On the Generalised Langevin Equation for Simulated Annealing

Martin Chak, Nikolas Kantas, Grigorios A. Pavliotis
Dept. of Mathematics,
Imperial College London

Abstract

In this paper, we consider the generalised (higher order) Langevin equation for the purpose of simulated annealing and optimisation 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 and overdamped Langevin dynamics with the same annealing schedule.

1 Introduction

Algorithms for optimisation have received 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 minimised, where use of optimisation techniques is heavily relied upon. We refer to [7, 62] for related discussions. Many models, for instance neural networks, use parameters that vary over a continuous space, where gradient-based optimisation methods can be used to find good parameters that generate effective predictive ability. As such, the design and analysis of such algorithms for global optimisation has been the subject of considerable research [60] and it has proved useful to study algorithms for global optimisation 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 optimisation is one of 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 $U : \mathbb{R}^n \to \mathbb{R}$:

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

Here W_t is a standard n-dimensional Wiener process and $T_t : \mathbb{R}_+ \to \mathbb{R}$ is an appropriate decreasing determinstic 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 [18]. 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 well-understood. We briefly mention classical references [15, 24, 25, 27, 28, 31, 32, 35], as well as the more recent [34, 41, 57], where one can find details and convergence results. In this paper we aim to consider generalised 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 $\nu_t(dx) \propto \exp(-\frac{U(x)}{T_t})dx$ converges weakly to the set of global minima based on the Laplace principle [33]. One can expect that if one replaces (1.1) with a stochastic process that mixes faster and maintains the same invariant distribution for constant temperatures, then the superior speed of convergence should improve

performance in optimisation 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 for finding 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, 39, 54], as well as for optimisation, e.g. interacting Langevin dynamics [65], consensus based optimisation [8, 9, 58], etc.

A natural candidate in this direction is to use the underdamped or kinetic 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 both the context of sampling and that of optimisation[1] is investigated in [16, 38] and [21] respectively. When $T_t = T$, (1.2) can converge faster than (1.1) to its invariant distribution

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

see [19] or Section 6.3 of [56] for particular comparisons and 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 [47], where the author established convergence to global minima using the generalised Γ-calculus [48] framework that is based on Bakry-Emery theory. Note that (1.2b) 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 [47].

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 optimisation. In particular we consider the Markovian approximation to the generalised Langevin equation:

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

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

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

with $A \in \mathbb{R}^{m\times m}$ being positive definite and $\Sigma \in \mathbb{R}^{m\times m}$ restricted to satisfying

$$\Sigma\Sigma^\top = A + A^\top.$$

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, $\lambda \in \mathbb{R}^{m\times n}$ is a rank n matrix with a left inverse $\lambda^{-1} \in \mathbb{R}^{n\times m}$.

Our aim is to establish convergence using similar techniques as [47] and investigate the improvements in performance. Equation (1.3) is related to the generalised Langevin equation, where memory is added to (1.2) by an integrating over past velocities with a kernel $\Gamma : \mathbb{R}_+ \to \mathbb{R}^{n\times n}$:

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

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 $T_t = T$, (1.4) is equivalent to (1.3) when setting

$$\Gamma(t) = \lambda^\top e^{-At}\lambda, \tag{1.5}$$

[1] under more restrictive conditions on the objective function or the initial condition than those considered here

and the invariant distribution becomes

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

see Section 8.2 of [56] for details[2]. 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 whilst preserving the invariant distribution in the x marginal. In the constant temperature context, (1.4) is natural from the point of statistical mechanics and has already been considered as a sampling tool in [10, 11, 12, 50] with considerable success. We will demonstrate numerically that the additional tuning parameters can improve performance; see also [49] 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 generalised Langevin dynamics (1.4) can be found in [52].

To motivate the use of (1.3), consider the quadratic case where $U = \alpha x^2$ and $0 < \alpha < 1$. This case allows for explicit or numerical calculation of the spectral gaps of the generators in (1.1)-(1.3) in order to compare the rate of convergence to equilibrium; see [53, 44] for details. For a given T, it is possible to choose λ and A, such that the spectral gap of the generator of (1.3) is much larger than that of (1.2) with the best choice of μ being used. The latter is already larger than that of the overdamped dynamics in (1.1). We will later demonstrate numerically that this will translate to better exploration in simulated annealing (when T_t is decreasing in time).

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 [64]. The latter is commonly preferred to discretising 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 optimisation; see Chapter 2 in [51] and also [26] 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, [20] has studied the long time behaviour and stability for different choices of a memory kernel as in (1.4). Finally, we note that generalised Langevin dynamics in (1.4) have additionally been studied in related areas such as sampling problems in molecular dynamics from chemical modelling [1, 10, 11, 12, 50, 68], see also [36] for work determining the kernel Γ in the generalised system (1.4) from data.

Our theoretical results will focus only on the continuous time dynamics and follow closely the approach in [47]. The main requirement in terms of assumptions are quadratic upper and lower bounds on U and bounded second derivatives. This is different to classical references such as [25], [27] or [32]. These works also rely on the Poincaré inequality, an approach which will be mirrored here (and in [47] for the underdamped case) using a log-Sobolev inequality; see also [31] 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 discretisation of (1.3), we will use a time discretisation scheme that appeared in [3], but will not present theoretical results on the time discretised dynamics; this is beyond the scope of this article. We refer instead the interested reader to [61] for a study on discretisation schemes for the system (1.3), [14] for recent results on (1.2) and its time-discretisation and [22, 23] for linking discrete time Markov chains with the overdamped Langevin system in (1.1).

1.1 Contributions and organisation of the paper

- To the best of the authors' knowledge, neither of the generalised Langevin systems (1.3) and (1.4) have been considered along with simulated annealing to solve a global optimisation problem. The main theoretical contribution consists of Theorem 2.4 that establishes convergence in probability of X_t in the higher order Markovian dynamics (1.3) to a global minimiser of U. For the optimal cooling schedule T_t, the rate of convergence is as the known rate for the Langevin system (1.2) presented in [47].

- On a more technical level, the assumptions and proofs here closely parallel those of [47] bar a number of differences. Due to the different dynamics, we use a different form of the distorted entropy, stated formally in

[2]To the best of the authors' knowledge, there is no known direct translation between (1.4) and (1.3) for a non-constant T_t; such a translation quite possibly exists and at the very least the intuition here is useful.

(C.8). In addition, we use different truncation arguments for establishing dissipation of this distorted entropy. We provide more details on differences in Remarks 2.1 and C.4. Also we make an effort to emphasise the role of the critical factor of the cooling schedule in the rate of convergence in Theorem 2.4. This can be seen in our assumptions for T_t and U below.

- Our results cover also convergence to equilibrium for the constant temperature case, which is relevant for sampling problems. In Proposition 2.5 we establish exponential convergence to equilibrium for (1.3) with $T_t = T > 0$, see also Remark 2.4. This is not surprising, see [52] for similar results.
- Detailed numerical experiments are provided to illustrate the performance of our approach. We also discuss thoroughly tuning issues. In particular, we investigate the role of matrix A and how it can be chosen to increase exploration of the state space. For the leapfrog time discretisation of [3], our results suggest that exploration of the state space is increased considerably compared to using an Euler discretisation of (1.2).

The paper is organised as follows. Section 2 will present the assumptions and main theoretical results. Detailed proofs can be found in the appendices. Section 3 presents numerical results demonstrating the effectiveness of our approach both in terms of reaching the global minimum and the exploration of the state space. In Section 4, 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 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 as $\nabla_x = (\partial_{x_1}, \ldots, \partial_{x_n})^\top$, the Hessian with D_x^2 and similarly for the y and z variables. 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 the operator norm $|M| = \sup\left\{ \frac{|Mv|}{|v|} : v \in \mathbb{R}^r \text{ with } x \neq 0 \right\}$. 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 μ_{T_t} be the instantaneous invariant law of the process

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

with $Z_{T_t} = \int \exp\big(-\frac{1}{T_t}\big(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\big)\big) dx\, dy\, dz$. Finally, define the density between the two laws:

$$h_t \quad := \quad \frac{dm_t}{d\mu_{T_t}}. \tag{2.3}$$

We proceed by stating our assumptions.

Assumption 1.

1. The potential U is smooth with bounded second derivatives

$$|D_x^2 U|_\infty := \sup_{x\in\mathbb{R}^n} |D_x^2 U(x)| < \infty \tag{2.4}$$

and satisfies

$$|\bar{a} \circ x|^2 + U_m \quad \le \quad U(x) \quad \le \quad |\bar{a} \circ x|^2 + U_M \tag{2.5}$$

$$\nabla_x U(x) \cdot x \ge r_1 |x|^2 - U_g \tag{2.6}$$

$$|\nabla_x U(x)|^2 \le r_2 |x|^2 + U_g \tag{2.7}$$

for some constants $\bar{a} \in \mathbb{R}_+^n$, $r_1, r_2, U_g > 0$, and $U_m, U_M \in \mathbb{R}$, where $\circ$ denotes the Hadamard product. In the rest of the paper, the smallest element of $\bar{a}$ is denoted as

$$a_m := \min_i \bar{a}_i,$$

where $\bar{a} = (\bar{a}_1, \ldots, \bar{a}_n)$.

2. The temperature $T_t : \mathbb{R}_+ \to \mathbb{R}_+$ is continuously differentiable and there exists some constant $t_0 > 0$ such that T_t satisfies for all $t > t_0$:
   (i) $\lim_{t\to\infty} T_t = 0$,
   (ii) $T_t \geq E(\ln t)^{-1}$ for some constant $E > U_M - U_m \geq 0$ and $\widetilde{T} > 0$,
   (iii) $-\widetilde{T}t^{-1} \leq T_t' \leq 0$ for some constant $\widetilde{T} > 0$. In addition, there exists a small $\delta > 0$ such that T_t is constant on $[0, \delta]$.
3. The initial density m_0 satisfies:
   (i) m_0 is smooth,
   (ii) $\int \frac{|\nabla m_0|^2}{m_0} dxdydz < \infty$,
   (iii) $\int (|x|^p + |y|^p + |z|^p) m_0 \, dxdydz < 0$ for $2 \leq p \leq \bar{p}$, $p \in \mathbb{N}$ and some $\bar{p} \in \mathbb{N}$.

Remark 2.1. Note that (2.5) and (2.7) deviate from [47]. The modification is useful for providing a clear characterisation of the annealing schedule (2. above) and the log-Sobolev constant in (C.17) found in the appendices. The relationship (C.17) between the log-Sobolev constant here and the critical value $U_M - U_m$ for E mirrors that between the spectral gap of the overdamped Langevin generator (of (1.1)) and the same critical value appearing in the annealing schedule in (1.1) as shown in [34] and [59]. For the overdamped case, more recent extenstions such as the Eyring-Kramers formula for the spectral gap and the log-Sobolev constant can be found in [43]. Future work could consider extension of these ideas for the underdamped and generalised Langevin case.

We present two key propositions.

Proposition 2.1. *For all $t > 0$, denote by $\left(X^{T_t}, Y^{T_t}, Z^{T_t}\right)$ a r.v. with distribution μ_{T_t}. For any $\delta, \alpha > 0$, there exists a constant $\hat{A} > 0$ such that*

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

Proposition 2.2. *Under Assumption 1, for all $t > 0$, X_t, Y_t, Z_t are well defined, $\mathbb{E}\left[|X_t|^2 + |Y_t|^2 + |Z_t|^2\right] < \infty$ and $m_t \in \mathcal{C}_+^\infty = \{m \in \mathcal{C}^\infty : m > 0\}$.*

Propositions 2.1 and 2.2 follow by modifications on the proofs of Lemma 3, Proposition 4 and Proposition 5 of [47]. The statements are restated as Propositions B.1 and B.2 along with proofs in the Appendices. The proof of Proposition 2.1 uses probabilistic arguments and the one for Proposition 2.2 Hörmander's condition and controllability.

Remark 2.2. Proposition 2.1 can be thought of as a Laplace principle; Proposition 2.2 says roughly that the process (1.3) does not blow up in finite time and and the noise in the dynamics (1.3c) for Z_t spreads throughout the system, that is to X_t and Y_t.

Proposition 2.3. *Under Assumption 1, for any $\alpha > 0$, there exists some constant $B > 0$ and $t_h > 0$, both depending on $|D_x^2 U|_\infty$, A, λ, T_t, α, U_M, U_m, U_g, E, r_2 and a_m, such that for all $t > t_h$,*

$$\int h_t \ln h_t d\mu_{T_t} \leq B\left(\frac{1}{t}\right)^{1-\frac{U_M-U_m}{E}-2\alpha}. \tag{2.8}$$

The full proof is the contained in the appendices and follows directly from Proposition C.8. Therein a similar statement is proved for the *distorted entropy* that has the following form:

$$H(t) \;:=\; \int \left(\frac{\langle S\nabla h_t, \nabla h_t\rangle}{h_t} + \beta(T_t^{-1}) h_t \ln(h_t) \right) d\mu_{T_t},$$

where S being a well chosen matrix (so that (C.14) holds) and $\beta(\cdot)$ is a particular polynomial (see (C.8) for the precise form of $H(t)$ and (C.11) for $\beta(\cdot)$). This construction of H compared to a standard definition of entropy compensates for the fact that the diffusion is degenerate (see [67] for a general discussion). The proof requires use of time derivatives of H, which is rigorously established using a truncation argument, whereby a sequence of compact functions with specific properties are multiplied onto the integrand. Then the problem is split into the partial time and partial temperature derivatives where, amongst other tools, (C.14) and a log-Sobolev inequality are used as in [47] to arrive at a bound that allows a Grönwall-type argument.

Remark 2.3. Proposition C.8 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 Theorem 2.4.

