 f(x)$ and h_{l_1} provides l. Therefore, we have coefficient $\rho\left(q, k-q-l, \{c_k\}_{k=0}^{+\infty}\right)$ similarly for the RHS.

G.1.5. PROOF OF LEMMA 36 AND THE RECURSIVE EXPRESSIONS OF $\{c_i\}$ AND $\{a_{i,j}\}$

Lemma 37. *Recall the recursive definition of $b_{s,m}$ in Equation (37),*

$$b_{s,m} = \sum_{n=2}^{s+m+2}\frac{1}{n!}\sum_{i=2}^{n}\sum_{q=0}^{n-i}\binom{n-i}{q}\rho\left(q+1, m-q, \{c_k\}_{k=0}^{+\infty}\right)\rho\left(n-1-q, s+q-n+2, \{c_k\}_{k=0}^{+\infty}\right),$$

where $s,m \ge 0$, $s+m\ge 1$, $c_k = -\frac{1}{k+1}$. We have $b_{s,m}+b_{m,s}=0$.

Proof. Let $c(x) = \sum_{k=0}^{+\infty} c_k x^k$, and we know that $c(x) = \frac{\log(1-x)}{x}$. Define

$$b(x,y) = \sum_{s,m\ge 0}^{+\infty} b_{s,m}x^s y^m.$$

Then we have

$$\begin{aligned}
&b(x,y)\\
&= \sum_{s,m\ge 0}^{+\infty} x^s y^m \sum_{n=2}^{s+m+2}\frac{1}{n!}\sum_{i=2}^{n}\sum_{q=0}^{n-i}\binom{n-i}{q}\rho\left(q+1, m-q, \{c_k\}_{k=0}^{+\infty}\right)\rho\left(n-1-q, s+q-n+2, \{c_k\}_{k=0}^{+\infty}\right)\\
&= \sum_{n=2}^{+\infty}\frac{1}{n!}\sum_{s+m\ge n-2}^{+\infty} x^s y^m \sum_{i=2}^{n}\sum_{q=0}^{n-i}\binom{n-i}{q}\rho\left(q+1, m-q, \{c_k\}_{k=0}^{+\infty}\right)\rho\left(n-1-q, s+q-n+2, \{c_k\}_{k=0}^{+\infty}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{n=2}^{+\infty} \frac{1}{n!} \sum_{s+m\geq n-2}^{+\infty} x^s y^m \sum_{q=0}^{n-2} \sum_{i=2}^{n-q} \binom{n-i}{q} \rho\left(q+1, m-q, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-1-q, s+q-n+2, \{c_k\}_{k=0}^{+\infty}\right) \\
&= \sum_{n=2}^{+\infty} \frac{1}{n!} \sum_{s+m\geq n-2}^{+\infty} x^s y^m \sum_{q=0}^{n-2} \binom{n-1}{q+1} \rho\left(q+1, m-q, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-1-q, s+q-n+2, \{c_k\}_{k=0}^{+\infty}\right) \\
&= \sum_{n=2}^{+\infty} \frac{1}{n!} \sum_{s+m\geq n-2}^{+\infty} x^s y^m \sum_{q=1}^{n-2} \binom{n-1}{q} \rho\left(q, m-q+1, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-q, s+q-n+1, \{c_k\}_{k=0}^{+\infty}\right) \\
&= \frac{1}{y} \sum_{n=2}^{+\infty} \frac{1}{n!} \sum_{s+m\geq n-2}^{+\infty} x^s y^{m+1} \sum_{q=1}^{n-2} \binom{n-1}{q} \rho\left(q, m-q+1, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-q, s+q-n+1, \{c_k\}_{k=0}^{+\infty}\right),
\end{aligned}
$$

where in the forth equality, we used hockey-stick identity [54], and in the fifth equality, we replace q with $q-1$. Note that

$$
\sum_{s+m\geq n-2}^{+\infty} x^s y^{m+1} \sum_{q=1}^{n-2} \binom{n-1}{q} \rho\left(q, m-q+1, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-q, s+q-n+1, \{c_k\}_{k=0}^{+\infty}\right) \tag{40}
$$

is equivalent to

$$
\left(yc(y)+xc(x)\right)^{n-1} c(x) - \left(xc(x)\right)^{n-1} c(x). \tag{41}
$$

To see why this is the case, let us first looked at

$$
\left(yc(y)+xc(x)\right)^{n-1} c(x).
$$

To have y^{m+1}, we can select $yc(y)$ from $\left(yc(y)+xc(x)\right)^{n-1}$ for q times, which results in coefficients $\binom{n-1}{q}$ and $\rho\left(q, m-q+1, \{c_k\}_{k=0}^{+\infty}\right)$. For x^s, we select $n-1-q$ of $xc(x)$ from $\left(yc(y)+xc(x)\right)^{n-1}$, and obtain $\rho\left(n-q, s+q-n+1, \{c_k\}_{k=0}^{+\infty}\right)$. Note that in Equation (40), the sum of q starts from 1, which means we have not considered the case where $yc(y)$ is never selected in $\left(yc(y)+xc(x)\right)^{n-1}$. That is why in Equation (41), we subtract $\left(xc(x)\right)^{n-1} c(x)$.

Next, we proceed with

$$
\begin{aligned}
b(x,y) &= \frac{1}{y} \sum_{n=2}^{+\infty} \frac{1}{n!} \left(\left(yc(y)+xc(x)\right)^{n-1} c(x) - \left(xc(x)\right)^{n-1} c(x) \right) \\
&= \frac{1}{y} \sum_{n=2}^{+\infty} \frac{1}{n!} \left(yc(y)+xc(x)\right)^{n-1} c(x) - \frac{1}{y} \sum_{n=2}^{+\infty} \frac{1}{n!} \left(xc(x)\right)^{n-1} c(x).
\end{aligned} \tag{42}
$$

For the first part, we have

$$
\begin{aligned}
&\frac{1}{y} \sum_{n=2}^{+\infty} \frac{1}{n!} \left(yc(y)+xc(x)\right)^{n-1} c(x) \\
&= \frac{c(x)}{y\left(yc(y)+xc(x)\right)} \sum_{n=2}^{+\infty} \frac{1}{n!} \left(yc(y)+xc(x)\right)^{n}
\end{aligned}
$$

$$= \frac{c(x)}{y\left(yc(y)+xc(x)\right)}\left(e^{yc(y)+xc(x)}-1-\left(yc(y)+xc(x)\right)\right),$$

where in the last equality we used Taylor expansion $e^x = \sum_{n=0}^{+\infty} \frac{1}{n!}x^n$. Plugging in the definition of $c(x)$, we have

$$\begin{aligned}
&\frac{c(x)}{y\left(yc(y)+xc(x)\right)}\left(e^{yc(y)+xc(x)}-1-\left(yc(y)+xc(x)\right)\right)\\
&= \frac{\log(1-x)}{xy\log\left((1-x)(1-y)\right)}\left((1-x)(1-y)-1-\log\left((1-x)(1-y)\right)\right)\\
&= \frac{\log(1-x)}{xy\log\left((1-x)(1-y)\right)}\left(xy-x-y-\log\left((1-x)(1-y)\right)\right).
\end{aligned}$$

For the second part of Equation (42),

$$\begin{aligned}
\frac{1}{y}\sum_{n=2}^{+\infty}\frac{1}{n!}\left(xc(x)\right)^{n-1}c(x) &= \frac{1}{xy}\sum_{n=2}^{+\infty}\frac{1}{n!}\left(xc(x)\right)^{n}\\
&= \frac{1}{xy}\left(e^{xc(x)}-1-xc(x)\right)\\
&= \frac{1}{xy}\left(1-x-1-\log(1-x)\right)\\
&= \frac{1}{xy}\left(-x-\log(1-x)\right).
\end{aligned}$$

Combining the two parts, we get

$$\begin{aligned}
b(x,y) &= \frac{\log(1-x)}{xy\log\left((1-x)(1-y)\right)}\left(xy-x-y-\log\left((1-x)(1-y)\right)\right)+\frac{1}{xy}\left(x+\log(1-x)\right)\\
&= \frac{\log(1-x)}{xy\log\left((1-x)(1-y)\right)}\left(xy-x-y\right)+\frac{1}{y}.
\end{aligned}$$

Now to prove the required result, we consider the generator function

$$\begin{aligned}
&b(x,y)+b(y,x)\\
&= \frac{\log(1-x)}{xy\log\left((1-x)(1-y)\right)}\left(xy-x-y\right)+\frac{1}{y}+\frac{\log(1-y)}{xy\log\left((1-x)(1-y)\right)}\left(xy-x-y\right)+\frac{1}{x}\\
&= \frac{1}{xy}(xy-x-y)+\frac{1}{x}+\frac{1}{y}\\
&= 1.
\end{aligned}$$

Note that the coefficients of $b(x,y)+b(y,x)$ for $x^s y^m$ with $s+m\geq 1$ is $b_{s,m}+b_{m,s}$, which means

$$b_{s,m}+b_{m,s}=0.$$

□

Lemma 38. *Define* $a_{0,0} = 1$ *and for* $s, m \geq 0$ *and* $s + m \geq 1$*, recall the recursive definition in Equation* (39)*,*

$$\begin{aligned} & a_{s,m} \\ = & -\sum_{n=2}^{s+m+1} \frac{1}{n!} \sum_{l=0}^{s} \sum_{r=0}^{m} a_{l,r} \sum_{q=0}^{n-1} \binom{n-1}{q} \rho\left(q, s-q-l, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-1-q, m+1-n+q-r, \{c_k\}_{k=0}^{+\infty}\right), \end{aligned}$$

where $c_k = -\frac{1}{1+k}$*. Then we have*

$$a(x,y) = \sum_{s,m\geq 0}^{+\infty} a_{s,m} x^s y^m = \frac{\log((1-x)(1-y))}{xy - x - y}.$$

Proof. We begin with

$$\begin{aligned} & a(x,y) \\ = & \sum_{s,m\geq 0}^{+\infty} a_{s,m} x^s y^m \\ = & 1 - \sum_{s,m\geq 1}^{+\infty} x^s y^m \sum_{n=2}^{s+m+1} \frac{1}{n!} \sum_{l=0}^{s} \sum_{r=0}^{m} a_{l,r} \sum_{q=0}^{n-1} \binom{n-1}{q} \rho\left(q, s-q-l, \{c_k\}_{k=0}^{+\infty}\right) \\ & \cdot \rho\left(n-1-q, m+1-n+q-r, \{c_k\}_{k=0}^{+\infty}\right) \\ = & 1 - \sum_{n=2}^{+\infty} \frac{1}{n!} \sum_{s+m\geq n-1}^{+\infty} x^s y^m \sum_{l=0}^{s} \sum_{r=0}^{m} a_{l,r} \sum_{q=0}^{n-1} \binom{n-1}{q} \rho\left(q, s-q-l, \{c_k\}_{k=0}^{+\infty}\right) \\ & \cdot \rho\left(n-1-q, m+1-n+q-r, \{c_k\}_{k=0}^{+\infty}\right) \end{aligned}$$

Similar to the proof technique used in Lemma 37, we note that

$$\sum_{s+m\geq n-1}^{+\infty} x^s y^m \sum_{l=0}^{s} \sum_{r=0}^{m} a_{l,r} \sum_{q=0}^{n-1} \binom{n-1}{q} \rho\left(q, s-q-l, \{c_k\}_{k=0}^{+\infty}\right) \rho\left(n-1-q, m+1-n+q-r, \{c_k\}_{k=0}^{+\infty}\right)$$

is equivalent to

$$a(x,y)\left(xc(x) + yc(y)\right)^{n-1},$$

where $c(x) = \sum_{k=0}^{+\infty} c_k x^k = \frac{\log(1-x)}{x}$. To constitute $x^s y^m$, we first select $a_{l,r} x^l y^r$ from $a(x,y)$, and then select q times $xc(x)$ from $\left(xc(x) + yc(y)\right)^{n-1}$.

Then we have

$$\begin{aligned} a(x,y) &= 1 - \sum_{n=2}^{+\infty} \frac{1}{n!} a(x,y)\left(xc(x) + yc(y)\right)^{n-1} \\ &= 1 - \frac{a(x,y)}{xc(x) + yc(y)} \sum_{n=2}^{+\infty} \frac{1}{n!} \left(xc(x) + yc(y)\right)^{n} \end{aligned}$$

$$= 1 - \frac{a(x,y)}{xc(x)+yc(y)} \left(e^{xc(x)+yc(y)} - 1 - (xc(x)+yc(y)) \right).$$

Plugging in the definition of $c(x)$, we obtain

$$\begin{aligned} a(x,y) &= 1 - \frac{a(x,y)}{\log\left((1-x)(1-y)\right)} \left((1-x)(1-y) - 1 - \log\left((1-x)(1-y)\right)\right) \\ &= 1 - \frac{a(x,y)}{\log\left((1-x)(1-y)\right)} \left(xy - x - y - \log\left((1-x)(1-y)\right)\right). \end{aligned}$$

Solving this equation for $a(x,y)$ gives us the desired result. □

G.1.6. PROOF OF THEOREM 13

Next, we need to show that $b(x)$ and $\mathcal{D}(x)$ is convergent. Solving for $b(x)$ is similar to the proof of Rosca et al. [53, Theorem A.2]. Here, we present the derivation of the diffusion term. We have

$$\mathcal{D}(x) = \sum_{p=0}^{+\infty} \eta^p h_p(x) = \sum_{p=0}^{+\infty} \eta^p \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot \left(\nabla^2 f(x)\right)^k \Sigma(x) \left(\nabla^2 f(x)\right)^{p-1-k},$$

where $a_{k,p-1-k}$ is determined in Lemma 12. The above equation implies

$$\begin{aligned} U^\intercal \mathcal{D} U &= U^\intercal \left(\sum_{p=0}^{+\infty} \eta^p \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot \left(\nabla^2 f(x)\right)^k \Sigma(x) \left(\nabla^2 f(x)\right)^{p-1-k} \right) U \\ &= \sum_{p=0}^{+\infty} \eta^p \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot \Lambda^k U^\intercal \Sigma(x) U \Lambda^{p-1-k}. \end{aligned}$$

Since Λ is a diagonal matrix, we have

$$\begin{aligned} \left[U^\intercal \mathcal{D} U\right]_{i,j} &= \sum_{p=0}^{+\infty} \eta^p \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot \lambda_i^k \left[U^\intercal \Sigma(x) U\right]_{i,j} \lambda_j^{p-1-k} \\ &= \eta \sum_{p=0}^{+\infty} \sum_{k=0}^{p-1} a_{k,p-1-k} \cdot (\eta\lambda_i)^k \left[U^\intercal \Sigma(x) U\right]_{i,j} (\eta\lambda_j)^{p-1-k} \\ &= \eta \sum_{k=0}^{+\infty} \sum_{p=k+1}^{+\infty} a_{k,p-1-k} \cdot (\eta\lambda_i)^k \left[U^\intercal \Sigma(x) U\right]_{i,j} (\eta\lambda_j)^{p-1-k} \\ &= \eta \left[U^\intercal \Sigma(x) U\right]_{i,j} \sum_{k=0}^{+\infty} \sum_{q=0}^{+\infty} a_{k,q} \cdot (\eta\lambda_i)^k (\eta\lambda_j)^q, \end{aligned}$$

where in the last equality we let $q = p - 1 - k$. Then according to Lemma 12, we have

$$\left[U^\intercal \mathcal{D} U\right]_{i,j} = \eta \left[U^\intercal \Sigma(x) U\right]_{i,j} a(\eta\lambda_i, \eta\lambda_j),$$

where $a(x,y) = \frac{\log(1-x)(1-y)}{xy-(x+y)}$.

G.1.7. PROOF OF THE WELL-POSEDNESS OF HA−SME

Proof for Lemma 14. In the first condition, the matrix $U^\intercal \Sigma U$ becomes diagonal. We can check that the diagonal elements are always positive definite, and by taking square root of the eigenvalues, we obtain

$$D = U\sqrt{\widetilde{\Lambda}\eta\frac{\log{(1-\eta\Lambda)}^2}{(1-\eta\Lambda)^2-1}}U^\intercal, \tag{43}$$

where $\Sigma = U\widetilde{\Lambda}U^\intercal$ and $\nabla^2 f(x) = U\Lambda U^\intercal$.

