Temperature and Friction Accelerated Sampling of Boltzmann-Gibbs Distribution

Molei Tao, Houman Owhadi, Jerrold E. Marsden

July 8, 2010

Abstract

This paper is concerned with tuning friction and temperature in Langevin dynamics for fast sampling from the canonical ensemble. We show that near-optimal acceleration is achieved by choosing friction so that the local quadratic approximation of the Hamiltonian is a critical damped oscillator. The system is also overheated and cooled down to its final temperature. The performances of different cooling schedules are analyzed as functions of total simulation time.

1 Introduction

We propose a method to accelerate the Langevin approach of sampling from Boltzmann-Gibbs (B-G) distribution. Specifically, consider the following Langevin Stochastic Differential Equations (SDE)

$$\begin{cases} Mdq &= pdt \\ dp &= -\nabla V(q)dt - cpdt + \sqrt{2c/\beta}dW \end{cases} \tag{1}$$

where $p, q \in \mathbb{R}^d$, M is the mass matrix, $V(\cdot)$ is potential energy, c is a positive semi-definite $d \times d$ matrix indicating the damping coefficient, $\beta \in \mathbb{R}^+$ is the inverse of temperature, and W is a standard Wiener process.

It is known that the stochastic process defined by (1) has an invariant distribution of Boltzmann-Gibbs distribution (also known as canonical ensemble) defined by:

$$d\mu = Z^{-1}\exp(-\beta H(q,p))dqdp. \tag{2}$$

where $Z = \int_{T^*\mathbb{R}^d}\exp(-\beta H(q,p))dqdp$ is the partition function, and $H(q,p) = p^T M^{-1}p/2 + V(q)$ is the Hamiltonian function.

When the solution of (1) is also geometrically ergodic with respect to μ (we refer to [1] and [2] for sufficient conditions on the potential V), it is then natural to use long-time trajectories of (1) as approximate samples of B-G distribution.

One important thing to notice is that being able to sample from B-G enables sampling an arbitrary probability density function. The trick is to set $V(q) = -\beta^{-1}\ln\pi(q)$, and then the marginal distribution on q from B-G will have the density function $\pi(\cdot)$.

This paper is concerned with the following questions:

- Although the friction parameter c does not affect the invariant distribution, it does affect the rate of convergence. How should c be chosen for faster convergence and hence accelerated sampling?

- If sampling from B-G is the objective, the inverse temperature β does not need to be kept constant over the total simulation time T. How to chose the cooling schedule $t \mapsto \beta(t), t \in [0, T]$ in order to minimize the distance between the distribution of $[q(T), p(T)]$ and the desired B-G?

Background: There is no need to repeat the importance of sampling the canonical ensembles of complicated systems, which is, however, a known computational challenge [3, 4, 5]. The nonlinearity of the potential and the curse of dimensionality, for instance, make sampling methods slowly convergent.

Classical sampling approaches include purely statistical methods such as Metropolis algorithm and importance sampling that are solely for sampling purposes (see for instance [6] and references therein for a review and comparison), stochastic molecular dynamics (primarily Langevin dynamics), deterministic dynamics plus an external thermostat (such as Nosé-Hoover [7, 8], Berendsen [9] or Andersen [10] thermostats), Hybrid Monte Carlo [11] (which introduces auxiliary dynamics to avoid random walks), etc. We also refer to [12] as an example that combines stochastic molecular dynamics and purely statistical approach.

Langevin dynamics adds friction and noise to mechanical equations to model energy exchange with a heat bath [13, 14, 15]. It has been shown in the context of classical molecular sampling that both stochastic dynamics and deterministic dynamics with thermostats outperform purely statistical methods in convergence rate as the size of the system grows (we refer to [16] for a linear alkane molecule). Since overdamped Langevin is a special case of Hybrid Monte Carlo [17], it is not surprising to observe cases in which Langevin dynamics is computationally more efficient than purely statistical methods. Moreover, if the system is stiff or multiscale, existing stiff or multiscale Langevin integrators such as SIM [18] or FLAVOR [19] can be directly employed for accelerated computation.

Annealing was first introduced in Simulated Annealing algorithm [20] for global optimization, which can also be viewed as (uniformly) sampling from the set of global minimizers of V. Temperature accelerated dynamics has been proposed in [21] for events simulations. The concept there is to raise temperature of the system to make rare events occur more frequently, intercept each attempted escape from potential wells and extrapolate time to low temperature. Another temperature approach has been used to calculate free energy [22]. In that method, overheated auxiliary variables are introduced to equilibrate the collective variables faster. We stay with the global annealing approach used in Simulated Annealing.

The proposed perspective of tuning friction and annealing temperature is distinct from prevailing accelerated sampling methods, such as conformational flooding [23], replica exchange [24], umbrella sampling [25], self-guided MD [26], hyperdynamics [27], affine invariant ensemble sampler [28], and many others reviewed in [29], and therefore

2

can be used concurrently with many of these methods. While tuning friction is mostly restricted to dynamics based methods, annealing may apply to any method that involves temperature. Note that temperature is a rather general notion because it can often be introduced artificially; for instance, see [30] for an example in which temperature is introduced in an MCMC algorithm for Bayesian updating.