Remark 2.4. Part of the analysis used in the proof of Proposition C.8 can be used for the sampling case and $T_t = T$, i.e. working only with the partial time derivatives mentioned above for the invariant distribution. Proposition 2.5 below shows exponential convergence to equilibrium for the generalised Langevin equation (1.3) with constant temperature.

We proceed with the statement of our main result.

Theorem 2.4. *Under Assumption 1, for any $\delta > 0$, as $t \to \infty$,*

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

If in addition $T_t = E(\ln t)^{-1}$, then for any $\alpha > 0$, there exists a constant $C > 0$ such that for all $t > 0$,

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

where the exponential rate $r : (U_M - U_m, \infty) \to \mathbb{R}$ is defined by

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

Proof. For all $t > 0$, denote by $\left(X^{T_t}, Y^{T_t}, Z^{T_t}\right)$ a random variable with distribution μ_{T_t}. For all $\delta > 0$, with the definition (2.3) of h_t and triangle inequality,

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

Pinsker's inequality gives

$$\int |h_t - 1| d\mu_{T_t} \le \left(2 \int h_t \ln h_t d\mu_{T_t} \right)^{\frac{1}{2}}, \tag{2.9}$$

which, by Proposition 2.3, together with Proposition 2.1 gives the result. □

Remark 2.5. The cooling schedule $T_t = E(\ln t)^{-1}$ is optimal with respect to the method of proof for Proposition C.8; see Proposition D.2 in the appendices. This is a consistent with most works in simulated annealing, e.g. [15, 24, 25, 27, 28, 31, 32, 35].

Remark 2.6. The 'mountain-like' shape of r indicates the bottleneck for the rate of convergence at low and high values of E: a small E means convergence to the instantaneous equilibrium μ_{T_t} is slow and a large E means the convergence of μ_{T_t} to the global minima of U is slow.

Proposition 2.5. *Let 1. and 3. of Assumption 1 hold and let $T_t = T$ for all t for some constant $T > 0$. It holds that*

$$\int |h_t - 1| d\mu_T \leq \sqrt{\frac{2H(0)}{\beta(T)}} e^{-\frac{C_*^{-1}}{2}t},$$

where C_* is the log-Sobolev constant (C.17)

$$C_* = A_* + \beta(T^{-1}) e^{(U_M - U_m)T^{-1}} \frac{T}{4} \max\left(2, a_m^{-2}\right) \tag{2.10}$$

for all $t > 0$, H is the distorted entropy defined in (C.8)*, $A_* = A_*(|D_x^2 U|_\infty, \lambda) > 0$ is a constant and β is a second order polynomial* (C.11) *depending on $|D_x^2 U|_\infty$, A and λ.*

Proof. See proof of Proposition D.1 in the Appendices. □

3 Numerical results

Here we investigate the numerical performance of (1.3) in terms of convergence to a global optimum and exploration capabilities and compare with (1.2). In Section 3.1, we will present the discretisation we use for both sets of dynamics and some details related to the annealing schedule and parameters. In Section 3.2 and 3.3, for different parameters and cost functions, we present results for the probability of convergence to the global minimum and transition rates between different regions of the state space. We will investigate thoroughly the effect of E appearing in the annealing schedule as well as the parameters in the dynamics (1.2) and (1.3).

3.1 Time discretisation

In order to simulate from (1.3), we will use the following time discretisation:

$$Y_{n+\frac{1}{2}} = Y_n - \frac{\Delta t}{2}\gamma \nabla U(X_n) + \frac{\Delta t}{2}\lambda^\top Z_n \tag{3.1a}$$

$$X_{n+1} = X_n + \Delta t \gamma Y_{n+\frac{1}{2}} \tag{3.1b}$$

$$Z_{n+1} = Z_n - \theta\lambda Y_{n+\frac{1}{2}} - \theta A Z_n + \alpha\sqrt{T_k}\,\Sigma\xi_n \tag{3.1c}$$

$$Y_{n+1} = Y_{n+\frac{1}{2}} - \frac{\Delta t}{2}\gamma \nabla U(X_{n+1}) + \frac{\Delta t}{2}\lambda^\top Z_{n+1} \tag{3.1d}$$

where Δt denotes the time incremements in the discretisation, ξ_n are i.i.d. standard m dimensional normal random variables with unit variance and $\theta = 1 - \exp(-\Delta t)$, and $\alpha = \sqrt{1 - \theta^2}$. Specifically this is method 2 of [3] applied on a slight modification of (1.3), where $\gamma Y_t dt$ and $\gamma \nabla U dt$ is used instead in the r.h.s. of (1.3a) and (1.3b). Tuning γ can improve numerical perfomance especially in high dimensional problems, but we note that this has no effect in terms of the instantaneous invariant density in (2.2); γ will not appear in (2.2) similar to λ and A. Unless stated otherwise, in the remainder we will use $\gamma = 1$.

As we will see below the choices for A is important. To illustrate this we will use different choices of the form $A = \mu A_i$; i here is an index for different designs of A. The first choice will be to set $m = n$ and set $A_1 = I_n$ where I_n is $n \times n$ identity matrix. For the rest, we will use $m = 2n$ and

$$A_2 = \begin{pmatrix} 0 & -I_n \\ I_n & I_n \end{pmatrix}, \quad A_3 = \begin{pmatrix} I_n & -I_n \\ I_n & I_n \end{pmatrix}, \quad A_4 = \begin{pmatrix} 1 & \dots & & 1 \\ -1 & \ddots & & \vdots \\ \vdots & \ddots & & \\ -1 & \dots & -1 & 1 \end{pmatrix}.$$

Similarly we will use in each case $\lambda = \bar{\lambda}\lambda_i$ with $\lambda_1 = I_n$ and

$$\lambda_i = \begin{pmatrix} I_n \\ 0 \end{pmatrix}$$

for $i = 2, 3, 4$. As a result $\bar{\lambda}, \mu > 0$ are the main tuning constants for (3.1) that do not involve the annealing schedule.

The Langevin system (1.2) will be approximated with the following Euler-Maruyama scheme,

$$X_{k+1} = X_k + \Delta t \gamma Y_k \tag{3.2a}$$

$$Y_{k+1} = Y_k - \Delta t \gamma \nabla U(X_k) - \Delta t \mu Y_k + \sqrt{\Delta t \mu T_k}\xi_k. \tag{3.2b}$$

To make valid comparisons, both (3.1) and (3.2) will use the same noise realisation ξ_k (or the first common n elements) and same step size Δt. Similarly, $\gamma = 1$ unless stated otherwise. Finally for both cases we will use following annealing schedule:

$$T_k = \left(\frac{1}{5} + \frac{\ln(1 + k\Delta t)}{E}\right)^{-1},$$

where E is an additional tuning parameter.

3.2 Sample path properties

Our first set of simulations focus on illustrating some properties of the sample paths generated by (3.1) and (3.2). We will use the following bivariate potential function

$$U(x_1, x_2) = \frac{x_1^2}{5} + \frac{x_2^2}{10} + 5e^{-x_1^2} - 7e^{-(x_1+5)^2-(x_2-3)^2} - 6e^{-(x_1-5)^2-(x_2+2)^2} + \frac{5x_1^2 e^{-\frac{x_1^2}{9}} \cos(x_1 + 2x_2)\cos(2x_1 - x_2)}{1 + \frac{x_2^2}{9}}. \tag{3.3}$$

The global minimum is located at $(-5, 3)$, but there are plenty of local minima where the process can get trapped. In addition, there is a barrier along the vertical line $\{x_1 = 0\}$ that makes crossing from each half plane less likely. Here we set $\Delta t = 0.1$, $E = 5$ and each sample is initialised at $(4, 2)$. As a result, it is harder to cross $\{x_1 = 0\}$ to reach the global minimum and it is quite common to get stuck in other local minima such as near $(5, -2)$.

To illustrate this, in Figure 3.1 we present contour plots of U together with typical realisation of sample paths (in the left panels) for (3.2) and (3.1) for the different choices of A_i. As expected, (3.1) generates smoother paths than those of (3.2). We also employ independent runs of each stochastic process for the same initialisation. The results are presented in the right panels of Figure 3.1, where we show heat maps for two dimensional histograms representing the frequency of visiting each (x_1, x_2) location over 15 independent realisations of each process. The heat maps in Figure 3.1 do not directly depict time dependence in the paths and only illustrate which areas are visited more frequently. Of course converging at the global minumum or the local one at $(5, -2)$ will result in more visits at these areas. The aim here is to investigate the exploration of the state space. A careful examination of the plots shows more visits for (3.1) near $\{x_1 = 0\}$. The increased number of crossings of the vertical line $\{x_1 = 0\}$ are also demonstrated in Table 3.1 for more independent runs.

3.3 Performance and tuning

As expected, the tuning parameters, E, $\bar{\lambda}$ and μ play significant roles in the performance of (3.1) and (3.2). As E is common to both, we wish to demontrate that the additional tuning variable for (3.1) will can improve performance.

We first comment on relative scaling of $\bar{\lambda}$ and μ 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

Method equation	Number of transitions across $x = 0$
(3.2)	20609
(3.1) with $A = A_1$	21532
(3.1) with $A = A_2$	38804
(3.1) with $A = A_3$	32745
(3.1) with $A = A_4$	38948

Table 3.1: Number of crossings across the vertical line $\{x_1 = 0\}$ for U defined in (3.3). The results are for $k = 10^5$ iterations over 10^4 independent runs.

distribution. It is observed numerically in [53] 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 μ thanks to Theorem 3.1 in [44]. For the rest of the comparison, we will use $\frac{\bar{\lambda}^2}{\mu}$ and μ as variables for (3.1) and (3.2) respectively as these quantities appear to have a distinct effect on the mixing in each case.

We will consider three different cost functions U and set $\Delta t = 0.02$. As before we will initialise at a point well separated from the global minimum and consider each method to be successful if it convergences at a particular tolerance region near the global minumum. The details are presented in Table 3.2. We choose the popular Alpine function in 12 dimensions (∇U_1 here is a subgradient) and two variants of (3.3). U_2 is modified to have the same quadratic confinement in x_1 and x_2 direction and there are several additional local minima due to the last term in the sum. More importantly, compared to (3.3) (and U_3) it has a narrow region near the origin that allows easier passage through $\{x_1 = 0\}$. On the other hand U_3 similar to (3.3) except that the well near the global minimum (and the dominant local minimum at $(5, -2)$) are elongated in the direction of x_2 (and x_1 respectively).

Cost function	Initial condition	Tolerance sets
$U_1(x) = \frac{1}{2}\sum_{i=1}^{12} \lvert x_i \sin(x_i) + 0.1x_i \rvert$.	$x_j = 6\ \forall j$	$x_j \in [-2, 2]\ \forall j$
$U_2(x_1, x_2) = \frac{x_1^2}{7} + \frac{x_2^2}{7} + 5\Big(1 - e^{-9x_2^2}\Big)e^{-x_1^2}$ $-7e^{-(x_1+5)^2-(x_2-3)^2} - 6e^{-(x_1-5)^2-(x_2+2)^2}$ $+\frac{5x_1^2 e^{-\frac{x_1^2}{9}} \cos(x_1+2x_2)\cos(2x_1-x_2)}{1+\frac{x_2^2}{9}}$	$x_1 = 4,\ x_2 = 2$	$x_1 \in [-6.5, -4.5],\ x_2 \in [1.5, 4.5]$
$U_3(x_1, x_2) = \frac{x_1^2}{5} + \frac{x_2^2}{10} + 5e^{-x_1^2} - 7e^{-2(x_1+5)^2-\frac{(x_2-3)^2}{5}}$ $-6e^{-\frac{(x_1-5)^2}{5}-2(x_2+2)^2}$	$x_1 = 4,\ x_2 = 2$	$x_1 \in [-6.5, -4.5],\ x_2 \in [1.5, 4.5]$

Table 3.2: Details of three different cost functions, initialisation and tolerance regions corresponding to region of attraction of global minimum.

In Figure 3.2 we present proportions of simulations converging at the region near the global minimum for $U = U_1$ depending on E and μ for (3.2) and on E and $\frac{\bar{\lambda}^2}{\mu}$ for (3.1) based on discussion above. To produce the figures related to (3.1) after setting $E, \frac{\bar{\lambda}^2}{\mu}$ we pick a random value of μ from a grid. The aim of this procedure is to ease visualisation, reduce computational cost and to emphasise that it is $\frac{\bar{\lambda}^2}{\mu}$ that is crucial for mixing and the performance here is not a product of a tedious tuning for μ. In addition, we only look at $A = A_1, A_2, A_3$; A_4 is omitted due to numerical instabilities when implementing (3.1) for such high dimensional dynamics. The right panels of Figure 3.2 are based on final state and the left on a time average over the last 5000 iterations. In this example it is clear (3.1) results to higher probability of reaching the global minumum. Another interesting observation is that for the generalised 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 μ proportional to $\bar{\lambda}^2$, makes it easier to find a configuration of the parameters $E, \mu, \bar{\lambda}$ that leads to good perfomance, compared to using (3.2) and tuning E, μ. We believe this is linked with the increased exploration demonstrated earlier in Figure 3.1.

In Figures 3.3 and 3.4 we present similar results for U_2 and U_3. The left panels show proportions of reaching near the correct global minimum calculated using time averages near the final point and the right panels present the number of jumps across $\{x_1 = 0\}$. All results are averaged over 20 independent runs. The aim here is to measure the extent of exploration of each process similar to Table 3.1. We observe that in both cases using (3.1) leads to higher number of jumps, and this registers as a marginal improvement in the probabilities of reaching the global minumum. We believe the benefit of the higher order dynamics here are the robustness of performance for different values of E and $\frac{\bar{\lambda}^2}{\mu}$. This is especially for using A_3 and A_4. Finally we note that despite similarities between U_2 and U_3 there are significant features that are different: the sharpness in the confinement, the shape and number of attracting wells and the shape of barriers that obstruct crossing regions in the state space. This will have a direct effect in performance, which can explain the difference in performance when comparing Figures 3.3 and 3.4; U_3 is a harder cost function to minimise. The generalised Langevin dynamics can improve performance in both cases and Figures 3.3 and 3.4 show that this is possible for a wide region in the tuning variables.