For the second condition, we consider small stepsize regime. We will proceed to show that the RHS matrix of Equation (24), denoted as M is positive semi-definite, so that such a square root always exists. Consider matrix K, whose (i,j)-th element is defined as

$$K_{i,j} = \frac{\log(1-\eta\lambda_i)+\log(1-\eta\lambda_j)}{(1-\eta\lambda_i)(1-\eta\lambda_j)-1}.$$

We consider small η, i.e., $\eta < \frac{1-\frac{\sqrt{2}}{2}}{\|\nabla^2 f(x)\|}$, therefore, K is real. The matrix M can then be written as

$$M = \eta U^\intercal\Sigma U \odot K = \eta U^\intercal\Sigma U \odot (\mathbf{1}+K-\mathbf{1}) = \eta\left(U^\intercal\Sigma U + U^\intercal\Sigma U\odot(K-\mathbf{1})\right),$$

where $\mathbf{1}$ is a all-one matrix. Now since $\eta > 0$, to determine the positive semi-definiteness of M, it is sufficient to determine positive semi-definiteness of $U^\intercal\Sigma U + U^\intercal\Sigma U\odot(K-\mathbf{1})$.

Note that for matrix $U^\intercal\Sigma U$, since U is orthogonal, it only changes the eigenspace of Σ, but does not change the eigenvalues. To see why this is the case, assume λ is a eigenvalue of Σ with v being the corresponding eigenvector. Then we have

$$\Sigma v = \lambda v.$$

Then $U^{-1}v$ is a eigenvector for $U^\intercal\Sigma U$ with eigenvalue λ, since

$$U^\intercal\Sigma U U^{-1}v = U^\intercal\Sigma v = U^\intercal\lambda v = \lambda U^{-1}v.$$

Going back to our problem, since $U^\intercal\Sigma U$ is positive definite, a sufficient condition for matrix $U^\intercal\Sigma U + U^\intercal\Sigma U\odot(K-\mathbf{1})$ to be PSD is that

$$\lambda_{\min}\left(U^\intercal\Sigma U\right) = \lambda_{\min}(\Sigma) \geq \|U^\intercal\Sigma U\odot(K-\mathbf{1})\|. \tag{44}$$

Let us look at $\|U^\intercal\Sigma U\odot(K-\mathbf{1})\|$, we know that

$$\begin{aligned}
\|U^\intercal\Sigma U\odot(K-\mathbf{1})\| &\leq \sqrt{d}\|U^\intercal\Sigma U\|\sup_{i,j}\left|[K-\mathbf{1}]_{i,j}\right| \\
&= \sqrt{d}\|U^\intercal\Sigma U\|\sup_{i,j}\left|\frac{\log\left((1-\eta\lambda_i)(1-\eta\lambda_j)\right)}{(1-\eta\lambda_i)(1-\eta\lambda_j)-1}-1\right| \\
&= \sqrt{d}\lambda_{\max}(\Sigma)\sup_{i,j}\left|\frac{\log\left((1-\eta\lambda_i)(1-\eta\lambda_j)\right)}{(1-\eta\lambda_i)(1-\eta\lambda_j)-1}-1\right|,
\end{aligned}$$

where we used Lemma 41 for the inequality.

According to Lemma 39,

$$\sup_{i,j}\left|\frac{\log\left((1-\eta\lambda_i)(1-\eta\lambda_j)\right)}{(1-\eta\lambda_i)(1-\eta\lambda_j)-1}-1\right|\leq \sup_{i,j} 1-(1-\max\{|\eta\lambda_i|,|\eta\lambda_j|\})^2 \leq 1-\left(1-\eta\left\|\nabla^2 f(x)\right\|\right)^2$$

Plugging in back to Equation (44), a sufficient condition for M to be PSD is

$$\lambda_{\min}(\Sigma)\geq\sqrt{d}\lambda_{\max}(\Sigma)\left(1-\left(1-\eta\left\|\nabla^2 f(x)\right\|\right)^2\right).$$

Rearranging the terms gives us the condition in the theorem. □

Proof of Theorem 18. To show the result, it is sufficient to show that the eigenvalues of D are lower bounded away from 0. For the first condition, we have an explicit form of D in Equation (43). Note that when the abstract value of all entries of $\eta\Lambda$ are smaller than 1, entries of $\log(1-\eta\Lambda)^2/\left((1-\eta\Lambda)^2-1\right)$ are lower bounded away from 0. Therefore, as long as eigenvalues of Σ are lower bounded, i.e., Σ is positive definite, the eigenvalues of D are lower bounded from 0. For the second condition, Similar to the proof for Lemma 14, we know that as long as the stepsize satisfies the condition, the eigenvalues of the diffusion coefficient are positive and lower bounded.

Now, we show that why lower boundedness of eigenvalues of D is sufficient to prove the result. We need to show that the drift term b and diffusion term D are Lipschitz. According to Lemma 42, the drift term is Lipschitz, and according to Lemma 43, we know that $DD^\intercal$ is Lipschitz. Therefore, $\partial[DD^\intercal]_{i,j}/\partial x_k$ is upper bounded. Then according to Lin and Maji [34, Equation (7)], we have

$$\operatorname{vec}\left(\frac{\partial D_{i,j}}{\partial DD^\intercal}\right)=(D\otimes I+I\otimes D)^{-1}\operatorname{vec}\left(\tilde{\mathbf{1}}_{i,j}\right),$$

where vec is the vectorization of matrices and $\otimes$ denotes the Kronecker product. $\tilde{\mathbf{1}}_{i,j}$ is a matrix whose (i,j)-th element is 1 and all other elements are 0. Next, we get

$$\begin{aligned}\frac{\partial D_{i,j}}{\partial x_k}&=\sum_{p,q}\frac{\partial D_{i,j}}{\partial[DD^\intercal]_{p,q}}\frac{\partial[DD^\intercal]_{p,q}}{\partial x_k}\\&=\sum_{p,q}\left[(D\otimes I+I\otimes D)^{-1}\operatorname{vec}\left(\tilde{\mathbf{1}}_{i,j}\right)\right]_{qd+p}\frac{\partial[DD^\intercal]_{p,q}}{\partial x_k}\\&\leq\sum_{p,q}\left\|(D\otimes I+I\otimes D)^{-1}\operatorname{vec}\left(\tilde{\mathbf{1}}_{i,j}\right)\right\|\left|\frac{\partial[DD^\intercal]_{p,q}}{\partial x_k}\right|\\&\leq\sum_{p,q}\left\|(D\otimes I+I\otimes D)^{-1}\right\|\left|\frac{\partial[DD^\intercal]_{p,q}}{\partial x_k}\right|.\end{aligned}$$

It remains to show that the eigenvalues of $(D\otimes I+I\otimes D)^{-1}$ is bounded. According to Steeb and Hardy [61, Theorem 2.15], the eigenvalues of $D\otimes I+I\otimes D$ are $\{a+b \mid a,b\in\{\lambda_i\}_{i=1}^d\}$, where λ_i are the eigenvalues of D. Therefore, the eigenvalues of $D\otimes I+I\otimes D$ are lower bounded since we have the eigenvalues of D being lower bounded by a positive constant. This implies that eigenvalues of $(D\otimes I+I\otimes D)^{-1}$ are upper bounded, which concludes the proof. □

G.1.8. HELPER LEMMAS

Lemma 39. *For any* $x, y \in \mathbb{R}$ *and* $\frac{\sqrt{2}}{2} - 1 \le x, y \le 1 - \frac{\sqrt{2}}{2}$*, it holds that*

$$\left| \frac{\log((1-x)(1-y))}{(1-x)(1-y)-1} - 1 \right| \le 1 - (1 - \max\{|x|, |y|\})^2 .$$

Proof. Let $z = (1-x)(1-y)$, then the LHS becomes $\left|\frac{\log z}{z-1} - 1\right|$. We first study the function

$$f(z) = \frac{\log z}{z-1} - 1,$$

whose gradient is given by

$$f'(z) = \frac{\frac{z-1}{z} - \log z}{(z-1)^2} = \frac{1 - \frac{1}{z} - \log z}{(z-1)^2} \le 0,$$

where the last inequality comes from the fact $\log z \ge 1 - \frac{1}{z}$ for $z > 0$. Also note that $f(1) = 0$, therefore, we know that

$$|f(z)| = \begin{cases} \frac{\log z}{z-1} - 1, & z \le 1 \\ 1 - \frac{\log z}{z-1}, & z > 1. \end{cases}$$

According to previous analysis, the maximum of $\left|\frac{\log z}{z-1} - 1\right|$ would be at the minimum or maximum possible value of z. Denoting $m := \max\{|x|, |y|\}$ with $0 \le m \le 1 - \frac{\sqrt{2}}{2}$, then we know that

$$\left| \frac{\log((1-x)(1-y))}{(1-x)(1-y)-1} - 1 \right| \le \max\left\{ \frac{\log(1-m)^2}{(1-m)^2 - 1} - 1, 1 - \frac{\log(1+m)^2}{(1+m)^2 - 1} \right\}$$

Next, we will show that for $0 \le m \le 1$,

$$\frac{\log(1-m)^2}{(1-m)^2 - 1} - 1 \ge 1 - \frac{\log(1+m)^2}{(1+m)^2 - 1}.$$

We start with

$$\begin{aligned} &\frac{\log(1-m)^2}{(1-m)^2 - 1} - 1 - \left(1 - \frac{\log(1+m)^2}{(1+m)^2 - 1}\right) \\ = &\frac{2(m+2)\log(1-m) + 2(m-2)\log(1+m) - 2m(m+2)(m-2)}{m(m+2)(m-2)}. \end{aligned}$$

Since $m(m+2)(m-2) < 0$, it suffice to show that the numerator is less or equal than 0. We let

$$g(m) = 2(m+2)\log(1-m) + 2(m-2)\log(1+m) - 2m(m+2)(m-2),$$

whose gradient is

$$g'(m) = \frac{m+2}{m-1} + \log(1-m) + \frac{m-2}{1+m} + \log(1+m) - 3m^2 + 4$$

$$\leq \frac{m+2}{m-1} + \frac{m-2}{1+m} - 4m^2 + 4$$
$$= \frac{3m^2(m^2-3)}{(1-m)(m+1)} \leq 0,$$

where the inequality holds because $\log z \leq z - 1$ for $z > 0$. Therefore, for $0 \leq m \leq 1$, we have $g(m) \leq g(0) = 0$. So far, we have proved that

$$\left|\frac{\log((1-x)(1-y))}{(1-x)(1-y)-1} - 1\right| \leq \frac{\log(1-m)^2}{(1-m)^2-1} - 1.$$

Next, we will prove that an upper bound is as follows

$$\frac{\log(1-m)^2}{(1-m)^2-1} - 1 \leq 1 - (1-m)^2,$$

for $0 \leq m \leq 1 - \frac{\sqrt{2}}{2}$. With a change of variable for $z = (1-m)^2$, to prove the above inequality is the same as showing the following for $\frac{1}{2} \leq z \leq 1$,

$$\frac{\log z}{z-1} - 1 - (1-z) = \frac{\log z}{z-1} + z - 2 = \frac{\log z + z^2 - 3z + 2}{z-1} \leq 0.$$

Then it is sufficient to show

$$h(z) := \log z + z^2 - 3z + 2 \geq 0.$$

We have

$$h'(z) = \frac{1}{z} + 2z - 3 = \frac{(z-1)(2z-1)}{z} \leq 0.$$

Therefore $h(z) \geq h(1) = 0$ for $\frac{1}{2} \leq z \leq 1$. The whole proof is finished. $\square$

Lemma 40. *Let $u : \mathbb{R}^d \to \mathbb{R}$ and $\{v_i(x)\}_{i=1}^n$ be vector fields, i.e., $v_i : \mathbb{R}^d \to \mathbb{R}^d$. The terms in the result of*

$$\nabla^2 \left(v_1 \cdot \nabla \left(v_2 \cdot \nabla \left(v_3 \cdot \nabla \left(\cdots v_n \cdot \nabla u\right)\right)\right)\right),$$

that contain only $\nabla^2 u$ and ∇v_i for $i \in [1, n]$ are

$$\sum_{S \subseteq [1,n]} \left(\prod_{i \in \widetilde{S}} \nabla v_i\right) \nabla^2 u \left(\prod_{i \in \widehat{[1,n] \setminus S}} \nabla v_i\right),$$

where $\cdot$ is the inner product between vectors, $\widetilde{S}$ is the ascending ordered set containing all elements from S, $\widehat{S}$ is the corresponding descending ordered set, and $\setminus$ is the set difference operator.

Proof. We will prove this by induction. First, let us consider the case of $n = 1$, then clearly, we have

$$\nabla^2 (v_1 \cdot \nabla u) = \nabla v_1 \nabla^2 u + \nabla^2 u \nabla v_1 + \mathcal{C},$$

where $\mathcal{C}$ contains irrelevant terms, i.e., terms with higher order gradients of v_i and u. Now assume the conclusion holds true for $n = 1, \ldots, m-1$, then when $n = m$, we have

$$\begin{aligned}
&\nabla^2 (v_1 \cdot \nabla (v_2 \cdot \nabla (v_3 \cdot \nabla (\cdots v_m \cdot \nabla u)))) \\
&= \nabla v_1 \nabla^2 (v_2 \cdot \nabla (v_3 \cdot \nabla (\cdots v_m \cdot \nabla u))) + \nabla^2 (v_2 \cdot \nabla (v_3 \cdot \nabla (\cdots v_m \cdot \nabla u))) \nabla v_1 + \mathcal{C} \\
&= \nabla v_1 \sum_{S \subseteq [2,m]} \left(\prod_{i \in \widetilde{S}} \nabla v_i \right) \nabla^2 u \left(\prod_{i \in \widehat{[2,m]} \setminus S} \nabla v_i \right) + \sum_{S \subseteq [2,m]} \left(\prod_{i \in \widetilde{S}} \nabla v_i \right) \nabla^2 u \left(\prod_{i \in \widehat{[2,m]} \setminus S} \nabla v_i \right) \nabla v_1 + \mathcal{C} \\
&= \sum_{S \subseteq [1,m]} \left(\prod_{i \in \widetilde{S}} \nabla v_i \right) \nabla^2 u \left(\prod_{i \in \widehat{[1,m]} \setminus S} \nabla v_i \right) + \mathcal{C},
\end{aligned}$$

which concludes the proof. □

Lemma 41. *We have for any $A, B \in \mathbb{R}^{d \times d}$,*

$$\|A \odot B\| \le \sqrt{d} \sup_{i,j} A_{i,j} \|B\|.$$

Proof. Let $e_1, e_2, \ldots, e_d$ be the standard basis of $\mathbb{R}^d$. Then we have for any $1 \le k \le d$,

$$\|(A \odot B) e_k\| \le \sup_{i,j} A_{i,j} \|B e_k\| \le \sup_{i,j} A_{i,j} \|B\|.$$

Then for any vector $v \in \mathbb{R}^d$, which can be written as $v = v_1 e_1 + v_2 e_2 + \cdots$, we have

$$\|(A \odot B) v\| \le \sum_{k=1}^{d} |v_k| \|(A \cdot B) e_k\| \le \sup_{i,j} A_{i,j} \|B\| \sum_{k=1}^{d} |v_k| \le \sup_{i,j} A_{i,j} \|B\| \sqrt{d} \|v\|,$$

where the last inequality we used Cauchy-Schwartz inequality. □

G.2. Approximation Error Analysis

G.2.1. PROOFS FOR SECTION E

Proof for Theorem 20. The proof follows the idea of Theorem 2.2 of Feng et al. [12], however our proof is more difficult since we need to deal with more complicated HA−SME. The assumption of $\|F(x;\xi)\|_{c^7} < \infty$ implies $\|f(x)\|_{c^7} < \infty$ and $\|\Sigma(x)\|_{c^6} < \infty$. According to Lemma 42, we have boundedness of $\|b(x)\|_{C^5}$, where $b(x)$ is the drift term of HA−SME. According to Lemma 42, we have boundedness of the diffusion term of HA−SME, i.e., $\frac{\|U(x)L(x)U(x)^\intercal\|_{C^5}}{\eta}$ is upper bounded. Let $u^n(x)$ defined the same as in Feng et al. [12], i.e., $u^n(x_0) = \mathbb{E}[u(x_n)]$. According to Theorem 2.1 of Feng et al. [12], for small enough η, we have boundedness of $\|u^n\|_{C^6}$.

