SIAM/ASA J. Uncertainty Quantification
Vol. 11, No. 1, pp. 139–167



# On the Generalized Langevin Equation for Simulated Annealing*

Martin Chak†, Nikolas Kantas†, and Grigorios A. Pavliotis†

**Abstract.** In this paper, we consider the generalized (higher order) Langevin equation for the purpose of simulated annealing and optimization of nonconvex functions. Our approach modifies the underdamped Langevin equation by replacing the Brownian noise with an appropriate Ornstein–Uhlenbeck process to account for memory in the system. Under reasonable conditions on the loss function and the annealing schedule, we establish convergence of the continuous time dynamics to a global minimum. In addition, we investigate the performance numerically and show better performance and higher exploration of the state space compared to the underdamped Langevin dynamics with the same annealing schedule.

**Key words.** nonconvex optimization, generalized Langevin equation, simulated annealing

**MSC codes.** 60J25, 46N10, 60J60

**DOI.** 10.1137/21M1462970

## 1. Introduction.

Algorithms for optimization have gained significant interest in recent years due to applications in machine learning, data science, and molecular dynamics. Models in machine learning are formulated to have some loss function and parameters with respect to which it is to be minimized, where use of optimization techniques is heavily relied upon. We refer the reader to [8, 67] for related discussions. Many models, such as neural networks, use parameters that vary over a continuous space, where gradient-based optimization methods can be used to find good parameters that generate effective predictive ability. As such, the design and analysis of such algorithms for global optimization has been the subject of considerable research [65], and it has proved useful to study algorithms for global optimization using tools from the theory of stochastic processes and dynamical systems. A paradigm of the use of stochastic dynamics for the design of algorithms for global optimization is simulated annealing, where overdamped Langevin dynamics with a time-dependent temperature (1.1) that decreases with an appropriate cooling schedule is used to guarantee the global minimum of a nonconvex loss function $U : \mathbb{R}^n \to \mathbb{R}$:

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2T_t}\, dW_t. \tag{1.1}$$

* Received by the editors December 2, 2021; accepted for publication (in revised form) August 22, 2022; published electronically March 3, 2023.
https://doi.org/10.1137/21M1462970
**Funding:** The first author was supported by an EPSRC studentship. The second and third authors were partially supported by JPMorgan Chase & Co. under J.P. Morgan A.I. Research Awards 2019. The third author was also partially supported by the EPSRC through grants EP/P031587/1, EP/L024926/1, and EP/L020564/1.

† Department of Mathematics, Imperial College London, London, SW7 2AZ, UK (mwc114@ic.ac.uk, n.kantas@imperial.ac.uk, g.pavliotis@imperial.ac.uk).

Here $W_t$ is a standard $n$-dimensional Wiener process, and $T_\cdot : (0,\infty) \to (0,\infty)$ is an appropriate deterministic function of time often referred to as the annealing or cooling schedule. For fixed $T_t = T > 0$, this is the dynamics used for the related problem of sampling from a possibly high dimensional probability measure, for example, in the unadjusted Langevin algorithm [21]. Gradually decreasing $T_t$ to zero balances the exploration-exploitation trade-off by allowing, at early times, larger noise to drive $X_t$ and hence sufficient mixing to escape local minima. Designing an appropriate annealing schedule is wellunderstood. We briefly mention classical references [16, 29, 30, 32–35, 39], as well as the more recent [38, 46, 61], where one can find details and convergence results. In this paper we aim to consider generalized versions of (1.1) for the same purpose.

Using dynamics such as (1.1) has clear connections with sampling. When $T_t = T$ is a constant function, the invariant distribution of $X$ is proportional to $\exp(-\frac{U(x)}{T})dx$. In addition, when $T_t$ decreases with time, the probability measure given by $\nu_t(dx) \propto \exp(-\frac{U(x)}{T_t})dx$ converges weakly to the set of global minima based on the Laplace principle [37]. If one replaces (1.1) with a stochastic process that mixes faster and maintains the same invariant distribution for constant temperatures, then one can expect the superior speed of convergence to improve performance in optimization due to the increased exploration of the state space. Indeed, it is well known that many different dynamics can be used in order to sample from a given probability distribution or to find the minima of a function when the dynamics is combined with an appropriate cooling schedule for the temperature. Different kinds of dynamics have already been considered for sampling, e.g., nonreversible dynamics, preconditioned unadjusted Langevin dynamics [2, 4, 44, 58], as well as for optimization, e.g., interacting Langevin dynamics [69] and consensus based optimization [10, 11, 62] to name a few.

A natural candidate in this direction is the underdamped Langevin dynamics:

$$dX_t = Y_t\,dt, \tag{1.2a}$$

$$dY_t = -\nabla U(X_t)\,dt - T_t^{-1}\mu Y_t\,dt + \sqrt{2\mu}\,dW_t. \tag{1.2b}$$

Here the reversibility property of (1.1) has been lost; the improvement from breaking reversibility in the context of both sampling and optimization is investigated in [19, 43] and [26], respectively. When $T_t = T$, (1.2) can converge faster than (1.1) to its invariant distribution

$$\rho(dx,dy) \propto \exp\left(-\frac{1}{T}\left(U(x) + \frac{|y|^2}{2}\right)\right)dx\,dy;$$

see [22] or section 6.3 of [59] for particular comparisons and see also [5, 6] for more applications using variants of (1.2). In the context of simulated annealing, using this set of dynamics has recently been studied rigorously in [51], where the author established convergence to global minima using the generalized $\Gamma$-calculus [52] framework that is based on Bakry–Emery theory. Note that (1.2) uses the temperature in the drift rather than the diffusion constant in the noise as in (1.1). Both formulations admit the same invariant measure when $T_t = T$. In the remainder of the paper, we adopt this formulation to be closer to that of [51].

In this paper we will consider an extension of the kinetic Langevin equation by adding an additional auxiliary variable that accounts for the memory in the system. To the best of the authors' knowledge, this has not been attempted before in the context of simulated annealing and global optimization. In particular, we consider the Markovian approximation to the generalized Langevin equation,

$$dX_t = Y_t\,dt, \tag{1.3a}$$

$$dY_t = -\nabla U(X_t)\,dt + \lambda^\top Z_t\,dt, \tag{1.3b}$$

$$dZ_t = -\lambda Y_t\,dt - T_t^{-1} A Z_t\,dt + \Sigma\,dW_t, \tag{1.3c}$$

where $A \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix, meaning that there exists a constant $A_c > 0$ such that $z^\top A z \geq A_c|z|^2$ for all $z \in \mathbb{R}^m$, $\Sigma \in \mathbb{R}^{m\times m}$ satisfies $\Sigma\Sigma^\top = 2A$, and $W_t$ is now $m$-dimensional. Here $X_t, Y_t \in \mathbb{R}^n$ and $Z_t \in \mathbb{R}^m$ (with $m \geq n$), $M^\top$ denotes the transpose of a matrix $M$, and $\lambda \in \mathbb{R}^{m\times n}$ is a rank $n$ matrix with a left inverse $\lambda^{-1} \in \mathbb{R}^{n\times m}$.

Our aim is to establish convergence using techniques similar to those in [51] and investigate the improvements in performance. Equation (1.3) is related to the generalized Langevin equation, where memory is added to (1.2) by integrating over past velocities with a kernel $\Gamma : (0,\infty) \to \mathbb{R}^{n\times n}$,

$$\ddot{x} = -\nabla U(x) - \int_0^t \Gamma(t-s)\dot{x}(s)\,ds + F_t, \tag{1.4}$$

with $F_t$ being a zero mean stationary Gaussian process with an autocorrelation matrix given by the fluctuation-dissipation theorem $\mathbb{E}(F_t F_s^\top) = T_t\Gamma(t-s)$. When[1] $T_t = T$, (1.4) is equivalent to (1.3), with $Z_0 \sim \mathcal{N}(0, TI)$ for identity matrix $I$ when setting $\Gamma(t) = \lambda^\top e^{-At}\lambda$; see Proposition 8.1 in [59]. In this case, the invariant distribution becomes

$$\rho(dx, dy, dz) \propto \exp\left(-\frac{1}{T}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)\right)dx\,dy\,dz.$$

In the spirit of adding a momentum variable in (1.1) to get (1.2), (1.3) adds an additional auxiliary variable to the Langevin system while preserving the invariant distribution in the $x$ marginal. In the constant temperature context, (1.4) is natural from the point of view of statistical mechanics and has already been considered as a sampling tool in [12, 13, 54] with considerable success. We will demonstrate numerically that the additional tuning parameters can improve performance; see also [53] for recent work demonstrating advantages of using (1.4) compared to using (1.2) when sampling from a log concave density. A detailed study of the Markovian approximation (1.3) of the generalized Langevin dynamics in (1.4) can be found in [56].

To motivate the use of (1.3), we consider the quadratic case where $U = \alpha x^2$ and $0 < \alpha < 1$. By Theorem 3.1 in [49], the calculation of the spectral gaps of the generators in (1.1)–(1.3) reduces in this case to finding roots of quadratic and cubic polynomials, respectively.

[1] To the best of our knowledge, there is no known direct translation between (1.4) and (1.3) for a nonconstant $T_t$; at the very least the intuition here is useful.

Straightforward numerical comparisons show that for these quadratic cases, the best choices of $\lambda, A$ yield an improvement in terms of the spectral gap compared to (1.2) with the best choice of $\mu$.

Use of (1.4) is also motivated by parallels with accelerated gradient descent algorithms. When the noise is removed from (1.2), the second order differential equation can be loosely considered as a continuous time version of Nesterov's algorithm [68]. The latter is commonly preferred to discretizing the first order differential equation given by the noiseless version of (1.1) because in the high dimensional and low iterations setting it achieves the optimal rate of convergence for convex optimization; see Chapter 2 in [55] and also see [31] for a nonconvex setting. Here we would like to investigate the effect of adding another auxiliary variable, which would correspond to a third order differential equation when noise is removed. When noise is added for the fixed temperature case, [25] has studied the long time behavior and stability for different choices of a memory kernel as in (1.4). Finally, we note that generalized Langevin dynamics in (1.4) have additionally been studied in related areas such as sampling problems in molecular dynamics from chemical modeling [1, 12, 13, 54, 71], see also [40] for work determining the kernel $\Gamma$ in the generalised system (1.4) from data.

Our theoretical results will focus only on the continuous time dynamics and follow the approach in [51]. The main requirements in terms of assumptions are quadratic upper and lower bounds on $U$ and bounded second derivatives. This is different from classical references such as [30], [32], or [35]. These works also rely on the Poincaré inequality, an approach which will be mirrored here (and is used in [51] for the underdamped case) using a log-Sobolev inequality; see also [34] for the relationship between such functional inequalities and the annealing schedule in the finite state space case. We will also present detailed numerical results for different choices of $U$. There are many possibilities for the method of discretization of (1.3); we will use a time discretization scheme that appeared in [3], but we will not present theoretical results on the time discretized dynamics; this is beyond the scope of this article. We refer the interested reader to [66] for a study on discretization schemes for the system (1.3), to [15] for a recent consideration on (1.2) and its time discretization, and to [27, 28] for linking discrete time Markov chains with the overdamped Langevin system in (1.1).

**1.1. Contributions and organization of the paper.** Here we summarize the main contributions of the paper.

- We provide a complete theoretical analysis of the simulated annealing algorithm for the generalized Langevin equation (1.3). The main theoretical contribution consists of Theorem 2.7 which establishes convergence in probability of $X_t$ in the higher order Markovian dynamics (1.3) to a global minimizer of $U$. For the optimal cooling schedule $T_t$, the rate of convergence is set as the known rate for the Langevin system (1.2) presented in [51].
- The initially non-Markovian property and pronounced degeneracy, in the sense of requiring a second commutator bracket for hypoellipticity by way of Hörmander, introduces additional difficulties that are overcome using techniques from [51]. As such, we use a different form of the distorted entropy that stated formally in (4.19). Additional technical improvements include a different truncation argument and a limiting sequence of nondegenerate SDEs for establishing dissipation of this distorted entropy.

These extensions also address certain technical issues in [51]; see Remarks 2.2, 4.2, and 4.11 for more details. Also we make an effort to emphasize the role of the critical factor of the cooling schedule in the rate of convergence in Theorem 2.7. This can be seen in our assumptions for $T_t$ and $U$.

- Numerical experiments are provided to illustrate the performance of our approach. We also discuss tuning issues. In particular, we investigate numerically the role of matrix $A$ and how it can be chosen to increase exploration of the state space. In regard to time discretization of (1.3) we use the leapfrog scheme of [3]. We compare this with a similar time discretization of (1.2) and observe that exploration of the state space is increased considerably.

The paper is organized as follows. Section 2 will present the assumptions and main theoretical results. Proofs can be found in section 4. Section 3 presents numerical results demonstrating the effectiveness of our approach in terms of reaching the global minimum. In section 5, we provide some concluding remarks.

**2. Main result.** Let $L_t$ denote the infinitesimal generator of the associated semigroup to (1.3) at $t > 0$ and temperature $T_t$. This is formally given by

$$L_t = (y \cdot \nabla_x - \nabla_x U(x) \cdot \nabla_y) + (z^\top \lambda \nabla_y - y^\top \lambda^\top \nabla_z) - T_t^{-1} z^\top A \nabla_z + A : D_z^2, \tag{2.1}$$

where we denote the gradient vector by $\nabla_x = (\partial_{x_1}, \dots, \partial_{x_n})^\top$, the Hessian by $D_x^2$, and the respective operators for the $y$ and $z$ variables similarly. For matrices $M, N \in \mathbb{R}^{r\times r}$ we denote $M : N = \sum_{i,j} M_{ij} N_{ij}$ for all $1 \le i, j \le r$ and denote the operator norm as $|M| = \sup\left\{ \frac{|Mv|}{|v|} : v \in \mathbb{R}^r \text{ with } v \neq 0 \right\}$. We will also use $|v|$ to denote Euclidean distance for a vector $v$. Let $m_t$ be the law of $(X_t, Y_t, Z_t)$ in (1.3), and, with slight abuse of notation, we will also denote as $m_t$ the corresponding Lebesgue density. Similarly, we define $\mu_{T_t}$ as the instantaneous invariant law of the process

$$\mu_{T_t}(dx, dy, dz) = \frac{1}{Z_{T_t}} \exp\left( -\frac{1}{T_t}\left( U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} \right)\right) dx\, dy\, dz \tag{2.2}$$

with $Z_{T_t} = \int \exp\left( -\frac{1}{T_t}(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2})\right) dx\, dy\, dz$. Finally, denote the density between the two laws as $h_t = \frac{dm_t}{d\mu_{T_t}}$. We proceed by stating our assumptions on the potential $U$.