4 Conclusions

We explored the possibility of using the generalised Langevin equations in the context of simulated annealing. Our main purpose was to establish convergence as 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, 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 the success proportion 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 discretisation 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 [37] for the Langevin equation for (1.3) and the extension of the work in [49] for non-identity matrices λ and A.

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) whilst retaining the appropriate equilibrium distribution. Our choice was motivated by a clear connection to the generalised Langevin equation (1.4) and link with accelerated gradient descent, but it could be the case that a different third or higher order equations could be used with possibly improved performance. Along these lines, one could consider adding skew-symmetric terms as in [17]. As regards to theory, an interesting extension could involve establishing how the results here can be extended to establish a comparison of optimisation and sampling in a nonconvex setting for an arbitrary number of dimensions similar to [40]. We leave for future work finding optimal constants in the convergence results, investigating dependance on parameters and how the limits of these parameters and constants relate to existing results for the Langevin equation in (1.2) in [47, 55]. Finally, one could also aim to extend large deviation results in [34, 41, 59] for the overdamped Langevin dynamics to the underdamped and generalised case.

Acknowledgements

The authors would like to thank Tony Lelievre, Gabriel Stoltz and Urbain Vaes for their helpful remarks. M.C. was funded under a EPSRC studentship. G.A.P. was partially supported by the EPSRC through grants EP/P031587/1, EP/L024926/1, and EP/L020564/1. N.K. and G.A.P. were funded in part by JPMorgan Chase & Co under a J.P. Morgan A.I. Research Awards 2019. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by JPMorgan Chase & Co. or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material does not constitute a solicitation or offer in any jurisdiction.

References

[1] S. A. Adelman and B. J. Garrison. Generalized Langevin theory for gas/solid processes: Dynamical solid models. *The Journal of Chemical Physics*, 65(9):3751–3761, 1976.

[2] H. AlRachid, L. Mones, and C. Ortner. Some remarks on preconditioning molecular dynamics. *SMAI J. Comput. Math.*, 4:57–80, 2018.

[3] A. D. Baczewski and S. D. Bond. Numerical integration of the extended variable generalized Langevin equation with a positive Prony representable memory kernel. *Journal of Chemical Physics*, 139(4):044107–044107, Jul 2013.

[4] C. H. Bennett. Mass tensor molecular dynamics. *Journal of Computational Physics*, 19(3):267 – 279, 1975.

[5] R. Biswas and D. R. Hamann. Simulated annealing of silicon atom clusters in Langevin molecular dynamics. *Phys. Rev. B*, 34:895–901, Jul 1986.

[6] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, and P. Gumbsch. Structural relaxation made simple. *Phys. Rev. Lett.*, 97:170201, Oct 2006.

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

[8] J. Carrillo, S. Jin, L. Li, and Y. Zhu. A consensus-based global optimization method for high dimensional machine learning problems. 09 2019.

[9] 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(6):1037–1066, 2018.

[10] M. Ceriotti. Generalized Langevin equation thermostats for ab initio molecular dynamics, 2014.

[11] M. Ceriotti, G. Bussi, and M. Parrinello. Langevin equation with colored noise for constant-temperature molecular dynamics simulations. *Phys. Rev. Lett.*, 102:020601, Jan 2009.

[12] M. Ceriotti, G. Bussi, and M. Parrinello. Colored-noise thermostats à la carte. *Journal of Chemical Theory and Computation*, 6(4):1170–1180, 2010.

[13] M. F. Chen and X. Y. Zhou. Applications of Malliavin calculus to stochastic differential equations with time-dependent coefficients. *Acta Math. Appl. Sinica (English Ser.)*, 7(3):193–216, 1991.

[14] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018.

[15] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu. Diffusion for global optimization in $\mathbf{R}^n$. *SIAM J. Control Optim.*, 25(3):737–753, 1987.

[16] A. B. Duncan, T. Lelièvre, and G. A. Pavliotis. Variance reduction using nonreversible Langevin samplers. *J. Stat. Phys.*, 163(3):457–491, 2016.

[17] A. B. Duncan, N. Nüsken, and G. A. Pavliotis. Using perturbed underdamped Langevin dynamics to efficiently sample from probability distributions. *J. Stat. Phys.*, 169(6):1098–1131, 2017.

[18] A. Durmus and E. Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

[19] A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *Ann. Probab.*, 47(4):1982–2010, 2019.

[20] S. Gadat and F. Panloup. Long time behaviour and stationary regime of memory gradient diffusions. *Ann. Inst. Henri Poincaré Probab. Stat.*, 50(2):564–601, 2014.

[21] X. Gao, M. Gurbuzbalaban, and L. Zhu. Breaking reversibility accelerates langevin dynamics for global non-convex optimization. *arXiv e-prints*, 12 2018. arXiv:1812.07725.

[22] S. B. Gelfand and S. K. Mitter. Recursive stochastic algorithms for global optimization in $\mathbf{R}^d$. *SIAM J. Control Optim.*, 29(5):999–1018, 1991.

[23] S. B. Gelfand and S. K. Mitter. Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions. *J. Optim. Theory Appl.*, 68(3):483–498, 1991.

[24] S. Gemam and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence - PAMI*, PAMI-6:721–741, 1984.

[25] S. Geman and C.-R. Hwang. Diffusions for global optimization. *SIAM J. Control Optim.*, 24(5):1031–1043, 1986.

[26] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[27] B. Gidas. Global optimization via the Langevin equation. In *1985 24th IEEE Conference on Decision and Control*, pages 774 – 778, Dec 1985.

[28] B. Gidas. Nonstationary Markov chains and convergence of the annealing algorithm. *J. Statist. Phys.*, 39(1-2):73–131, 1985.

[29] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[30] A. Guionnet and B. Zegarlinski. Lectures on logarithmic Sobolev inequalities. In *Séminaire de Probabilités, XXXVI*, volume 1801 of *Lecture Notes in Math.*, pages 1–134. Springer, Berlin, 2003.

[31] R. Holley and D. Stroock. Simulated annealing via Sobolev inequalities. *Comm. Math. Phys.*, 115(4):553–569, 1988.

[32] 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(2):333–347, 1989.

[33] C.-R. Hwang. Laplace's method revisited: weak convergence of probability measures. *Ann. Probab.*, 8(6):1177–1182, 1980.

[34] 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(3):253–295, 1990.

[35] 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(1):169–185, 1987.

[36] H. Lei, N. A. Baker, and X. Li. Data-driven parameterization of the generalized Langevin equation. *Proc. Natl. Acad. Sci. USA*, 113(50):14183–14188, 2016.

[37] B. Leimkuhler and C. Matthews. *Molecular dynamics*, volume 39 of *Interdisciplinary Applied Mathematics*. Springer, Cham, 2015. With deterministic and stochastic numerical methods.

[38] 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(2):237–274, 2013.

[39] 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, pages 1788–1794. AAAI Press, 2016.

[40] 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(42):20881–20885, 2019.

[41] D. Márquez. Convergence rates for annealing diffusion processes. *Ann. Appl. Probab.*, 7(4):1118–1139, 1997.

[42] J. C. Mattingly and A. M. Stuart. Geometric ergodicity of some hypo-elliptic diffusions for particle motions. volume 8, pages 199–214. 2002. Inhomogeneous random systems (Cergy-Pontoise, 2001).

[43] G. Menz and A. Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *Ann. Probab.*, 42(5):1809–1884, 2014.

[44] 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(1):40–60, 2002.

[45] D. Michel. Conditional laws and Hörmander's condition. In *Stochastic analysis (Katata/Kyoto, 1982)*, volume 32 of *North-Holland Math. Library*, pages 387–408. North-Holland, Amsterdam, 1984.

[46] L. Miclo. Recuit simulé sur $\mathbf{R}^n$. Étude de l'évolution de l'énergie libre. *Ann. Inst. H. Poincaré Probab. Statist.*, 28(2):235–266, 1992.

[47] P. Monmarché. Hypocoercivity in metastable settings and kinetic simulated annealing. *Probab. Theory Related Fields*, 172(3-4):1215–1248, 2018.

[48] P. Monmarché. Generalized Γ calculus and application to interacting particles on a graph. *Potential Anal.*, 50(3):439–466, 2019.

[49] W. Mou, Y.-A. Ma, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan. High-Order Langevin Diffusion Yields an Accelerated MCMC Algorithm. *arXiv e-prints*, page arXiv:1908.10859, Aug 2019.

[50] M. Nava, M. Ceriotti, C. Dryzun, and M. Parrinello. Evaluating functions of positive-definite matrices using colored-noise thermostats. *Phys. Rev. E*, 89:023302, Feb 2014.

[51] Y. Nesterov. *Lectures on convex optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, Cham, 2018. Second edition of [MR2142598].

[52] M. Ottobre and G. A. Pavliotis. Asymptotic analysis for the generalized Langevin equation. *Nonlinearity*, 24(5):1629–1653, 2011.

[53] M. Ottobre, G. A. Pavliotis, and K. Pravda-Starov. Exponential return to equilibrium for hypoelliptic quadratic systems. *J. Funct. Anal.*, 262(9):4000–4039, 2012.

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

[55] G. Pavliotis, G. Stoltz, and U. Vaes. The generalized Langevin equation: long-time behavior and diffusive transport in a periodic potential. preprint, 2020.

[56] G. A. Pavliotis. *Stochastic processes and applications*, volume 60 of *Texts in Applied Mathematics*. Springer, New York, 2014. Diffusion processes, the Fokker-Planck and Langevin equations.

[57] M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Probab.*, 8(1):10–44, 1998.

[58] 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(1):183–204, 2017.

[59] G. Royer. A remark on simulated annealing of diffusion processes. *SIAM J. Control Optim.*, 27(6):1403–1408, 1989.

[60] S. Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016.

[61] M. Sachs. *The generalised Langevin equation: asymptotic properties and numerical analysis*. PhD thesis, The University of Edinburgh, 2017.

[62] H. Song, I. Triguero, and E. Özcan. A review on the self and dual interactions between machine learning and optimisation. *Progress in Artificial Intelligence*, 8(2):143–165, 2019.

[63] D. W. Stroock and S. R. S. Varadhan. On the support of diffusion processes with applications to the strong maximum principle. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. III: Probability theory*, pages 333–359, 1972.

[64] 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:Paper No. 153, 43, 2016.

[65] Y. Sun and A. Garcia. Interactive diffusions for global optimization. *J. Optim. Theory Appl.*, 163(2):491–509, 2014.

[66] G. Teschl. *Ordinary differential equations and dynamical systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[67] C. Villani. Hypocoercivity. *Mem. Amer. Math. Soc.*, 202(950):iv+141, 2009.

[68] X. Wu, B. R. Brooks, and E. Vanden-Eijnden. Self-guided Langevin dynamics via generalized Langevin equation. *J Comput Chem*, 37(6):595–601, Mar 2016.

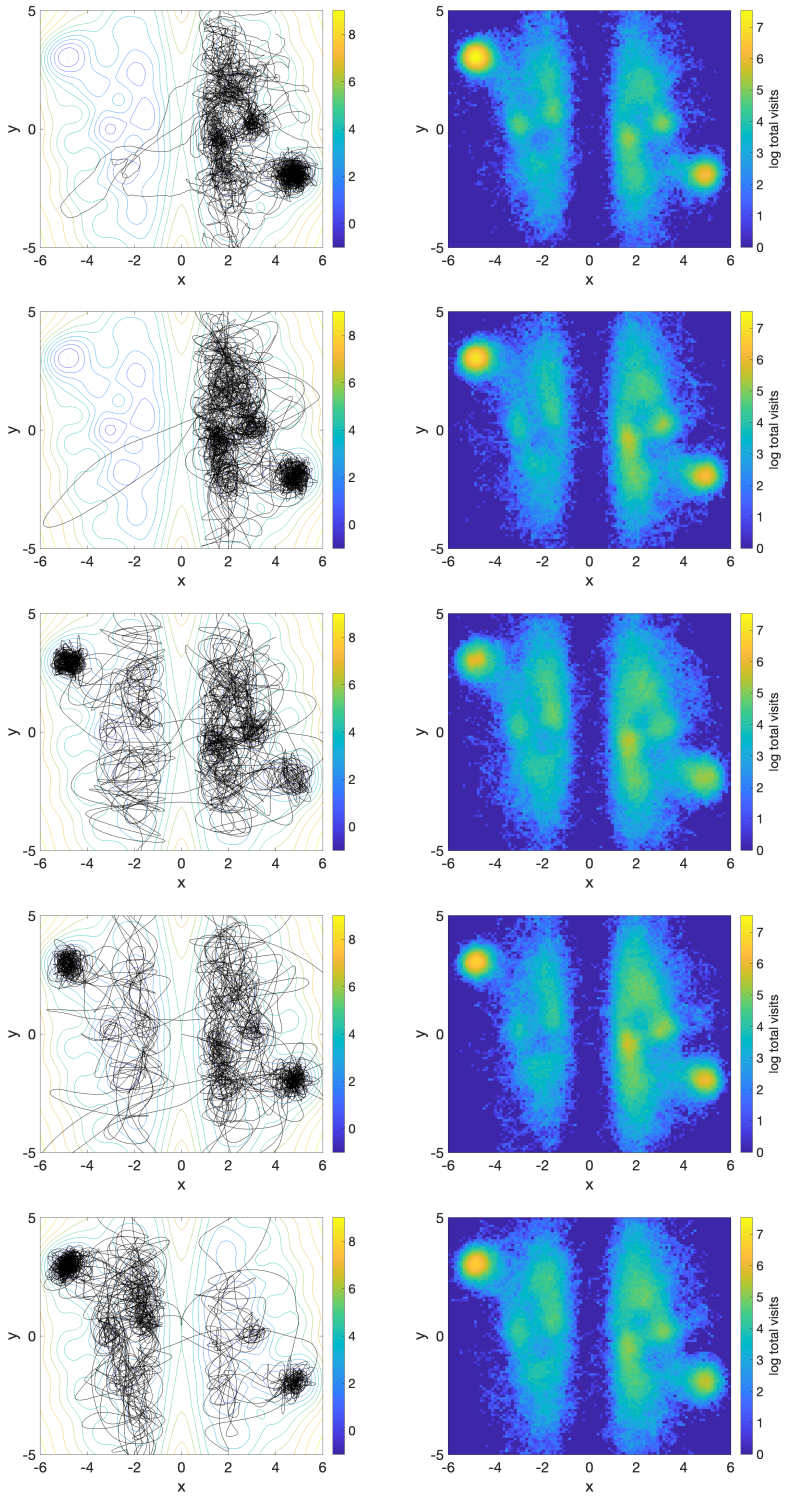

Figure 3.1: Dynamics in order from top: (3.2), (3.1) with $A = A_1, \ldots, A_4$. Left: One instance of noise realisation. Right: Log histogram of 20 independent runs.