Starting both from $u^n(x)$, we first measure the error between the discrete-time SGD and continuous-time iterates after one-step. For the discrete-time SGD, by Taylor expansion, and boundedness of $\nabla f(x;\xi)$ and $\|u^n(x)\|_{C^6}$, we have

$$\begin{aligned}
&\left|u^{n+1}(x) - u^n(x) + \eta\langle\nabla f(x), \nabla u^n(x)\rangle - \frac{1}{2}\eta^2\mathbb{E}\left[\nabla f(x;\xi)\nabla f(x;\xi)^\intercal\right] : \nabla^2 u^n(x)\right| \\
=& \left|u^{n+1}(x) - u^n(x) + \eta\langle\nabla f(x), \nabla u^n(x)\rangle - \frac{1}{2}\eta^2\left(\nabla f(x)\nabla f(x)^\intercal + \Sigma(x)\right) : \nabla^2 u^n(x)\right| \\
\leq& \mathcal{O}\left(\eta^3\right).
\end{aligned} \tag{45}$$

In continuous time, after time η, according to Lemma 44, we have

$$\left|e^{\eta\mathcal{L}} - u^n(x) - \eta\mathcal{L}u^n(x) - \frac{\eta^2}{2}\mathcal{L}^2 u^n(x)\right| \leq \mathcal{O}(\eta^3). \tag{46}$$

Note that different from previous SDEs (Equations (4) and (7)), now $\eta\mathcal{L}u^n(x)$ and $\frac{\eta^2}{2}\mathcal{L}^2 u^n(x)$ contains infinite many terms and we need to consider their errors. We know that

$$\eta\mathcal{L}u^n(x) = \eta b(x)\cdot\nabla u^n(x) + \frac{\eta}{2}D(x)D(x)^\intercal : \nabla^2 u^n(x),$$

where $b(x)$ is the drift term of `HA-SME`. By Taylor expansion of $\mathcal{L}u^n$ w.r.t. η and Lemma 45, we have

$$\begin{aligned}
&\Bigg|\eta b(x)\cdot\nabla u^n(x) + \eta\nabla f(x)\cdot\nabla u^n(x) + \frac{\eta^2}{2}\nabla^2 f(x)\nabla f(x)\cdot\nabla u^n(x) \\
&\quad + \frac{\eta}{2}U(x)L(x)U(x)^\intercal : \nabla^2 u^n(x) - \frac{1}{2}\eta^2\Sigma(x) : \nabla^2 u^n(x)\Bigg| \leq \mathcal{O}(\eta^3).
\end{aligned}$$

Then we have

$$\left|\eta\mathcal{L}u^n(x) + \eta\nabla f(x)\cdot\nabla u^n(x) + \frac{\eta^2}{2}\nabla^2 f(x)\nabla f(x)\cdot\nabla u^n(x) - \frac{1}{2}\eta^2\Sigma(x) : \nabla^2 u^n(x)\right| \leq \mathcal{O}\left(\eta^3\right). \tag{47}$$

Next, we look at the errors in $\frac{\eta^2}{2}\mathcal{L}^2 u^n(x)$. By Taylor expansion of $\mathcal{L}^2 u^n$ w.r.t. η and Lemma 45, we have

$$\left|\frac{\eta^2}{2}\mathcal{L}^2 u^n(x) - \frac{\eta^2}{2}\nabla^2 f(x)\nabla f(x)\cdot\nabla u^n(x) - \frac{\eta^2}{2}\nabla f(x)\nabla f(x)^\intercal : \nabla^2 u^n(x)\right| \leq \mathcal{O}\left(\eta^3\right). \tag{48}$$

Combining Equations (46) to (48), we get

$$\left|e^{\eta\mathcal{L}}u^n(x) - u^n(x) + \eta\nabla f(x)\cdot\nabla u^n(x) - \frac{\eta^2}{2}\Sigma(x) : \nabla^2 u^n(x) - \frac{\eta^2}{2}\nabla f(x)\nabla f(x)^\intercal : \nabla^2 u^n(x)\right| \leq \mathcal{O}\left(\eta^3\right).$$

Combining the above equation and Equation (45), we have

$$\left|e^{\eta\mathcal{L}}u^n(x) - u^{n+1}(x)\right| \leq \mathcal{O}\left(\eta^3\right).$$

Denote

$$E^n = \|u^n(x) - u(x, n\eta)\|_{L^\infty}.$$

We obtain

$$\begin{aligned}
E^{n+1} &= \left\|u^{n+1}(x) - u(x,(n+1)\eta) + e^{\eta\mathcal{L}}u^n(x) - e^{\eta\mathcal{L}}u^n(x)\right\|_{L^\infty} \\
&\le \left\|e^{\eta\mathcal{L}}\left(u^n(x) - u(x,n\eta)\right)\right\|_{L^\infty} + \left\|u^{n+1}(x) - e^{\eta\mathcal{L}}u^n(x)\right\|_{L^\infty} \\
&= \left\|e^{\eta\mathcal{L}}\left(u^n(x) - u(x,n\eta)\right)\right\|_{L^\infty} + \mathcal{O}\left(\eta^3\right) \\
&\le \|u^n(x) - u(x,n\eta)\|_{L^\infty} + \mathcal{O}\left(\eta^3\right) \\
&= E^n + \mathcal{O}\left(\eta^3\right),
\end{aligned}$$

where the last inequality comes from the L^∞ contraction of $e^{t\mathcal{L}}$ (see Lemma 2.1 of Feng et al. [12]). Then we know that

$$E^n \le n\mathcal{O}\left(\eta^3\right) = \frac{T}{\eta}\mathcal{O}\left(\eta^3\right) \le \mathcal{O}\left(\eta^2\right).$$

□

Proof of Lemma 21. The lemma follows by Lemmas 42 and 43 □

Proof of Lemma 25. We will show the results for both the convex and strongly-convex (corresponding to our Remark 28) settings. According to Feng et al. [12], for any $n \ge 0$, $u^{n+1}(x)$ can be written as

$$u^{n+1}(x) = \mathbb{E}\left[u^n\left(x - \eta\nabla F(x;\xi)\right)\right].$$

Let us check the first-order derivatives of u^{n+1}. Denoting $y = x - \eta\nabla F(x;\xi)$, we obtain

$$\nabla u^{n+1}(x) = \mathbb{E}\left[\left(I - \eta\nabla^2 F(x;\xi)\right)\nabla u^n(y)\right].$$

1. If $f(\cdot)$ is strongly-convex, we obtain

$$\sup_x \left\|\nabla u^{n+1}(x)\right\| \le (1 - \eta\mu)\sup_x \|\nabla u^n(x)\|.$$

   With small η, we have $\eta\sup_{x,\xi}\left\|\nabla^2 F(x;\xi)\right\| < 1$, and the recursion form is a contraction:

$$\begin{aligned}
\sup_x \|\nabla u^n(x)\| &\le (1 - \mu\eta)^n \sup_x \|\nabla u(x)\| \\
&\le e^{-\mu\eta n}\sup_x \|\nabla u(x)\| \\
&\le \mathcal{O}\left(e^{-\mu\eta n}\|u\|_{C^1}\right).
\end{aligned}$$

2. If $f(\cdot)$ is convex, we have

$$\sup_x \left\|\nabla u^{n+1}(x)\right\| \le \sup_x \|\nabla u^n(x)\| \le \mathcal{O}\left(\|u\|_{C^1}\right).$$

The notation $\mathcal{O}\left(\cdot\right)$ does not depend on n, η or upper bounds for derivatives of F in $\mathcal{O}(\cdot)$ with orders higher than the second order.

Next, we consider the second-order gradients:

$$\nabla^2 u^{n+1}(x) = \mathbb{E}\left[\left(I - \eta \nabla^2 F(x;\xi)\right) \nabla^2 u^n(y) \left(I - \eta \nabla^2 F(x;\xi)\right) - \eta \nabla u^n(y)^\intercal \nabla^3 f(x)\right].$$

Then we also obtain a recursion for the 2-norm of its vector form:

$$\begin{aligned}
&\underbrace{\sup_x \left\|\operatorname{vec}\left(\nabla^2 u^{n+1}(x)\right)\right\|_2}_{a_{n+1}} &(49)\\
&= \sup_x \left\|\nabla^2 u^{n+1}(x)\right\|_F &(50)\\
&\le \sup_x \left\|\left(I - \eta \nabla^2 F(x;\xi)\right) \nabla^2 u^n(y) \left(I - \eta \nabla^2 F(x;\xi)\right)\right\|_F + \eta \sup_x \left\|\nabla u^n(y)^\intercal \nabla^3 f(x)\right\|_F\\
&= \sup_x \left\|\operatorname{vec}\left(\left(I - \eta \nabla^2 F(x;\xi)\right) \nabla^2 u^n(y) \left(I - \eta \nabla^2 F(x;\xi)\right)\right)\right\|_2 + \eta \sup_x \left\|\nabla u^n(y)^\intercal \nabla^3 f(x)\right\|_F\\
&= \sup_x \left\|\left(I - \eta \nabla^2 F(x;\xi)\right)^{\otimes 2} \operatorname{vec}\left(\nabla^2 u^n(y)\right)\right\|_2 + \eta \sup_x \left\|\nabla u^n(y)^\intercal \nabla^3 f(x)\right\|_F\\
&\le \sup_x \left\|\left(I - \eta \nabla^2 F(x;\xi)\right)^{\otimes 2}\right\|_2 \left\|\operatorname{vec}\left(\nabla^2 u^n(y)\right)\right\|_2 + \eta \sup_x \left\|\nabla u^n(y)^\intercal \nabla^3 f(x)\right\|_F\\
&\le \sup_x \left\|I - \eta \nabla^2 F(x;\xi)\right\|_2^2 \left\|\operatorname{vec}\left(\nabla^2 u^n(y)\right)\right\|_2 + \eta \sup_x \left\|\nabla u^n(y)^\intercal \nabla^3 f(x)\right\|_F,
\end{aligned}$$

where $\otimes$ is the Kronecker product, and we used that $\|A \otimes B\|_2 \le \|A\|_2 \|B\|_2$ for matrices A and B.

1. If $f(\cdot)$ is strongly-convex, the recursion becomes

$$\underbrace{\sup_x \left\|\operatorname{vec}\left(\nabla^2 u^{n+1}(x)\right)\right\|_2}_{a_{n+1}} \le \underbrace{(1-\eta\mu)^2}_{c} \underbrace{\sup_x \left\|\operatorname{vec}\left(\nabla^2 u^n(x)\right)\right\|_2}_{a_n} + \underbrace{\eta \sup_x \left\|\nabla u^n(y)^\intercal \nabla^3 f(x)\right\|_F}_{b_n}. \tag{51}$$

According to our assumption and previous results, $b_n \le \mathcal{O}(\eta e^{-\mu\eta n}\|u\|_{C^1})$. Also we choose small η such that $1 - \eta\mu > 0$. It holds for the recursion that

$$a_{n+1} \le c^{n+1} a_0 + \sum_{s=0}^{n} c^{n-s} b_s.$$

In our case, it becomes

$$\begin{aligned}
\sup_x \left\|\operatorname{vec}\left(\nabla^2 u^n(x)\right)\right\|_2 &\le (1-\eta\mu)^{2n} \sup_x \left\|\operatorname{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \sum_{s=0}^{n} (1-\eta\mu)^{2(n-s)} \mathcal{O}(\eta e^{-\mu\eta s}\|u\|_{C^1})\\
&\le (1-\eta\mu)^{n} \sup_x \left\|\operatorname{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \sum_{s=0}^{n} e^{-\mu\eta 2(n-s)} \mathcal{O}(\eta e^{-\mu\eta s}\|u\|_{C^1})\\
&\le e^{-\mu\eta n} \sup_x \left\|\operatorname{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \sum_{s=0}^{n} e^{\mu\eta s} \mathcal{O}(\eta e^{-2\mu\eta n}\|u\|_{C^1})
\end{aligned}$$

$$
\begin{aligned}
&= e^{-\mu\eta n} \sup_x \left\|\mathrm{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \frac{e^{\eta\mu(n+1)}-1}{e^{\eta\mu}-1}\mathcal{O}(\eta e^{-2\mu\eta n}\|u\|_{C^1}) \\
&\le e^{-\mu\eta n} \sup_x \left\|\mathrm{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \frac{e^{\eta\mu(n+1)}}{e^{\eta\mu}-1}\mathcal{O}(\eta e^{-2\mu\eta n}\|u\|_{C^1}) \\
&\le e^{-\mu\eta n} \sup_x \left\|\mathrm{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \frac{e^{\eta\mu n}}{e^{\eta\mu}-1}\mathcal{O}(\eta e^{-2\mu\eta n}\|u\|_{C^1}) \\
&= e^{-\mu\eta n} \sup_x \left\|\mathrm{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \frac{1}{e^{\eta\mu}-1}\mathcal{O}(\eta e^{-\mu\eta n}\|u\|_{C^1}) \\
&\le e^{-\mu\eta n} \sup_x \left\|\mathrm{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \frac{1}{\eta\mu}\mathcal{O}(\eta e^{-\mu\eta n}\|u\|_{C^1}) \\
&= e^{-\mu\eta n} \sup_x \left\|\mathrm{vec}\left(\nabla^2 u(x)\right)\right\|_2 + \mathcal{O}(e^{-\mu\eta n}\|u\|_{C^1}) \\
&\le \mathcal{O}(e^{-\mu\eta n}\|u\|_{C^2}), \qquad (52)
\end{aligned}
$$

where we used that $e^x \ge x+1$.

2. If $f(\cdot)$ is convex, we have

$$
\sup_x \left\|\mathrm{vec}\left(\nabla^2 u^{n+1}(x)\right)\right\|_2 \le \sup_x \left\|\mathrm{vec}\left(\nabla^2 u^{n}(x)\right)\right\|_2 + \mathcal{O}(\eta).
$$

Therefore, we obtain

$$
\sup_x \left\|\mathrm{vec}\left(\nabla^2 u^{n}(x)\right)\right\|_2 \le \mathcal{O}\left(T\|u\|_{C^2}\right).
$$

Next, we proceed with induction.

1. If $f(\cdot)$ is strongly-convex, assume it holds that for any $0 < s \le p$, we have $\sup_x \|\mathrm{vec}\left(\nabla^s u^n(x)\right)\|_F \le \mathcal{O}(e^{-\mu\eta n}\|u\|_{C^s})$. Then, we study the $p+1$-th order gradient of u^{n+1}:

$$
\left\|\mathrm{vec}\left(\nabla^{p+1} u^{n+1}(x)\right)\right\|_2 = \left\|\left(I-\eta\nabla^2 F(x;\xi)\right)^{\otimes p+1}\mathrm{vec}\left(\nabla^{p+1}u^n(y)\right)\right\|_2 + \eta\mathcal{O}(e^{-\mu\eta n}\|u\|_{C^p}).
$$

The first term is the result by applying the ∇ operator $p+1$ times on u. For other terms, at least one ∇ is applied on $\left(I-\eta\nabla^2 F(x;\xi)\right)$, which results in a η and gradients of F higher than the second-order (which is hidden in $\mathcal{O}(\cdot)$). Also, the gradients of u are at most the p-th order, which by induction are already upper bounded. Therefore, similarly to Equation (51), we obtain

$$
\begin{aligned}
&\sup_x \left\|\mathrm{vec}\left(\nabla^{p+1} u^{n+1}(x)\right)\right\|_2 \\
&\le \sup_x \left\|\left(I-\eta\nabla^2 F(x;\xi)\right)^{\otimes p+1}\mathrm{vec}\left(\nabla^{p+1}u^n(y)\right)\right\|_2 + \eta\mathcal{O}(e^{-\mu\eta n}\|u\|_{c^p}) \\
&\le \sup_x \left\|\left(I-\eta\nabla^2 F(x;\xi)\right)\right\|_2^{p+1} \sup_x \left\|\mathrm{vec}\left(\nabla^{p+1}u^n(x)\right)\right\|_2 + \eta\mathcal{O}(e^{-\mu\eta n}\|u\|_{c^p}) \\
&\le (1-\eta\mu)^{p+1} \sup_x \left\|\mathrm{vec}\left(\nabla^{p+1}u^n(x)\right)\right\|_2 + \eta\mathcal{O}(e^{-\mu\eta n}\|u\|_{c^p}),
\end{aligned}
$$

which can be then recursively bounded similarly to Equation (52), i.e.,

$$
\sup_x \left\|\mathrm{vec}\left(\nabla^{p+1} u^{n}(x)\right)\right\|_2
$$

$$
\begin{aligned}
&\leq e^{-\mu\eta n} \sup_x \left\|\operatorname{vec}\left(\nabla^{p+1} u(x)\right)\right\|_2 + \mathcal{O}(e^{-\mu\eta n}\|u\|_{c^p}) \\
&\leq \mathcal{O}(e^{-\mu\eta n}\|u\|_{c^{p+1}}).
\end{aligned}
$$