*Assumption* 1. The function $U$ belongs in $\mathcal{C}^\infty(\mathbb{R}^n)$, and its second derivatives satisfy

$$|D_x^2 U|_\infty := \sup_{x\in\mathbb{R}^n} \max\left( \sup_{ij} |\partial_i \partial_j U(x)|, |D_x^2 U(x)| \right) < \infty. \tag{2.3}$$

Its first derivatives satisfy

$$\nabla_x U(x) \cdot x \ge r_1 |x|^2 - U_g, \tag{2.4}$$

$$|\nabla_x U(x)|^2 \le r_2 |x|^2 + U_g \tag{2.5}$$

for some constants $r_1, r_2 \in \mathbb{R}$, $U_g > 0$. Moreover, either

(a)

$$|\bar{a} \circ x|^2 + U_m \le U(x) \le |\bar{a} \circ x|^2 + U_M \tag{2.6}$$

for some $U_m, U_M \in \mathbb{R}$, $\bar{a} \in (0, \infty)^n$, where $\circ$ denotes the Hadamard product; or

(b)

- $U$ is a nonnegative Morse function, in the sense that there exists $1 \leq C_H < \infty$ such that if $x \in \mathbb{R}^n$ satisfies $\nabla_x U(x) = 0$, then

$$\frac{1}{C_H} \leq \|D_x^2 U(x)\| \leq C_H;$$

- $U$ is nondegenerate in the sense that
  - For any two local minima $m_i, m_j \in \mathbb{R}^n$, there exists a unique (communicating saddle) point $s_{i,j} \in \mathbb{R}^n$ such that
    - $\nabla_x U(s_{i,j}) = 0$,
    - $U(s_{i,j}) = \inf\{\max_{s\in[0,1]} U(\gamma(s)) : \gamma \in C([0,1], \mathbb{R}^n), \gamma(0) = m_i, \gamma(1) = m_j\}$,
    - the dimension of the unstable subspace of $D_x^2 U(s_{i,j})$ is equal to 1.
  - Setting $m_1$ to be the global minimum of $U$, we see there exists $\delta > 0$ and an ordering of the local minima $\{m_2, m_3, \dots\}$ such that $U(s_{1,2}) - U(m_2) \geq U(s_{1,i}) - U(m_i) + \delta$ for all $i \geq 3$.

Note that (2.4) and (2.5) imply

$$a_m|x|^2 + U_m \leq U(x) \leq a_M|x|^2 + U_M \tag{2.7}$$

for some $a_m, a_M > 0$, $U_m, U_M \in \mathbb{R}$. In the rest of the paper, if (2.6) holds, then the smallest and largest elements of $\bar{a}$ are denoted $a_m = \min_i \bar{a}_i$ and $a_M = \max_i \bar{a}_i$, respectively, where $\bar{a} = (\bar{a}_1, \dots, \bar{a}_n)$.

*Assumption* 2. The temperature $T_t$ satisfies $\lim_{t\to\infty} T_t = 0$.

Before we proceed with further assumptions on the annealing schedule $T_t$ and on the initial distribution, note that under Assumptions 1 and 2, a log-Sobolev inequality holds.

Proposition 2.1. *Under Assumptions 1 and 2, there exist constants $t_{ls}^{(0)}$, $\hat{E}$, and $A_*^{(0)} > 0$ and a finite order polynomial $r^{(0)} : (0,\infty) \to (0,\infty)$ with coefficients depending on $U$ such that for all $0 < h \in C^\infty(\mathbb{R}^{2n+m})$, satisfying $\int h d\mu_{T_t} = 1$, it holds that*

$$\int h \ln h d\mu_{T_t} \leq C_t^{(0)} \int \frac{|\nabla h|^2}{h} d\mu_{T_t}, \tag{2.8}$$

*where for $t > t_{ls}^{(0)}$,*

$$C_t^{(0)} = r^{(0)}\left(T_t^{-\frac{1}{2}}\right) e^{\hat{E}T_t^{-1}}. \tag{2.9}$$

The proof is deferred to section SM5 of the supplementary material, and the constant $\hat{E}$ from the above proposition will be used not only in stating the following assumption about $T_t$ but also in what follows. In the case of Assumption 1(a), $\hat{E}$ can be taken as $U_M - U_m$; otherwise, for Assumption 1(b) it is the critical depth [48] of $U$.

*Assumption* 3. The cooling schedule $T_\cdot : [0,\infty) \to (0,\infty)$ is continuously differentiable and bounded above, and there exists some constant $t_0 > 1$ such that $T_t$ satisfies, for all $t > t_0$,

(i) $T_t \geq E(\ln t)^{-1}$ for some constant $E > \hat{E} \geq 0$, where $\hat{E}$ is the constant in Proposition 2.1;

(ii) $|T_t'| \le \widetilde{T} t^{-1}$ for some constant $\widetilde{T} > 0$.

*Assumption* 4. The initial law $m_0$ admits a bounded density with respect to the Lebesgue measure on $\mathbb{R}^{2n+m}$, also denoted $m_0$, satisfying

(i) $m_0 \in \mathcal{C}^\infty(\mathbb{R}^{2n+m})$,
(ii) $\int \frac{|\nabla m_0|^2}{m_0} dxdydz < \infty$,
(iii) $\int (|x|^2 + |y|^2 + |z|^2) m_0 \, dxdydz < \infty$.

*Remark* 2.2. Note that (2.5) and (2.6) deviate from [51]. Condition (2.6) is useful for a self-contained exposition of the log-Sobolev constant in (4.28); it is satisfied, for instance, by a multivariate Gaussian after a rotation of the $x$ coordinates. The alternative condition that $U$ is a nondegenerate Morse function allows us to conveniently apply the results of [48], in which case $\hat{E}$ is given as the critical depth of $U$.

We present two key propositions.

Proposition 2.3. *Under Assumptions 1 and 3, for all $t > 0$, denote by $(X^{T_t}, Y^{T_t}, Z^{T_t})$ a random variable with distribution $\mu_{T_t}$. For any $\delta, \alpha > 0$, there exists a constant $\hat{A} > 0$ such that*

$$\mathbb{P}(U(X^{T_t}) > \min U + \delta) \le \hat{A} e^{-\frac{\delta - \alpha}{T_t}}$$

*holds for all $t > 0$.*

*Proof.* The result follows exactly as in Lemma 3 in [51]. ∎

Proposition 2.4. *Under Assumptions 1, 3, and 4, for all $t > 0$, $(X_t, Y_t, Z_t)$ are well defined as the unique strong solution to (1.3), $\mathbb{E}[|X_t|^2 + |Y_t|^2 + |Z_t|^2] < \infty$, and the law $m_t$ admits an everywhere positive density with respect to the Lebesgue measure on $\mathbb{R}^{2n+m}$.*

For the proof of Proposition 2.4, see Proposition 4.1 in section 4.

Proposition 2.3 can be thought of as a Laplace principle; Proposition 2.4 asserts that the process (1.3) does not blow up in finite time and that the noise in the dynamics (1.3) for $Z_t$ spreads throughout the system, that is to $X_t$ and $Y_t$.

Proposition 2.5. *Under Assumptions 1, 3, and 4, for any $0 < \alpha \le \frac{1}{2} - \frac{\hat{E}}{2E}$, there exists some constant $B > 0$ such that for all $t \ge 0$,*

$$\int h_t \ln h_t d\mu_{T_t} \le B \left(\frac{1}{t}\right)^{1 - \frac{\hat{E}}{E} - 2\alpha}. \tag{2.10}$$

The full proof is contained in section 4 and follows from Proposition 4.12. It uses an approximating sequence of SDEs, in which all of the elements have nondegenerate noise. The problem is split into the partial time and partial temperature derivatives where, among other tools, (4.23) and a log-Sobolev inequality are used as in [51] to arrive at a bound that allows a Grönwall-type argument.

*Remark* 2.6. Proposition 4.12 is a statement about the distorted entropy $H(t)$, which bounds the entropy $\int h_t \ln h_t d\mu_{T_t}$. In fact, this is achieved in such a way that the bound

becomes less sharp as $t$ becomes large but without consequences for our main theorem, Theorem 2.7.

We proceed with the statement of our main result, using $t_h$ from Proposition 2.5.

Theorem 2.7. *Under Assumptions 1, 2, 3, and 4, for any $\delta > 0$, as $t \to \infty$,*

$$\mathbb{P}(U(X_t) \leq \min U + \delta) \to 1.$$

*If in addition $T_t = E(\ln t)^{-1}$, then for any $0 < \alpha \leq \min\left(\frac{1}{2} - \frac{\hat{E}}{2E}, \delta\right)$, there exists a constant $C > 0$ such that for all $t \geq 0$,*

$$\mathbb{P}(U(X_t) > \min U + \delta) \leq C\left(\frac{1}{t}\right)^{r^e(E)},$$

*where the rate $r^e : (\hat{E}, \infty) \to \mathbb{R}$ is defined by*

$$\begin{aligned} r^e(E) &:= \min\left(\frac{1 - \frac{\hat{E}}{E} - 2\alpha}{2}, \frac{\delta - \alpha}{E}\right) \\ &= \begin{cases} \frac{1}{2}\left(1 - \frac{\hat{E}}{E} - 2\alpha\right) & \text{if } E < \frac{\hat{E} + 2(\delta - \alpha)}{1 - 2\alpha}, \\ \frac{\delta - \alpha}{E} & \text{otherwise.} \end{cases} \end{aligned}$$

*Proof.* For all $t > 0$, denote by $(X^{T_t}, Y^{T_t}, Z^{T_t})$ a random variable with distribution $\mu_{T_t}$. For all $\delta > 0$, with the definition of $h_t$ and the triangle inequality, we have

$$\mathbb{P}(U(X_t) > \min U + \delta) \leq \mathbb{P}(U(X^{T_t}) > \min U + \delta) + \int |h_t - 1| d\mu_{T_t}.$$

Pinsker's inequality gives

$$\int |h_t - 1| d\mu_{T_t} \leq \left(2 \int h_t \ln h_t d\mu_{T_t}\right)^{\frac{1}{2}}, \tag{2.11}$$

which, by Proposition 2.5 together with Proposition 2.3 gives the result. ∎

The cooling schedule $T_t = E(\ln t)^{-1}$ is optimal with respect to the method of proof for Proposition 4.12; see Proposition SM8.2. This is consistent with works in simulated annealing, e.g., [16, 29, 30, 32, 33, 34, 35, 39].

The 'mountain-like' shape of $r^e$ indicates the bottleneck for the rate of convergence at low and high values of $E$: a small $E$ means the convergence to the instantaneous equilibrium $\mu_{T_t}$ is slow, and a large $E$ means the convergence of $\mu_{T_t}$ to the global minima of $U$ is slow.

Although the focus in Theorem 2.7 is on decaying $T_t$, it is only for convergence to the global minimum where Assumption 2 is used. In particular, the convergence result in Proposition 2.5 is valid for temperature schedules that are not converging to zero. This includes the instance of using a variable temperature in order to tackle the problem of metastability in the sampling problem.

**3. Numerical results.** Here we investigate the numerical performance of (1.3) in terms of convergence to a global optimum and of exploration capabilities and make comparisons with (1.2). The details of the discretizations we use for both sets of dynamics and some details related to the annealing schedule and parameters can be found in supplementary section SM7.1. Rates of transition between different regions of the state space can also be found in supplementary section SM7.2. In section 3.1, for different parameters and cost functions, we present results for the probability of convergence to the global minimum. We investigate the effect of $E$ appearing in the annealing schedule and also study the effect of the parameters in the dynamics (1.2) and (1.3). In particular, we consider different $\lambda = \bar{\lambda}\lambda_i$ and $A = \mu A_i$ in the generalized Langevin dynamics for $\bar{\lambda}, \mu > 0$; the specific forms of $\lambda_i$ and $A_i$ are given in supplementary section SM7.1. Note that $\mu$ is used also as the friction parameter in (1.2), which makes notational sense because $\mu$ determines the relative strength of the Ornstein–Uhlenbeck part of the respective dynamics. In addition, we introduce a coefficient $\gamma > 0$ in front of the terms in (1.2) and (1.3) corresponding to the part in the respective generators given by $y \cdot \nabla_x - \nabla_x U(x) \cdot \nabla_y$ (see supplementary section SM7.1 for details); unless otherwise stated, we keep $\gamma = 1$.

**3.1. Performance and tuning.** As expected, the tuning parameters $E$, $\bar{\lambda}$, and $\mu$ play significant roles in the performance of the discretizations. As $E$ is common to both (1.2) and (1.3), we wish to demonstrate that the additional tuning variable for the generalized Langevin dynamics will improve performance.