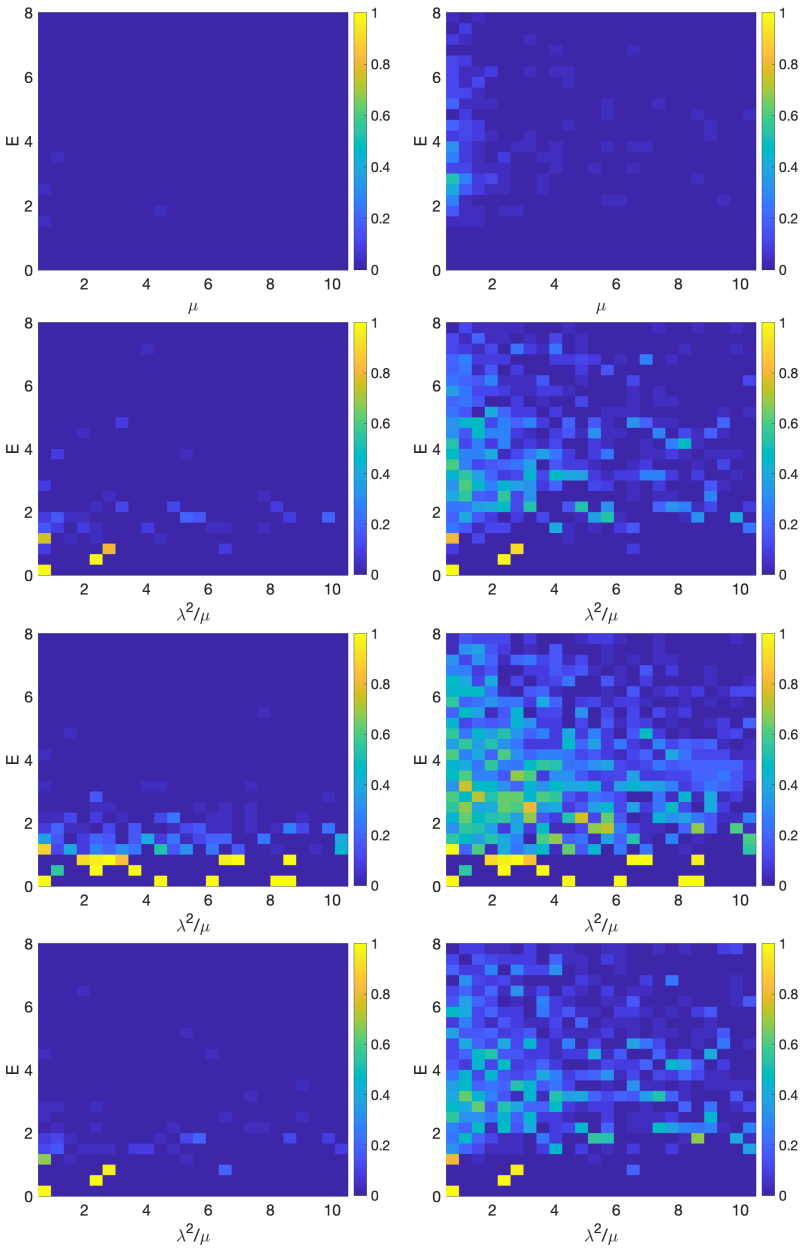

Figure 3.2: Proportion of simulations satisfying optimality tolerance for $U = U_1$. Panels from top to bottom: (3.2), (3.1) with $A = A_1, A_2, A_3$ (A_4 is omitted due to numerical instabilities when implementing (3.1)). Left: Final position, right: time-average of last 5000 iterations. We use $\gamma = 3$ for improving visualisation, the results and improvement in using (3.1) are similar for the case of $\gamma = 1$. Results here are for 20 independent runs and $k \leq 5 \cdot 10^4$ -iterations.

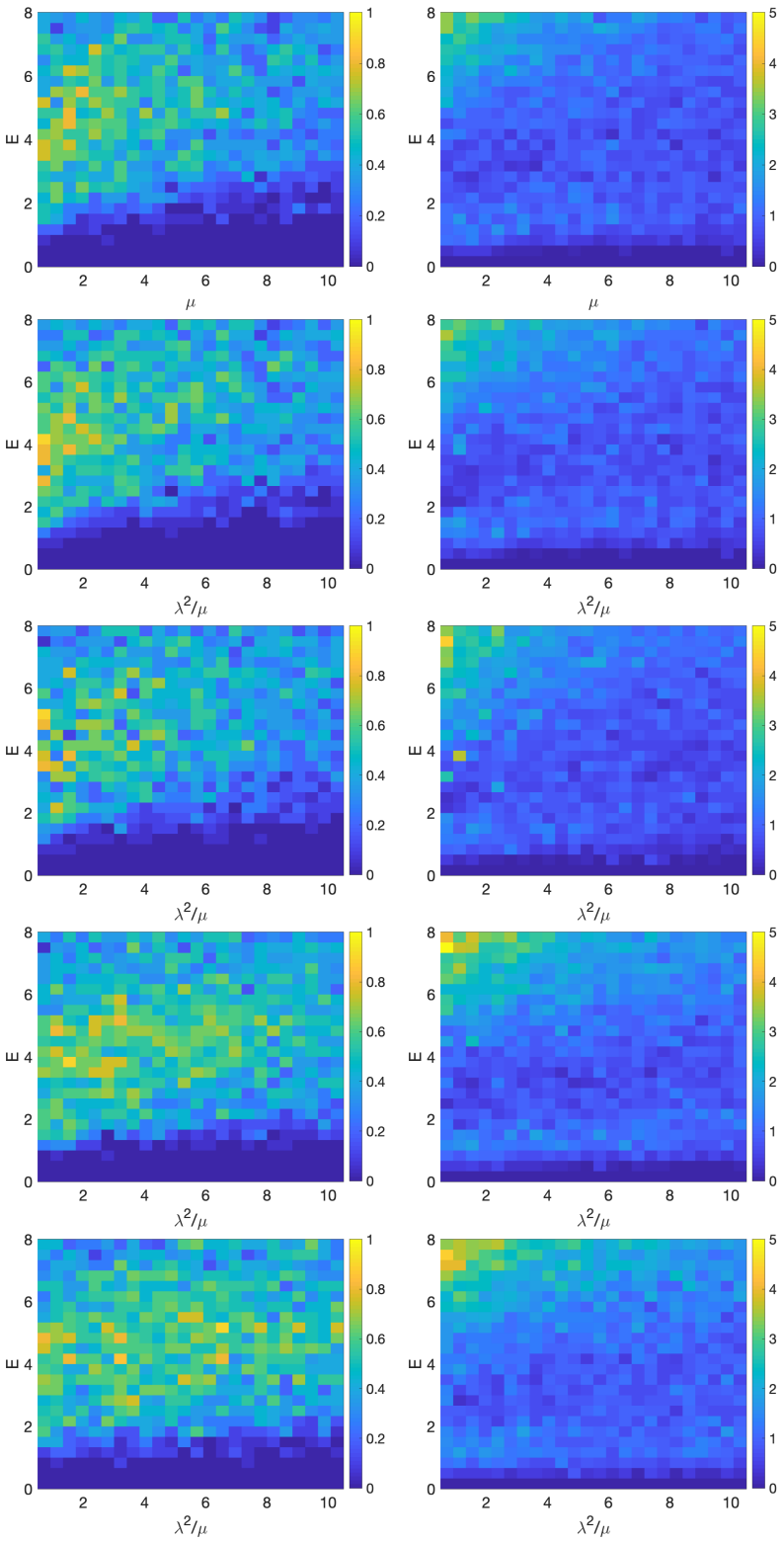

Figure 3.3: Both proportion of success and numerical transition rates for $U = U_2$. Panels from top to bottom: (3.2), (3.1) with $A = A_1, A_2, A_3, A_4$. Left: Proportion satisfying the optimality tolerance for time-average of last 5000 iterations. Right: Average number of crossings across $\{x_1 = 0\}$ for each independent run. The remaining details are as in caption of Figure 3.2.

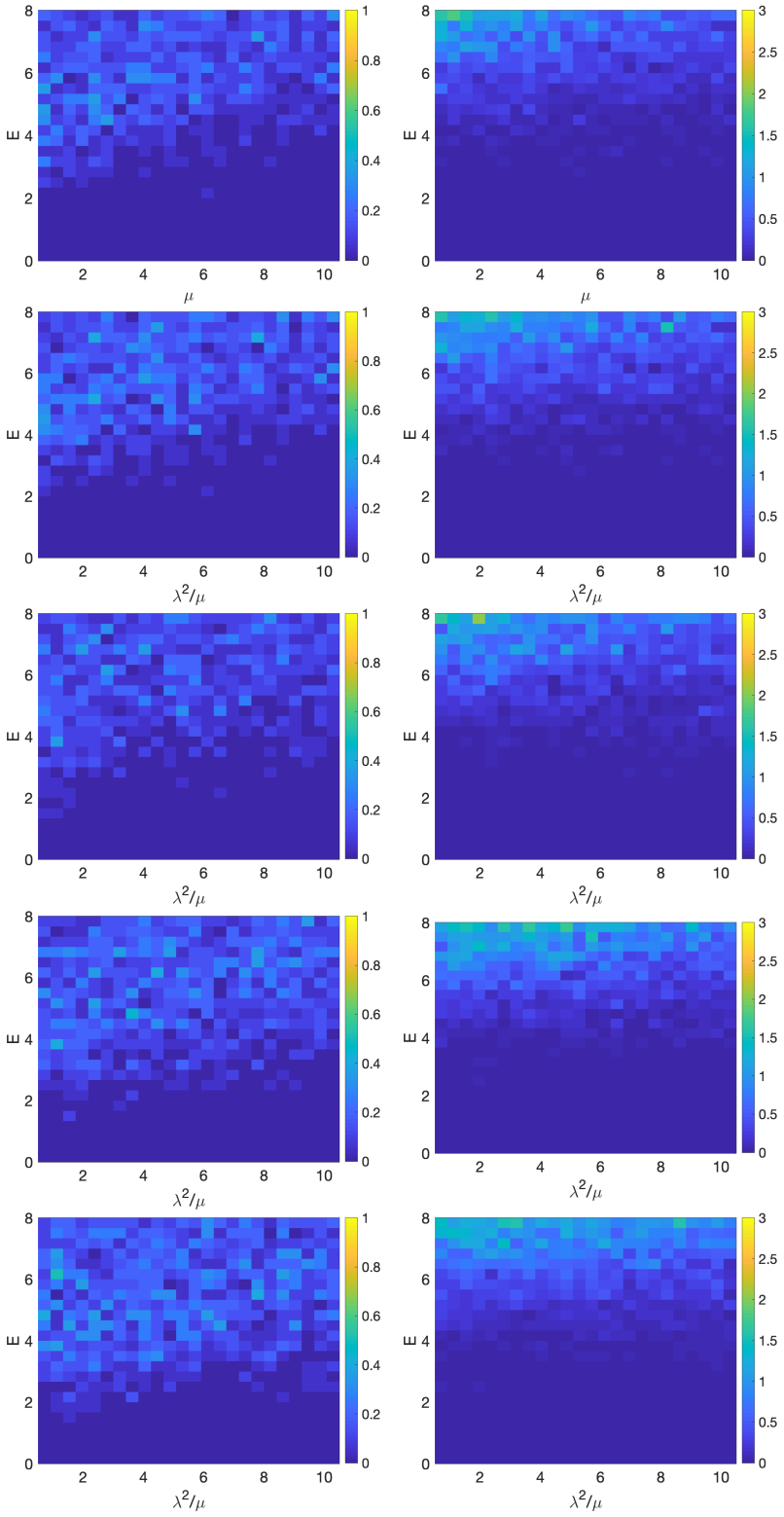

Figure 3.4: Results for $U = U_3$. Details are as in caption of Figure 3.3.