Now we know that

$$
\sum_{1\leq |J|\leq p} \left|D^J u^n(x)\right| \leq \mathcal{O}\left(e^{-\mu\eta n}\|u\|_{C^p}\right).
$$

It follows

$$
\begin{aligned}
\eta \sum_{k=0}^{\lfloor T/\eta\rfloor-1} \sum_{1\leq |J|\leq p} \left|D^J u^k(x)\right| &\leq \eta \sum_{k=0}^{\lfloor T/\eta\rfloor-1} \mathcal{O}\left(e^{-\mu\eta k}\|u\|_{C^p}\right) \\
&\leq \eta \frac{1-e^{-\eta\mu\lfloor T/\eta\rfloor}}{1-e^{-\mu\eta}} \mathcal{O}\left(\|u\|_{C^p}\right) \\
&\leq \frac{\eta}{1-e^{-\mu\eta}} \mathcal{O}\left(\|u\|_{C^p}\right) \\
&\leq \mathcal{O}\left(\|u\|_{C^p}\right).
\end{aligned}
$$

2. If $f(\cdot)$ is convex, we assume for any $0 < s \leq p$, $\sup_x \|\operatorname{vec}(\nabla^s u^n(x))\|_F \leq \mathcal{O}(T^{s-1}\|u\|_{C^s})$. Similarly to the strongly-convex case, we obtain

$$
\left\|\operatorname{vec}\left(\nabla^{p+1} u^{n+1}(x)\right)\right\|_2 = \left\|\left(I - \eta\nabla^2 F(x;\xi)\right)^{\otimes p+1} \operatorname{vec}\left(\nabla^{p+1} u^n(y)\right)\right\|_2 + \eta\mathcal{O}(T^{p-1}\|u\|_{C^p}),
$$

which implies

$$
\sup_x \left\|\operatorname{vec}\left(\nabla^{p+1} u^n(x)\right)\right\|_2 \leq \mathcal{O}(T^p\|u\|_{C^{p+1}}).
$$

Further it holds that

$$
\eta \sum_{k=0}^{\lfloor T/\eta\rfloor-1} \sum_{1\leq |J|\leq p} \left|D^J u^k(x)\right| \leq \mathcal{O}\left(T^{p+1}\|u\|_{C^p}\right).
$$

□

Proof for Theorem 23. The assumption of $\|F(x;\xi)\|_{c^8} < \infty$ implies $\|f(x)\|_{c^8} < \infty$. According to Lemma 42, we have boundedness of $\|b(x)\|_{C^6}$, where $b(x)$ is the drift term of HA−SME. According to Lemma 42, we have boundedness of the diffusion term of HA−SME, i.e., $\frac{\|U(x)L(x)U(x)^\intercal\|_{C^6}}{\eta}$ is upper bounded. Denote $M_p^n := \sum_{1\leq|J|\leq p} \left|D^J u^n(x)\right|$.

Similar to the proof for Theorem 20, we first have

$$
\left\|u^{n+1} - u^n + \eta\nabla f\cdot\nabla u - \frac{1}{2}\eta^2\Sigma : \nabla^2 u^n\right\|_{L^\infty} \leq \mathcal{O}\left(\eta^3 s^3 M_3^n\right).
$$

First, we expand $e^{\eta\mathcal{L}}u^n$ using Taylor's expansion and Lemma 44,

$$
\left\|e^{\eta\mathcal{L}}u^n - u^n - \eta\mathcal{L}u^n - \frac{1}{2}\eta^2\mathcal{L}^2 u^n - \frac{1}{3!}\eta^3\mathcal{L}^3 u^n\right\|_{L^\infty} \leq \mathcal{O}\left(\eta^4\lambda^3 s^4 M_8^n\right).
$$

Then using Lemma 45, we find the expansion of $\mathcal{L}u^n$, $\mathcal{L}^2u^n$ and $\mathcal{L}^3u^n$, respectively.

$$\begin{aligned}\mathcal{L}u^n &= -\nabla f \cdot \nabla u^n + \eta\left(-\frac{1}{2}\nabla^2 f\nabla f \cdot \nabla u^n + \frac{1}{2}\Sigma : \nabla^2 u^n\right) \\ &\quad + \frac{\eta^2}{2}\left(-\frac{2}{3}\left(\nabla^2 f\right)^2 \nabla f \cdot \nabla u^n + \frac{1}{2}\left(\Sigma\nabla^2 f + \nabla^2 f\Sigma\right) : \nabla^2 u^n\right) + \mathcal{O}\left(\eta^3\lambda^3 s M_2^n\right). \quad (53)\end{aligned}$$

For $\mathcal{L}^2u^n$, we have

$$\begin{aligned}\mathcal{L}^2u^n &= \nabla f(x)^\intercal\nabla^2 f\nabla u^n + \nabla f^\intercal\nabla^2 u^n\nabla f \\ &\quad + \eta\left(\nabla f^\intercal\left(\nabla^2 f\right)^2\nabla u^n + \nabla f^\intercal\nabla^2 f\nabla^2 u^n\nabla f - \frac{1}{2}\Sigma : \left(\nabla^2 u^n\nabla^2 f + \nabla^2 f\nabla^2 u^n\right) + \mathcal{O}(M_3^n)\right) \\ &\quad + \mathcal{O}\left(\eta^2\lambda^3 s^2 M_4^n\right).\end{aligned}$$

For $\mathcal{L}^3u^n$, we have

$$\mathcal{L}^3u^n = -\nabla f^\intercal\left(\nabla^2 f\right)^2\nabla u^n - 3\nabla f^\intercal\nabla^2 f\left(\nabla u^n\right)^2 + \mathcal{O}\left(M_3^n\right) + \mathcal{O}\left(\eta\lambda^3 s^3 M_6^n\right).$$

Summarizing the above, we obtain

$$\left\|e^{\eta\mathcal{L}}u^n - u^{n+1}\right\|_{L^\infty} \le \mathcal{O}\left(\eta^3 s^3 M_3^n + \eta^4\lambda^3 s^4 M_8^n\right).$$

Following the last few steps in the proof of Theorem 20, we obtain

$$E^n \le \sum_{k=0}^{n-1}\mathcal{O}\left(\eta^3 s^3 M_3^k + \eta^4\lambda^3 s^4 M_8^k\right)$$

finish the proof.

□

Proof for Theorem 24. The proof follows similarly to the proof of Theorem 23. For SME-2, in the expansion of $\mathcal{L}u^n$, i.e., Equation (53), the term

$$\frac{\eta^2}{2}\left(-\frac{2}{3}\left(\nabla^2 f\right)^2\nabla f \cdot \nabla u^n + \frac{1}{2}\left(\Sigma\nabla^2 f + \nabla^2 f\Sigma\right) : \nabla^2 u^n\right)$$

is missing. Therefore, after cancellation of terms, there will be error terms of $\lambda^2 s\|\nabla u^n\|$ and $\lambda\left\|\nabla^2 u^n\right\|$. For SPF, in Equation (53), the term

$$\frac{\eta^2}{2}\left(\frac{1}{2}\left(\Sigma\nabla^2 f + \nabla^2 f\Sigma\right) : \nabla^2 u^n\right)$$

is missing. Therefore, it results in additional error of $\lambda\left\|\nabla^2 u^n\right\|$. □

Proof for Theorem 26. The proof is simply to combine the results from Theorem 23 and Lemma 25.
□

G.2.2. HELPER LEMMAS

Lemma 42. *For $n \geq 0$, assume $f \in C_b^{n+2}(\mathbb{R}^d)$ and denote $\lambda := \sup_{x\in\mathbb{R}^d} \left\|\nabla^2 f(x)\right\|$ and $s := \sup_{x\in\mathbb{R}^d} \|\nabla f(x)\|$. There exist constants η_0 and $C > 0$, both independent of λ and s, such that for any η satisfying $\eta < \eta_0$ and $\eta\lambda < C$, the drift term of* `HA-SME`*, defined in in Equation (23), satisfies*

$$\max_{0\leq i\leq d} |[b(x)]_i| < \mathcal{O}\left(s\right) \text{ and } \max_{0\leq i\leq d} \|[b(x)]_i\|_{C^n} < \mathcal{O}\left(s+\lambda\right),$$

where $[b(x)]_i$ is the i-th entry of $b(x)$, and $\mathcal{O}(\cdot)$ hides the dependence on the upper bounds for the derivatives of f higher than the second-order derivatives.

Proof. Recall the definition of the drifting term of `HA-SME` in Equation (23)

$$b(x) = U(x)\frac{\log\left(I - \eta\Lambda(x)\right)}{\eta\Lambda(x)}U(x)^\intercal \nabla f(x).$$

We first note that $b(x)$ can be written as

$$b(x) = -\sum_{p=0}^{\infty} \frac{1}{p+1}\eta^p \left(\nabla^2 f(x)\right)^p \nabla f(x).$$

The (i,j)-th element of A^p for matrix $A \in \mathbb{R}^{d\times d}$ can be represented as

$$[A^p]_{i,j} = \sum_{s_1=1}^{d}[A]_{i,s_1}\sum_{s_2=1}^{d}[A]_{s_1,s_2}\sum_{s_3=1}^{d}[A]_{s_2,s_3}\cdots\sum_{s_{p-1}}[A]_{s_{p-2},s_{p-1}}[A]_{s_{p-1},j}.$$

Therefore, for any i, the i-th element of $b(x)$ can be written as

$$\begin{aligned}
[b(x)]_i &= -\sum_{p=0}^{\infty} \frac{1}{p+1}\eta^p \left[\left(\nabla^2 f(x)\right)^p \nabla f(x)\right]_i \\
&= -\sum_{p=0}^{\infty} \frac{1}{p+1}\eta^p \sum_{s_1=1}^{d}\left[\nabla^2 f(x)\right]_{i,s_1}\sum_{s_2=1}^{d}\left[\nabla^2 f(x)\right]_{s_1,s_2}\cdots\sum_{s_p=1}^{d}\left[\nabla^2 f(x)\right]_{s_{p-1},s_p}[\nabla f(x)]_{s_p} \\
&= -\sum_{p=0}^{\infty} \frac{1}{p+1}\eta^p \underbrace{\sum_{s_1=1}^{d}\partial_i\partial_{s_1}f(x)\sum_{s_2=1}^{d}\partial_{s_1}\partial_{s_2}f(x)\cdots\sum_{s_p=1}^{d}\partial_{s_{p-1}}\partial_{s_p}f(x)\partial_{s_p}f(x)}_{\text{(A)}}.
\end{aligned}$$

Notice that Term (A) can be expanded as a summation of d^p terms, each of which contains $p+1$ factors. Note that the first-order and second-order partial derivatives are upper bounded by s and λ respectively, i.e., $|D^\alpha f| \leq s$ for $|a| = 1$ and $|D^\alpha f| \leq \lambda$ for $|a| = 2$. Therefore, we have

$$\begin{aligned}
|[b(x)]_i| &\leq \sum_{p=0}^{\infty}\frac{1}{p+1}\eta^p d^p \lambda^p s \\
&= s\sum_{p=0}^{\infty}\frac{1}{p+1}\left(\eta d\lambda\right)^p.
\end{aligned}$$

We get a power series in the above equation, and according to Cauchy–Hadamard theorem [7], the convergence radius for $g(x) = \sum_{p=0}^{\infty} \frac{1}{p+1} x^p$ is

$$\frac{1}{\limsup_{p\to\infty} \left|\frac{1}{p+1}\right|^{\frac{1}{p}}} = \frac{1}{1} = 1.$$

Then we know that $|[b(x)]_i|$ is upper bounded by $\mathcal{O}(s)$ as long as

$$\eta d\lambda < 1 \quad \Longrightarrow \quad \eta\lambda < \frac{1}{d}.$$

Next, we consider the first-order derivative of $b(x)$. We denote the upper bound for higher order partial derivatives as B, i.e., $\max_{2<|a|\le n+2} |D^\alpha f| \le B$. Recall that in $b(x)$, Term (A) contains d^p terms and each term has p factors of second-order derivatives of $f(x)$ and one factor of the first-order derivative of $f(x)$. If we take gradient of Term (A) w.r.t. x, each term, according to the product rule of gradients and by taking derivatives w.r.t. each factor, will result in: (1) p terms, each consisting of one factors of $\partial^3 f$, $p-1$ factor of $\partial^2 f$, and one factor of ∂f; (2) 1 term consisting of $p+1$ factors of $\partial^2 f$. Therefore, we can bound

$$\begin{aligned}
|\partial_j [b(x)]_i| &\le \sum_{p=0}^{\infty} \frac{1}{p+1} \eta^p d^p \left(pBs\lambda^{p-1} + \lambda^{p+1}\right) \\
&= s \sum_{p=0}^{\infty} \frac{p}{p+1} d^p \eta B (\eta\lambda)^{p-1} + \lambda \sum_{p=0}^{\infty} \frac{1}{p+1} (d\eta\lambda)^p
\end{aligned}$$

For the second summation, the convergence radius of the power series is the same as the upper bound for $[b(x)]_i$, and it is bounded by $\mathcal{O}(\lambda)$ for sufficiently small η and $\eta\lambda$. For the first summation, comparing to the upper bound of $[b(x)]_i$, the coefficients of the power series are multiplied by p. However, this will not change the convergence radius. It is known that for sequences a_p and b_p, if $\limsup_{p\to\infty} a_p = A$ and $\lim_{p\to\infty} b_p = B$, then $\limsup_{p\to\infty} (a_p b_p) = AB$. Note that

$$\lim_{p\to\infty} |p|^{\frac{1}{p}} = 1.$$

Therefore, by applying Cauchy–Hadamard theorem, multiplying the coefficients with p would not change the radius of convergence. Therefore, if we have $\eta\lambda$ and ηB smaller than the convergence radius of $\sum_p \frac{1}{1+p} x^p$, that is 1, the first summation in the upper bound of $|\partial_j [b(x)]_i|$ can be upper bounded by some constants. As a result, we can conclude that

$$|\partial_j [b(x)]_i| \le \mathcal{O}(s+\lambda).$$

Following the same idea, we identify that after applying twice partial derivatives on $b(x)$, i.e., $\partial_j \partial_k [b(x)]_i$, each term in Term (A) becomes:

$$p\partial^4 f \left(\partial^2 f\right)^{p-1} \partial f + p(p-1) \left(\partial^3 f\right)^2 \left(\partial^2 f\right)^{p-2} \partial f + (2p+1)\partial^3 f \left(\partial^2 f\right)^p$$

Thus, we have

$$
\begin{aligned}
|\partial_j\partial_k[b(x)]_i| &\le \sum_{p=0}^{\infty}\frac{1}{p+1}\eta^p d^p\left(p(p-1)B^2 s\lambda^{p-2}+pBs\lambda^{p-1}+(2p+1)B\lambda^p\right)\\
&= s\sum_{p=0}^{\infty}\frac{p(p-1)}{p+1}d^p(B\eta)^2(\lambda\eta)^{p-2}+s\sum_{p=0}^{\infty}\frac{p}{p+1}d^pB\eta(\lambda\eta)^{p-1}+B\sum_{p=0}^{\infty}\frac{2p+1}{p+1}d^p(\lambda\eta)^p.
\end{aligned}
$$

By the same reason as what we did for the $\partial_j[b(x)]_i$, the three summations converge, and are upper bounded by some constants not relying on λ.

For derivatives higher than the second ones, we can follow the same logic to show that as long as ηB and $\eta\lambda$ is small enough, the derivatives are upper bounded and such bound does not rely on λ. Specifically, for the k-th order derivative of $b(x)$ with $k \ge 2$, we will have terms of the following form derived from each term in Term (A):

$$
\mathcal{O}\left(p^k\right)\left(\prod_{j=1}^{m}\partial^{q_j}f\right)\left(\partial^2 f\right)^s\left(\partial f\right)^r,
$$

with $m+s+r=p+1$ and $q_j \ge 3$, $s \le p$ and $r \le 1$. As we only consider derivatives of $b(x)$ up to some finite order. We have finite such terms and they can be bounded using the idea presented previously.