2 Method for friction and temperature accelerated Boltzmann-Gibbs sampling

Background algorithms: Although any Langevin integrator can serve as a background algorithm and be tuned and annealed, in our numerical simulations we base on the 1st-order B-G preserving Geometric Langevin Algorithm (GLA) introduced in [31], which is recapped as follows:

$$
\begin{cases}
\hat{p}_n &= e^{-c_n h} p_n + \sqrt{\frac{1-e^{-2c_n h}}{\beta_n}} \xi_n \\
q_{n+1} &= q_n + h\hat{p}_n \\
p_{n+1} &= \hat{p}_n - h\nabla V(q_{n+1})
\end{cases}
\tag{3}
$$

where h is the timestep length, ξ_n's are i.i.d. standard normal random variables, and $c_n = c$ and $\beta_n = \beta$ in absence of friction tuning or temperature annealing.

The choice of GLA is motivated by its conformal-symplecticity and long-time properties [31]. Specifically, under certain conditions, GLA is not only pathwise accurate but also convergent towards B-G up to a diminishing numerical error. It is worth mentioning that similar properties are shown to hold under weaker conditions for a Metropolized version of GLA [12, 32], which can also be tuned and annealed for accelerated samplings.

For multiscale or stiff systems (where $V(q) = V_0(q) + \epsilon^{-1}V_1(q)$ for instance), FLA-VORS [19] are possible alternative background algorithms that are also conformal-symplectic (we also refer to SIMS [18] for quadratic stiff potentials).

Choice of friction: If V is quadratic (of the form $V = \frac{q^T K q}{2}$), we show in Appendix 5.1 that optimal acceleration is achieved by choosing $c = 2K^{\frac{1}{2}}$ so that all degrees of freedom of the harmonic oscillator are critically damped. Based on this observation, we heuristically propose to tune the friction c_n at each time step of the simulation according to the Hessian of the potential V:

$$
\begin{cases}
k_n &= \begin{cases} \frac{1}{2}\frac{\partial^2 V}{\partial q^2}(q_n), & \frac{\partial^2 V}{\partial q^2}(q_n) \succ 0 \\ \alpha^2/4I, & \text{otherwise} \end{cases} \\
c_n &= 2\sqrt{k_n}
\end{cases}
\tag{4}
$$

where α is a fixed real parameter, preassigned to handle the case of negative curvature; for instance, it could be equal to 0 or to the original value of c.

3

Choice of temperature: Annealing has successfully been applied to optimization problems [20]. A cooling schedule describes how to choose $T(n) = 1/\beta_n$ as a function of n. For optimization based cooling schedules, one requires $\lim_{i \to \infty} T(i) = 0$. We refer to [33, 34, 35] for general reviews of optimization based cooling schedules, and to [36, 37] for theoretical bounds on convergence. In this paper we are interested in situations where the total number of steps N is finite and fixed, the final temperature $T(N) = T_f = 1/\beta > 0$ is strictly positive and is the temperature at which one wishes to sample the B-G distribution.

It is then natural to seek to minimize the distance between the distribution of (q_N, p_N) and B-G at temperature $1/\beta$ using $T(1), \ldots, T(N-1)$ as optimization variables. In Appendix 5.2 we derive a bound on this distance using transition state theory and convergence rates of Markov chains. A numerical minimization of that bound suggests the following near-optimal cooling schedule for $T_f > 0$ (for $T_f = 0$ we refer to [34] and references therein) and $N < N_0$ (N_0 is the number of steps needed for sampling by a naive Langevin simulation; see Appendix 5.3 for details):

$$\beta_n = \frac{n}{N}\frac{1}{T_f} + (1 - \frac{n}{N})\frac{1}{T_i}, \quad T(n) = 1/\beta_n \tag{5}$$

where N is the total-number of simulation steps, T_f the temperature at which the Gibbs distribution needs to be sampled, and the initial temperature $T_i > T_f$ is a free parameter chosen to overcome the maximal potential barrier, i.e., $T_i \gg \Delta V/k$ (for simplicity we let the Boltzmann constant k be equal to one in our setting; ΔV can be intuitively interpreted as the maximum elevation in potential landscape, and we refer to [38] for a rigorous definition).

Friction and temperature accelerated sampling: Put together, annealed and tuned GLA (AnnealTuneGLA) for accelerated B-G sampling is the following:

$$\begin{cases} k_n &= \begin{cases} \frac{1}{2}\frac{\partial^2 V}{\partial q^2}(q_n) & \frac{\partial^2 V}{\partial q^2}(q_n) \succ 0 \\ \alpha^2/4 & otherwise \end{cases} \\ c_n &= 2\sqrt{k_n} \\ \beta_n &= \frac{n}{N}\frac{1}{T_f} + (1 - \frac{n}{N})\frac{1}{T_i} \\ \hat{p}_n &= e^{-c_n h}p_n + \sqrt{\frac{1-e^{-2c_n h}}{\beta_n}}\xi_n \\ q_{n+1} &= q_n + h\hat{p}_n \\ p_{n+1} &= \hat{p}_n - h\nabla V(q_{n+1}) \end{cases} \tag{6}$$

Comparing to the background GLA, the distribution of the accelerated trajectory at a fixed time is closer to the desired B-G in the total variation sense. A possible exact preservation of a near-by distribution is, however, not yet proved for AnnealTuneGLA.