We first comment on relative scaling of $\bar{\lambda}$ and $\mu$ based on earlier work for quadratic $U$ and $T_t = T$ being constant. A quadratic $U$ satisfies the bounds in Assumption 1 and is of particular interest because analytical calculations are possible for the spectral gap of $L_t$, which in turn gives the (exponential) rate of convergence to the equilibrium distribution. It is observed numerically in [57] that in this case, (1.3) has a spectral gap that is approximately a function of $\frac{\bar{\lambda}^2}{\mu}$. On the other hand, the spectral gap of (1.2) with quadratic $U$ is a function of $\mu$ thanks to Theorem 3.1 in [49]. For the rest of the comparison, we will use $\frac{\bar{\lambda}^2}{\mu}$ and $\mu$ as variables for the respective discretizations as these quantities appear to have a distinct effect on the mixing in each case. We mention that these choices of variables also allow one to adjust the global Lipschitz constant of the drift coefficient for free in the generalized Langevin equation (1.3) up to that of $\nabla U$ and 1, while in (1.2), this grows as $\mu$ grows. Therefore one can expect to be able to take a stepsize in the simplest (Euler–Maruyama) discretization of (1.3) that is at least that of (1.2); note, however, that such benefits related to the stepsize and the Lipschitz constant disappear for more commonly used schemes as in [41]. A detailed stability analysis is beyond the scope of this paper and in the following comparisons we do not mention any influence from the numerical discretizations on the continuous time dynamics.

We will mainly consider the popular Alpine function in 12 dimensions (see supplementary Table SM2; $\nabla U_1$ here is a subgradient), with additional cases presented in supplementary section SM7.3, setting $\Delta t = 0.02$ (see supplementary section SM7.1). Note the Alpine function does not strictly satisfy Assumption 1, but since drift conditions for Lyapunov functions are typically available even for weakly growing potentials [18] for the dynamics considered here, the trajectories are expected (and are observed) to remain, in a loose sense, close to 0. Therefore

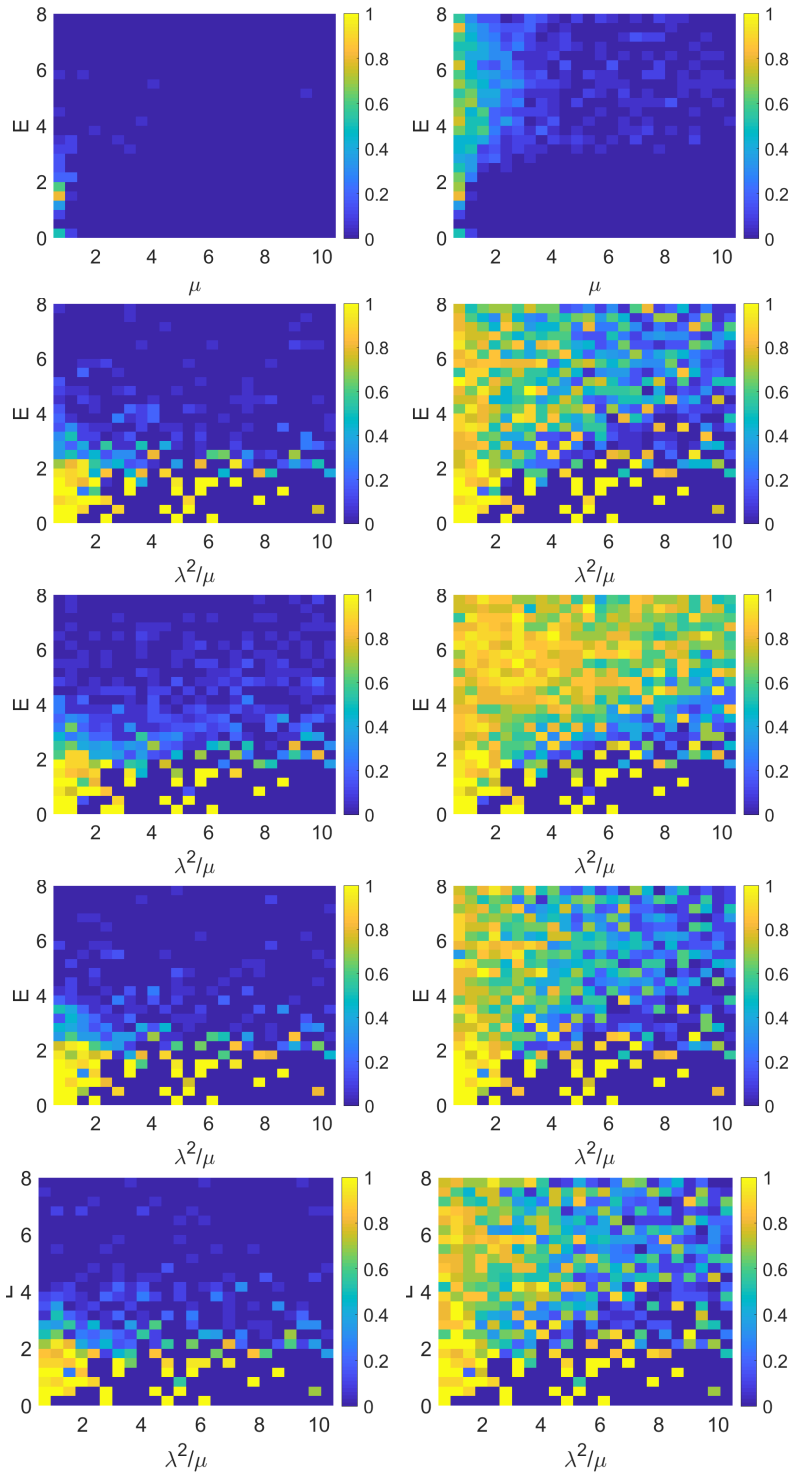


**Figure 3.1.** *Proportion of simulations close to the global minimum for the Alpine function as $U$. Panels from top to bottom: Langevin (SM7.2); generalized Langevin (SM7.1) with $A = A_1, A_2, A_3, A_4$. Left: Final position. Right: time average of last 5000 iterations. We use $\gamma = 3$ for improving visualisation; similar comparisons hold for $\gamma = 1$.*

we may mollify or modify the behavior at infinity of $U$ to satisfy Assumption 1 with no real observable consequence.

We will initialize at a point well separated from the global minimum and consider each method to be successful if, at the end of the simulation, either the endpoint or an average of the last points is contained within a tolerance region around the global minimum.

In Figure 3.1 we present proportions of 20 independent simulations converging at the region near the global minimum for $U = U_1$ (see supplementary Table SM2) depending on $E$ and $\mu$ for the discretization of the Langevin dynamics and on $E$ and $\frac{\bar{\lambda}^2}{\mu}$ for that of the generalized Langevin dynamics based on the discussion above. Each simulation is run for $k = 5 \cdot 10^4$ iterations. The left panels of Figure 3.1 are based on the final state, and the right panels are based on an average of the positions (of $X$) over the last 5000 iterations. In this example it is clear empirically that the generalized Langevin dynamics results in a higher probability of reaching the global minimum. Another interesting observation is that for the generalized Langevin dynamics, good performance is more robust to the chosen value of $E$. In this example, this means that adding an additional tuning variable and scaling $\mu$ proportional to $\bar{\lambda}^2$ make it easier to find a configuration of the parameters $E, \mu, \bar{\lambda}$ that lead to good performance, compared to using the Langevin dynamics and tuning $E, \mu$. It's also worth noting the cases of small $E$ where the generalized Langevin dynamics performs significantly better than the Langevin dynamics in the top plots and even better than the case of the same dynamics and larger $E$. This is an improvement that is not completely encapsulated by the analytic results here; it indicates that the deterministic dynamics ($E = 0$) can be inherently much more successful at climbing out of local minima, which translates into better convergence rates in the $E > 0$ cases.

The selection $A = A_2$, shown as the third row in each column of Figure 3.1 and Figures SM3 and SM4 in the supplementary material, does not satisfy the, probably superfluous, symmetry assumption as stated in the introduction, but it is noteworthy that the performance varies to such a large extent for different $U$ and that any optimality of $A$, which we leave for future work, could change depending on whether or not the symmetry assumption is in place.

## 4. Proofs.

### 4.1. Notation and preliminaries.

Unless stated otherwise, $\partial_t$ is used to denote the partial derivative with respect to $t$ with $T_t$ fixed (whenever its operand depends on $T_t$), whereas $\frac{d}{dt}$ denotes the full derivative in $t$. In addition, $\nabla$ denotes the gradient in $\mathbb{R}^{2n+m}$ space, and $d\zeta$ will be used for the Lebesgue measure on $\mathbb{R}^{2n+m}$. The notation $\mathbb{1}_S$ will be used for the indicator function on the set $S$.

For all $k > 0$, recall the standard mollifier $\varphi : \mathbb{R} \to \mathbb{R}$ and the associated $\varphi_k : \mathbb{R} \to \mathbb{R}$ to be given by

$$\varphi_k(x) := \frac{1}{k}\varphi\left(\frac{x}{k}\right), \qquad \varphi(x) := \begin{cases} e^{\frac{1}{x^2-1}}\left(\int_{-1}^{1} e^{\frac{1}{y^2-1}} dy\right)^{-1} & \text{if } -1 < x \le 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

For existence and uniqueness of (1.3), we will use the setting in [63]. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space, and let $\mathcal{F}_t$, $t \in [0, \infty)$ be a normal filtration. Here $(W_t)_{t \ge 0}$ is a standard Wiener process on $\mathbb{R}^m$ with respect to $\mathcal{F}_t$, $t \in [0, \infty)$.

The formal[2] $L^2(\mu_{T_t})$-adjoint $L_t^*$ of $L_t$ is given by

$$L_t^* = -(y\cdot\nabla_x - \nabla_x U(x)\cdot\nabla_y) - (z^\top\lambda\nabla_y - y^\top\lambda^\top\nabla_z) - T_t^{-1}z^\top A\nabla_z + A : D_z^2. \tag{4.2}$$

Let $\epsilon \geq 0$, and consider the perturbed system

$$dX_t^\epsilon = Y_t^\epsilon\,dt + \epsilon(-T_t^{-1}\nabla_x U(X_t^\epsilon)\,dt + dW_t^1), \tag{4.3a}$$

$$dY_t^\epsilon = -\nabla_x U(X_t^\epsilon)\,dt + \lambda^\top Z_t^\epsilon\,dt + \epsilon(-T_t^{-1}Y_t^\epsilon\,dt + dW_t^2), \tag{4.3b}$$

$$dZ_t^\epsilon = -\lambda Y_t^\epsilon\,dt - T_t^{-1}AZ_t^\epsilon\,dt + \Sigma\,dW_t^3, \tag{4.3c}$$

with $(X_0^\epsilon, Y_0^\epsilon, Z_0^\epsilon) = (X_0, Y_0, Z_0)$ restricted as in Assumption 4, where $W_t^1, W_t^2, W_t^3$ are independent $n$-dimensional and $m$-dimensional Wiener processes. As before, the law and density of (4.3a) will be denoted by $m_t^\epsilon$ along with $h_t^\epsilon = \frac{dm_t^\epsilon}{d\mu_{T_t}}$. Let the linear differential operators $S_t^x$, $S_t^y$ and their respective formal $L^2$-adjoints $S_t^{x\top}$ and $S_t^{y\top}$ be given by

$$\begin{aligned} S_t^x &= -T_t^{-1}\nabla_x U\cdot\nabla_x + \Delta_x, & S_t^y &= -T_t^{-1}y\cdot\nabla_y + \Delta_y,\\ S_t^{x\top} &= T_t^{-1}\nabla_x U\cdot\nabla_x + T_t^{-1}\Delta_x U + \Delta_x, & S_t^{y\top} &= T_t^{-1}y\cdot\nabla_y + T_t^{-1}n + \Delta_y. \end{aligned}$$

Note that the formal $L^2(\mu_{T_t})$-adjoints of $S_t^x$ and $S_t^y$ coincide with $S_t^x$ and $S_t^y$, so that the generator, denoted $L_t^\epsilon$, associated to (4.3a) and its formal $L^2(\mu_{T_t})$-adjoint are given by the formal operators

$$L_t^\epsilon = L_t + \epsilon(S_t^x + S_t^y), \qquad L_t^{\epsilon *} = L_t^* + \epsilon(S_t^x + S_t^y).$$

For any $\phi \in \mathcal{C}^\infty$ and $f : \mathbb{R}^{2n+m} \to \mathbb{R}$ smooth enough,

$$L_t^\epsilon(\phi(f)) = \phi'(f)L_t^\epsilon(f) + \phi''(f)\Gamma_t^\epsilon(f), \tag{4.4}$$

where $\Gamma_t^\epsilon$ is the carré du champ operator for $L_t^\epsilon$ given by

$$\Gamma_t^\epsilon(f) = \frac{1}{2}L_t^\epsilon(f^2) - fL_t^\epsilon(f) = \nabla f\cdot(A^\epsilon\nabla f), \tag{4.5}$$

$A^\epsilon \in \mathbb{R}^{(2n+m)\times(2n+m)}$ denotes the matrix with entries

$$A_{ij}^\epsilon := \begin{cases} \epsilon & \text{if } 1 \leq i = j \leq 2n,\\ A_{i-2n,j-2n} & \text{if } 2n+1 \leq i,j \leq 2n+m,\\ 0 & \text{otherwise,} \end{cases}$$

and $A_{i,j}$ denotes the $(i,j)$th entry of $A$. Let $\mathcal{C}_+^\infty = \{f \in \mathcal{C}^\infty : f > 0\}$. For $\Phi : \mathcal{C}_+^\infty \to \mathcal{C}^\infty$ differentiable in the sense that for any $f \in \mathcal{C}_+^\infty$, $g \in \mathcal{C}^\infty$,

$$(d\Phi(f).g)(\zeta) := \lim_{s\to 0}\frac{(\Phi(f+sg))(\zeta) - (\Phi(f))(\zeta)}{s}$$

[2]See, for instance, Appendix B in [23]. In the present paper the infinitesimal generators and their adjoints are considered as honest differential operators acting on smooth functions.

exists for all $\zeta \in \mathbb{R}^{2n+m}$, the $\Gamma_\Phi$ operator for $L_t^{\epsilon *}$ is defined by

$$\Gamma_{L_t^{\epsilon *},\Phi}(h) := \frac{1}{2}(L_t^{\epsilon *}\Phi(h) - d\Phi(h).(L_t^{\epsilon *}h)). \tag{4.6}$$

As is well known, $L_t^{\epsilon *}$ does not satisfy the standard chain and product rules due to the additional term from the second derivatives in $L_t^{\epsilon *}$; straightforward calculations give

$$L_t^{\epsilon *}(\psi(f)) = \psi'(f)L_t^{\epsilon *}f + \psi''(f)\nabla f \cdot (A^\epsilon \nabla f) \tag{4.7}$$

$$L_t^{\epsilon *}(fg) = fL_t^{\epsilon *}(g) + gL_t^{\epsilon *}(f) + \nabla f \cdot (2A^\epsilon \nabla g) \tag{4.8}$$

for all $f, g \in \mathcal{C}^\infty$ and $\psi \in \mathcal{C}^\infty$. Note $\nabla f \cdot (A^\epsilon \nabla f)$ and $\nabla f \cdot (2A^\epsilon \nabla g)$ are, respectively, the *carré du champ* and its symmetric bilinear operator via polarization for $L_t^{\epsilon *}$.