□

Lemma 43. *For $n \ge 0$, assume $f \in C_b^{n+2}\left(\mathbb{R}^d\right)$ and $[\Sigma]_{i,j} \in C_b^n\left(\mathbb{R}^d\right)$ for all $1 \le i,j \le d$ where $[A]_{i,j}$ is the (i,j)-th entry of matrix A, then there exists $\eta_0, C > 0$, such that for any $\eta < \eta_0$ and $\eta\lambda < C$, where $\lambda := \sup_{x\in\mathbb{R}^d}\left\|\nabla^2 f(x)\right\|$ and C, η_0 does not depend on λ, we have that the diffusion term $D(x)$ of stochastic principle flow satisfy*

$$
\max_{0\le i,j\le d}\frac{\left\|\left[D(x)D(x)^\intercal\right]_{i,j}\right\|_{C^n}}{\eta} < \mathcal{O}\left(1\right),
$$

where $\mathcal{O}(\cdot)$ hides the dependence on the upper bounds for the derivatives of f higher than the second-order derivatives and $\max_{0\le i,j\le d}\|[\Sigma]_{i,j}\|_{C^n}$.

Proof. The proof idea is similar to the proof of Lemma 42, however, in this case, we do not have an explicit form for the coefficients of the power series. The diffusion term can be represented as

$$
D(x)D(x)^\intercal = \sum_{p=1}^{\infty}\eta^p\sum_{k=0}^{p-1}a_{k,p-1-k}\cdot\left(\nabla^2 f(x)\right)^k\Sigma(x)\left(\nabla^2 f(x)\right)^{p-1-k},
$$

where $a_{k,p-1-k}$ are absolute constants such that by Taylor expansion at $(0,0)$,

$$
g(x,y) := \frac{\log(1-x)(1-y)}{xy-(x+y)} = \sum_{s,m\ge 0}^{+\infty}a_{s,m}x^s y^m.
$$

For the power series above, we know that its convergence radius is greater or equal than 1 for both x and y, meaning that $x < 1$ and $y < 1$ is a sufficient condition for it to converge. Also, according to the Cauchy–Hadamard theorem for multiple variables [58], the signs of the coefficients do affect the convergence radius, therefore the convergence radius of

$$\widetilde{g}(x, y) \coloneqq \sum_{s,m \geq 0}^{+\infty} |a_{s,m}| x^s y^m$$

is also greater or equal than 1. Now we look back into $D(x)D(x)^\intercal$. We note that the (i, j)-th element of $A^p B A^k$ for $A, B \in \mathbb{R}^{d \times d}$ can be written as

$$\left[A^p B A^k\right]_{i,j} = \sum_{s_1=1}^{d} [A]_{i,s_1} \sum_{s_2=1}^{d} [A]_{s_1,s_2} \cdots \sum_{s_p=1}^{d} [A]_{s_{p-1},s_p} \sum_{m=1}^{d} [S]_{s_p,m} \sum_{q_1=1}^{d} [A]_{m,q_1} \sum_{q_2=1}^{d} [A]_{q_1,q_2} \cdots \\ \sum_{q_{k-1}=1}^{d} [A]_{q_{k-2},q_{k-1}} [A]_{q_{k-1},j}\,.$$

Therefore, we can derive that

$$\begin{aligned}
&\left[D(x)D(x)^\intercal\right]_{i,j} \\
&= \sum_{p=1}^{\infty} \eta^p \sum_{k=0}^{p-1} a_{k,p-1-k} \sum_{s_1=1}^{d} \left[\nabla^2 f(x)\right]_{i,s_1} \sum_{s_2=1}^{d} \left[\nabla^2 f(x)\right]_{s_1,s_2} \cdots \sum_{s_k=1}^{d} \left[\nabla^2 f(x)\right]_{s_{k-1},s_k} \sum_{m=1}^{d} [\Sigma(x)]_{s_k,m} \\
&\quad \sum_{q_1=1}^{d} \left[\nabla^2 f(x)\right]_{m,q_1} \sum_{q_2=1}^{d} \left[\nabla^2 f(x)\right]_{q_1,q_2} \cdots \sum_{q_{p-k-2}=1}^{d} \left[\nabla^2 f(x)\right]_{q_{p-k-3},q_{p-k-2}} \left[\nabla^2 f(x)\right]_{q_{p-k-2},j} \\
&= \sum_{p=1}^{\infty} \eta^p \sum_{k=0}^{p-1} a_{k,p-1-k} \sum_{s_1=1}^{d} \partial_i \partial_{s_1} f(x) \sum_{s_2=1}^{d} \partial_{s_1} \partial_{s_2} f(x) \cdots \sum_{s_k=1}^{d} \partial_{s_{k-1}} \partial_{s_k} f(x) \sum_{m=1}^{d} [\Sigma(x)]_{s_k,m} \\
&\quad \sum_{q_1=1}^{d} \partial_m \partial_{q_1} f(x) \sum_{q_2=1}^{d} \partial_{q_1} \partial_{q_2} f(x) \cdots \sum_{q_{p-k-2}=1}^{d} \partial_{q_{p-k-3}} \partial_{q_{p-k-2}} f(x) \partial_{q_{p-k-2}} \partial_j f(x).
\end{aligned}$$

Let us look at the term associated with $a_{k,p-1-k}$, i.e., everything after $a_{k,p-1-k}$ in the above equation. It contains $p - 1$ summations, thus resulting in d^{p-1} terms. Each term has p factors consisting of one element of $\Sigma(x)$ and others being second-order derivatives of $f(x)$. Denoting $S \coloneqq \max_{i,j} \|[\Sigma(x)]_{i,j}\|_{C^6}$, we can derive

$$\begin{aligned}
\left|\left[D(x)D(x)^\intercal\right]_{i,j}\right| &\leq \sum_{p=1}^{\infty} \eta^p \sum_{k=0}^{p-1} |a_{k,p-1-k}| d^{p-1} S \lambda^{p-1} \\
&= \eta S \sum_{p=1}^{\infty} \sum_{k=0}^{p-1} |a_{k,p-1-k}| \left(\eta d \lambda\right)^{p-1} \\
&= \eta S \sum_{p=1}^{\infty} \sum_{k=0}^{p-1} |a_{k,p-1-k}| \left(\eta d \lambda\right)^{k} \left(\eta d \lambda\right)^{p-1-k}
\end{aligned}$$

$$= \eta S \sum_{s,m \geq 0}^{\infty} |a_{s,m}| \left(\eta d \lambda\right)^{s} \left(\eta d \lambda\right)^{m}.$$

According to our previous reasoning, the power series is convergent if

$$\eta d \lambda < 1 \quad \Longrightarrow \quad \eta \lambda < \frac{1}{d},$$

and we have

$$\max_{i,j} \frac{\left\| \left[D(x)D(x)^{\intercal}\right]_{i,j} \right\|_{C^0}}{\eta} < \infty.$$

Next, we consider the first-order derivative of $D(x)D(x)^{\intercal}$. Similar to the reasoning in the proof of Lemma 42, after taking derivatives there are pd^{p-1} terms associated with coefficients $a_{k,p-1-k}$, and each term has p factors consisting of one factor being derivatives of $\Sigma(x)$ or $\Sigma(x)$ and others being up to third-order derivatives of $f(x)$. Specifically, we have the following terms associated with $a_{k,p-1-k}$:

$$(p-1)\partial^3 f \left(\partial^2 f\right)^{p-2} \Sigma + \left(\partial^2 f\right)^{p-1} \partial \Sigma.$$

Let B be the constant such that $\max_{2<|a|\leq n+2 \text{and} |a| \neq 2} |D^{\alpha} f| \leq B$, and let $m \coloneqq \max\left\{B\eta, \lambda\eta\right\}$. We have

$$\left| \partial_k \left[D(x)D(x)^{\intercal}\right]_{i,j} \right| \leq \eta S \sum_{p=1}^{\infty} \sum_{k=0}^{p-1} |a_{k,p-1-k}|(p-1)d^{p-1} B\eta \left(\lambda\eta\right)^{p-2} + \eta S \sum_{p=1}^{\infty} \sum_{k=0}^{p-1} |a_{k,p-1-k}| \left(d\lambda\eta\right)^{p-1}.$$

The second summation can be bounded the same way as before. For the first summation, it is upper bounded by

$$\begin{aligned} &\eta S \sum_{p=1}^{\infty} \sum_{k=0}^{p-1} |a_{k,p-1-k}|(p-1) \left(dm\right)^{k} \left(dm\right)^{p-1-k} \\ &= \eta S \sum_{s,m \geq 0}^{\infty} (s+m)|a_{s,m}| \left(dm\right)^{s} \left(dm\right)^{m}. \end{aligned}$$

Now we have a power series with coefficients multiplied by $(s+m)$ compared to the previous one. However this would not change the convergence radius. Denote the convergence radius of $\widetilde{g}(x,y)$ as r_1 and r_2. According to Cauchy–Hadamard theorem for multiple variables, we have

$$\limsup_{s+m \to \infty} |a_{s,m} r_1^s r_2^m|^{\frac{1}{s+m}} = 1.$$

We note that r_1 and r_2 are also convergence radius for our new power series, since

$$\limsup_{s+m \to \infty} |(s+m) a_{s,m} r_1^s r_2^m|^{\frac{1}{s+m}} = \limsup_{s+m \to \infty} |a_{s,m} r_1^s r_2^m|^{\frac{1}{s+m}} \underbrace{\lim_{s+m \to \infty} |s+m|^{\frac{1}{s+m}}}_{=1}$$

$$= \limsup_{s+m\to\infty} |a_{s,m} r_1^s r_2^m|^{\frac{1}{s+m}} = 1.$$

Therefore a sufficient condition for the convergence of the power series is

$$\max\{\eta\lambda, \eta B\} < \frac{1}{d}.$$

The proof is similar for higher-order derivatives. Every time we take derivative, the coefficients would be multiplied by a factor of order $s+m$, however, that would not change the radius of convergence. Specifically, for the k-th derivative of $D(x)D(x)^\intercal$ with $k \geq 2$, terms associated with $a_{k,p-1-k}$ have the following form:

$$\mathcal{O}\left(p^k\right)\left(\prod_{j=1}^{m} \partial^{q_j} f\right)\left(\partial^2 f\right)^s \partial^r \Sigma,$$

with $m+s = p-1$, $q_j \geq 3$, $s \leq p-1$ and $r \leq k$. These terms can be bounded using similar idea of the previous proof.

□

Lemma 44. *Let $\mathcal{L}$ be the infinitesimal generator for an SDE defined as*

$$dX_t = b(X_t)dt + D(X_t)dW_t,$$

where $b : \mathbb{R}^d \to \mathbb{R}^d$ and $D : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. We have for any $n \geq 1$ and $t > 0$, it holds

$$\left\| e^{t\mathcal{L}} u - \sum_{p=0}^{n-1} \frac{t^p}{p!} \mathcal{L}^p u \right\|_{L^\infty} \leq \mathcal{O}\left(t^n \max_i \|b\|_{C^0} \|b\|_{C^{2(n-1)}}^{n-1} \left\|DD^\intercal\right\|_{C^{2(n-1)}}^n \|u\|_{C^{2n}} \right),$$

where $\|b\|_{C^m} := \max_{0\leq i\leq d} \|[b]_i\|_{C^m}$ and $\left\|DD^\intercal\right\|_{C^m} := \max_{0\leq i\leq d} \left\| \left[DD^\intercal\right]_{i,j} \right\|_{C^m}$.

Proof. According to the Taylor expansion of $e^{t\mathcal{L}}$ on t, we obtain for any x

$$\left| e^{t\mathcal{L}} u(x) - \sum_{p=0}^{n-1} \frac{t^p}{p!} \mathcal{L}^p u(x) \right| = \left| \frac{t^n}{n!} \left.\frac{\partial^n e^{t\mathcal{L}} u(x)}{(\partial t)^n}\right|_{t=s} \right|,$$

for some $0 \leq s \leq t$. According to Feng et al. [12, Lemma 2.1], $e^{t\mathcal{L}}$ is a contraction, therefore, we have

$$\begin{aligned} \left\| e^{t\mathcal{L}} u(x) - \sum_{p=0}^{n-1} \frac{t^p}{p!} \mathcal{L}^p u(x) \right\|_{L^\infty} &= \left\| \frac{t^n}{n!} \left.\frac{\partial^n e^{t\mathcal{L}} u(x)}{(\partial t)^n}\right|_{t=s} \right\|_{L^\infty} \\ &\leq \frac{t^n}{n!} \|\mathcal{L}^n u\|_{L^\infty}. \end{aligned}$$

Note that $\mathcal{L}$ is linear operator and it will apply ∇^2 operator Therefore, we obtain gradients of b and $DD^\intercal$ up to the $2(n-1)$-th order and gradients of u up to the $2n$-th order. Note that we bound the first b using $\|b\|_{C^0}$ and the following $n-1$ occurrences of b using $\|b\|_{C^{2(n-1)}}$, since they have different dependence on λ for `HA-SME` according to Lemma 42. It would be more convenient for fine-grained analysis later. □

Lemma 45. *For $n \geq 0$, assume $f \in C_b^{2n}(\mathbb{R}^d)$ and $[\Sigma(x)]_{i,j} \in C_b^{2n-2}\left(\mathbb{R}^d\right)$ for any $0 \leq i,j \leq d$. Let $\mathcal{L}$ be the infinitesimal generator for `HA-SME`, which depends on η. We have for any function $u : \mathbb{R}^d \to \mathbb{R}$, and integer $n, p > 0$,*

$$\left\| \frac{\partial^p \mathcal{L}^n u}{(\partial \eta)^p} \right\|_{L^\infty} \leq \mathcal{O}\left(\lambda^{n+p-1} s^n \|u\|_{C^{2n}}\right),$$

where $\lambda := \sup_{x\in\mathbb{R}^d} \left\|\nabla^2 f(x)\right\|$, $s := \sup_{x\in\mathbb{R}^d} \|\nabla f(x)\|$ and $\mathcal{O}(\cdot)$ hides the dependence on the upper bounds for the derivatives of f other than the second-order derivatives and $\max_{0\leq i,j\leq d} \|[\Sigma]_{i,j}\|_{C^n}$.

Proof. We note that when taking partial gradient of b w.r.t. η, the dependence on λ on the upper bounds for the derivatives, i.e., $\left\|\left[\frac{\partial b}{\partial \eta}\right]_i\right\|_{C^m}$, would increase its order by 1. That is to say, if $\left\|\left[\frac{\partial^k b}{(\partial\eta)^k}\right]_i\right\|_{C^m} \leq \mathcal{O}(\lambda^q)$, then $\left\|\left[\frac{\partial^{k+1} b}{(\partial\eta)^{k+1}}\right]_i\right\|_{C^m} \leq \mathcal{O}(\lambda^{q+1})$. The reason is that gradients w.r.t. x of b can be written as power series of η and $\eta\lambda$ (see the proof of Lemma 42). Whenever we taking partial derivative w.r.t. η, the order of η compared to λ in each term is decreased by 1. Then we can take a λ factor out and the remaining series still converges (the same reasoning as in the proof of Lemma 42). The same logic applies to $DD^\intercal$. In $\mathcal{L}^n u$, we will have at most $n-1$ occurrences of gradients of b. Therefore, before taking derivatives w.r.t. η, the upper bounds would be at most λ^{n-1}. After taking the derivative p times, the order of λ is at most $n+p-1$. Similar logic applies to $DD^\intercal$ and corresponding Lemma 43. □

G.3. Exact Match of SGD on Quadratics

Before diving into the proof, we will introduce some basics about complex normal distribution and complex linear SDEs.