Appendices

Appendix A Notation and preliminaries

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

$$L_t(\phi(f)) = \phi'(f)L_t(f) + \phi''(f)\Gamma_t(f), \tag{A.1}$$

where Γ_t is the carré du champ operator for L_t given by

$$\Gamma_t(f) = \frac{1}{2}L_t(f^2) - fL_t(f) = \nabla_z f \cdot (A\nabla_z f). \tag{A.2}$$

The $L^2(\mu_{T_t})$ adjoint L_t^* of L_t is

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

Recall from Monmarché [47, 48] for $\Phi : \mathcal{A}_+ \to \mathcal{A}$, where $\mathcal{A}$ and $\mathcal{A}_+$ are appropriate spaces (namely $\mathcal{A}$ is assumed to be an algebra contained in the domain $\mathcal{D}(L_t^*)$ of L_t^* fixed by L_t^* and $\mathcal{A}_+ = \{f \in \mathcal{A} : f \geq 0\}$), differentiable in the sense that for any $f, g \in \mathcal{A}_+$,

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

exists for all $x \in \mathbb{R}^{2n+m}$, the Γ_Φ operator for L_t^* is defined by

$$\Gamma_{L_t^*,\Phi}(h) \quad := \quad \frac{1}{2}(L_t^*\Phi(h) - d\Phi(h).(L_t^* h)). \tag{A.3}$$

It will be helpful to keep in mind that L_t^* is only a term away from satisfying the standard chain and product rules:

$$L_t^*(\psi(f)) \;=\; \psi'(f)L_t^* f + \psi''(f)\nabla_z f \cdot (A\nabla_z f) \tag{A.4}$$

$$L_t^*(fg) \;=\; fL_t^*(g) + gL_t^*(f) + \frac{1}{2}\nabla_z f \cdot ((A + A^\top)\nabla_z g) \tag{A.5}$$

for all $f, g \in \mathcal{A}$ and $\psi \in \mathcal{C}^\infty$. $\nabla_z f \cdot (A\nabla_z f)$ and $\frac{1}{2}\nabla_z f \cdot ((A + A^\top)\nabla_z g)$ are respectively the carré du champ and its symmetric bilinear operator via polarisation for L_t^*.
In addition, square brackets on a scalar-valued D_1 and a vector-valued operator D_2 both acting on scalar-valued functions denote the commutator bracket as follows:

$$[D_1, D_2]h = (D_1(D_2 h)_1 - (D_2 D_1 h)_1, \; \dots) \tag{A.6}$$

for h smooth enough; this will be used as in (C.12).

Appendix B Auxiliary results

Proposition B.1. *For all $t > 0$, denote by $\left(X^{T_t}, Y^{T_t}, Z^{T_t}\right)$ a r.v. with distribution μ_{T_t}. For any $\delta, \alpha > 0$, there exists a constant $\hat{A} > 0$ such that*

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

Proof. The result follows exactly as in Lemma 3 in [47]. □

Proposition B.2. *Under Assumption 1, for all $t > 0$, X_t, Y_t, Z_t are well defined, $\mathbb{E}\big[|X_t|^2 + |Y_t|^2 + |Z_t|^2\big] < \infty$ and $m_t \in \mathcal{C}_+^\infty = \{m \in \mathcal{C}^\infty : m > 0\}$.*

Proof. Nonexplosiveness and finiteness of second moments follow as in Proposition 4 in [47] with the modification

$$G(x,y,t) = \frac{1}{T_t}\bigg(U(x) - \min_{\mathbb{R}^n} U + \frac{|y|}{2} + \frac{|z|}{2}\bigg).$$

One can establish that $(\partial_t + L_t)G \leq CG$ for some constant C depending on t, T_t. Nonexplosiveness and $\mathbb{E}[G(X_t, Y_t, Z_t]] < \infty$ follow by Markov's inequality and Ito's formula, which implies finite second moments for the process.

For smoothness of the law m_t of (X_t, Y_t, Z_t), Theorem 1.2 in [13] can be used.[3]
For positivity of m_t, the steps in Lemma 3.4 of [42] can be followed: for simplicity, consider the case $m = n$, $\lambda = A = I_n$ and $T_t = 1$, where the associated control problem becomes

$$\dot{X}_t + \ddot{X}_t + \dddot{X}_t + \nabla U(X_t) + D_x^2 U(X_t)\dot{X}_t = \dot{V}_t, \qquad X_0 = x_0 \in \mathbb{R}^n \tag{B.1}$$

where $V_\cdot : \mathbb{R}_+ \to \mathbb{R}^n$ is a time-varying control and dots indicate partial time derivatives. Given an arbitrary point $X^* \in \mathbb{R}^n$, set for some fixed $T > 0$ the control

$$V_t = \int \bigg(\frac{X^* - x_0}{T} + \nabla U\bigg(\frac{tX^* + (T-t)x_0}{T}\bigg) + D_x^2 U\bigg(\frac{tX^* + (T-t)x_0}{T}\bigg)\frac{X^* - x_0}{T}\bigg)dt.$$

By the boundedness assumption on the second derivatives of U, the unique solution to the control problem (B.1) is

$$X_t = \frac{tX^* + (T-t)x_0}{T}. \tag{B.2}$$

Now since with non-zero probability Brownian motion stays within an ϵ-neighbourhood of any continuously differentiable path, and in particular of V_t, then positivity of m_t follows by the support theorem of Stroock and Varadhan (Theorem 5.2 in [63]).
For the general case the initial and final values for $Y_\cdot$ and $Z_\cdot$ are not as above, that is, when $Y_0 = \dot{X}_0$, $Z_0 - \nabla U(X_0) = \ddot{X}_0$, $Y_T = \dot{X}_T$ and $Z_T - \nabla U(X_T) = \ddot{X}_T$ are arbitrary values. Then it is easy to initialise and finalise $\dot{X}_t$ and $\ddot{X}_t$ at some small $\hat{t} > 0$ and $T - \hat{t}$ respectively and extend the previous argument using a piecewise definition of V_t, e.g. see Lemma 4.2 and Appendix of [20]. □

Appendix C Proof of Proposition 2.3

The following effort up to Proposition C.8 is towards showing dissipation of a distorted entropy as required in the proof of Theorem 2.4.

C.1 Lyapunov function

Lemma C.1. *Let $\delta < 0$ be a small enough constant and $R : \mathbb{R}^{2n+m+1} \to \mathbb{R}$ be defined as*

$$R(x,y,z,T_t) \quad := \quad U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} + \delta T_t\bigg(y^\top \lambda^{-1} z + \frac{1}{2}x \cdot y\bigg). \tag{C.1}$$

Then there exist constants $a, b, c, d > 0$ such that

$$a(|x|^2 + |y|^2 + |z|^2) - d \quad \leq \quad R(x,y,z,T_t) \quad \leq \quad b(|x|^2 + |y|^2 + |z|^2) + d \tag{C.2}$$

and

$$L_t(R) \quad \leq \quad -cT_t R + d. \tag{C.3}$$

[3] This work caters to the unbounded coefficients in the stochastic differential equation, the time-dependence of one of the coefficients and the need for using the general Hörmander's condition involving the 'X_0' operator rather than the restrictive one, see [45] for definitions. Indeed, the constant assumption 2.(iii) on the annealing schedule for a small interval at the beginning is used so that this theorem can be applied.

Proof. The first statement is clear by the quadratic assumption (2.5) on U for small enough δ. For the second statement, fix δ to be small enough for the first statement and additionally to satisfy

$$\delta \le \frac{A_c}{2}\left[\left(\frac{|\lambda|^2}{2r_1}+1+\frac{r_2}{r_1}\left|\lambda^{-1}\right|^2\right)\left(\max_{t\ge 0}T_t\right)^2+2|A|^2\left|\lambda^{-1}\right|^2\right]^{-1}, \tag{C.4}$$

where $|\cdot|$ is the operator norm and $A_c>0$ is the coercivity constant of the positive definite matrix A.
Consider each of the terms of $L_t(R)$ seperately,

$$L_t\left(U(x)+\frac{|y|^2}{2}+\frac{|z|^2}{2}\right)=-\frac{1}{T_t}z^\top Az+\mathrm{Tr}A. \tag{C.5}$$

Using the quadratic bound (2.7) on $\nabla_x U$,

$$\begin{aligned}
L_t(y^\top\lambda^{-1}z) &= -\nabla_x U\lambda^{-1}z+|z|^2-|y|^2-T_t^{-1}z^\top A(\lambda^{-1})^\top y\\
&\le \frac{r_1}{4r_2}|\nabla_x U|^2+\frac{r_2}{r_1}\left|\lambda^{-1}\right|^2|z|^2+|z|^2-|y|^2+\frac{|y|^2}{4}\\
&\quad +T_t^{-2}|A|^2\left|\lambda^{-1}\right|^2|z|^2\\
&\le \frac{r_1}{4}|x|^2+\frac{r_1}{4r_2}U_g-\frac{3}{4}|y|^2\\
&\quad +\left(1+\frac{r_2}{r_1}\left|\lambda^{-1}\right|^2+T_t^{-2}|A|^2\left|\lambda^{-1}\right|^2\right)|z|^2.
\end{aligned} \tag{C.6}$$

Then also with the bound (2.6) for $\nabla_x U\cdot x$,

$$\begin{aligned}
L_t(x\cdot y) &= |y|^2-\nabla_x U\cdot x+z^\top\lambda x\\
&\le |y|^2-r_1|x|^2+U_g+\frac{|\lambda|^2}{r_1}|z|^2+\frac{r_1}{4}|x|^2.
\end{aligned} \tag{C.7}$$

Combining (C.5), (C.6), (C.7), given a large enough $C>0$,

$$\begin{aligned}
L_t(R(x,y,z,T_t)) &= L_t\left(U(x)+\frac{|y|^2}{2}+\frac{|z|^2}{2}\right)+\delta T_t L_t(y^\top\lambda z)+\frac{\delta T_t}{2}L_t(x\cdot y)\\
&\le -\delta T_t\frac{r_1}{8}|x|^2-\delta T_t\frac{1}{4}|y|^2-\frac{1}{T_t}z^\top Az+C\\
&\quad +\delta T_t\left[\frac{|\lambda|^2}{2r_1}+\left(1+\frac{r_2}{r_1}\left|\lambda^{-1}\right|^2+T_t^{-2}|A|^2\left|\lambda^{-1}\right|^2\right)\right]|z|^2.
\end{aligned}$$

Therefore for δ satisfying the bound (C.4), the $|z|^2$ term can be bounded,

$$\begin{aligned}
L_t(R(x,y,z,T_t)) &\le -\delta T_t\frac{r_1}{8}|x|^2-\delta T_t\frac{1}{4}|y|^2-\frac{A_c}{2\max_{t\ge 0}T_t}|z|^2+C\\
&\le -\frac{DT_t}{b}R+C',
\end{aligned}$$

where $D>0$ is small enough, $C'>0$ is large enough and the right inequality of (C.2) has been used.

□

Lemma C.2. *For $2\le p\le\bar{p}$ with $p,\bar{p}\in\mathbb{N}$ from Assumption 1 on m_0, $\frac{\mathbb{E}[R(X_t,Y_t,Z_t,T_t)^p]}{(\ln(e+t))^p}$ is bounded uniformly in time.*

Proof. It is equivalent to prove the result for $R+d>0$ in place of R for any $p\le\bar{p}$. Use induction w.r.t. p. The case for $p=0$ is obvious. Let

$$R_t \quad := \quad R(X_t,Y_t,Z_t,T_t).$$

Consider the following terms separately

$$\frac{d}{dt}\mathbb{E}[(R_t+d)^p] = \partial_t \mathbb{E}[(R_t+d)^p] + T_t' \partial_{T_t}\mathbb{E}[(R_t+d)^p].$$

Firstly, by definition (C.1) of R and the left bound of (C.2),

$$\begin{aligned} T_t'\partial_{T_t}\mathbb{E}[(R_t+d)^p] &= T_t'\mathbb{E}\left[p(R_t+d)^{p-1}\delta\left(y^\top\lambda^{-1}z+\frac{1}{2}x\cdot y\right)\right] \\ &\le |T_t'|\mathbb{E}\left[p(R_t+d)^{p-1}\delta\left|y^\top\lambda^{-1}z+\frac{1}{2}x\cdot y\right|\right] \\ &\le \frac{B_p}{t}\mathbb{E}[(R_t+d)^p] \end{aligned}$$

for a constant $B_p \ge 0$.

With (A.1) and (A.2), for $p \ge 1$, using property (C.3) from Lemma C.1 and again the left bound of (C.2), consider the auxiliary expression

$$\begin{aligned} \partial_t\mathbb{E}[(R_t+d)^p] &= \mathbb{E}[(L_t((R+d)^p))(X_t,Y_t,Z_t,T_t)] \\ &= \mathbb{E}[p(R_t+d)^{p-1}(L_t(R))(X_t,Y_t,Z_t,T_t)) \\ &\quad + p(p-1)(R_t+d)^{p-2}(\Gamma_t(R))(X_t,Y_t,Z_t,T_t)] \\ &\le \mathbb{E}[p(R_t+d)^{p-1}(-cT_tR_t+d) \\ &\quad + p(p-1)(R_t+d)^{p-2}(Z_t+\delta T_t(\lambda^{-1})^\top Y_t)\cdot(A(Z_t+\delta T_t(\lambda^{-1})^\top Y_t))] \\ &\le -cpT_t\mathbb{E}[(R_t+d)^p] + A_p\mathbb{E}[(R_t+d)^{p-1}] \end{aligned}$$

for a constant $A_p \ge 0$. Then with the induction assumption, the slow-decay assumption on T_t, for large enough t,

$$\begin{aligned} \frac{d}{dt}(e^{\frac{c}{2}p\int_0^t T_s ds}\mathbb{E}[(R_t+d)^p]) &\le A_p e^{\frac{c}{2}p\int_0^t T_s ds}(\ln(e+t))^{p-1}, \\ \mathbb{E}[(R_t+d)^p] &\le \mathbb{E}[(R_0+d)^p]e^{-\frac{c}{2}p\int_0^t T_s ds} + \int_0^t A_p e^{-\frac{c}{2}p\int_s^t T_u du}(\ln(e+s))^{p-1}ds \\ &\le \mathbb{E}[(R_0+d)^p] + A_p(\ln(e+t))^{p-1}\int_0^t e^{-\frac{c}{2}pT_t(t-s)}ds \\ &\le \mathbb{E}[(R_0+d)^p] + A_p(\ln(e+t))^{p-1}\frac{1-e^{-\frac{c}{2}pT_t t}}{\frac{c}{2}pT_t}. \end{aligned}$$

Using the assumptions on m_0 and again the assumption $T_t \ge E(\ln t)^{-1}$, the result follows.