In addition, for a scalar-valued $D_1$ and a vector-valued operator $D_2$ both acting on scalar-valued functions, denote the commutator bracket as follows:

$$[D_1, D_2]h = (D_1(D_2h)_1 - (D_2D_1h)_1, \ \ldots, D_1(D_2h)_{d_{D_2}} - (D_2D_1h)_{d_{D_2}}) \tag{4.9}$$

for $h \in \mathcal{C}^\infty$, where $d_{D_2} \in \mathbb{N}$ is the number of elements in the output of $D_2$.

**4.2. Auxiliary results.** For the next result, the space of smooth functions to be used is that from [14]: let $\mathcal{C}_{b,c}^\infty = \mathcal{C}_{b,c}^\infty((0,\infty)\times\mathbb{R}^{2n+m})$ be the space of real-valued functions $f : (0,\infty)\times\mathbb{R}^{2n+m} \to \mathbb{R}$ such that

1. $f$ is measurable with respect to $\mathcal{B}((0,\infty)) \otimes \mathcal{B}(\mathbb{R}^{2n+m})$,
2. for all $t > 0$, $f(t,\cdot)$ is smooth and $f$ is bounded on compact subsets of $\mathbb{R}_{>0} \times \mathbb{R}^{2n+m}$.

Proposition 4.1. *Under Assumptions 1, 3, and 4, for all $t > 0$ and $\epsilon \geq 0$, the unique strong solution $(X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon)$ to (4.3) is well defined and there exists some constant $\kappa > 0$ such that*

$$\mathbb{E}[|X_t^\epsilon|^2 + |Y_t^\epsilon|^2 + |Z_t^\epsilon|^2] \leq e^{\kappa t}\mathbb{E}[|X_0|^2 + |Y_0|^2 + |Z_0|^2] < \infty. \tag{4.10}$$

*Furthermore, for all time $t > 0$, the law of the process $(X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon)$*

- *admits an almost-everywhere finite strictly positive density, also denoted $m_t^\epsilon$, with respect to the Lebesgue measure on $\mathbb{R}^{2n+m}$;*
- *is the unique integrable distributional solution to*

$$\begin{cases} \partial_t m_t^\epsilon = (L_t^\top + \epsilon(S_t^{x\top} + S_t^{y\top}))m_t^\epsilon, \\ m_0^\epsilon = m_0, \end{cases} \tag{4.11}$$

  *where $L_t^\top$ is the formal $L^2$-adjoint of $L_t$.*

*Finally, when $\epsilon > 0$, $m_\bullet$ and its partial derivative in time belong in $\mathcal{C}_{b,c}^\infty$.*

For the notion of integrable distributional solutions, see page 338 in [7].

*Proof.* Existence and uniqueness of an almost surely continuous $\mathcal{F}_t$-adapted processes follows by conditions (2.3) and (2.5) using Theorem 3.1.1 in [63]; in addition, (4.10) holds by the same theorem. For the claim that the law admits a density, we will apply Theorem 1 in

[36] for the case of an arbitrary deterministic starting point. First, condition (H1) in the same article is verified. Take the sets '$K_n$' to be

$$K_p = \prod_{i=1}^{2n+m} [-p, p]$$

for all $p \in \mathbb{N}$. The unique solution to (4.3a) with a deterministic starting point $(X_0, Y_0, Z_0) = (x_0, y_0, z_0) \in \mathbb{R}^{2n+m}$ satisfies the same bound (4.10) as before when initializing from $m_0$. Moreover, for the random sets

$$\Xi_p = \{s > 0 : (X_u^\epsilon, X_u^\epsilon, X_u^\epsilon) \in K_p, 0 \le u \le s\},$$

for $p \in \mathbb{N}$, the solution $(\hat{X}_t^\epsilon, \hat{Y}_t^\epsilon, \hat{Z}_t^\epsilon)$ to the stopped stochastic differential equation

$$d\hat{X}_t^{\epsilon,p} = \mathbb{1}_{\Xi_p}(t)(\hat{Y}_t^{\epsilon,p}\, dt + \epsilon(-T_t^{-1}\nabla_x U(\hat{X}_t^{\epsilon,p})\, dt + dW_t^1)), \tag{4.12a}$$

$$d\hat{Y}_t^{\epsilon,p} = \mathbb{1}_{\Xi_p}(t)(-\nabla_x U(\hat{X}_t^{\epsilon,p})\, dt + \lambda^\top \hat{Z}_t^{\epsilon,p}\, dt + \epsilon(-T_t^{-1}\hat{Y}_t^{\epsilon,p}\, dt + dW_t^2)), \tag{4.12b}$$

$$d\hat{Z}_t^{\epsilon,p} = \mathbb{1}_{\Xi_p}(t)(-\lambda\hat{Y}_t^{\epsilon,p}\, dt - T_t^{-1}A\hat{Z}_t^{\epsilon,p}\, dt + \Sigma\, dW_t^3), \tag{4.12c}$$

is well defined by Theorem 3.1.1 in [63], and the corresponding bound

$$\mathbb{E}[|\hat{X}_t^{\epsilon,p}|^2 + |\hat{Y}_t^{\epsilon,p}|^2 + |\hat{Z}_t^{\epsilon,p}|^2] \le e^{\kappa t}(|x_0|^2 + |y_0|^2 + |z_0|^2) < \infty$$

holds. Identifying $(\hat{X}_t^{\epsilon,p}, \hat{Y}_t^{\epsilon,p}, \hat{Z}_t^{\epsilon,p}) = (X^\epsilon_{t\wedge\sup\Xi_p}, Y^\epsilon_{t\wedge\sup\Xi_p}, Z^\epsilon_{t\wedge\sup\Xi_p})$ almost surely yields that[3] for any $\tau > 0$,

$$\begin{aligned}\mathbb{P}(\inf\{t \ge 0 : (X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon) \notin K_p\} \le \tau) &\le \frac{1}{p^2}\mathbb{E}[|X^\epsilon_{\tau\wedge\sup\Xi_p}|^2 + |Y^\epsilon_{\tau\wedge\sup\Xi_p}|^2 + |Z^\epsilon_{\tau\wedge\sup\Xi_p}|^2] \\ &\le \frac{e^{\kappa\tau}}{p^2}(|x_0|^2 + |y_0|^2 + |z_0|^2)\end{aligned}$$

and, in particular, that for any $\tau > 0$,

$$\mathbb{P}(\inf\{t \ge 0 : (X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon) \notin K_p\} \le \tau) \to 0 \quad \text{as } p \to \infty. \tag{4.13}$$

Suppose for contradiction that with nonzero probability, the increasing-in-$p$ random variable $\inf\{t \ge 0 : (X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon) \notin K_p\}$ converges to a real value as $p \to \infty$. Then there exists a time $\hat{\tau} > 0$ such that with nonzero probability,

$$\inf\{t \ge 0 : (X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon) \notin K_p\} \le \hat{\tau} \quad \forall p \in \mathbb{N},$$

which contradicts (4.13). Therefore condition (H1) in [36] holds for (4.3a). Condition (H2) in the same article holds due to the $K_p$ being compact and to the smoothness assumption on $U$. It can be readily checked that the local weak Hörmander condition (LWH) in [36] also

[3] Alternatively, Corollary 1.2 of section 5 in [24] can be used.

holds at any $(t, y_0)$ for any $r \in (0,t)$ and $R > 0$. Therefore by Theorem 1 in [36], due to our Assumptions 1 and 3, the solution to (4.3a) with a deterministic starting point $\zeta_0 \in \mathbb{R}^{2n+m}$ admits a smooth density $p_t^{\zeta_0} \in \mathcal{C}^\infty(\mathbb{R}^{2n+m})$ for all $t > 0$. Moreover by Theorem 2 in [36], for any fixed $\zeta \in \mathbb{R}^{2n+m}$, $\mathbb{R}^{2n+m} \ni \zeta_0 \mapsto p_t^{\zeta_0}(\zeta)$ is lower semicontinuous and hence measurable, so that the $\mathbb{R} \cup \{\pm\infty\}$-valued function on $\mathbb{R}^{2n+m}$,

$$\int_{\mathbb{R}^{2n+m}} p_t^{\zeta_0} m_0(d\zeta_0), \tag{4.14}$$

is integrable by Fubini's theorem and so is almost-everywhere $\mathbb{R}$-valued on $\mathbb{R}^{2n+m}$. By Itô's rule, (4.14) solves (4.11) in the distributional sense. In addition, (4.11) is the unique integrable solution by Theorem 9.6.3 in [7], which requires for any $T > 0$ that there exist $V \in C^2(\mathbb{R}^{2n+m})$ such that

1. $V(x) \to \infty$ as $|x| \to \infty$, and
2. for some constant $C_V > 0$ and all $(x,t) \in \mathbb{R}^{2n+m} \times (0,T)$, it holds that $L_t^\epsilon V \geq -C_V V$ and $|\nabla V| \leq C_V V$.

Setting $V(x,y,z) = 1 + U(x) - U_m + \frac{|y|^2}{2} + \frac{|z|^2}{2}$ and calculating

$$L_t^\epsilon\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right) = \epsilon\left(-\frac{1}{T_t}|\nabla_x U|^2 + \Delta_x U - \frac{1}{T_t}|y|^2 + n\right) - \frac{1}{T_t} z^\top A z + \mathrm{Tr} A, \tag{4.15}$$

it is clear from assumptions (2.3) and (2.5) and either (2.6) or (2.7) on $U$ that these conditions are satisfied since $T$ is finite; therefore there is a unique integrable solution to (4.11) in the sense of the definition on page 338 in [7]. The expression in (4.14) is thus the density for the law of the solution to (4.3a) with initial law $m_0$ at time $t$.

For $\epsilon > 0$, the time-dependent law of $(X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon)$ and its partial derivative with respect to time belongs in $\mathcal{C}_{b,c}^\infty$ by Theorem 1.1 in [14] because (4.14) satisfies (4.11).

For positivity of the density where $\epsilon = 0$, the steps in Lemma 3.4 of [47] involving the solution to an associated control problem can be followed. A detailed proof can be found in section SM1 of the supplementary material. ∎

*Remark* 4.2. For smoothness of the density, the results in [70] can also be considered, but there the assumptions are slightly mismatched. First, the statement assumes boundedness of $\partial^\alpha V$ for any multi-index $\alpha$, where $V$ would in this case be any of the coefficients appearing in (1.3), which fails for $|\alpha| = 0$. Second in case of (A.1) in [70], condition (i) fails, and in the case of (A.2), condition (i) fails due to $V_0$. Both of these assumptions seem possibly unnecessary in the proofs, but we avoid this in favor of the more recent work [36].

The results below up to Proposition 4.12 are directed towards showing dissipation of a distorted entropy as required in the proof of Theorem 2.7.

### 4.3. Lyapunov function.

Lemma 4.3. *Under Assumptions 1, 3, and 4, there exist constants $a, b, c, d, \delta > 0$ independent of $\epsilon$ such that $R : \mathbb{R}^{2n+m+1} \to \mathbb{R}$ defined as*

$$R(x,y,z,T_t) := U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} + \delta T_t\left(y^\top \lambda^{-1} z + \frac{1}{2} x \cdot y\right) \tag{4.16}$$

*satisfies*

$$a(|x|^2+|y|^2+|z|^2)-d \le R(x,y,z,T_t) \le b(|x|^2+|y|^2+|z|^2)+d, \tag{4.17}$$

*and there exists* $0<\epsilon' \le 1$ *for which* $\epsilon \le \epsilon'$ *implies*

$$L_t^\epsilon R \le -cT_t R + \frac{d}{T_t}. \tag{4.18}$$

*Proof.* By the quadratic assumption (2.7) on $U$ and boundedness Assumption 3 on $T_t$, it is clear that there exists $\hat{\delta}>0$ such that the first statement (4.17) holds with $d=\max(|U_m|,|U_M|)$ for all $\delta \in (0,\hat{\delta}]$. Inequality (4.18) follows by a calculation using our assumptions on $U$, $T_t$ and applications of Young's inequality. A detailed proof can be found in section SM2 of the supplementary material. ∎

Lemma 4.4. *Under Assumptions* 1, 3, *and* 4 *and for* $0 \le \epsilon \le \epsilon'$, *the solution* $(X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon)$ *to* (4.3) *is such that* $\frac{\mathbb{E}[R(X_t^\epsilon,Y_t^\epsilon,Z_t^\epsilon,T_t)]}{(\ln(e+t))^2}$ *is bounded uniformly in time* $t$ *and in* $\epsilon$.

*Proof.* See section SM3 in the supplementary material. ∎

Corollary 4.5. *Under Assumptions* 1, 3, *and* 4 *and for* $0 \le \epsilon \le \epsilon'$, *the solution* $(X_t^\epsilon, Y_t^\epsilon, Z_t^\epsilon)$ *to* (4.3) *is such that* $\frac{\mathbb{E}[|X_t^\epsilon|^2+|Y_t^\epsilon|^2+|Z_t^\epsilon|^2]}{(\ln(e+t))^2}$ *is bounded uniformly in time and in* $\epsilon$.