G.3.1. Complex Normal Distribution

Since our proposed SDE may operate in complex space, similar to PF, we provide here some basics about the complex distribution, specifically the complex normal distribution. A complex random vector is defined by two real random variables:

$$z = x + iy,$$

where x and y are two real random vectors, which may be correlated. The random vector z is called a complex normal vector if $\begin{bmatrix} x \\ y \end{bmatrix}$ is a normal vector. In contrast to real-valued random vector, there are three parameters that define a complex normal variable $\mathcal{CN}(\mu, \Gamma, C)$,

$$\mu := \mathbb{E}[z] \qquad \text{(Expectation)}$$
$$\Gamma := \mathbb{E}\left[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^H\right] \qquad \text{(Covariance)}$$
$$C := \mathbb{E}\left[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^\intercal\right] \qquad \text{(Pseudo-Covariance)}$$

Similar to the real case, for $z \sim \mathcal{CN}(\mu, \Gamma, C)$, the following property holds

$$Az + b \sim \mathcal{CN}\left(A\mu + b, A\Gamma A^H, ACA^\intercal\right).$$

If we look at the real and imaginary part separately, we have

$$\begin{aligned} \Gamma &= \mathrm{Cov}\left[x, x\right] + \mathrm{Cov}\left[y, y\right] + i\left(\mathrm{Cov}\left[y, x\right] - \mathrm{Cov}\left[x, y\right]\right), \\ C &= \mathrm{Cov}\left[x, x\right] - \mathrm{Cov}\left[y, y\right] + i\left(\mathrm{Cov}\left[y, x\right] + \mathrm{Cov}\left[x, y\right]\right). \end{aligned} \tag{54}$$

For the covariance of x and y, we have

$$\begin{aligned} \mathrm{Cov}\left[x, x\right] &= \frac{1}{2}\mathrm{Re}\left(\Gamma + C\right), \\ \mathrm{Cov}\left[y, y\right] &= \frac{1}{2}\mathrm{Re}\left(\Gamma - C\right). \end{aligned} \tag{55}$$

G.3.2. COMPLEX LINEAR STOCHASITC DIFFERENTIAL EQUATION

We consider the linear stochastic differential equation

$$dz = Azdt + DdW_t, \tag{56}$$

where $z \in \mathbb{C}^d$, $A \in \mathbb{C}^{d\times d}$, $D \in \mathbb{C}^{d\times m}$, and $W_t \in \mathbb{C}^m$ is a standard real Brownian motion. Note that there are also notions of complex-valued Brownian motion [51] defined as

$$W_t^c = W_t^r + iW_t^i,$$

where W_t^r and W_t^i are independent real Brownian motions. Diffusion process with complex-valued Brownian motion can be re-written as a SDE with real-valued Brownian motions:

$$dz = Azdt + D\begin{bmatrix} I \\ iI \end{bmatrix}^\intercal d\begin{bmatrix} W_t^r \\ W_t^i \end{bmatrix}.$$

The complex-valued SDE defined in Equation (56) can be re-written as a real-valued SDE.

$$d\begin{bmatrix} \mathrm{Re}(z) \\ \mathrm{Im}(z) \end{bmatrix} = \begin{bmatrix} \mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A) \end{bmatrix}\begin{bmatrix} \mathrm{Re}(z) \\ \mathrm{Im}(z) \end{bmatrix} dt + \begin{bmatrix} \mathrm{Re}(D) \\ \mathrm{Im}(D) \end{bmatrix} dW_t.$$

G.3.3. HELPER LEMMAS

Lemma 46 (Topsøe [63, Equation (22)])**.** *For $x \in \mathbb{R}$ and $x \geq 0$, it holds that*

$$\log(1+x) \leq \frac{x(6+x)}{6+4x}.$$

Lemma 47. *For $x \in \mathbb{R}$ and $1 \leq x \leq 2$, it holds that*

$$(5-8x)\log(x-1) \geq 16 - 8x.$$

Proof. Denote $f(x) = (5-8x)\log(x-1) - 16 + 8x$, and we have

$$f'(x) = \frac{-8(x-1)\log(x-1) - 3}{x-1}.$$

Let $g(x) = (x-1)\log(x-1)$, then we get

$$g'(x) = 1 + \log(x-1).$$

We know that the minimum of $g(x)$ is obtained at $x = 1/e + 1$, i.e., $g(x) \geq -1/e$. Therefore,

$$f'(x) \leq \frac{8/e - 3}{x - 1} \leq 0.$$

Hence $f(x)$ is decreasing, and $f(x) \geq f(2) \geq 0$, which concludes the proof. $\square$

Lemma 48. *For $x \in \mathbb{R}$ and $1 \leq x \leq 2$, it holds that*

$$-4x^4 \log^2(3/2) \leq 5/2 - 3x.$$

Proof. Denote $f(x) = -4x^4 \log^2(3/2) - 2.5 + 3x$. We have

$$f'(x) = -16x^3 \log^2(3/2) + 3,$$

and

$$f''(x) = -48x^2 \log^2(3/2) < 0.$$

Therefore, $f'(x)$ is decreasing and is 0 when $x = \sqrt[3]{\frac{3}{16\log^2(3/2)}}$. Hence, $f(x)$ obtains its maximum at this point and $f\left(\sqrt[3]{\frac{3}{16\log^2(3/2)}}\right) < 0$, which completes the proof. $\square$

Lemma 49. *For $x \in \mathbb{R}$ and $1 \leq x \leq 2$, it holds that*

$$x^3 \left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) \leq 0.09 - 0.1x.$$

Proof. Denote $f(x) = x^3 \left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) - 0.09 + 0.1x$. We have

$$f'(x) = 3x^2 \left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) + 0.1 \leq f'(1) \leq 0,$$

since $4\log^2(3/2) - 6\log(3/2)\log(4/3) < 0$. Therefore, $f(x)$ is non-increasing and $f(x) \leq f(1) \leq 0$. $\square$

Lemma 50. *Consider complex-valued linear SDE*

$$dz = Azdt + DdW_t,$$

where $z \in \mathbb{C}^d$, $A \in \mathbb{C}^{d\times d}$ is diagonalizable with orthogonal basis, i.e., $A = USU^\intercal$ with diagonal matrix S and orthogonal matrix U, $D \in \mathbb{C}^{d\times m}$, and $W_t \in \mathbb{C}^m$ is a standard real Brownian motion. The covariance and pseudo-covariance at time t are

$$\Gamma(t) = U \int_0^t \exp\left(\tau S\right) U^\intercal DD^H U \exp\left(\tau \overline{S}\right) d\tau U^\intercal$$
$$C(t) = U \int_0^t \exp\left(\tau S\right) U^\intercal DD^\intercal U \exp\left(\tau S\right) d\tau U^\intercal.$$

Proof. Recall that the covariance and pseudo-covariance are closely related to 4 matrices, $\mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Re}(z)\right]$, $\mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Im}(z)\right]$, $\mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Im}(z)\right]$ $\mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Re}(z)\right]$. Therefore, we study the dynamics of these matrices.

We assume $z(0)$ is a deterministic variable, i.e., its covariance is 0, then according to Equation (6.20) of Särkkä and Solin [55],

$$\begin{aligned}&\mathbb{E}\left[\begin{bmatrix}\mathrm{Re}(z(t)) - \mathbb{E}\left[\mathrm{Re}(z(t))\right] \\ \mathrm{Im}(z(t)) - \mathbb{E}\left[\mathrm{Im}(z(t))\right]\end{bmatrix}\begin{bmatrix}\mathrm{Re}(z(t)) - \mathbb{E}\left[\mathrm{Re}(z(t))\right] \\ \mathrm{Im}(z(t)) - \mathbb{E}\left[\mathrm{Im}(z(t))\right]\end{bmatrix}^\intercal\right] = \begin{bmatrix}\mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Re}(z)\right] & \mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Im}(z)\right] \\ \mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Re}(z)\right] & \mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Im}(z)\right]\end{bmatrix} \\ &= \int_0^t \exp\left(\tau\begin{bmatrix}\mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A)\end{bmatrix}\right)\begin{bmatrix}\mathrm{Re}(D) \\ \mathrm{Im}(D)\end{bmatrix}\left(\exp\left(\tau\begin{bmatrix}\mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A)\end{bmatrix}\right)\begin{bmatrix}\mathrm{Re}(D) \\ \mathrm{Im}(D)\end{bmatrix}\right)^\intercal d\tau,\end{aligned} \tag{57}$$

where the exponential operator is matrix exponential.

Next, let us look at the matrix $\exp\left(\tau\begin{bmatrix}\mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A)\end{bmatrix}\right)$, which has some special properties. Note that we have

$$\mathrm{Re}(A) = U\,\mathrm{Re}(S)U^\intercal, \quad \text{and} \quad \mathrm{Im}(A) = U\,\mathrm{Im}(S)U^\intercal.$$

Therefore, we have

$$\begin{aligned}\begin{bmatrix}\mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A)\end{bmatrix} &= \begin{bmatrix}U\,\mathrm{Re}(S)U^\intercal & -U\,\mathrm{Im}(S)U^\intercal \\ U\,\mathrm{Im}(S)U^\intercal & U\,\mathrm{Re}(S)U^\intercal\end{bmatrix} \\ &= \begin{bmatrix}U & \\ & U\end{bmatrix}\begin{bmatrix}\mathrm{Re}(S) & -\mathrm{Im}(S) \\ \mathrm{Im}(S) & \mathrm{Re}(S)\end{bmatrix}\begin{bmatrix}U^\intercal & \\ & U^\intercal\end{bmatrix}.\end{aligned}$$

Note that the matrix $\begin{bmatrix}U & \\ & U\end{bmatrix}$ is orthogonal, so by the definition of matrix exponential, we have

$$\begin{aligned}\exp\left(\tau\begin{bmatrix}\mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A)\end{bmatrix}\right) &= \sum_{p=0}^{\infty}\frac{\tau^p}{p!}\begin{bmatrix}\mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A)\end{bmatrix}^p \\ &= \begin{bmatrix}U & \\ & U\end{bmatrix}\sum_{p=0}^{\infty}\frac{\tau^p}{p!}\begin{bmatrix}\mathrm{Re}(S) & -\mathrm{Im}(S) \\ \mathrm{Im}(S) & \mathrm{Re}(S)\end{bmatrix}^p\begin{bmatrix}U^\intercal & \\ & U^\intercal\end{bmatrix} \\ &= \begin{bmatrix}U & \\ & U\end{bmatrix}\exp\left(\tau\begin{bmatrix}\mathrm{Re}(S) & -\mathrm{Im}(S) \\ \mathrm{Im}(S) & \mathrm{Re}(S)\end{bmatrix}\right)\begin{bmatrix}U^\intercal & \\ & U^\intercal\end{bmatrix}.\end{aligned}$$

By the property of matrix exponential, if for matrices X and Y, we have $XY = YX$, then $\exp(X+Y) = \exp(X)\exp(Y)$. We verify that this property holds for our case:

$$\begin{bmatrix}\mathrm{Re}(S) & -\mathrm{Im}(S) \\ \mathrm{Im}(S) & \mathrm{Re}(S)\end{bmatrix} = \begin{bmatrix}\mathrm{Re}(S) & \\ & \mathrm{Re}(S)\end{bmatrix} + \begin{bmatrix} & -\mathrm{Im}(S) \\ \mathrm{Im}(S) & \end{bmatrix},$$

and

$$\begin{bmatrix}\mathrm{Re}(S) & \\ & \mathrm{Re}(S)\end{bmatrix}\begin{bmatrix} & -\mathrm{Im}(S) \\ \mathrm{Im}(S) & \end{bmatrix} = \begin{bmatrix} & -\mathrm{Re}(S)\,\mathrm{Im}(S) \\ \mathrm{Re}(S)\,\mathrm{Im}(S) & \end{bmatrix}$$

$$
= \begin{bmatrix} & -\operatorname{Im}(S) \\ \operatorname{Im}(S) & \end{bmatrix} \begin{bmatrix} \operatorname{Re}(S) & \\ & \operatorname{Re}(S) \end{bmatrix}.
$$

Therefore, we have

$$
\begin{aligned}
\exp\left(\tau \begin{bmatrix} \operatorname{Re}(S) & -\operatorname{Im}(S) \\ \operatorname{Im}(S) & \operatorname{Re}(S) \end{bmatrix}\right) &= \exp\left(\tau \begin{bmatrix} \operatorname{Re}(S) & \\ & \operatorname{Re}(S) \end{bmatrix}\right) \exp\left(\tau \begin{bmatrix} & -\operatorname{Im}(S) \\ \operatorname{Im}(S) & \end{bmatrix}\right) \\
&= \begin{bmatrix} \exp(\tau \operatorname{Re}(S)) & \\ & \exp(\tau \operatorname{Re}(S)) \end{bmatrix} \exp\left(\tau \begin{bmatrix} & -\operatorname{Im}(S) \\ \operatorname{Im}(S) & \end{bmatrix}\right),
\end{aligned}
$$

Where the last equality holds because $\operatorname{Re}(S)$ is diagonal. Next, we will show that

$$
\exp\left(\tau \begin{bmatrix} & -\operatorname{Im}(S) \\ \operatorname{Im}(S) & \end{bmatrix}\right) = \begin{bmatrix} \cos(\tau \operatorname{Im}(S)) & -\sin(\tau \operatorname{Im}(S)) \\ \sin(\tau \operatorname{Im}(S)) & \cos(\tau \operatorname{Im}(S)) \end{bmatrix},
$$

where the $\sin$ and $\cos$ operators are applied element-wise. To see this, let us write down the expansion of the matrix exponential.

$$
\begin{aligned}
&\exp\left(\tau \begin{bmatrix} & -\operatorname{Im}(S) \\ \operatorname{Im}(S) & \end{bmatrix}\right) \\
&= I + \begin{bmatrix} & -\tau \operatorname{Im}(S) \\ \tau \operatorname{Im}(S) & \end{bmatrix} + \frac{1}{2} \begin{bmatrix} -\tau^2 \operatorname{Im}(S)^2 & \\ & -\tau^2 \operatorname{Im}(S)^2 \end{bmatrix} \\
&\quad + \frac{1}{3!} \begin{bmatrix} & \tau^3 \operatorname{Im}(S)^3 \\ -\tau^3 \operatorname{Im}(S)^3 & \end{bmatrix} + \frac{1}{4!} \begin{bmatrix} \tau^4 \operatorname{Im}(S)^4 & \\ & \tau^4 \operatorname{Im}(S)^4 \end{bmatrix} + \cdots \\
&= \begin{bmatrix} \cos(\tau \operatorname{Im}(S)) & -\sin(\tau \operatorname{Im}(S)) \\ \sin(\tau \operatorname{Im}(S)) & \cos(\tau \operatorname{Im}(S)) \end{bmatrix}.
\end{aligned}
$$

Now going back to Equation (57), we have

$$
\begin{aligned}
&\begin{bmatrix} \operatorname{Cov}[\operatorname{Re}(z), \operatorname{Re}(z)] & \operatorname{Cov}[\operatorname{Re}(z), \operatorname{Im}(z)] \\ \operatorname{Cov}[\operatorname{Im}(z), \operatorname{Re}(z)] & \operatorname{Cov}[\operatorname{Im}(z), \operatorname{Im}(z)] \end{bmatrix} \\
&= \begin{bmatrix} U & \\ & U \end{bmatrix} \int_0^t \begin{bmatrix} \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) & -\exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\ \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) & \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \end{bmatrix} \begin{bmatrix} U^\intercal \operatorname{Re}(D) \\ U^\intercal \operatorname{Im}(D) \end{bmatrix} \\
&\quad \left(\begin{bmatrix} \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) & -\exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\ \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) & \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \end{bmatrix} \begin{bmatrix} U^\intercal \operatorname{Re}(D) \\ U^\intercal \operatorname{Im}(D) \end{bmatrix} \right)^\intercal d\tau \begin{bmatrix} U^\intercal & \\ & U^\intercal \end{bmatrix} \\
&= \begin{bmatrix} U & \\ & U \end{bmatrix} \int_0^t \begin{bmatrix} P(\tau) & Q(\tau) \\ M(\tau) & N(\tau) \end{bmatrix} d\tau \begin{bmatrix} U^\intercal & \\ & U^\intercal \end{bmatrix},
\end{aligned}
$$

where

$$
\begin{aligned}
P(\tau) = {} & \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \\
& - \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\
& - \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \\
& + \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)), \\
Q(\tau) = {} & \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S))
\end{aligned}
$$

$$
\begin{aligned}
&+ \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \\
&- \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\
&- \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)), \\
M(\tau) = {} & \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \\
&- \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\
&+ \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \\
&- \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)), \\
N(\tau) = {} & \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\
&+ \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) \\
&+ \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S)) \\
&+ \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)) U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U \exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S)).
\end{aligned}
$$

Now we have

$$
\begin{bmatrix} U^\intercal \operatorname{Cov}[\operatorname{Re}(z), \operatorname{Re}(z)]\, U & U^\intercal \operatorname{Cov}[\operatorname{Re}(z), \operatorname{Im}(z)]\, U \\ U^\intercal \operatorname{Cov}[\operatorname{Im}(z), \operatorname{Re}(z)]\, U & U^\intercal \operatorname{Cov}[\operatorname{Im}(z), \operatorname{Im}(z)]\, U \end{bmatrix} = \begin{bmatrix} \int_0^t P(\tau) d\tau & \int_0^t Q(\tau) d\tau \\ \int_0^t M(\tau) d\tau & \int_0^t N(\tau) d\tau \end{bmatrix}.
$$