□

Corollary C.3. *For any $2 \le p \le \bar{p}$, $\frac{\mathbb{E}\left[\left(|X_t|^2+|Y_t|^2+|Z_t|^2\right)^p\right]}{(\ln(e+t))^p}$ is bounded uniformly in time.*

Proof. By the lower bound on R in (C.2),

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

which concludes by Lemma C.2 after expanding. □

C.2 Form of Distorted Entropy

Let $H(t)$ be the distorted entropy

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

where $S_0, S_1 > 0$ are the constants

$$S_0 := (1 + |D_x^2 U|_\infty^2)^{\frac{1}{2}}, \tag{C.9}$$

$$S_1 := 2 + 156S_0^2 + 1024S_0^4 \tag{C.10}$$

and β is a second order polynomial

$$\beta(T_t^{-1}) := 1 + \beta_0 + \beta_1 T_t^{-1} + \beta_2 T_t^{-2} \tag{C.11}$$

with large enough coefficients $\beta_0, \beta_1, \beta_2 > 0$ depending on $|D_x^2 U|_\infty$, $\left|A^\top\right|$ and

$$\hat{\lambda}^2 := \max\left(|\lambda|^2, \left|\lambda^\top\right|^2, \left|\lambda^{-1}\right|^2, \left|\lambda^{-1}\right|\left|\lambda^\top\right|\right).$$

Remark C.1. This particular expression for H is not necessarily the best choice and it is quite possible to have only ∇_x and ∇_y and no ∇_z appearing in the first integrand for instance. However the above is a working expression and optimality of this is left as future work.

First, an auxiliary result which can be found as Lemma 12 of [47] is stated along with its proofs from [47] since the proof is not too long. Notice the Φ^* appearing in Lemma C.4 appears in the first two terms of $H(t)$.

Lemma C.4. *For*

$$\Phi^*(h) = \frac{|M\nabla h|^2}{h},$$

where M is matrix-valued, we have

$$\Gamma_{L_t^*,\Phi^*}(h) \geq \frac{(M\nabla h)\cdot[L_t^*, M\nabla]h}{h}$$

for all $0 < h \in \mathcal{A}_+$, where the square bracket denotes the commutator vector (A.6)*, i.e.*

$$[L_t^*, M\nabla]h = (L_t^*(M\nabla h)_1 - (M\nabla L_t^* h)_1, \dots). \tag{C.12}$$

Proof. The second term in definition (A.3) of $\Gamma_{L_t^*,\Phi}(h)$ can be calculated to be

$$-d\Phi^*(h).L_t^* h = -\frac{2}{h}(M\nabla h)\cdot(M\nabla L_t^* h) + \frac{L_t^* h}{h^2}|M\nabla h|^2. \tag{C.13}$$

Using the adjusted product and chain rules (A.5) and (A.4) for L_t^*, the first term in the definition (A.3) of $\Gamma_{L_t^*,\Phi}$ can be calculated to be

$$\begin{aligned}
L_t^*(\Phi^*(h)) &= \frac{1}{h}L_t^*(|M\nabla h|^2) + |M\nabla h|^2 L_t^*\left(\frac{1}{h}\right) \\
&\quad - \sum_i \frac{2}{h^2}(M\nabla h)_i \nabla_z h \cdot ((A + A^\top)\nabla_z (M\nabla h)_i) \\
&= \frac{2}{h}\left((M\nabla h)\cdot L_t^* M\nabla h + \sum_i \nabla_z (M\nabla h)_i \cdot (A\nabla_z (M\nabla h)_i)\right) \\
&\quad + |M\nabla h|^2 L_t^*\left(\frac{1}{h}\right) - \sum_i \frac{2}{h^2}(M\nabla h)_i \nabla_z h \cdot ((A + A^\top)\nabla_z (M\nabla h)_i),
\end{aligned}$$

where the last summands can be bounded below with the positive definite property of A,

$$\begin{aligned}
&-\frac{2}{h^2}(M\nabla h)_i\nabla_z h\cdot((A+A^\top)\nabla_z(M\nabla h)_i)\\
&\quad=-\frac{2}{h^2}(M\nabla h)_i(\nabla_z h)^\top A(\nabla_z(M\nabla h)_i)-\frac{1}{h^2}(\nabla_z(M\nabla h)_i)^\top A((M\nabla h)_i\nabla_z h)\\
&\quad\geq-\frac{2}{h^3}(M\nabla h)_i^2(\nabla_z h)^\top A\nabla_z h-\frac{2}{h}(\nabla_z(M\nabla h)_i)^\top A\nabla_z(M\nabla h)_i,
\end{aligned}$$

which produces the bound

$$L_t^*(\Phi^*(h))\geq\frac{2}{h}(M\nabla h)\cdot L_t^*M\nabla h+|M\nabla h|^2L_t^*\left(\frac{1}{h}\right)-\frac{2}{h^3}|M\nabla h|^2\nabla_z h\cdot(A\nabla_z h).$$

Combining this with (C.13) then using the adjusted chain rule (A.4),

$$\begin{aligned}
\Gamma_{L_t^*,\Phi^*}(h)&\geq\frac{1}{h}(M\nabla h)\cdot[L_t^*,M\nabla]h+\frac{1}{2}|M\nabla h|^2\left(L_t^*\left(\frac{1}{h}\right)+\frac{L_t^*h}{h^2}\right)\\
&\quad-\frac{1}{h^3}|M\nabla h|^2\nabla_z h\cdot(A\nabla_z h)\\
&=\frac{1}{h}(M\nabla h)\cdot[L_t^*,M\nabla]h.
\end{aligned}$$

□

With Lemma C.4, the distorted entropy (C.8) can be shown to be a correct one with the following proposition.

Proposition C.5. *Let Ψ_{T_t} be the operator appearing in the integrand of the distorted entropy H, that is*

$$\begin{aligned}
\Psi_{T_t}(h)\ :=\ &\frac{\left|2\nabla_x h+8S_0(\nabla_y h+\lambda^{-1}\nabla_z h)\right|^2}{h}+\frac{\left|\nabla_y h+S_1\lambda^{-1}\nabla_z h\right|^2}{h}\\
&+\beta(T_t^{-1})h\ln(h)
\end{aligned}$$

for $h\in\mathcal{A}_+$. It holds that

$$\Gamma_{L_t^*,\Psi_{T_t}}(h)\geq\frac{|\nabla h|^2}{h}.\tag{C.14}$$

Remark C.2. H satisfying this property is crucial for proving dissipation in Proposition C.8 and was the main consideration when making remark C.1.

Proof. Let Φ_1,Φ_2,Φ_3 be defined by

$$\Psi_{T_t}=:\Phi_1+\Phi_2+\beta(T_t^{-1})\Phi_3.$$

Note that the Γ_Φ operator is linear in the second operator argument by linearity of L_t^*, so that (C.14) is

$$\Gamma_{L_t^*,\Phi_1}(h)+\Gamma_{L_t^*,\Phi_2}(h)+\beta(T_t^{-1})\Gamma_{L_t^*,\Phi_3}(h)\geq\frac{|\nabla h|^2}{h}.$$

Consider $\Gamma_{L_t^*,\Phi_3}$ first. Using the definition (A.3) of $\Gamma_{L_t^*,\Phi}$ and the product and chain rule (A.5) and (A.4) for L_t^*,

$$\begin{aligned}
\Gamma_{L_t^*,\Phi_3}(h)&=\frac{1}{2}\left(\ln hL_t^*h+hL_t^*\ln h+\frac{2|\nabla_z h|^2}{h}-(1+\ln h)L_t^*h\right)\\
&=\frac{1}{2}\left(\ln hL_t^*h+L_t^*h+\frac{|\nabla_z h|^2}{h}-(1+\ln h)L_t^*h\right)\\
&=\frac{|\nabla_z h|^2}{2h}.
\end{aligned}\tag{C.15}$$

Since the goal is to show (C.14), the availability of the term (C.15) means that regardless of how negative of a contribution $\Gamma_{L_t^*,\Phi_1}$ and $\Gamma_{L_t^*,\Phi_2}$ makes in the z-derivative term in (C.15), it is not a concern; this is reflected in the factor β.

For $\Gamma_{L_t^*,\Phi_1}$ and $\Gamma_{L_t^*,\Phi_2}$ we use $S_0, S_1 > 0$ defined as before in (C.9)-(C.10).

Lemma C.4 gives

$$\begin{aligned}
h\Gamma_{L_t^*,\Phi_1}(h) &\geq (2\nabla_x + 8S_0(\nabla_y + \lambda^{-1}\nabla_z))h \cdot [L_t^*, 2\nabla_x + 8S_0(\nabla_y + \lambda^{-1}\nabla_z)]h \\
&= (2\nabla_x + 8S_0(\nabla_y + \lambda^{-1}\nabla_z))h \cdot (-2(D_x^2U)\nabla_y + 8S_0(\nabla_x - \lambda^\top\nabla_z \\
&\quad + \nabla_y + T_t^{-1}\lambda^{-1}A^\top\nabla_z)h \\
&= 16S_0|\nabla_x h|^2 + 2\nabla_x h \cdot ((-2D_x^2U + 8S_0)\nabla_y h) \\
&\quad + \nabla_x h \cdot (8S_0(-\lambda^\top + T_t^{-1}\lambda^{-1}A^\top)\nabla_z h) + 64S_0^2\nabla_x h \cdot \nabla_y h \\
&\quad + 8S_0\nabla_y h \cdot ((-2D_x^2U + 8S_0)\nabla_y h) \\
&\quad + 8S_0\nabla_y h \cdot (8S_0(-\lambda^\top + T_t^{-1}\lambda^{-1}A^\top)\nabla_z h) + 64S_0^2\nabla_x h \cdot (\lambda^{-1}\nabla_z h) \\
&\quad + ((-2D_x^2U + 8S_0)\nabla_y h) \cdot (8S_0\lambda^{-1}\nabla_z h) \\
&\quad + 8S_0(\lambda^{-1}\nabla_z h) \cdot (8S_0(-\lambda^\top + T_t^{-1}\lambda^{-1}A^\top)\nabla_z h).
\end{aligned}$$

In order to get a bound in terms of $(\partial_i h)^2$ terms rather than $\partial_i h \partial_j h$ terms, bounding $\partial_i h \partial_j h$ terms with some care and using the boundedness assumption (2.4) on the second derivatives of U yield

$$\begin{aligned}
h\Gamma_{L_t^*,\Phi_1}(h) &\geq 16S_0|\nabla_x h|^2 - \Big(2|\nabla_x h|^2 + 2|D_x^2U|_\infty^2|\nabla_y h|^2 + 8|\nabla_x h|^2 + 8S_0^2|\nabla_y h|^2\Big) \\
&\quad - \Big(2|\nabla_x h|^2 + 8S_0^2\hat{\lambda}^2\Big(1 + T_t^{-2}\big|A^\top\big|^2\Big)|\nabla_z h|^2\Big) \\
&\quad - \Big(|\nabla_x h|^2 + 1024S_0^4|\nabla_y h|^2\Big) - \Big(16S_0|D_x^2U|_\infty|\nabla_y h|^2 \\
&\quad + 64S_0^2|\nabla_y h|^2\Big) - \Big(32S_0^2|\nabla_y h|^2 + 32S_0^2\hat{\lambda}^2\Big(1 + T_t^{-2}\big|A^\top\big|^2\Big)|\nabla_z h|^2\Big) \\
&\quad - \Big(|\nabla_x h|^2 + 1024S_0^2\hat{\lambda}^2|\nabla_z h|^2\Big) - \Big(\Big(2|D_x^2U|_\infty^2 + 32S_0^2\Big)|\nabla_y h|^2 \\
&\quad + 32S_0^2\hat{\lambda}^2|\nabla_z h|^2\Big) - 64S_0^2\hat{\lambda}^2\Big(1 + T_t^{-2}\big|A^\top\big|^2\Big)|\nabla_z h|^2 \\
&\geq 2|\nabla_x h|^2 + S_0^2(-156 - 1024S_0^2)|\nabla_y h|^2 \\
&\quad + S_0^2\hat{\lambda}^2\Big(-1160 - 104T_t^{-2}\big|A^\top\big|^2\Big)|\nabla_z h|^2.
\end{aligned}$$