*Proof.* By the lower bound on $R$ in (4.17),

$$\mathbb{E}[|X_t^\epsilon|^2+|Y_t^\epsilon|^2+|Z_t^\epsilon|^2] \le \mathbb{E}\left[\frac{R(X_t^\epsilon,Y_t^\epsilon,Z_t^\epsilon,T_t)+d}{a}\right],$$

which concludes the proof by Lemma 4.4. ∎

**4.4. Form of distorted entropy.** For $\epsilon \ge 0$, let $H^\epsilon(t)$ be the distorted entropy

$$H^\epsilon(t) = \int \Bigg( \frac{|2\nabla_x h_t^\epsilon + 8S_0(\nabla_y h_t^\epsilon + \lambda^{-1}\nabla_z h_t^\epsilon)|^2}{h_t^\epsilon} + \frac{|\nabla_y h_t^\epsilon + S_1\lambda^{-1}\nabla_z h_t^\epsilon|^2}{h_t^\epsilon} + \beta(T_t^{-1}) h_t^\epsilon \ln(h_t^\epsilon) \Bigg) d\mu_{T_t}, \tag{4.19}$$

where $S_0, S_1 > 0$ are the constants

$$S_0 := (1+|D_x^2 U|_\infty^2)^{\frac{1}{2}}, \qquad S_1 := 2+28S_0^2+1024S_0^4, \tag{4.20}$$

and $\beta$ is a second order polynomial (see (4.21) and the end of the proof for Proposition 4.7) to be determined by Proposition 4.7 and independent of $\epsilon$.

*Remark* 4.6. This particular expression for $H$ is not necessarily the best possible choice. However, the above is a working expression, and optimality is left as future work; see also [60].

Using Lemma SM4.1 from the supplementary material, the following proposition shows that the distorted entropy (4.19) is useful .

Proposition 4.7. *There exist* $\beta_0, \beta_1, \beta_2 > 0$ *independent of* $\epsilon$ *such that for* $\beta : \mathbb{R} \to \mathbb{R}$ *given by*

$$\beta(x) := 1+\beta_0+\beta_1 x+\beta_2 x^2, \tag{4.21}$$

*the operator* $\Psi_{T_t}$,

$$\Psi_{T_t}(h) := \frac{|2\nabla_x h + 8S_0(\nabla_y h + \lambda^{-1}\nabla_z h)|^2}{h} + \frac{|\nabla_y h + S_1\lambda^{-1}\nabla_z h|^2}{h} + \beta(T_t^{-1})h\ln(h), \tag{4.22}$$

*for* $h \in \mathcal{C}_+^\infty$ *satisfies*

$$\Gamma_{L_t^{\epsilon*},\Psi_{T_t}}(h) \geq \frac{|\nabla h|^2}{h} \tag{4.23}$$

*for all* $0 \leq \epsilon \leq 1$.

*Remark* 4.8. $\beta_0$, $\beta_1$, $\beta_2$ depend on $\hat{\lambda}^2 := \max(|\lambda|^2, |\lambda^\top|^2, |\lambda^{-1}|^2, |\lambda^{-1}||\lambda^\top|)$, $|D_x^2 U|_\infty$, and $|A|$. $H$ satisfying property (4.23) is crucial for proving dissipation in Proposition 4.12.

*Proof.* Let $\Phi_1$,$\Phi_2$,$\Phi_3$ be the terms in $\Psi_{T_t}$,

$$\Phi_1(h) := \frac{|2\nabla_x h + 8S_0(\nabla_y h + \lambda^{-1}\nabla_z h)|^2}{h}, \quad \Phi_2(h) := \frac{|\nabla_y h + S_1\lambda^{-1}\nabla_z h|^2}{h}, \quad \Phi_2(h) := h\ln(h). \tag{4.24}$$

Note that the $\Gamma_\Phi$ operator is linear in the $\Phi$ argument by linearity of $L_t^{\epsilon*}$, so that (4.23) can be written as $\Gamma_{L_t^{\epsilon*},\Phi_1}(h) + \Gamma_{L_t^{\epsilon*},\Phi_2}(h) + \beta(T_t^{-1})\Gamma_{L_t^{\epsilon*},\Phi_3}(h) \geq \frac{|\nabla h|^2}{h}$. Consider $\Gamma_{L_t^{\epsilon*},\Phi_3}$ first. Using the definition (4.6) of $\Gamma_{L_t^{\epsilon*},\Phi}$, the product and chain rule (4.8) and (4.7) for $L_t^{\epsilon*}$, and the coercivity property of $A$, we get

$$\begin{aligned}\Gamma_{L_t^{\epsilon*},\Phi_3}(h) &= \frac{1}{2}\left((\ln h + 1)L_t^{\epsilon*}h + \frac{1}{h}\nabla h\cdot(A^\epsilon\nabla h) - (1+\ln h)L_t^{\epsilon*}h\right)\\ &= \frac{1}{2h}\nabla h\cdot(A^\epsilon\nabla h) \geq \frac{1}{2h}(\epsilon|\nabla_x h|^2 + \epsilon|\nabla_y h|^2 + A_c|\nabla_z h|^2).\end{aligned} \tag{4.25}$$

Since the goal is to show (4.23), the availability of (4.25) counteracts any negative contributions in the $z$-derivative term, and any order $\epsilon$ contributions in the $x$- and $y$-derivatives, from $\Gamma_{L_t^{\epsilon*},\Phi_1}$ and $\Gamma_{L_t^{\epsilon*},\Phi_2}$; this counterweight materializes as $\beta$.

For $\Gamma_{L_t^{\epsilon*},\Phi_1}$ and $\Gamma_{L_t^{\epsilon*},\Phi_2}$, $S_0 > 0$ and $S_1 > 0$ as in (4.20) are used. Detailed derivations for the following inequalities are explicitly stated in section SM4 of the supplementary material; by use of Lemma SM4.1, repeated applications of Young's inequality, and the definition of $\hat{\lambda}$ (see Remark 4.8), it holds that

$$\begin{aligned}h\Gamma_{L_t^{\epsilon*},\Phi_2}(h) > &-|\nabla_x h|^2 + (1 + 28S_0^2 + 1024S_0^4)|\nabla_y h|^2\\ &- \frac{1}{2}\hat{\lambda}^2\left(3 + 2S_1 + S_1^2 + 3S_1^4 + S_1^2T_t^{-1}\left(|A| + \frac{\epsilon}{2}\right) + 3S_1^2T_t^{-2}|A|^2\right)|\nabla_z h|^2.\end{aligned} \tag{4.26}$$

The last term $\Gamma_{L_t^{\epsilon*},\Phi_1}$ compensates for the negative $x$-derivative:

$$\begin{aligned}h\Gamma_{L_t^{\epsilon*},\Phi_1}(h) \geq &\left(2 - 4(2 + (1+4S_0^2)S_0)\epsilon T_t^{-1}\right)|\nabla_x h|^2 + \Big(S_0^2(-28 - 1024S_0^2)\\ &- 8S_0(1+5S_0)\epsilon T_t^{-1}\Big)|\nabla_y h|^2 - \Big(S_0^2\hat{\lambda}^2(160 + 128T_t^{-2}|A|^2 + 1024S_0^2)\\ &+ 8S_0\hat{\lambda}^2(1+4S_0)\epsilon T_t^{-1}\Big)|\nabla_z h|^2.\end{aligned}$$

Matching powers in $T_t^{-1}$ to take

$$\beta_0 = \frac{1}{A_c}\left(S_0^2\hat{\lambda}^2(160+1024S_0^2)+\frac{1}{2}\hat{\lambda}^2(3+2S_1+S_1^2+3S_1^4)\right),$$
$$\beta_1 = \frac{1}{A_c}\left(4(2+(1+4S_0^2)S_0)+8S_0(1+5S_0)+8S_0\hat{\lambda}^2(1+4S_0)+\frac{1}{2}\hat{\lambda}^2\left(S_1^2\left(|A|+\frac{1}{2}\right)\right)\right),$$
$$\beta_2 = \frac{1}{A_c}\left(128S_0^2\hat{\lambda}^2|A|^2+\frac{3}{2}\hat{\lambda}^2S_1^2|A|^2\right),$$

using $\epsilon \leq 1$ and putting together the bounds for $\Gamma_{L_t^{\epsilon^*},\Phi_3}, \Gamma_{L_t^{\epsilon^*},\Phi_2}, \Gamma_{L_t^{\epsilon^*},\Phi_1}$ gives (4.23). ∎

### 4.5. Log-Sobolev inequality.

Proposition 4.9. *Under Assumptions 1 and 2 and for $\epsilon \geq 0$, there exist constants $t_{ls}, A_* > 0$ and a finite order polynomial $r : (0,\infty) \to (0,\infty)$ with coefficients depending on $U$ and $\lambda$ but independent of $\epsilon$ such that the distorted entropy (4.19) satisfies*

$$H^\epsilon(t) \leq C_t \int \frac{|\nabla h_t^\epsilon|^2}{h_t^\epsilon} d\mu_{T_t}, \tag{4.27}$$

*where for $t > t_{ls}$,*

$$C_t = A_* + r\left(T_t^{-\frac{1}{2}}\right)e^{\hat{E}T_t^{-1}}. \tag{4.28}$$

*Proof.* Given Proposition 2.1, only the first two terms in the integrand of $H^\epsilon(t)$ are left, which leads directly to the inequality corresponding to $A_*$. ∎

**4.6. Proof of dissipation.** Lemma 4.10 constructs a sequence of compactly supported functions that are multiplied with the integrand in $H(t)$. It gives sufficient properties for retrieving a bound on $\partial_t H(t)$ after passing the derviative under the integral sign and passing the limit in the sequence of approximating initial densities. The key sufficient property turns out to be (4.29).

Let $\varphi_k$ be given as in (4.1), and let $\nu_k := \varphi_k * \mathbb{1}_{(-\infty,k^2]} \leq 1$ for $k > 0$.

Lemma 4.10. *For $k > 0$, define the smooth functions $\eta_k : \mathbb{R}^{2n+m+1} \to \mathbb{R}$,*

$$\eta_k = \nu_k(-\ln(R+2d)),$$

*where $d > 0$ is the same as in (4.17). The following properties hold:*

1. *$\eta_k$ is compactly supported;*
2. *$\eta_k$ converges to 1 pointwise as $k \to \infty$;*
3. *for some constant $C > 0$ independent of $k$, $t$, and $0 \leq \epsilon \leq \min(1, \epsilon')$,*

$$L_t^\epsilon \eta_k \leq \frac{CT_t^{-1}}{k}. \tag{4.29}$$

*Proof.* By the quadratic assumption (2.7) on $U$ and the bound (4.17) on $R$, $R$ grows quadratically, and in particular for an arbitrarily large constant $R_{(0)} > 0$, a compact set $K$

can be chosen such that $R > R_{(0)}$ in $\mathbb{R}^{2n+m} \setminus K$; along with the support of $\nu_m$ being bounded below, the first statement is clear. The second statement is also trivial to check. The third statement is an application of (4.4), (4.5), and (4.18); detailed calculations can be found in section SM2 in the supplementary material. ∎

*Remark* 4.11. Lemma 4.10 is different from Lemma 16 in [51]. We believe the first few equations in the proof of Lemma 16 in [51] contain a sign error; as a consequence the proofs in [51] beyond that point require significant modifications. Here we address this by modifying the truncation arguments we require, proving (4.29) instead of Lemma 17 of [51]. In addition, the finiteness of the distorted entropy is required. This is the reason for using the perturbed dynamics in (4.3a), so that Theorem 7.4.1 in [7] can be used.

The proof of Proposition 4.12 follows in the direction of Lemma 19 of [51].

Proposition 4.12. *Under Assumption 1, 2, 3, and 4 and for* $0 < \epsilon \leq \min(1, \epsilon')$, *it holds that for any* $0 < \alpha \leq \frac{1}{2}(1 - \frac{\hat{E}}{E})$, *there exists some constant* $B > 0$ *and some* $t_H > 0$ *both independent of* $\epsilon$, *such that for all* $t > t_H$,

$$H^\epsilon(t) \leq B\left(\frac{1}{t}\right)^{1-\frac{\hat{E}}{E}-2\alpha}. \tag{4.30}$$

*Proof.* Consider for $t \geq 0$ the auxiliary distorted entropies

$$\begin{aligned} H_k^\epsilon(t) &= \int \eta_k \Bigg( \frac{|2\nabla_x h_t^\epsilon + 8S_0(\nabla_y h_t^\epsilon + \lambda^{-1}\nabla_z h_t^\epsilon)|^2}{h_t^\epsilon} + \frac{|\nabla_y h_t^\epsilon + S_1\lambda^{-1}\nabla_z h_t^\epsilon|^2}{h_t^\epsilon} \\ &\qquad + \beta(T_t^{-1}) h_t^\epsilon \ln(h_t^\epsilon) \Bigg) d\mu_{T_t} \\ &= \int \eta_k(\Phi_1(h_t^\epsilon) + \Phi_2(h_t^\epsilon) + \beta(T_t^{-1})\Phi_3(h_t^\epsilon)) d\mu_{T_t} = \int \eta_k \Psi_{T_t}(h_t^\epsilon) d\mu_{T_t}, \end{aligned} \tag{4.31}$$

where we recall that $h_t^\epsilon = m_t^\epsilon \mu_{T_t}^{-1}$, $\Phi_1$, $\Phi_2$, $\Phi_3$ are as in (4.24), and $\eta_k$ is as in Lemma 4.10. Due to the appearance of $\eta_k$, the function $H_k^\epsilon$ is differentiable, and the order between the time derivative and the integral can be exchanged:

$$\frac{d}{dt} H_k^\epsilon(t) = \int \eta_k \partial_t(\Psi_{T_t}(h_t^\epsilon)) d\mu_{T_t} + T_t' \int \eta_k \partial_{T_t}(\Psi_{T_t}(h_t^\epsilon)\mu_{T_t}) dx dy dz. \tag{4.32}$$