Denote the covariance of $z(t)$ as $\Gamma(t)$. According to the definition of covariance for complex random variables, we have

$$
U^\intercal \Gamma(t) U = \int_0^t P(\tau) d\tau + \int_0^t N(\tau) d\tau + i \left(\int_0^t M(\tau) d\tau - \int_0^t Q(\tau) d\tau \right).
$$

Let us denote the i-th eigenvalue of A as a_i, i.e., the i-th diagonal element of S. Since $\exp(\tau \operatorname{Re}(S)) \cos(\tau \operatorname{Im}(S))$ and $\exp(\tau \operatorname{Re}(S)) \sin(\tau \operatorname{Im}(S))$ are both diagonal, we have the (i,j)-th element of $U^\intercal \Gamma(t) U$ satisfies

$$
\begin{aligned}
&\left[U^\intercal \Gamma(t) U\right]_{i,j} \\
= &\int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&+ \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&+ i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&- i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&- \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau \\
&+ \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau \\
&- i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau
\end{aligned}
$$

$$
\begin{aligned}
&- i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau \\
&- \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&+ \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&+ i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&+ i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} d\tau \\
&+ \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau \\
&+ \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau \\
&- i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i)\right) \sin\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau \\
&+ i \int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i)\right) \cos\left(\tau \operatorname{Im}(a_j)\right) \left[U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} d\tau.
\end{aligned}
$$

Let us look at the terms associated with $\left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j}$. According to properties of trigonometric functions, we have

$$
\begin{aligned}
\cos(x)\cos(y) + \sin(x)\sin(y) &= \cos(x-y) \\
\sin(x)\cos(y) - \cos(x)\sin(y) &= \sin(x-y),
\end{aligned}
$$

which implies the factor associated with $\left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j}$ is

$$
\begin{aligned}
&\int_0^t \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \cos\left(\tau \operatorname{Im}(a_i) - \tau \operatorname{Im}(a_j)\right) d\tau \\
&+ \int_0^t i \exp\left(\tau \operatorname{Re}(a_i) + \operatorname{Re}(a_j)\right) \sin\left(\tau \operatorname{Im}(a_i) - \tau \operatorname{Im}(a_j)\right) d\tau.
\end{aligned} \tag{58}
$$

Note that for complex variable z, $\operatorname{Re}(\exp(z)) = \exp(\operatorname{Re}(z))\cos(\operatorname{Im}(z))$ and $\operatorname{Im}(\exp(z)) = \exp(\operatorname{Re}(z))\sin(\operatorname{Im}(z))$. To see why this is the case, by Euler's formula,

$$
\exp(z) = \exp\left(\operatorname{Re}(z) + i\operatorname{Im}(z)\right) = \exp\left(\operatorname{Re}(z)\right)\exp\left(i\operatorname{Im}(z)\right) = \exp\left(\operatorname{Re}(z)\right)\left(\cos\left(\operatorname{Im}(z)\right) + i\sin\left(\operatorname{Im}(z)\right)\right).
$$

Therefore, going back to Equation (58), this factor is equal to

$$
\begin{aligned}
&\int_0^t \operatorname{Re}\left(\exp\left(\tau(\operatorname{Re}(a_i) + \operatorname{Re}(a_j) + i(\operatorname{Im}(a_i) - \operatorname{Im}(a_j)))\right)\right) d\tau \\
&\qquad + \int_0^t i\operatorname{Im}\left(\exp\left(\tau(\operatorname{Re}(a_i) + \operatorname{Re}(a_j) + i(\operatorname{Im}(a_i) - \operatorname{Im}(a_j)))\right)\right) d\tau \\
&= \int_0^t \exp\left(\tau(\operatorname{Re}(a_i) + \operatorname{Re}(a_j) + i(\operatorname{Im}(a_i) - \operatorname{Im}(a_j)))\right) d\tau
\end{aligned}
$$

$$= \int_0^t \exp\left(\tau(a_i + \overline{a}_j)\right) d\tau,$$

where $\overline{x}$ is the complex conjugate of x.

Similarly, we can calculate the factors associated with $\left[U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j}$, $\left[U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j}$, and $\left[U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j}$.

$$\begin{aligned}
&\left[U^\intercal \Gamma(t) U\right]_{i,j} \\
=& \left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} \int_0^t \exp\left(\tau(a_i + \overline{a}_j)\right) d\tau - i \left[U^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} \int_0^t \exp\left(\tau(a_i + \overline{a}_j)\right) d\tau \\
&+ i \left[U^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U\right]_{i,j} \int_0^t \exp\left(\tau(a_i + \overline{a}_j)\right) d\tau + \left[U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} \int_0^t \exp\left(\tau(a_i + \overline{a}_j)\right) d\tau \\
=& \left[U^\intercal \operatorname{Re}(D) \operatorname{Re}(D)^\intercal U + U^\intercal \operatorname{Im}(D) \operatorname{Im}(D)^\intercal U + iU^\intercal \operatorname{Im}(D) \operatorname{Re}(D)^\intercal U - iU^\intercal \operatorname{Re}(D) \operatorname{Im}(D)^\intercal U\right]_{i,j} \int_0^t \exp(\tau(a_i + \\
=& \left[U^\intercal D D^H U\right]_{i,j} \int_0^t \exp\left(\tau(a_i + \overline{a}_j)\right) d\tau.
\end{aligned}$$

Therefore, we can obtain that

$$\Gamma(t) = U \int_0^t \exp\left(\tau S\right) U^\intercal D D^H U \exp\left(\tau \overline{S}\right) d\tau U^\intercal.$$

Similarly, for pseudo-covariance C, we have

$$C(t) = U \int_0^t \exp\left(\tau S\right) U^\intercal D D^\intercal U \exp\left(\tau S\right) d\tau U^\intercal.$$

□

G.3.4. PROOFS FOR SECTION F

Proof for Proposition 29. **Solution for SGD** Let us consider more general covariance matrix Σ instead of $\sigma^2 I$. The iterates of SGD can be written as

$$\begin{aligned}
x_k &= x_{k-1} - \eta\left(A x_{k-1} + \xi_{k-1}\right) \\
&= \left(I - \eta A\right) x_{k-1} - \eta \xi_{k-1} \\
&= \left(I - \eta A\right)^2 x_{k-2} - \eta \left(I - \eta A\right) \xi_{k-2} - \eta \xi_{k-1} \\
&\dots \\
&= \left(I - \eta A\right)^k x_0 - \eta \sum_{m=0}^{k-1} \left(I - \eta A\right)^m \xi_{k-1-m}.
\end{aligned}$$

Clearly, it is a linear combination of independent Gaussian variables, therefore x_{k+1} is also has a Gaussian distribution. Moreover,

$$\mathbb{E}\left[x_k\right] = \left(I - \eta A\right)^k x_0 = U\left(I - \eta \Lambda\right)^k U^\intercal x_0$$

$$\mathrm{Cov}\left[x_k, x_k\right] = \eta^2 \sum_{m=0}^{k-1} \left(I - \eta A\right)^m \Sigma \left(I - \eta A\right)^m = \eta^2 \sum_{m=0}^{k-1} U \left(I - \eta \Lambda\right)^m U^\intercal \Sigma U \left(I - \eta \Lambda\right)^m U^\intercal.$$

Plugging in $\Sigma = \sigma^2 I$ gives us the desired result.

The three SDEs considered applied on quadratics with Assumption 2 are OU processes. According to Särkkä and Solin [55, Section 6.2], for a linear time-invariant SDE

$$dX_t = FX_t dt + LdW_t,$$

the solution would be a Gaussian variable and its mean and covariance satisfy

$$\mathbb{E}\left[X(x_0, t)\right] = \exp\left(Ft\right) x_0$$
$$\mathrm{Cov}\left[X(x_0,t), X(x_0,t)\right] = \int_0^t \exp\left(F(t-\tau)\right) LL^\intercal \exp\left(F(t-\tau)\right)^\intercal d\tau.$$

In our case, since we assume isotropic noise, i.e., the noise covariance is $\sigma^2 I$, the diffusion coefficients of the three SDEs we consider is also isotropic, i.e., $\sqrt{\eta}\sigma I$. Under this condition, F and L commute, and the covariance has closed form solution.

Solution for `SME-1` The SDE `SME`-1 applied on the problem we consider is

$$dX_t = -AX_t dt + \sqrt{\eta}\sigma I dW_t.$$

Then the mean is

$$\mathbb{E}\left[X(x_0, t)\right] = \exp\left(-At\right) x_0,$$

and the covariance is

$$\begin{aligned}
\mathrm{Cov}\left[X(x_0,t), X(x_0,t)\right] &= \int_0^t \exp\left(A(\tau - t)\right) \eta\sigma^2 \exp\left(A(\tau - t)\right) d\tau \\
&= \eta\sigma^2 \int_0^t U \exp\left(\Lambda(\tau - t)\right) U^\intercal U \exp\left(\Lambda(\tau - t)\right) U^\intercal d\tau \\
&= \eta\sigma^2 U \int_0^t \exp\left(\Lambda(\tau - t)\right) \exp\left(\Lambda(\tau - t)\right) d\tau U^\intercal \\
&= \eta\sigma^2 U \int_0^t \exp\left(2\Lambda(\tau - t)\right) d\tau U^\intercal \\
&= \eta\sigma^2 U \frac{I - \exp(-2\Lambda t)}{2\Lambda} U^\intercal.
\end{aligned}$$

Solution for `SME-2` The SDE `SME`-2 applied on the problem we consider is

$$dX_t = -(A + \frac{\eta}{2}A^2) X_t dt + \sqrt{\eta}\sigma I dW_t.$$

Then the mean is

$$\mathbb{E}\left[X(x_0, t)\right] = \exp\left(-\left(A + \frac{\eta}{2}A^2\right) t\right) x_0,$$

and the covariance is

$$
\begin{aligned}
\mathrm{Cov}\left[X(x_0,t),X(x_0,t)\right] &= \int_0^t \exp\left(\left(A+\frac{\eta}{2}A^2\right)(\tau-t)\right)\eta\sigma^2\exp\left(\left(A+\frac{\eta}{2}A^2\right)(\tau-t)\right)d\tau \\
&= \eta\sigma^2\int_0^t U\exp\left(\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)(\tau-t)\right)U^\intercal U\exp\left(\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)(\tau-t)\right)U^\intercal d\tau \\
&= \eta\sigma^2 U\int_0^t \exp\left(\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)(\tau-t)\right)\exp\left(\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)(\tau-t)\right)d\tau U^\intercal \\
&= \eta\sigma^2 U\int_0^t \exp\left(2\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)(\tau-t)\right)d\tau U^\intercal \\
&= \eta\sigma^2 U\frac{I-\exp(-2\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)t)}{2\left(\Lambda+\frac{\eta}{2}\Lambda^2\right)}U^\intercal .
\end{aligned}
$$

Solution for `SPF` The SDE `SPF`applied on the problem we consider is

$$
dX_t = U\frac{\log\left(1-\eta\Lambda\right)}{\eta}U^\intercal X_t dt+\sqrt{\eta}\sigma I dW_t .
$$

Then the mean is

$$
\begin{aligned}
\mathbb{E}\left[X(x_0,t)\right] &= \exp\left(U\frac{\log\left(1-\eta\Lambda\right)}{\eta}U^\intercal t\right)x_0 \\
&= U\exp\left(\log\left(1-\eta\Lambda\right)^{t/\eta}\right)U^\intercal x_0 \\
&= U\left(1-\eta\Lambda\right)^{t/\eta}U^\intercal x_0,
\end{aligned}
$$

and the covariance is

$$
\begin{aligned}
\mathrm{Cov}\left[X(x_0,t),X(x_0,t)\right] &= \int_0^t \exp\left(U\frac{\log\left(1-\eta\Lambda\right)}{\eta}U^\intercal(\tau-t)\right)\eta\sigma^2\exp\left(U\frac{\log\left(1-\eta\Lambda\right)}{\eta}U^\intercal(\tau-t)\right)d\tau \\
&= \eta\sigma^2 U\int_0^t \exp\left(2\frac{\log\left(1-\eta\Lambda\right)}{\eta}(\tau-t)\right)d\tau U^\intercal \\
&= \eta\sigma^2 U\left(\frac{\eta}{-2\log\left(1-\eta\Lambda\right)}\left(1-\exp\left(\frac{2\log\left(1-\eta\Lambda\right)}{\eta}t\right)\right)\right)U^\intercal \\
&= \eta^2\sigma^2 U\left(\frac{1}{-2\log\left(1-\eta\Lambda\right)}\left(1-\left(1-\eta\Lambda\right)^{2t/\eta}\right)\right)U^\intercal .
\end{aligned}
$$

□

Proof for Desideratum 1. If the distribution of complex variable matches the real variable, then by the definition of matching, we have

$$
\begin{gathered}
\mathbb{E}\left[\mathrm{Re}(z)\right]=\mathbb{E}\left[\widetilde{z}\right] \quad \mathbb{E}\left[\mathrm{Im}(z)\right]=0 \\
\mathrm{Cov}\left[\mathrm{Re}(z),\mathrm{Re}(z)\right]=\mathrm{Cov}\left[\widetilde{z},\widetilde{z}\right] \quad \mathrm{Cov}\left[\mathrm{Im}(z),\mathrm{Im}(z)\right]=0.
\end{gathered}
$$

Then according to Equation (54), we get

$$
\Gamma(z)=\mathrm{Cov}\left[\mathrm{Re}(z),\mathrm{Re}(z)\right]+\mathrm{Cov}\left[\mathrm{Im}(z),\mathrm{Im}(z)\right]+i\left(\mathrm{Cov}\left[\mathrm{Im}(z),\mathrm{Re}(z)\right]-\mathrm{Cov}\left[\mathrm{Re}(z),\mathrm{Im}(z)\right]\right)
$$

$$
\begin{aligned}
&= \mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Re}(z)\right] = \mathrm{Cov}\left[\widetilde{z}, \widetilde{z}\right] \\
\Gamma(z) &= \mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Re}(z)\right] - \mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Im}(z)\right] + i\left(\mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Re}(z)\right] + \mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Im}(z)\right]\right) \\
&= \mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Re}(z)\right] = \mathrm{Cov}\left[\widetilde{z}, \widetilde{z}\right],
\end{aligned}
$$

as $\mathrm{Im}(z) = 0$ deterministically. Next, we consider the opposite direction. The means match trivially. According to Equation (55), we have

$$
\begin{aligned}
\mathrm{Cov}\left[\mathrm{Re}(z), \mathrm{Re}(z)\right] &= \frac{1}{2}\left(\Gamma(z) + C(z)\right) = \mathrm{Cov}\left[\widetilde{z}, \widetilde{z}\right] \\
\mathrm{Cov}\left[\mathrm{Im}(z), \mathrm{Im}(z)\right] &= \frac{1}{2}\left(\Gamma(z) - C(z)\right) = 0,
\end{aligned}
$$

which completes the proof. □

G.3.5. PROOF FOR PROPOSITION 30

Proof. Consider function

$$
f(x) = \frac{1}{2} x^\intercal \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} x,
$$

and

$$
\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix} \quad \text{with} \quad \Sigma_{11}, \Sigma_{12}, \Sigma_{22} > 0 \quad \text{and} \quad \Sigma_{11}\Sigma_{22} - \Sigma_{12}^2 = 0.
$$

Note that the last condition implies that Σ is positive semi-definite and not positive definite. It is easy to come up with such a Σ, e.g., $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$.

According to Lemma 31, a necessary condition for matching the discrete-time iterates is that the RHS of Equation (29) is positive semi-definite, otherwise it cannot be decomposed to $\left(U^\intercal D\right)\left(U^\intercal D\right)^H$. Therefore, as long as we can show that the determinant of Σ is negative, i.e., it contains a negative eigenvalue, then we can claim that there is no such a decomposition, i.e., no linear SDE can match the discrete-time iterates.