$\Gamma_{L_t^*,\Phi_2}$ compensates for the negative y derivatives:

$$\begin{aligned}
h\Gamma_{L_t^*,\Phi_2}(h) &\geq (\nabla_y + S_1\lambda^{-1}\nabla_z)h \cdot [L_t^*, \nabla_y + S_1\lambda^{-1}\nabla_z]h \\
&= (\nabla_y + S_1\lambda^{-1}\nabla_z)h \cdot (\nabla_x - \lambda^\top\nabla_z + S_1\nabla_y + S_1T_t^{-1}\lambda^{-1}A^\top\nabla_z)h \\
&= \nabla_x h \cdot \nabla_y h - \nabla_y h \cdot (\lambda^\top\nabla_z h) + S_1|\nabla_y h|^2 + S_1T_t^{-1}\nabla_y h \cdot (\lambda^{-1}A^\top\nabla_z h) \\
&\quad + S_1\nabla_x h \cdot (\lambda^{-1}\nabla_z h) - S_1(\lambda^{-1}\nabla_z h) \cdot (\lambda^\top\nabla_z h) \\
&\quad + S_1^2\nabla_y h \cdot (\lambda^{-1}\nabla_z h) + S_1^2T_t^{-1}(\lambda^{-1}\nabla_z h) \cdot (\lambda^{-1}A^\top\nabla_z h),
\end{aligned}$$

where using the inequalities

$$\begin{aligned}
\nabla_x h \cdot \nabla_y h &\geq -\frac{1}{2}|\nabla_x h|^2 - \frac{1}{2}|\nabla_y h|^2 \\
-\nabla_y h \cdot (\lambda^\top \nabla_z h) &\geq -\frac{1}{6}|\nabla_y h|^2 - \frac{3}{2}\left|\lambda^\top\right|^2 |\nabla_z h|^2 \\
S_1 T_t^{-1} \nabla_y h \cdot (\lambda^{-1} A^\top \nabla_z h) &\geq -\frac{1}{6}|\nabla_y h|^2 - \frac{3}{2} S_1^2 T_t^{-2} \left|\lambda^{-1}\right|^2 \left|A^\top\right|^2 |\nabla_z h|^2 \\
S_1 \nabla_x h \cdot (\lambda^{-1} \nabla_z h) &\geq -\frac{1}{2}|\nabla_x h|^2 - \frac{1}{2} S_1^2 \left|\lambda^{-1}\right|^2 |\nabla_z h|^2 \\
S_1^2 \nabla_y h \cdot (\lambda^{-1} \nabla_z h) &\geq -\frac{1}{6}|\nabla_y h|^2 - \frac{3}{2} S_1^4 \left|\lambda^{-1}\right|^2
\end{aligned}$$

gives the bound

$$\begin{aligned}
h\Gamma_{L_t^*,\Phi_2}(h) \geq -|\nabla_x h|^2 + (1 + 156 S_0^2 + 1024 S_0^2)|\nabla_y h|^2 - \frac{1}{2}\hat{\lambda}^2 \Big(3 + S_1 + S_1^2 \\
+ 3S_1^4 + S_1^2 T_t^{-1} \left|A^\top\right| + 3 S_1^2 T_t^{-2} \left|A^\top\right|^2 \Big) |\nabla_z h|^2 .
\end{aligned}$$

Putting together the bounds gives (C.14) given the coefficents (C.11) in β are large enough. □

C.3 Log-Sobolev Inequality

Proposition C.6. *The distorted entropy* (C.8) *satisfies*

$$H(h_t) \leq C_t \int \frac{|\nabla h_t|^2}{h_t} d\mu_{T_t}, \tag{C.16}$$

where

$$C_t = A_* + \beta(T_t^{-1}) e^{(U_M - U_m) T_t^{-1}} \frac{T_t}{4} \max\left(2, a_m^{-2}\right) \tag{C.17}$$

for all $t > 0$ and some constant $A_* > 0$ depending on $|D_x^2 U|_\infty$ and λ.

Proof. The first two terms in the integrand of $H(t)$ after expanding lead directly to the inequality result for A_*. For the last term of $H(t)$, the standard log-Sobolev inequality for a Gaussian measure [29] alongside the properties that log-Sobolev inequalities tensorises and are stable under perturbations, which can be found as Theorem 4.4 and Property 4.6 in [30] respectively, yields the result. Since the proof of Property 4.6 in [30] is not too long, it is repeated for U satisfying the quadratic assumption (2.5) in order to get the precise form of C_t:

$$\begin{aligned}
\int h_t \ln h_t d\mu_{T_t} &= \int (h_t \ln h_t - h_t + 1) d\mu_{T_t} \\
&\leq \int (h_t \ln h_t - h_t + 1) Z_{T_t}^{-1} e^{-\frac{U_m}{T_t}} e^{-\frac{1}{T_t}\left(|\bar{a}\circ x|^2 + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} dx dy dz \\
&= e^{-\frac{U_m}{T_t}} Z_{T_t}^{-1} \int h_t \ln h_t e^{-\frac{1}{T_t}\left(|\bar{a}\circ x|^2 + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} dx dy dz \\
&\leq e^{-\frac{U_m}{T_t}} \max\left(\frac{T_t}{2}, \max_i \frac{T_t}{4\bar{a}_i^2}\right) Z_{T_t}^{-1} \int \frac{|\nabla h_t|^2}{h_t} e^{-\frac{1}{T_t}\left(|\bar{a}\circ x|^2 + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} dx dy dz \\
&\leq e^{\frac{U_M - U_m}{T_t}} \max\left(\frac{T_t}{2}, \frac{T_t}{4a_m^2}\right) \int \frac{|\nabla h_t|^2}{h_t} d\mu_{T_t},
\end{aligned}$$

where the first inequality follows by (2.5) since $x \ln x - x + 1 \geq 0$ for all $x \geq 0$. □

C.4 Proof of Dissipation

To see that $H(t)$ dissipates over time, it will helpful to be able to pass a time derivative under the integral sign in $H(t)$, which would be straightforward for compactly supported smooth integrands but is not so for $H(t)$. Lemma C.7 constructs compactly supported functions that when multiplied with the integrand in $H(t)$ gives sufficient properties for retrieving a bound on $\partial_t H(t)$ after passing the derviative under the integral sign.

The key nontrivial sufficient property turns out to be (C.18) below.

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be the mollifier

$$\varphi(x) := \begin{cases} e^{\frac{1}{x^2-1}} \left(\int_{-1}^{1} e^{\frac{1}{y^2-1}} dy \right)^{-1} & \text{if } -1 < x \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

$$\varphi_m(x) := \frac{1}{m} \varphi\left(\frac{x}{m}\right)$$

and

$$\nu_m := \varphi_m * \mathbb{1}_{(-\infty, m^2]} \leq 1$$

for $m > 0$ where $\mathbb{1}_{(-\infty,m^2]}$ is the indicator function on $(-\infty, m^2]$.

Lemma C.7. *The smooth functions $\eta_m : \mathbb{R}^{2n+m+1} \to \mathbb{R}^n$*

$$\eta_m = \nu_m(\ln(R + 2d)),$$

where $d > 0$ is the negative of the lower bound of R as in (C.2)*,*

1. *are compactly supported,*
2. *converge to 1 pointwise,*
3. *satisfy for some constant $C > 0$ independent of m and t*

$$L_t \eta_m \leq \frac{C}{m}. \tag{C.18}$$

Remark C.3. Lemma C.7 is different to Lemma 16 in [47]. We believe the few first equations in the proof of Lemma 16 [47] is incorrect and contradicts with equation (4) in the same paper. As a result we require proving (C.18) instead of Lemma 17 of [47].

Proof. By the quadratic assumption (2.5) on U and the bound (C.2) on R, R grows quadratically and in particular is bounded below by an arbitrarily large constant at infinity; along with the support of ν_m being bounded above, the first statement is clear. The second statement is also easy to check.
With the chain rule (A.1) and (A.2) for L_t,

$$L_t \eta_m = \nu_m'(\ln(R + 2d)) L_t \ln(R + 2d) + \nu_m''(\ln(R + 2d))(\nabla_z \ln(R + 2d))^\top A \nabla_z \ln(R + 2d).$$

It can be seen that ν_m' and ν_m'' are of at most order m^{-1}, explicitly:

$$\begin{aligned} \nu_m(x) &= \int_{-\infty}^{m^2} \varphi_m(x - y) dy = \int_{-\infty}^{x+m^2} \varphi_m(z) dz, \\ \nu_m'(x) &= \varphi_m(x + m^2) \leq m^{-1} \max \varphi, \\ \nu_m''(x) &= \varphi_m'(x + m^2) \leq m^{-2} \max \varphi'. \end{aligned}$$

Therefore for a constant $\bar{C} > 0$,

$$L_t \eta_m \leq \bar{C} \Big(m^{-1} L_t \ln(R + 2d) + m^{-2} (\nabla_z \ln(R + 2d))^\top A \nabla_z \ln(R + 2d) \Big).$$

A quick calculation using the Lyapunov property (C.3) of R and again the chain rule (A.1) and (A.2) for L_t reveals

$$\begin{aligned} L_t \ln(R+2d) &= \frac{L_t R}{R+2d} - \frac{(\nabla_z R)^\top A \nabla_z R}{(R+2d)^2} \\ &\leq \frac{-cT_t R + d}{R+2d} - \frac{A_c |\nabla_z R|^2}{(R+2d)^2} \\ &\leq \frac{-cT_t(R+d) + cT_t d + d}{R+2d} - \frac{A_c |\nabla_z R|^2}{(R+2d)^2} \\ (\nabla_z \ln(R+2d))^\top A \nabla_z \ln(R+2d) &\leq |A| |\nabla_z \ln(R+2d)|^2 \\ &= |A| \left| \frac{\nabla_z R}{R+2d} \right|^2, \end{aligned}$$

which are bounded over $\mathbb{R}^{2n+m+1}$ considering R grows quadratically in space. □

The proof of Proposition C.8 follows very closely to Lemma 19 of [47] and its preceding lemmas.

Remark C.4. As explained in Remark C.3, Lemma C.7 uses a different smooth compact function for the truncation than the one found in Section 5.2 in [47]. Lemma C.7 is designed to establish (C.21) below, which is pivotal for the proof.

Proposition C.8. *For any $\alpha > 0$, there exists some constant $B > 0$ and some $t_H > 0$ both depending on $|D_x^2 U|_\infty$, A, λ, T_t, α, U_M, U_m,U_g, E, r_2 and a_m, such that for all $t > t_H$,*

$$H(t) \leq B \left(\frac{1}{t} \right)^{1 - \frac{U_M - U_m}{E} - 2\alpha}. \tag{C.19}$$

Proof. Consider the auxiliary distorted entropies

$$\begin{aligned} H_{\eta_m}(t) &= \int \eta_m \Bigg(\frac{\left| 2\nabla_x h_t + 8 S_0 (\nabla_y h_t + \lambda^{-1} \nabla_z h_t) \right|^2}{h_t} + \frac{\left| \nabla_y h_t + S_1 \lambda^{-1} \nabla_z h_t \right|^2}{h_t} \\ &\quad + \beta(T_t^{-1}) h_t \ln(h_t) \Bigg) d\mu_{T_t} \\ &= \int \eta_m (\Phi_1(h_t) + \Phi_2(h_t) + \beta(T_t^{-1}) \Phi_3(h_t)) d\mu_{T_t} = \int \eta_m \Psi_{T_t}(h_t) d\mu_{T_t}, \end{aligned}$$

where recall $h_t = m_t \mu_{T_t}^{-1}$ is the ratio (2.3) between the law of the process and its instantaneous equilibrium. The point of Lemma C.7 was so that a time derivative can be pushed under the integral:

$$\frac{d}{dt} H_{\eta_m}(t) = \int \eta_m \partial_t (\Psi_{T_t}(h_t)) d\mu_{T_t} + T_t' \int \eta_m \partial_{T_t} (\Psi_{T_t}(h_t) \mu_{T_t}) dx dy dz, \tag{C.20}$$

where ∂_t is fixed with respect to T_t and vice versa for ∂_{T_t}.
Consider the terms separately. Since m_t is the law of the process (1.3) and L_t^* is the $L^2(\mu_{T_t})$ adjoint of L_t,

$$\int f \partial_t m_t = \int f L_t^T m_t = \int L_t f m_t = \int L_t f \frac{m_t}{\mu_{T_t}} \mu_{T_t} = \int f L_t^* \left(\frac{m_t}{\mu_{T_t}} \right) \mu_{T_t},$$

where L_t^T is the $L^2(\mathbb{R}^{2n+m})$ adjoint of L_t, yielding $\partial_t m_t = L_t^* h_t \mu_{T_t}$. This allows the first term in (C.20) to be

bounded as follows.

$$\begin{aligned}\int \eta_m \partial_t(\Psi_{T_t}(h_t))d\mu_{T_t} &= \int \eta_m d\Psi_{T_t}(h_t).\partial_t h_t d\mu_{T_t} \\ &= \int \eta_m d\Psi_{T_t}(h_t).\frac{\partial_t m_t}{\mu_{T_t}} d\mu_{T_t} \\ &= \int \eta_m d\Psi_{T_t}(h_t).L_t^* h_t d\mu_{T_t} \\ &= -\int \eta_m 2\Gamma_{L_t^*,\Psi_{T_t}}(h_t)d\mu_{T_t} + \int \eta_m L_t^*(\Psi_{T_t}(h_t))d\mu_{T_t} \\ &= -\int \eta_m 2\Gamma_{L_t^*,\Psi_{T_t}}(h_t)d\mu_{T_t} \\ &\quad + \int L_t \eta_m \Big(\Psi_{T_t}(h_t) + \beta(T_t^{-1})e^{-1}\Big)d\mu_{T_t} \\ &\le -2\int \eta_m \frac{|\nabla h_t|^2}{h_t} d\mu_{T_t} + \frac{C}{m}\int \Big(\Psi_{T_t}(h_t) + \beta(T_t^{-1})e^{-1}\Big)d\mu_{T_t} \end{aligned} \tag{C.21}$$

by Proposition C.5 and Lemma C.7 where $\beta(T_t^{-1})e^{-1}$ is added to force

$$\beta(T_t^{-1})(h_t \ln h_t + e^{-1}) \ge 0 \implies \Psi_{T_t}(h_t) + \beta(T_t^{-1})e^{-1} \ge 0.$$

(C.21) is a satisfactory bound for now - after taking the limit $m \to \infty$, the log-Sobolev inequality (C.16) will bound the first term on the right hand side of (C.21) such that a Grönwall-type argument can be applied.