The terms will be considered separately. Since $m_t^\epsilon$ is the density of the law of (4.3a) and $L_t^{\epsilon *}$ is the $L^2(\mu_{T_t})$-adjoint of $L_t^\epsilon$, by Itô's rule for smooth compactly supported $f$ on $\mathbb{R}^{2n+m}$,

$$\int f \partial_t m_t^\epsilon = \partial_t \int f m_t^\epsilon = \int L_t^\epsilon f m_t^\epsilon = \int L_t^\epsilon f \frac{m_t^\epsilon}{\mu_{T_t}} \mu_{T_t} = \int f L_t^{\epsilon *}\left(\frac{m_t^\epsilon}{\mu_{T_t}}\right)\mu_{T_t}. \tag{4.33}$$

The first term in (4.32) is then bounded as

$$\begin{aligned} \int \eta_k \partial_t(\Psi_{T_t}(h_t^\epsilon)) d\mu_{T_t} &= \int \eta_k d\Psi_{T_t}(h_t^\epsilon).\partial_t h_t^\epsilon d\mu_{T_t} = \int \eta_k d\Psi_{T_t}(h_t^\epsilon).\frac{\partial_t m_t^\epsilon}{\mu_{T_t}} d\mu_{T_t} \\ &= \int \eta_k d\Psi_{T_t}(h_t^\epsilon).L_t^{\epsilon *} h_t^\epsilon d\mu_{T_t} \end{aligned}$$

$$
\begin{aligned}
&= -\int 2\eta_k \Gamma_{L_t^{\epsilon *},\Psi_{T_t}}(h_t^\epsilon) d\mu_{T_t} + \int \eta_k L_t^{\epsilon *}(\Psi_{T_t}(h_t^\epsilon)) d\mu_{T_t} \\
&= -\int 2\eta_k \Gamma_{L_t^{\epsilon *},\Psi_{T_t}}(h_t^\epsilon) d\mu_{T_t} + \int L_t^\epsilon \eta_k \Big(\Psi_{T_t}(h_t^\epsilon) + \beta(T_t^{-1}) e^{-1}\Big) d\mu_{T_t} \\
&\le -2\int \eta_k \frac{|\nabla h_t^\epsilon|^2}{h_t^\epsilon} d\mu_{T_t} + \frac{C T_t^{-1}}{k} \int \Big(\Psi_{T_t}(h_t^\epsilon) + \beta(T_t^{-1}) e^{-1}\Big) d\mu_{T_t}
\end{aligned} \tag{4.34}
$$

using Proposition 4.7 and Lemma 4.10, where $\beta(T_t^{-1})e^{-1} \int L_t^{\epsilon *} \eta_k d\mu_{T_t} = 0$ is added to enforce

$$
\beta(T_t^{-1})(h_t^\epsilon \ln h_t^\epsilon + e^{-1}) \ge 0, \quad \text{so that} \quad \Psi_{T_t}(h_t^\epsilon) + \beta(T_t^{-1}) e^{-1} \ge 0.
$$

For the second term in (4.32), consider the $\Phi_1$ and $\Phi_2$ terms in the integrand $\eta_k \partial_{T_t}(\Psi_{T_t} \mu_{T_t}) = \eta_k \partial_{T_t}((\Phi_1 + \Phi_2 + \beta(T_t^{-1})\Phi_3)\mu_{T_t})$ of $H_k(t)$ with the forms

$$
\partial_{T_t}(\Phi_i(h_t^\epsilon)\mu_{T_t}) = \partial_{T_t} |M_i \nabla \ln \left(\frac{m_t^\epsilon}{\mu_{T_t}}\right)|^2 m_t^\epsilon, \qquad i = 1, 2,
$$

for the corresponding matrices $M_1$ and $M_2$ depending on $S_0$, $S_1$, and $\lambda$. Applying the partial derivative in $T_t$,

$$
\partial_{T_t}(\Phi_i(h_t^\epsilon)\mu_{T_t}) = -2(M_i \nabla \ln h_t^\epsilon \cdot M_i \nabla \partial_{T_t} \ln \mu_{T_t}) m_t^\epsilon, \tag{4.35}
$$

and using definition (2.2) for $\mu_{T_t}$ and $Z_{T_t} = \int_{\mathbb{R}^{2n+m}} e^{-\frac{1}{T_t}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} dx dy dz$, gives

$$
\begin{aligned}
\partial_{T_t} \ln \mu_{T_t} &= \mu_{T_t}^{-1} \partial_{T_t} \left( Z_{T_t}^{-1} e^{-\frac{1}{T_t}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} \right) \\
&= \mu_{T_t}^{-1} \left( - Z_{T_t}^{-2} \partial_{T_t} Z_{T_t} + \frac{Z_{T_t}^{-1}}{T_t^2} \left( U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} \right) \right) e^{-\frac{1}{T_t}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} \\
&= \mu_{T_t}^{-1} \left( - \mu_{T_t} Z_{T_t}^{-1} \partial_{T_t} Z_{T_t} + \frac{\mu_{T_t}}{T_t^2} \left( U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} \right) \right) \\
&= -\int \frac{1}{T_t^2} \left( U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} \right) d\mu_{T_t} + \frac{1}{T_t^2} \left( U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} \right).
\end{aligned} \tag{4.36}
$$

Note that the exchange in differentiation and integration is justified by the quadratic bounds (2.7) on $U$. Integrating by parts in $y$ and $z$ (or simply using formulae for second moments) gives $\frac{n+m}{2T_t}$ for the $|y|^2$ and $|z|^2$ terms in the first integral. The integral over $U$ can be dealt with using assumptions (2.4) and (2.5); more specifically,

$$
\begin{aligned}
\int U d\mu_{T_t} &\le \int (a_M^2 |x|^2 + U_M) d\mu_{T_t} \le \int \left( \frac{a_M^2}{r_1} (\nabla U \cdot x + U_g) + U_M \right) d\mu_{T_t} \\
&= \frac{a_M^2}{r_1} (n T_t + U_g) + U_M \\
\int U d\mu_{T_t} &\ge \int (a_m^2 |x|^2 + U_m) d\mu_{T_t} \ge \int \left( \frac{a_m^2}{r_2 + 1} (|\nabla U|^2 - U_g + |x|^2) + U_m \right) d\mu_{T_t} \\
&\ge \int \left( \frac{a_m^2}{r_2 + 1} (2 \nabla U \cdot x - U_g) + U_m \right) d\mu_{T_t} = \frac{a_m^2}{r_2 + 1} (2 n T_t - U_g) + U_m.
\end{aligned}
$$

Plugging into (4.36) gives

$$p_1\Big(T_t^{-1}\Big) \le \partial_{T_t} \ln \mu_{T_t} - \frac{1}{T_t^2}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} - \frac{n+m}{2}T_t\right) \le p_2\Big(T_t^{-1}\Big), \tag{4.37}$$

where $p_1(x) = -\frac{a_M^2 n}{r_1}x - \Big(\frac{a_M^2 U_g}{r_1} + U_M\Big)x^2$ and $p_2(x) = -\frac{2a_m^2 n}{r_2+1}x + \Big(\frac{a_m^2 U_g}{r_2+1} - U_m\Big)x^2$.Substituting (4.36) back into (4.35), we get

$$\begin{aligned}\partial_{T_t}(\Phi_i(h_t^\epsilon)\mu_{T_t}) &\le \left(|M_i\nabla \ln h_t^\epsilon|^2 + T_t^{-4}|M_i\nabla\Big(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\Big)|^2\right)m_t^\epsilon \\ &\le \Phi_i(h_t^\epsilon)\mu_{T_t} + \widetilde{C}T_t^{-4}\Big(1 + |x|^2 + |y|^2 + |z|^2\Big)m_t^\epsilon \end{aligned} \tag{4.38}$$

for a constant $\widetilde{C} \ge 0$ independent of $k$ and $\epsilon$ by the quadratic assumption (2.5) on $|\nabla_x U|^2$ and $\eta_m \le 1$.

For the last integrand in the last term of the right-hand side of (4.32), namely the derivative over $\Phi_3(h_t^\epsilon)\mu_{T_t} = \frac{m_t^\epsilon}{\mu_{T_t}} \ln \frac{m_t^\epsilon}{\mu_{T_t}}\mu_{T_t}$, the left inequality of (4.37) gives

(4.39)

$$\begin{aligned}&\partial_{T_t}(\beta(T_t^{-1})\Phi_3(h_t^\epsilon)\mu_{T_t}) \\ &= -T_t^{-2}\beta'(T_t^{-1})\Phi_3(h_t^\epsilon)\mu_{T_t} + \beta(T_t^{-1})\partial_{T_t}\ln\frac{m_t^\epsilon}{\mu_{T_t}}m_t^\epsilon \\ &= -T_t^{-2}\beta'(T_t^{-1})(\Phi_3(h_t^\epsilon) + e^{-1})\mu_{T_t} + T_t^{-2}\beta'(T_t^{-1})e^{-1}\mu_{T_t} - \beta(T_t^{-1})\partial_{T_t}\ln\mu_{T_t}m_t^\epsilon \\ &\le T_t^{-2}\beta'(T_t^{-1})e^{-1}\mu_{T_t} + \beta(T_t^{-1})|p_1\Big(T_t^{-1}\Big) + \frac{1}{T_t^2}\left(-\frac{n+m}{2}T_t + U_M + a_M|x|^2 + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)|m_t^\epsilon,\end{aligned}$$

where in the last step $\Phi_3 + e^{-1} \ge 0$, $\beta_1, \beta_2 > 0$, and (2.7) have been used. Putting together the bounds (4.38) and (4.39) and applying Corollary 4.5 yields

$$\begin{aligned}\int \eta_k \partial_{T_t}(\Psi_{T_t}(h_t^\epsilon)\mu_{T_t})d\zeta &\le q\Big(T_t^{-1}\Big)\Big(H_k^\epsilon(t) + \mathbb{E}\Big[1 + |X_t^\epsilon|^2 + |Y_t^\epsilon|^2 + |Z_t^\epsilon|^2\Big]\Big) \\ &\le p\Big(T_t^{-1}\Big)\Big(H_k^\epsilon(t) + \hat{C}\Big),\end{aligned} \tag{4.40}$$

where $p$ and $q$ are some finite order polynomials with nonnegative coefficients, $\hat{C} > 0$, both independent of $k$ and $\epsilon$.

Returning to (4.32), collecting (4.34) and (4.40), and then integrating from any $s \ge 0$ to $t > s$ gives

$$\begin{aligned}H_k^\epsilon(t) - H_k^\epsilon(s) \le 2\int_s^t\Bigg(&-\int \eta_k \frac{|\nabla h_u^\epsilon|^2}{h_u}d\mu_{T_u} + \frac{CT_u^{-1}}{k}(H^\epsilon(u) + \beta(T_u^{-1})e^{-1}) \\ &+ |T_u'|p\Big(T_u^{-1}\Big)\Big(H_k^\epsilon(u) + \hat{C}\Big)\Bigg)du.\end{aligned} \tag{4.41}$$

Fix an arbitrary $S > 0$. By the square integrability theorem, Theorem 7.4.1 in [7], the log-Sobolev inequality (4.27), (2.5), and the finiteness of second moments (4.10), it holds that

$$\begin{aligned}\int_0^S H^\epsilon(u)du &\le \int_0^S C_u \int \frac{|\nabla h_u^\epsilon|^2}{h_u^\epsilon} d\mu_{T_u} du \\ &= \int_0^S C_u \int \frac{|\nabla m_u^\epsilon + T_u^{-1} m_u^\epsilon (\nabla_x U + y + z)|^2}{m_u^\epsilon} dx dy dz du < \infty. \end{aligned} \tag{4.42}$$

Then in (4.41) the $k \to \infty$ limit can be taken. Due to (4.42), the term denominated by $k$ goes to zero. Applying Fatou's lemma (adding and subtracting $\beta(T_t^{-1})e^{-1}\int \eta_m d\mu_{T_t}$ wherever necessary for positivity) and using $\eta_m \le 1$, it holds that for $s < t$,

$$H^\epsilon(t) - H^\epsilon(s) \le -2\int_s^t \int \frac{|\nabla h_u^\epsilon|^2}{h_u^\epsilon} d\mu_{T_u} du + \int_s^t |T_u'| p\Big(T_u^{-1}\Big)\Big(H^\epsilon(u) + \hat{C}\Big) du, \tag{4.43}$$

and for[4] $t_{ls} < s < t$,

$$H^\epsilon(t) - H^\epsilon(s) \le \int_s^t \Big(\Big(|T_u'| p\Big(T_u^{-1}\Big) - 2C_u^{-1}\Big) H^\epsilon(u) + \hat{C}|T_u'| p\Big(T_u^{-1}\Big)\Big) du. \tag{4.44}$$

Since $t^\alpha \gg (\ln t)^{\frac{\rho}{2}}$ for any $\rho, \alpha > 0$ and large enough $t > 0$, for any $\alpha > 0$, there exists $t_1 > \max(t_{ls}, t_0)$, where $t_0$ is as in Assumption 3, and $c_1, c_2 > 0$ independent of $k, \epsilon$ such that for all $t \ge t_1$,

$$|T_t'| p\Big(T_t^{-1}\Big) \le c_1 \Big(\frac{1}{t}\Big)^{1-\alpha}, \tag{4.45}$$

$$-2C_t^{-1} \le -c_2 \Big(\frac{1}{t}\Big)^{\frac{\hat{E}}{E}+\alpha}, \tag{4.46}$$

where the assumption $T_t \ge \frac{E}{\ln t}$ and (4.28) have been used. Using further that $E > \hat{E}$ by Assumption 3, then taking $\alpha < \frac{1}{2}(1 - \frac{\hat{E}}{E})$, there exists $t_2 \ge t_1$ independent of $\epsilon$ such that for $t \ge t_2$,

$$|T_t'| p\Big(T_t^{-1}\Big) - 2C_t^{-1} \le -c_3 \Big(\frac{1}{t}\Big)^{\frac{\hat{E}}{E}+\alpha}, \tag{4.47}$$

and from (4.44), for $t_2 < s < t$,

$$H^\epsilon(t) - H^\epsilon(s) \le \int_s^t \Big( - c_3 \Big(\frac{1}{u}\Big)^{\frac{\hat{E}}{E}+\alpha} H^\epsilon(u) + \hat{C} c_1 \Big(\frac{1}{u}\Big)^{1-\alpha}\Big) du. \tag{4.48}$$