The RHS matrix of Equation (29) in our case is

$$
\begin{bmatrix} \frac{2\Sigma_{11}\,\mathrm{Re}(\log(1-\eta))}{\eta-2} & \frac{\Sigma_{12}(\log(1-\eta)+\log(1+\eta))}{-\eta} \\ \frac{\Sigma_{12}\left(\log(1+\eta)+\overline{\log(1-\eta)}\right)}{-\eta} & \frac{2\Sigma_{22}\log(1+\eta)}{\eta+2} \end{bmatrix},
$$

where we used $x + \overline{x} = 2\,\mathrm{Re}(x)$ to simplify the result. Then, our goal is to show

$$
\frac{4\Sigma_{11}\Sigma_{22}\,\mathrm{Re}\left(\log(1-\eta)\right)\log(1+\eta)}{(\eta-2)(\eta+2)} - \frac{\Sigma_{12}^2\left(\log(1-\eta)+\log(1+\eta)\right)\left(\overline{\log(1-\eta)}+\log(1+\eta)\right)}{\eta^2} < 0.
$$

It is then sufficient to show the following:

$$
\frac{\left(\log(1-\eta)+\log(1+\eta)\right)\left(\overline{\log(1-\eta)}+\log(1+\eta)\right)(\eta-2)(\eta+2)}{4\eta^2\,\mathrm{Re}\left(\log(1-\eta)\right)\log(1+\eta)} > \frac{\Sigma_{11}\Sigma_{22}}{\Sigma_{12}^2} = 1. \tag{59}
$$

To see why this is the case, notice that for all for all $\eta > 0$, it holds that

$$\frac{\operatorname{Re}(\log(1-\eta))}{\eta - 2} > 0.$$

Next, we will discuss in three different cases, i.e., $0 < \eta < 1$, $1 < \eta < 2$ and $\eta > 2$, to prove the above inequality. For the boundary case, one can check that the inequality holds when $\eta \to 2$.

When $0 < \eta < 1$ In this case, we have $\operatorname{Re}(\log(1-\eta)) < 0$, therefore, Equation (59) is equivalent to

$$\begin{aligned}
&4\eta^2 \operatorname{Re}\left(\log(1-\eta)\right)\log(1+\eta) > \left(\log(1-\eta)+\log(1+\eta)\right)\left(\overline{\log(1-\eta)}+\log(1+\eta)\right)(\eta-2)(\eta+2)\\
\iff &4(\log(1-\eta)+\log(1+\eta))(\overline{\log(1-\eta)}+\log(1+\eta)) > \eta^2(\log(1-\eta)-\log(1+\eta))(\overline{\log(1-\eta)}-\log(1+\eta)).
\end{aligned} \tag{60}$$

When $0 < \eta < 1$, we have $\overline{\log(1-\eta)} = \log(1-\eta)$, which implies the following equivalence of Equation (60):

$$\begin{aligned}
&4\left(\log(1-\eta)+\log(1+\eta)\right)^2 > \eta^2\left(\log(1-\eta)-\log(1+\eta)\right)^2\\
\iff\quad &4\left(\log\left(1-\eta^2\right)\right)^2 > \eta^2\left(\log(1-\eta)-\log(1+\eta)\right)^2\\
\iff\quad &-2\log\left(1-\eta^2\right) > \eta\left(\log(1+\eta)-\log(1-\eta)\right)\\
\iff\quad &\underbrace{\eta\log(1+\eta)-\eta\log(1-\eta)+2\log\left(1-\eta^2\right)}_{g(\eta)} < 0.
\end{aligned}$$

Now we analyze the function $g(\eta)$.

$$\begin{aligned}
g'(\eta) &= \log(1+\eta)+\frac{\eta}{1+\eta}-\log(1-\eta)+\frac{\eta}{1-\eta}-\frac{4\eta}{1-\eta^2}\\
&= \log\left(\frac{1+\eta}{1-\eta}\right)-\frac{2\eta}{1-\eta^2}\\
&= \log\left(1+\frac{2\eta}{1-\eta}\right)-\frac{2\eta}{1-\eta^2}\\
&\le \frac{\frac{2\eta}{1-\eta}\left(6+\frac{2\eta}{1-\eta}\right)}{6+\frac{8\eta}{1-\eta}}-\frac{2\eta}{1-\eta^2}\\
&= \frac{12\eta-8\eta^2}{(1-\eta)(6+2\eta)}-\frac{2\eta}{1-\eta^2}\\
&= -\frac{8\eta^3}{(1-\eta)(1+\eta)(6+2\eta)} < 0,
\end{aligned}$$

where the first inequality is according to Lemma 46. Now we know that $g(\eta)$ is strictly decreasing over $0 < \eta < 1$, therefore $\sup_{0<\eta<1} g(\eta) < g(0) = 0$.

When $1 < \eta < 2$ In this case, we also obtain Equation (60). In addition, since $\log(1-\eta) = \pi i + \log(\eta - 1)$, we have

$$\log(1-\eta)\overline{\log(1-\eta)} = \log^2(1-\eta) + \pi^2.$$

Therefore, Equation (60) is equivalent to

$$\begin{aligned}
& 4\left(\log^2(\eta-1) + \pi^2 + 2\log(\eta-1)\log(1+\eta) + \log^2(1+\eta)\right) \\
& \qquad > \eta^2\left(\log^2(\eta-1) + \pi^2 - 2\log(\eta-1)\log(1+\eta) + \log^2(1+\eta)\right) \\
\iff \quad & 4\left(\log(\eta-1) + \log(1+\eta)\right)^2 - \eta^2\left(\log(\eta-1) - \log(1+\eta)\right)^2 + \pi^2(4-\eta^2) > 0 \qquad (61) \\
\iff \quad & (4-\eta^2)\left(\log^2(\eta-1) + \log^2(\eta+1) + \pi^2\right) + (8+2\eta^2)\log(\eta-1)\log(1+\eta) > 0.
\end{aligned}$$

To prove the above inequality, it is sufficient to prove a lower bound of LHS is greater than 0.

$$\begin{aligned}
& \left(4-\eta^2\right)\left(\log^2(\eta-1) + \log^2(\eta+1) + \pi^2\right) + (8+2\eta^2)\log(\eta-1)\log(1+\eta) \\
> & \left(4-\eta^2\right)\left(\log^2(\eta-1) + \log^2(\eta+1) + \pi^2\right) + (8+2\eta^2)\log(\eta-1)\eta \\
\geq & \left(4-\eta^2\right)\left(\log^2(\eta-1) + \log^2(\eta+1) + \pi^2\right) + (22\eta-12)\log(\eta-1) \\
\geq & (6-3\eta)\log^2(\eta-1) + \left(4-\eta^2\right)\left(\log^2(1+\eta) + \pi^2\right) + (22\eta-12)\log(\eta-1) \\
\geq & \underbrace{(6-3\eta)\log^2(\eta-1) + \left(4-\eta^2\right)\left((\eta\log(3/2) + \log(4/3))^2 + \pi^2\right) + (22\eta-12)\log(\eta-1)}_{h(\eta)},
\end{aligned}$$

where the first inequality is because $\log(1+\eta) < \eta$ for $1 < \eta < 2$, and the second inequality holds as $2\eta^3 + 2\eta \leq 22\eta - 12$. This can be seen by noticing that $2\eta^3 + 2\eta$ is convex on this interval and $22\eta - 12$ is the linear interpolation of the endpoints. Similarly, $4 - \eta^2$ is concave, and lower bounded by linear interpolation $6 - 3\eta$, which gives us the third inequality. The final inequality comes from the linear interpolation of endpoints of concave function $\log(1+\eta)$.

Next, we study the lower bound of $h(\eta)$. By taking its derivative, we obtain

$$\begin{aligned}
h'(\eta) = & \frac{1}{\eta-1}(3(1-\eta)\log^2(\eta-1) + (16\eta-10)\log(\eta-1) - 4\eta^4\log^2(3/2) \\
& + \eta^3\left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) + \eta^2\left(6\log(3/2)\log(4/3) - 2\log^2(4/3) - 2\pi^2 + 8\log^2(3/2)\right) \\
& + \eta(2\log^2(4/3) + 2\pi^2 - 8\log^2(3/2) + 8\log(4/3)\log(3/2) + 22) - 8\log(4/3)\log(3/2) - 12) \\
\leq & \frac{1}{\eta-1}((16\eta-10)\log(\eta-1) - 4\eta^4\log^2(3/2) + \eta^3\left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) \\
& + \eta^2\left(6\log(3/2)\log(4/3) - 2\log^2(4/3) - 2\pi^2 + 8\log^2(3/2)\right) \\
& + \eta(2\log^2(4/3) + 2\pi^2 - 8\log^2(3/2) + 8\log(4/3)\log(3/2) + 22) - 8\log(4/3)\log(3/2) - 12) \\
\leq & \frac{1}{\eta-1}(16\eta - 32 - 4\eta^4\log^2(3/2) + \eta^3\left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) \\
& + \eta^2\left(6\log(3/2)\log(4/3) - 2\log^2(4/3) - 2\pi^2 + 8\log^2(3/2)\right) \\
& + \eta(2\log^2(4/3) + 2\pi^2 - 8\log^2(3/2) + 8\log(4/3)\log(3/2) + 22) - 8\log(4/3)\log(3/2) - 12) \\
\leq & \frac{1}{\eta-1}(16\eta - 32 + 2.5 - 3\eta + \eta^3\left(4\log^2(3/2) - 6\log(3/2)\log(4/3)\right) \\
& + \eta^2\left(6\log(3/2)\log(4/3) - 2\log^2(4/3) - 2\pi^2 + 8\log^2(3/2)\right)
\end{aligned}$$

$$
\begin{aligned}
&+ \eta(2\log^2(4/3) + 2\pi^2 - 8\log^2(3/2) + 8\log(4/3)\log(3/2) + 22) - 8\log(4/3)\log(3/2) - 12) \\
\leq& \frac{1}{\eta - 1}(16\eta - 32 + 2.5 - 3\eta + 0.09 - 0.1\eta \\
&+ \eta^2\left(6\log(3/2)\log(4/3) - 2\log^2(4/3) - 2\pi^2 + 8\log^2(3/2)\right) \\
&+ \eta(2\log^2(4/3) + 2\pi^2 - 8\log^2(3/2) + 8\log(4/3)\log(3/2) + 22) - 8\log(4/3)\log(3/2) - 12).
\end{aligned}
$$

where the second, third and forth inequalities are according to Lemmas 47 to 49 respectively. The remaining numerator is a quadratic function of η, and one can easily verify that it is less than 0. Therefore, we conclude that $h'(\eta) < 0$. Hence $h(x) \geq h(2) = 0$.

When $\eta > 2$ In this case, our target becomes to prove Equation (61), however with the opposite sign of the inequality.

$$
\begin{aligned}
&4\left(\log(\eta - 1) + \log(1 + \eta)\right)^2 - \eta^2\left(\log(\eta - 1) - \log(1 + \eta)\right)^2 + \pi^2\left(4 - \eta^2\right) \\
\leq& 4\left(\log(\eta - 1) + \log(1 + \eta)\right)^2 - 4\left(\log(\eta - 1) - \log(1 + \eta)\right)^2 + \pi^2\left(4 - \eta^2\right) \\
=& 16\log(\eta - 1)\log(1 + \eta) + \pi^2\left(4 - \eta^2\right) \\
<& 16(\eta - 2)\log(1 + \eta) + \pi^2\left(4 - \eta^2\right) \\
\leq& 16(\eta - 2)\frac{\eta(6 + \eta)}{6 + 4\eta} + \pi^2\left(4 - \eta^2\right) \\
=& \frac{(8 - 2\pi^2)\eta^3 + (32 - 3\pi^2)\eta^2 + (8\pi^2 - 96)\eta + 12\pi^2}{3 + 2\eta} \\
=& \frac{-(x - 2)\left(\left(2\pi^2 - 8\right)x^2 + \left(7\pi^2 - 48\right)x + 6\pi^2\right)}{3 + 2\eta},
\end{aligned}
$$

where the second inequality holds as $\log(x) < x - 1$ for $x > 1$ and the third inequality is according to Lemma 46. In addition, $\left(2\pi^2 - 8\right)x^2 + \left(7\pi^2 - 48\right)x + 6\pi^2 \geq 0$, therefore, we have

$$
4\left(\log(\eta - 1) + \log(1 + \eta)\right)^2 - \eta^2\left(\log(\eta - 1) - \log(1 + \eta)\right)^2 + \pi^2\left(4 - \eta^2\right) < 0,
$$

which concludes the proof.

□

Proof for Lemma 31. To match the expectation, the argument is similar to the proof in Rosca et al. [53, Section A.6]. Here, we proceed to match the covariance and pseudo-covariance. Let us first look at the continuous-time process. We denote the covairance of X as $\Gamma(X)$. According to Lemma 50, we have

$$
\begin{aligned}
&\left[U^\intercal \Gamma(t) U\right]_{i,j} \\
=& \frac{\left[U^\intercal \Sigma U\right]_{i,j}\left(\log\left(1 - \eta\lambda_i\right) + \overline{\log\left(1 - \eta\lambda_i\right)}\right)}{\eta\lambda_i\lambda_j - (\lambda_i + \lambda_j)} \int_0^t \exp\left(\tau\frac{\log\left(1 - \eta\lambda_i\right)}{\eta}\right)\exp\left(\tau\frac{\overline{\log\left(1 - \eta\lambda_j\right)}}{\eta}\right) d\tau \\
=& \frac{\eta\left[U^\intercal \Sigma U\right]_{i,j}\left(\log\left(1 - \eta\lambda_i\right) + \overline{\log\left(1 - \eta\lambda_i\right)}\right)}{\left(1 - \eta\lambda_i\right)\left(1 - \eta\lambda_j\right) - 1} \int_0^t \exp\left(\frac{\tau}{\eta}\left(\log\left(1 - \eta\lambda_i\right) + \overline{\log\left(1 - \eta\lambda_j\right)}\right)\right) d\tau
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\eta \left[U^{\intercal}\Sigma U\right]_{i,j} \left(\log\left(1-\eta\lambda_i\right) + \overline{\log\left(1-\eta\lambda_i\right)}\right)}{\left(1-\eta\lambda_i\right)\left(1-\eta\lambda_j\right)-1} \frac{\eta}{\log\left(1-\eta\lambda_i\right) + \overline{\log\left(1-\eta\lambda_i\right)}} \left(\exp\left(\frac{t}{\eta}\left(\log\left(1-\eta\lambda_i\right) + \overline{\log\left(1-\eta\right.}\right.\right.\right. \\
&= \frac{\eta^2 \left[U^{\intercal}\Sigma U\right]_{i,j}}{\left(1-\eta\lambda_i\right)\left(1-\eta\lambda_j\right)-1} \left(\exp\left(\frac{t}{\eta}\left(\log\left(1-\eta\lambda_i\right) + \overline{\log\left(1-\eta\lambda_j\right)}\right)\right) - 1\right) \\
&= \frac{\eta^2 \left[U^{\intercal}\Sigma U\right]_{i,j}}{\left(1-\eta\lambda_i\right)\left(1-\eta\lambda_j\right)-1} \left(\left((1-\eta\lambda_i)(1-\eta\lambda_j)\right)^{\frac{t}{\eta}} - 1\right),
\end{aligned}
$$

where we used the fact that for complex-valued variable z, it holds that $\exp\left(\overline{z}\right) = \overline{\exp(z)}$. Now we consider time stamp $t = k\eta$ for $k = 1, 2, \ldots$, where we have

$$
\begin{aligned}
\left[U^{\intercal}\Gamma(k\eta)U\right]_{i,j} &= \frac{\eta^2 \left[U^{\intercal}\Sigma U\right]_{i,j}}{\left(1-\eta\lambda_i\right)\left(1-\eta\lambda_j\right)-1} \left(\left((1-\eta\lambda_i)(1-\eta\lambda_j)\right)^{k} - 1\right) \\
&= \eta^2 \left[U^{\intercal}\Sigma U\right]_{i,j} \sum_{m=0}^{k-1} \left((1-\eta\lambda_i)(1-\eta\lambda_j)\right)^{m}.
\end{aligned}
$$

Then we have

$$
U^{\intercal}\Gamma(k\eta)U = \eta^2 \sum_{m=0}^{k-1} \left(I-\eta\Lambda\right)^m U^{\intercal}\Sigma U \left(I-\eta\Lambda\right)^m,
$$

which implies

$$
\begin{aligned}
\Gamma(k\eta) &= \eta^2 \sum_{m=0}^{k-1} U\left(I-\eta\Lambda\right)^m U^{\intercal}\Sigma U \left(I-\eta\Lambda\right)^m U^{\intercal} \\
&= \eta^2 \sum_{m=0}^{k-1} \left(I-\eta A\right)^m \Sigma \left(I-\eta A\right)^m,
\end{aligned}
$$

which matches the covariance of discrete-time variable. The proof for matching pseudo-covariance is almost the same. □