For the second term in (C.20), consider the Φ_1 and Φ_2 terms in the integrand $\eta_m \partial_{T_t}(\Psi_{T_t}\mu_{T_t}) = \eta_m \partial_{T_t}((\Phi_1 + \Phi_2 + \beta(T_t^{-1})\Phi_3)\mu_{T_t})$ of $H_{\eta_m}(t)$ as

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

for the corresponding matrices M_1 and M_2. With the form (2.2) of the equilibrium μ_{T_t}, this yields

$$\eta_m \partial_{T_t}(\Phi_i(h_t)\mu_{T_t}) = -2\eta_m (M_i \nabla \ln h_t \cdot M_i \nabla \partial_{T_t} \ln \mu_{T_t}) m_t, \tag{C.22}$$

where the trickiest part, using $Z_{T_t} = \int_{\mathbb{R}^{2n+m}} e^{-\frac{1}{T_t}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} dxdydz$, gives

$$\begin{aligned}&\partial_{T_t} \ln \mu_{T_t} \\ &= \mu_{T_t}^{-1} \partial_{T_t}\left(Z_{T_t}^{-1} e^{-\frac{1}{T_t}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)}\right) \\ &= \mu_{T_t}^{-1}\left(-Z_{T_t}^{-2}\partial_{T_t} Z_{T_t} - \frac{Z_{T_t}^{-1}}{T_t^2}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)\right) e^{-\frac{1}{T_t}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)} \\ &= \mu_{T_t}^{-1}\left(-\mu_{T_t} Z_{T_t}^{-1}\partial_{T_t} Z_{T_t} - \frac{\mu_{T_t}}{T_t^2}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right)\right) \\ &= \int_{\mathbb{R}^{2n+m}} \frac{1}{T_t^2}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right) d\mu_{T_t} - \frac{1}{T_t^2}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2}\right). \end{aligned} \tag{C.23}$$

Integrating by parts in each of the variables x, y, z, where for the x variable $\nabla_x U(x) \cdot x$ is used, after some manipulation using the quadratic assumptions (2.5) and (2.6) for $U(x)$:

$$p_1\Big(T_t^{-1}\Big) \le \partial_{T_t} \ln \mu_{T_t} + \frac{1}{T_t^2}\left(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} - \frac{n+m}{2}T_t\right) \le p_2\Big(T_t^{-1}\Big). \tag{C.24}$$

where $p_1(x) = \frac{2a_m^2 n}{r_2+1}x - \Big(\frac{a_m^2 U_g}{r_2+1} + U_m\Big)x^2$ and $p_2(x) = \frac{a_M^2 n}{r_1}x + \Big(\frac{a_M^2 U_g}{r_1} + U_M\Big)x^2$ are found from bounding the integral over $U(x)$.
Substituting (C.23) back into (C.22),

$$\begin{aligned}\eta_m \partial_{T_t}(\Phi_i(h_t)\mu_{T_t}) &\le \eta_m \left(|M_i \nabla \ln h_t|^2 + T_t^{-4} \left| M_i \nabla \Big(U(x) + \frac{|y|^2}{2} + \frac{|z|^2}{2} \Big) \right|^2 \right) m_t \\ &\le \Phi_i(h_t)\mu_{T_t} + \widetilde{C} T_t^{-4} \Big(1 + |x|^2 + |y|^2 + |z|^2 \Big) m_t \end{aligned} \tag{C.25}$$

for constant $\widetilde{C} \ge 0$ by the quadratic assumption (2.7) on $|\nabla_x U|^2$ and $\eta_m \le 1$.
In the second term of (C.20), $\Phi_3(h_t) = \frac{m_t}{\mu_{T_t}} \ln \frac{m_t}{\mu_{T_t}}$, using the left inequality of (C.24), gives

$$\begin{aligned}&\eta_m \partial_{T_t}(\beta(T_t^{-1})\Phi_3(h_t)\mu_{T_t}) \\ &\quad = -\eta_m T_t^{-2}\beta'(T_t^{-1})\Phi_3(h_t)\mu_{T_t} - \eta_m \beta(T_t^{-1})\partial_{T_t} \ln \mu_{T_t} m_t \\ &\quad = -\eta_m T_t^{-2}\beta'(T_t^{-1})(\Phi_3(h_t) + e^{-1})\mu_{T_t} + \eta_m T_t^{-2}\beta'(T_t^{-1})e^{-1}\mu_{T_t} \\ &\qquad - \eta_m \beta(T_t^{-1})\partial_{T_t} \ln \mu_{T_t} m_t \\ &\quad \le T_t^{-2}\beta'(T_t^{-1})e^{-1}\mu_{T_t} \\ &\qquad + \beta(T_t^{-1}) \left| p_1\Big(T_t^{-1}\Big) + \frac{1}{T_t^2}\left(\frac{n+m}{2} T_t - U_M - |\bar{a} \circ x|^2 - \frac{|y|^2}{2} - \frac{|z|^2}{2} \right) \right| m_t, \end{aligned} \tag{C.26}$$

where in the last step $\eta_m \le 1$ and (2.5) have been used.

Putting together the bounds (C.25) and (C.26) and applying Corollary C.3 yields

$$\begin{aligned}\int \eta_m \partial_{T_t}(\Psi_{T_t}(h_t)\mu_{T_t}) dx dy dz &\le q\Big(T_t^{-1}\Big)\Big(H(t) + \mathbb{E}\Big[1 + |X_t|^2 + |Y_t|^2 + |Z_t|^2\Big]\Big) \\ &\le p\Big(T_t^{-1}\Big)\Big(H(t) + \hat{C}\Big), \end{aligned} \tag{C.27}$$

where p and q are some finite order polynomial with nonnegative coefficients and $\hat{C} \ge 0$.

Returning to (C.20), collecting (C.21) and (C.27) then integrating from any s to t gives

$$\begin{aligned}H_{\eta_m}(t) - H_{\eta_m}(s) \le \int_s^t \Big(&- 2\int \eta_m \frac{|\nabla h_u|^2}{h_u} d\mu_{T_u} + |T_u'| p\Big(T_u^{-1}\Big)\Big(H(u) + \hat{C}\Big) \Big) du \\ &+ \mathcal{O}(m^{-1}). \end{aligned}$$

After taking $m \to \infty$ then applying Fatou's lemma for the first terms of both sides and $\eta_m \le 1$ for the second term on the left, this becomes

$$H(t) - H(s) \le \int_s^t - 2\int \frac{|\nabla h_u|^2}{h_u} d\mu_{T_u} du + \int_s^t |T_u'| p\Big(T_u^{-1}\Big)\Big(H(u) + \hat{C}\Big) du \tag{C.28}$$

$$\le \int_s^t \Big(\Big(|T_u'| p\Big(T_u^{-1}\Big) - 2C_u^{-1} \Big) H(u) + \hat{C} |T_u'| p\Big(T_u^{-1}\Big) \Big) du, \tag{C.29}$$

where the last inequality follows by the log-Sobolev inequality (C.16).
Taking $s \to t$, the derivative of H can be seen to be bounded in such a way that Theorem 2.17 of [66] can be applied to see that

$$H(t) < \infty \quad \forall t \ge 0. \tag{C.30}$$

After taking $s \to t$, it remains to show that

$$2C_t^{-1} \gg |T_t'| p\Big(T_t^{-1}\Big) \tag{C.31}$$

for large t.
Here the assumptions on T_t are imposed. Considering the form (C.17) of C_t and that $t^\alpha \gg (\ln t)^p$ for any $p > 0$ and $\alpha > 0$ given large enough t, this yields that for any $\alpha > 0$, there exists $t_1 > 0$ and $c_i > 0$, $i = 1, 2, 3$ such that for all $t \geq t_1$,

$$\frac{d}{dt}H(t) \leq \left(c_1\left(\frac{1}{t}\right)^{1-\alpha} - c_2\left(\frac{1}{t}\right)^{\frac{U_M-U_m}{E}+\alpha}\right)H(t) + c_3\left(\frac{1}{t}\right)^{1-\alpha}.$$

Since $E > U_M - U_m$ by assumption, taking α small enough, there exists $c_4 > 0$ and $t_2 > 0$ such that for all $t \geq t_2$,

$$\frac{d}{dt}H(t) \leq -c_4 H(t)\left(\frac{1}{t}\right)^{\frac{U_M-U_m}{E}+\alpha} + c_3\left(\frac{1}{t}\right)^{1-\alpha}. \tag{C.32}$$

Setting

$$\gamma_1(t) := c_4\left(\frac{1}{t}\right)^{\frac{U_M-U_m}{E}+\alpha}, \qquad \gamma_2(t) := c_3\left(\frac{1}{t}\right)^{1-\alpha}$$

and following the argument as per [47] from Lemma 6 in [46], (C.32) becomes

$$\frac{d}{dt}\left(H(t)\ e^{\int_{t_2}^t \gamma_1(s)ds}\right) \leq e^{\int_{t_2}^t \gamma_1(s)ds}\gamma_2(t) = \frac{\gamma_2(t)}{\gamma_1(t)}\gamma_1(t)e^{\int_{t_2}^t \gamma_1(s)ds}, \tag{C.33}$$

but for any $t \geq t_2$ and $t^* \leq t$, taking $\alpha \leq \frac{1}{2}\Big(1 - \frac{U_M-U_m}{E}\Big)$,

$$\frac{\gamma_2(t)}{\gamma_1(t)} = \frac{c_3}{c_4}\left(\frac{1}{t}\right)^{1-\frac{U_M-U_m}{E}-2\alpha} \leq \frac{c_3}{c_4}\left(\frac{1}{t^*}\right)^{1-\frac{U_M-U_m}{E}-2\alpha}, \tag{C.34}$$

which allows (C.33) to give

$$H(t) \leq H(t_2)e^{-\int_{t_2}^t \gamma_1(s)ds} + \frac{c_3}{c_4}\left(\frac{1}{t^*}\right)^{1-\frac{U_M-U_m}{E}-2\alpha}(1 - e^{-\int_{t_2}^t \gamma_1(s)ds}).$$

The proof is finished after substituting back in $t^* = t$ since $H(t_2)$ is finite due to (C.30) and the expression

$$\int_{t_2}^t \gamma_1(s)ds = c_5\left(\left(\frac{1}{t}\right)^{-1+\frac{U_M-U_m}{E}+\alpha} - \left(\frac{1}{t_2}\right)^{-1+\frac{U_M-U_m}{E}+\alpha}\right),$$

grows to infinity as $t \to \infty$, where $c_5 > 0$ is a constant. □

Remark C.5. The annealing schedule T_t is chosen to satisfy the relationship (C.31) between C_t^{-1} and $|T_t'| p\Big(T_t^{-1}\Big)$. (C.31) has two purposes - for the coefficient of $H(u)$ on the right hand side of (C.29) to be negative and for this coefficient to be much stronger in magnitude for large t than the last term in (C.29). These allow (C.34) and consequently for (C.33) to be a fruitful step.

Appendix D Additional Results

We conclude the appendices by presenting the analog of Proposition C.8 for the $T_t = T > 0$ sampling case and a result about the choice of the annealing schedule.

Proposition D.1. *Let 1. and 3. of Assumption 1 hold and let $T_t = T$ for all t for some constant $T > 0$. It holds that*

$$\int |h_t - 1| d\mu_T \leq \sqrt{\frac{2H(0)}{\beta(T)}} e^{-\frac{C_*^{-1}}{2}t}$$

where C_* is the log-Sobolev constant

$$C_* = A_* + \beta(T^{-1}) e^{(U_M - U_m)T^{-1}} \frac{T}{4} \max\left(2, a_m^{-2}\right). \tag{D.1}$$

Proof. After Pinsker's inequality (2.9) and consideration of the definition (C.8) of H, what remains is the partial time derivative part of the proof of Proposition C.8. The proof concludes by the same calculations, keeping in mind $T_t' = 0$, until (C.29) followed by the Grönwall argument. □

Proposition D.2. *The schedule $T_t = \frac{E}{\ln(e+t)}$, $E > U_M - U_m$ is optimal in the sense that for any differentiable $f : \mathbb{R}_+ \to \mathbb{R}_+$, if*

$$T_t = \frac{1}{f(t)} \left(\frac{U_M - U_m}{\ln(e+t)} \right), \tag{D.2}$$

C_t is the log-Sobolev factor (C.17) *and p is the polynomial with nonnegative coefficients from the proof of Proposition C.8, then the relation* (C.31)*:*

$$2C_t^{-1} \gg |T_t'| p\left(T_t^{-1}\right)$$

holds for large times only if $\limsup_{t\to\infty} f(t) \leq 1$.

Proof. Suppose there exists a constant $\delta > 0$ and times $(t_i)_{i\in\mathbb{N}}$ such that $0 < t_i \to \infty$ and

$$f(t_i) \geq 1 + \delta \quad \forall i.$$

From its definition (C.17),

$$C_t^{-1} \sim \mathcal{O}(e^{-(U_M - U_m)T_t^{-1}} T_t^{-1}),$$

which after substituting in the form (D.2) for T_t is

$$e^{-(U_M - U_m)T_t^{-1}} T_t^{-1} = (e+t)^{-f(t)} \frac{f(t)\ln(e+t)}{U_M - U_m} \sim \mathcal{O}(t^{-f(t)} f(t) \ln t). \tag{D.3}$$

Compare this to

$$|T_t'| p\left(T_t^{-1}\right) \propto \frac{p(f(t)\ln(e+t))}{(f(t)\ln(e+t))^2} \left(\frac{f(t)}{e+t} + |f'(t)| \ln(e+t) \right), \tag{D.4}$$

which has order at least $(tf(t))^{-1}(\ln t)^{-2}$. For $t = t_i$ large enough, $f(t) \geq 1 + \delta$ and so

$$t^{-f(t)} f(t) \ln t \ll (tf(t))^{-1} (\ln t)^{-2}, \tag{D.5}$$

violating (C.31). □

Remark D.1. One can strengthen the proposition by making precise the form of p from Proposition C.8, which will determine how slowly $f(t)$ is allowed to converge to 1; in fact p should be at least fifth order. This appears inconsequential with respect to optimality and so is omitted.