To obtain the corresponding differential inequality for all time, (4.48) can be divided by $t - s$ and mollified with (4.1) for $0 < k < 1$, and the limit $s \to t$ can be taken:

$$\begin{aligned}&\lim_{\hat{\epsilon}\to 0} \frac{1}{2\hat{\epsilon}} \int_{t-1}^{t+1} \varphi_k(t-u)(H^\epsilon(u+\hat{\epsilon}) - H^\epsilon(u-\hat{\epsilon}))du \\ &\qquad \le \lim_{\hat{\epsilon}\to 0} \frac{1}{2\hat{\epsilon}} \int_{t-1}^{t+1} \varphi_k(t-u) \int_{u-\hat{\epsilon}}^{u+\hat{\epsilon}} \Big( - c_3 \Big(\frac{1}{u'}\Big)^{\frac{\hat{E}}{E}+\alpha} H^\epsilon(u') + \hat{C} c_1 \Big(\frac{1}{u'}\Big)^{1-\alpha}\Big) du' du \end{aligned}$$

[4] $t_{ls}$ from Proposition 4.9.

$$\leq \int_{t-1}^{t+1} \varphi_k(t-u) \lim_{\hat{\epsilon}\to 0} \frac{1}{2\hat{\epsilon}} \int_{u-\hat{\epsilon}}^{u+\hat{\epsilon}} \left( -c_3 \left(\frac{1}{u'}\right)^{\frac{\hat{E}}{E}+\alpha} H^\epsilon(u') + \hat{C} c_1 \left(\frac{1}{u'}\right)^{1-\alpha} \right) du' du$$
$$= \int_{t-1}^{t+1} \varphi_k(t-u) \left( -c_3 \left(\frac{1}{u}\right)^{\frac{\hat{E}}{E}+\alpha} H^\epsilon(u) + \hat{C} c_1 \left(\frac{1}{u}\right)^{1-\alpha} \right) du$$

for $t \geq t_2 + 2$, where the second-to-last line follows from Fatou's lemma and dominated convergence (adding and subtracting $\beta(T_{u'}^{-1})e^{-1}$ to/from $H^\epsilon$ for Fatou); the last equality follows by the Lebesgue differentiation theorem. Therefore

$$\frac{d}{dt}(\varphi_k * H^\epsilon)(t) \leq -c_3 \left(\frac{1}{t+1}\right)^{\frac{\hat{E}}{E}+\alpha} (\varphi_k * H^\epsilon)(t) + \hat{C}' \left(\frac{1}{t-1}\right)^{1-\alpha}$$

for some constant $\hat{C}' > 0$ independent of $k, \epsilon$. Setting

$$\gamma_1(t) := c_3 \left(\frac{1}{t+1}\right)^{\frac{\hat{E}}{E}+\alpha}, \qquad \gamma_2(t) := \hat{C}' \left(\frac{1}{t-1}\right)^{1-\alpha}$$

and following the argument as per [51] from Lemma 6 in [50], there exists $t_3 \geq t_2 + 2$, $c_4, c_5, c_6 > 0$ independent of $k$ and $\epsilon$ such that for $t \geq t_3$,

$$\frac{d}{dt}\left(\frac{\gamma_2}{\gamma_1}\right)(t) = \frac{(t+1)^{\frac{\hat{E}}{E}+\alpha}}{(t-1)^{1-\alpha}} \left(\frac{c_4}{t+1} - \frac{c_5}{t-1}\right) \geq -c_6 t^{-1},$$

so that there exists $t_4 \geq t_3$ independent of $k$ and $\epsilon$ such that for $t \geq t_4$,

$$\frac{d}{dt}\left(\varphi_k * H^\epsilon - \frac{2\gamma_2}{\gamma_1}\right)(t) \leq -\gamma_1(t) \left(\varphi_k * H^\epsilon(t) - \frac{2\gamma_2(t)}{\gamma_1(t)}\right),$$

and consequently,

$$\varphi_k * H^\epsilon(t) \leq \frac{2\gamma_2(t)}{\gamma_1(t)} + \varphi_k * H^\epsilon(t_4) e^{-\int_{t_4}^t \gamma_1(u)du}. \tag{4.49}$$

Finally, from (4.48) (adding and subtracting $\beta(T_{u'}^{-1})e^{-1}$ to/from $H^\epsilon$), it holds that for $t \geq t_4 + 2$,

$$H^\epsilon(t) = \int_{t-2k}^t \varphi_k(t-k-s) ds H^\epsilon(t) \leq \int_{t-2k}^t \varphi_k(t-k-s) H^\epsilon(s) ds + \tilde{g}(2k) \tag{4.50}$$

for some $\tilde{g} : \mathbb{R} \to \mathbb{R}$ satisfying $\tilde{g}(k') \to 0$ as $k' \to 0$, so that (4.49) yields

$$H^\epsilon(t) \leq \frac{2\gamma_2(t-k)}{\gamma_1(t-k)} + \varphi_k * H^\epsilon(t_4) e^{-\int_{t_4}^{t-k} \gamma_1(u)du} + \tilde{g}(2k),$$

where $\varphi_k * H^\epsilon(t_4)$ can be bounded independently of $k$ in a similar spirit to (4.50), and taking $k \to 0$ concludes the proof. ∎

*Remark* 4.13. The annealing schedule $T_t$ is chosen to satisfy the relationship (4.47) between $C_t^{-1}$ and $|T_t'|p(T_t^{-1})$.

**4.7. Degenerate noise limit.** After taking advantage of the square integrability theorem, Theorem 7.4.1 in [7], we see that for the case with a nondegenerate diffusion term in the proof of Proposition 4.12, the $\epsilon \to 0$ limit is taken to obtain the same dissipation inequality in this section.

*Proof of Proposition 2.5.* From (4.43), for any $0 \le s < t$ and $0 < \epsilon \le \min(1, \epsilon')$, it holds that

$$H^\epsilon(t) - H^\epsilon(s) \le \int_s^t |T_u'| p\Big(T_u^{-1}\Big)\Big(H^\epsilon(u) + \hat{C}\Big) du,$$

where $p$ is a finite order polynomial with nonnegative coefficients, $\hat{C} > 0$ is a constant, and both are independent of $\epsilon$. Therefore, mollifying in time and taking $s \to t$ as in the end of the proof of Proposition 4.12, it is straightforward that $H^\epsilon$ is uniformly bounded in[5] $0 \le t \le t_H$ and $0 < \epsilon \le \min(1, \epsilon')$. Moreover, by Proposition 4.12 the entropy $\int h_t^\epsilon \ln h_t^\epsilon d\mu_{T_t}$ is bounded uniformly in $t > t_H$ and $0 < \epsilon \le \min(1, \epsilon')$. Therefore for any $t \ge 0$ by the de la Vallée-Poussin criterion (see, for example, [17]), the subset $\{h_t^\epsilon : 0 < \epsilon \le \min(1, \epsilon')\} \subset L^1(\mu_{T_t})$ is uniformly integrable, and consequently the Dunford–Pettis theorem imposes the existence of a weak limit $g_t \in L^1(\mu_{T_t})$ for a (sub)sequence $(\epsilon_i)_{i\in\mathbb{N}}$ such that $\epsilon_i \to 0$:

$$h_t^{\epsilon_i} \rightharpoonup g_t, \quad \text{in } L^1(\mu_{T_t}) \quad \text{as } i \to \infty.$$

For any $S > 0$ and any compactly supported smooth test function $\phi : [0, S) \times \mathbb{R}^{2n+m} \to \mathbb{R}$, omitting the dependence on the space variable $\zeta = (x, y, z)$ wherever convenient, denoting $D_S := (0, S) \times \mathbb{R}^{2n+m}$, and using Itô's rule, we get

$$\begin{aligned} 0 &= \lim_{i\to\infty} \int_{D_S} (m_t^{\epsilon_i} - g_t \mu_{T_t})(-\partial_t - L_t)\phi dt d\zeta \\ &= \lim_{i\to\infty} \int_{D_S} \epsilon_i m_t^{\epsilon_i} (S_t^x + S_t^y)\phi dt d\zeta + \int_{D_S} g_t \mu_{T_t} (\partial_t + L_t)\phi dt d\zeta + \int_{\mathbb{R}^{2n+m}} m_0 \phi(0, \zeta) dt d\zeta \\ &= \int_{D_S} g_t \mu_{T_t} (\partial_t + L_t)\phi dt d\zeta + \int_{\mathbb{R}^{2n+m}} m_0 \phi(0, \zeta) dt d\zeta, \end{aligned} \tag{4.51}$$

so that in the distributional sense of [7],

$$\begin{cases} \partial_t (g_t \mu_{T_t}) = L_t^\top (g_t \mu_{T_t}) & \text{on } \mathbb{R}^{2n+m} \quad \forall t > 0, \\ (g_0 \mu_{T_0}) = m_0. \end{cases} \tag{4.52}$$

By Proposition 4.1, the solution to (4.52) is unique in the class of integrable solutions, and since $m_t$ belongs in this same class, it holds that

$$g_t \mu_{T_t} = m_t$$

for all $t \in [0, S]$, that is,

$$m_t^{\epsilon_i} \rightharpoonup m_t, \quad \text{in } L^1(\mu_{T_t}) \quad \text{as } i \to \infty$$

[5] $t_H$ from Proposition 4.12.

for all $0 \leq t < S$. By Corollary 3.8 in [9], there exists a sequence $(\hat{m}_t^i)_{i\in\mathbb{N}}$ made up of convex combinations of $m_t^{\epsilon_i}$ that converge strongly to $m_t$ in $L^1$; hence we have a subsequence $(\hat{m}_t^{i_j})_{j\in\mathbb{N}}$ that converges pointwise almost everywhere. By Fatou's lemma, convexity of $f(x) = x \ln x \geq e^{-1}$ for $x > 0$, and Proposition 4.12, for $t > t_H$, we get

$$\int h_t \ln h_t d\mu_{T_t} = \int m_t \ln\left(\frac{m_t}{\mu_{T_t}}\right) \leq \liminf_{j\to\infty} \int \hat{m}_t^{i_j} \ln\left(\frac{\hat{m}_t^{i_j}}{\mu_{T_t}}\right) \leq B\left(\frac{1}{t}\right)^{1-\frac{\hat{E}}{E}-2\alpha}.$$

**5. Conclusions.** We explored the possibility of using the generalized Langevin equations in the context of simulated annealing. Our main purpose was to establish convergence for the underdamped Langevin equation and provide a proof of concept in terms of performance improvement. Although the theoretical results hold for any scaling matrix $A$ given the stated restrictions, we saw in our numerical results that its choice has great impact on the performance. In section 3, $A_2, A_3$, or $A_4$ seemed to improve the exploration on the state space and/or the success of the algorithm. There is plenty of work still required in terms of providing a more complete methodology for choosing $A$. This is left as future work and is also closely linked with time discretization issues as a poor choice for $A$ could lead to numerical integration stiffness. This motivates the development and study of improved numerical integration schemes, in particular the extension of the conception and analysis on numerical schemes such as BAOAB [41] for the Langevin equation for (1.3) and the extension of the work in [53] for nonidentity matrices $\lambda$ and $A$. See [42] for work in this direction.

In addition, the system in (1.3) is not the only way to add an auxiliary variable to the underdamped Langevin equations in (1.2) while retaining the appropriate equilibrium distribution. Our choice was motivated by a clear connection to the generalized Langevin equation (1.4) and a link with accelerated gradient descent, but it could be the case that a different third or higher order equation could be used with possibly improved performance. Along these lines, one could consider adding skew-symmetric terms as in [20]. In regard to theory, an interesting extension could involve establishing how the results here can be extended to establish a comparison of optimization and sampling in a nonconvex setting for an arbitrary number of dimensions similar to [45]. We leave for future work finding optimal constants in the convergence results and investigating dependence on parameters and how the limits of these parameters and constants relate to existing results for the Langevin equation in (1.2) in [51, 60]. Finally, one could also aim to extend large deviation results in [38, 46, 64] for the overdamped Langevin dynamics to the underdamped and generalized case.

**Acknowledgments.** The authors would like to thank Tony Lelievre, Gabriel Stoltz, Urbain Vaes and the anonymous referees for their helpful remarks.

## REFERENCES

[1] S. A. Adelman and B. J. Garrison, *Generalized Langevin theory for gas/solid processes: Dynamical solid models*, J. Chem. Phys., 65 (1976), pp. 3751–3761, https://doi.org/10.1063/1.433564.

[2] H. AlRachid, L. Mones, and C. Ortner, *Some remarks on preconditioning molecular dynamics*, SMAI J. Comput. Math., 4 (2018), pp. 57–80, https://doi.org/10.5802/smai-jcm.29.

[3] A. D. Baczewski and S. D. Bond, *Numerical integration of the extended variable generalized Langevin equation with a positive Prony representable memory kernel*, J. Chem. Phys., 139 (2013), 044107, https://doi.org/10.1063/1.4815917.
[4] C. H. Bennett, *Mass tensor molecular dynamics*, J. Comput. Phys., 19 (1975), pp. 267–279, https://doi.org/10.1016/0021-9991(75)90077-7.
[5] R. Biswas and D. R. Hamann, *Simulated annealing of silicon atom clusters in Langevin molecular dynamics*, Phys. Rev. B, 34 (1986), pp. 895–901, https://doi.org/10.1103/PhysRevB.34.895.
[6] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch, *Structural relaxation made simple*, Phys. Rev. Lett., 97 (2006), 170201, https://doi.org/10.1103/PhysRevLett.97.170201.
[7] V. I. Bogachev, N. V. Krylov, M. Röckner, and S. V. Shaposhnikov, *Fokker–Planck–Kolmogorov Equations*, Math. Surveys Monogr. 207, American Mathematical Society, Providence, RI, 2015, https://doi.org/10.1090/surv/207.
[8] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Rev., 60 (2018), pp. 223–311, https://doi.org/10.1137/16M1080173.
[9] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer, New York, 2011.
[10] J. Carrillo, S. Jin, L. Li, and Y. Zhu, A *Consensus-Based Global Optimization Method for High Dimensional Machine Learning Problems*, preprint, https://arxiv.org/abs/1909.09249, 2019.
[11] J. A. Carrillo, Y.-P. Choi, C. Totzeck, and O. Tse, *An analytical framework for consensus-based global optimization method*, Math. Models Methods Appl. Sci., 28 (2018), pp. 1037–1066, https://doi.org/10.1142/S0218202518500276.
[12] M. Ceriotti, G. Bussi, and M. Parrinello, *Langevin equation with colored noise for constant-temperature molecular dynamics simulations*, Phys. Rev. Lett., 102 (2009), 020601, https://doi.org/10.1103/PhysRevLett.102.020601.
[13] M. Ceriotti, G. Bussi, and M. Parrinello, *Colored-noise thermostats à la carte*, J. Chem. Theory Comput., 6 (2010), pp. 1170–1180, https://doi.org/10.1021/ct900563s .
[14] M. Chaleyat-Maurel and D. Michel, *Hypoellipticity theorems and conditional laws*, Z. Wahrsch. Verw. Gebiete, 65 (1984), pp. 573–597.
[15] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *Underdamped Langevin MCMC: A non-asymptotic analysis*, in Proceedings of the 31st Conference On Learning Theory, S. Bubeck, V. Perchet, and P. Rigollet, eds., Proc. Mach. Learn. Res. 75, PMLR, 2018, pp. 300–323.
[16] T.-S. Chiang, C.-R. Hwang, and S.-J. Sheu, *Diffusion for global optimization in* $\mathbf{R}^n$, SIAM J. Control Optim., 25 (1987), pp. 737–753, https://doi.org/10.1137/0325042.
[17] J. Diestel, *Uniform integrability: An introduction*, in School on Measure Theory and Real Analysis (Grado, 1991) .Rend. Istit. Mat. Univ. Trieste 23, Università degli Studi di Trieste, 1993, pp. 41–80.
[18] R. Douc, G. Fort, and A. Guillin, *Subgeometric rates of convergence of f-ergodic strong Markov processes*, Stochastic Process. Appl., 119 (2009), pp. 897–923, https://doi.org/10.1016/j.spa.2008.03.007.
[19] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis, *Variance reduction using nonreversible Langevin samplers*, J. Stat. Phys., 163 (2016), pp. 457–491, https://doi.org/10.1007/s10955-016-1491-2.
[20] A. B. Duncan, N. Nüsken, and G. A. Pavliotis, *Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions*, J. Stat. Phys., 169 (2017), pp. 1098–1131, https://doi.org/10.1007/s10955-017-1906-8.
[21] A. Durmus and É. Moulines, *High-dimensional Bayesian inference via the unadjusted Langevin algorithm*, Bernoulli, 25 (2019), pp. 2854–2882, https://doi.org/10.3150/18-BEJ1073.
[22] A. Eberle, A. Guillin, and R. Zimmer, *Couplings and quantitative contraction rates for Langevin dynamics*, Ann. Probab., 47 (2019), pp. 1982–2010, https://doi.org/10.1214/18-AOP1299.
[23] J.-P. Eckmann and M. Hairer, *Non-equilibrium statistical mechanics of strongly anharmonic chains of oscillators*, Comm. Math. Phys., 212 (2000), pp. 105–164.
[24] A. Friedman, *Stochastic Differential Equations and Applications*, Vol. 1 , Probab. Math. Statist. 28, Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975.
[25] S. Gadat and F. Panloup, *Long time behaviour and stationary regime of memory gradient diffusions*, Ann. Inst. H. Poincaré Probab. Statist., 50 (2014), pp. 564–601, https://doi.org/10.1214/12-AIHP536.

[26] X. Gao, M. Gurbuzbalaban, and L. Zhu, *Breaking Reversibility Accelerates Langevin Dynamics for Global Non-convex Optimization*, preprint, https://arxiv.org/abs/1812.07725, 2018.
[27] S. B. Gelfand and S. K. Mitter, *Recursive stochastic algorithms for global optimization in* $\mathbf{R}^d$, SIAM J. Control Optim., 29 (1991), pp. 999–1018, https://doi.org/10.1137/0329055.
[28] S. B. Gelfand and S. K. Mitter, *Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions*, J. Optim. Theory Appl., 68 (1991), pp. 483–498, https://doi.org/10.1007/BF00940066.
[29] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images*, IEEE Trans. Pattern Anal. Mach. Intell., PAMI-6 (1984), pp. 721–741, https://doi.org/10.1109/TPAMI.1984.4767596.
[30] S. Geman and C.-R. Hwang, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031–1043, https://doi.org/10.1137/0324060.
[31] S. Ghadimi and G. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program., 156 (2016), pp. 59–99, https://doi.org/10.1007/s10107-015-0871-8.
[32] B. Gidas, *Global optimization via the Langevin equation*, in Proceedings of the 1985 24th IEEE Conference on Decision and Control, 1985, pp. 774–778, https://doi.org/10.1109/CDC.1985.268602.
[33] B. Gidas, *Nonstationary Markov chains and convergence of the annealing algorithm*, J. Stat. Phys., 39 (1985), pp. 73–131, https://doi.org/10.1007/BF01007975.
[34] R. Holley and D. Stroock, *Simulated annealing via Sobolev inequalities*, Comm. Math. Phys., 115 (1988), pp. 553–569.
[35] R. A. Holley, S. Kusuoka, and D. W. Stroock, *Asymptotics of the spectral gap with applications to the theory of simulated annealing*, J. Funct. Anal., 83 (1989), pp. 333–347, https://doi.org/10.1016/0022-1236(89)90023-2.
[36] R. Höpfner, E. Löcherbach, and M. Thieullen, *Strongly degenerate time inhomogeneous SDEs: Densities and support properties: Application to Hodgkin–Huxley type systems*, Bernoulli, 23 (2017), pp. 2587–2616, https://doi.org/10.3150/16-BEJ820.
[37] C.-R. Hwang, *Laplace's method revisited: Weak convergence of probability measures*, Ann. Probab., 8 (1980), pp. 1177–1182, https://www.jstor.org/stable/2243019.
[38] C.-R. Hwang and S. J. Sheu, *Large-time behavior of perturbed diffusion Markov processes with applications to the second eigenvalue problem for Fokker-Planck operators and simulated annealing*, Acta Appl. Math., 19 (1990), pp. 253–295, https://doi.org/10.1007/BF01321859.
[39] H. J. Kushner, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo*, SIAM J. Appl. Math., 47 (1987), pp. 169–185, https://doi.org/10.1137/0147010.
[40] H. Lei, N. A. Baker, and X. Li, *Data-driven parameterization of the generalized Langevin equation*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 14183–14188, https://doi.org/10.1073/pnas.1609587113.
[41] B. Leimkuhler and C. Matthews, *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, Interdiscip. Appl. Math. 39, Springer, Cham, 2015.
[42] B. Leimkuhler and M. Sachs, *Efficient Numerical Algorithms for the Generalized Langevin Equation*, preprint, https://arxiv.org/abs/2012.04245, 2020.
[43] T. Lelièvre, F. Nier, and G. A. Pavliotis, *Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion*, J. Stat. Phys., 152 (2013), pp. 237–274, https://doi.org/10.1007/s10955-013-0769-x.
[44] C. Li, C. Chen, D. Carlson, and L. Carin, *Preconditioned stochastic gradient Langevin dynamics for deep neural networks*, in Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, Washington, DC, 2016, pp. 1788–1794.
[45] Y.-A. Ma, Y. Chen, C. Jin, N. Flammarion, and M. I. Jordan, *Sampling can be faster than optimization*, Proc. Natl. Acad. Sci. USA, 116 (2019), pp. 20881–20885, https://doi.org/10.1073/pnas.1820003116.
[46] D. Márquez, *Convergence rates for annealing diffusion processes*, Ann. Appl. Probab., 7 (1997), pp. 1118–1139, https://doi.org/10.1214/aoap/1043862427.

[47] J. C. Mattingly and A. M. Stuart, *Geometric ergodicity of some hypo-elliptic diffusions for particle motions*, Inhomogeneous Random Systems (Cergy-Pontoise, 2001), Markov Process. Related Fields, 8 (2002), pp. 199–214.

[48] G. Menz and A. Schlichting, *Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape*, Ann. Probab., 42 (2014), pp. 1809–1884, https://doi.org/10.1214/14-AOP908.

[49] G. Metafune, D. Pallara, and E. Priola, *Spectrum of Ornstein-Uhlenbeck operators in $L^p$ spaces with respect to invariant measures*, J. Funct. Anal., 196 (2002), pp. 40–60, https://doi.org/10.1006/jfan.2002.3978.

[50] L. Miclo, *Recuit simulé sur $\mathbf{R}^n$. Étude de l'évolution de l'énergie libre*, Ann. Inst. H. Poincaré Probab. Statist., 28 (1992), pp. 235–266, http://www.numdam.org/item?id=AIHPB_199228_2_235_0.

[51] P. Monmarché, *Hypocoercivity in metastable settings and kinetic simulated annealing*, Probab. Theory Related Fields, 172 (2018), pp. 1215–1248, https://doi.org/10.1007/s00440-018-0828-y.

[52] P. Monmarché, *Generalized $\Gamma$ calculus and application to interacting particles on a graph*, Potential Anal., 50 (2019), pp. 439–466, https://doi.org/10.1007/s11118-018-9689-3.

[53] W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, *High-order Langevin diffusion yields an accelerated MCMC algorithm*, J. Mach. Learn. Res., 22 (2021), pp. 1–41, http://jmlr.org/papers/v22/20-576.html.

[54] M. Nava, M. Ceriotti, C. Dryzun, and M. Parrinello, *Evaluating functions of positive-definite matrices using colored-noise thermostats*, Phys. Rev. E, 89 (2014), 023302, https://doi.org/10.1103/PhysRevE.89.023302.

[55] Y. Nesterov, *Lectures on Convex Optimization*, 2nd ed., Springer Optim. Appl. 137, Springer, Cham, 2018, https://doi.org/10.1007/978-3-319-91578-4.

[56] M. Ottobre and G. A. Pavliotis, *Asymptotic analysis for the generalized Langevin equation*, Nonlinearity, 24 (2011), pp. 1629–1653, https://doi.org/10.1088/0951-7715/24/5/013.

[57] M. Ottobre, G. A. Pavliotis, and K. Pravda-Starov, *Exponential return to equilibrium for hypoelliptic quadratic systems*, J. Funct. Anal., 262 (2012), pp. 4000–4039, https://doi.org/10.1016/j.jfa.2012.02.008.

[58] S. Patterson and Y. W. Teh, *Stochastic gradient Riemannian Langevin dynamics on the probability simplex*, in Advances in Neural Information Processing Systems 26, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds., Curran Associates, Inc., Red Hook, NY, 2013, pp. 3102–3110.

[59] G. A. Pavliotis, *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, Texts Appl. Math. 60, Springer, New York, 2014, https://doi.org/10.1007/978-1-4939-1323-7.

[60] G. A. Pavliotis, G. Stoltz, and U. Vaes, *Scaling limits for the generalized Langevin equation*, J. Nonlinear Sci., 31 (2021), 8, https://doi.org/10.1007/s00332-020-09671-4.

[61] M. Pelletier, *Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing*, Ann. Appl. Probab., 8 (1998), pp. 10–44, https://doi.org/10.1214/aoap/1027961032.

[62] R. Pinnau, C. Totzeck, O. Tse, and S. Martin, *A consensus-based model for global optimization and its mean-field limit*, Math. Models Methods Appl. Sci., 27 (2017), pp. 183–204, https://doi.org/10.1142/S0218202517400061.

[63] C. Prévôt and M. Röckner, *A Concise Course on Stochastic Partial Differential Equations*, Lecture Notes in Math. 1905, Springer, Berlin, 2007.

[64] G. Royer, *A remark on simulated annealing of diffusion processes*, SIAM J. Control Optim., 27 (1989), pp. 1403–1408, https://doi.org/10.1137/0327072.

[65] S. Ruder, *An Overview of Gradient Descent Optimization Algorithms*, preprint, https://arxiv.org/abs/1609.04747, 2016.

[66] M. Sachs, *The Generalised Langevin Equation: Asymptotic Properties and Numerical Analysis*, Ph.D. thesis, The University of Edinburgh, 2017.

[67] H. Song, I. Triguero, and E. Özcan, *A review on the self and dual interactions between machine learning and optimisation*, Prog. Artif. Intell., 8 (2019), pp. 143–165, https://doi.org/10.1007/s13748-019-00185-z.

[68] W. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, J. Mach. Learn. Res., 17 (2016), 153.
[69] Y. Sun and A. Garcia, *Interactive diffusions for global optimization*, J. Optim. Theory Appl., 163 (2014), pp. 491–509, https://doi.org/10.1007/s10957-013-0394-5.
[70] S. Taniguchi, *Applications of Malliavin's calculus to time-dependent systems of heat equations*, Osaka J. Math., 22 (1985), pp. 307–320.
[71] X. Wu, B. R. Brooks, and E. Vanden-Eijnden, *Self-guided Langevin dynamics via generalized Langevin equation*, J. Comput. Chem., 37 (2016), pp. 595–601, https://doi.org/10.1002/jcc.24015.