It is worth mentioning that 1st-order GLA is not unconditionally stable, nor is AnnealTuneGLA. Therefore, h or α should not be chosen to be too large.
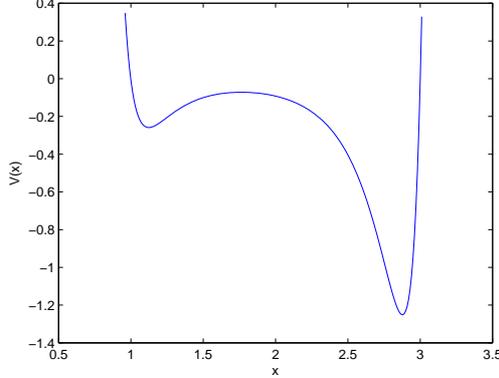
3 Numerical experiments

Figure 1: Potential energy landscape.

Consider a one dimensional nonlinear molecular system consisting of two distinct heavy (fixed) atoms and a light atom between them. It is modeled as a single degree of freedom Hamiltonian system with a Lennard-Jones potential function $V(q) = \left(q^{-12} - q^{-6}\right) + 5\left((4-q)^{-12} - (4-q)^{-6}\right)$ (Figure 1). The energy landscape consists of a local potential barrier and two potential wells. The attraction due to the right atom is larger than the left one. If one starts the dynamics with zero initial momentum and position in the left basin, the asymptotic (long time) position distribution will be a marginal of B-G and concentrated in the right basin. Therefore, the expectation of position q at a fixed time can be used as an indicator of the convergence rate for this nonlinear system.
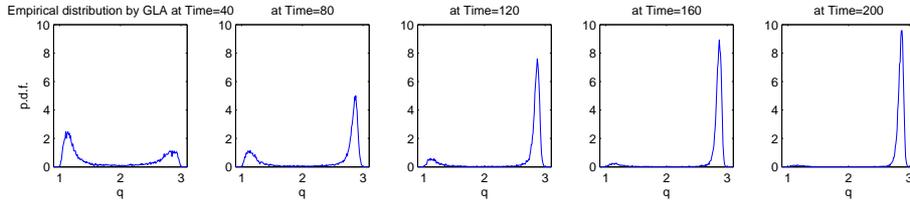

Figure 2: Evolution of the empirical distribution obtained by GLA (Eq. 3) with $c = 0.1$. The Markov process is converging as the distribution peaks more and more in the right potential basin. Simulation is done with a step length $h = 0.01$ and distributions are approximated empirically by an ensemble of 10000 trajectories.

Throughout this section we use parameters $\beta = 10$, $q(0) = 1.1$ and $p(0) = 0$. With an arbitrarily chosen $c = 0.1$, Langevin dynamics integrated with a B-G preserving method GLA (Eq. 3) takes more than 200 time units before indiscernible convergence (Figure 2). Enumerating c values for fixed β (and hence temperature T), one obtains different values of $\mathbb{E}[q(\text{TotalTime})]$ for a fixed total simulation time (Figure 3). This confirms that the value of c affects the convergence rate. The optimal fixed value is $c = 0.7$ in
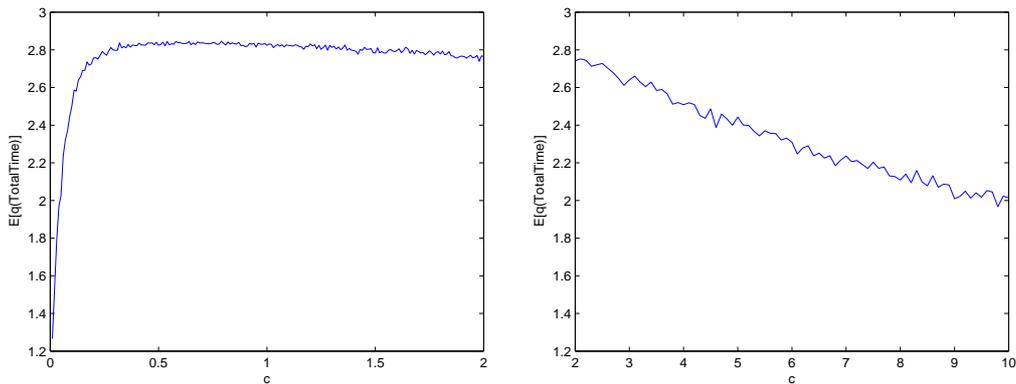
5

Figure 3: Expectations of position at a fixed time for different frictions obtained by GLA (Eq. 3). Larger expectation implies better convergence in this problem, and therefore this indicates the relationship between choice of c and convergence rate. The fixed time is TotalTime=100, step length is $h = 0.01$, expectations are calculated by an empirical average over an ensemble of 1000 trajectories. c values are enumerated from 0.01, 0.02, ..., 1.99, 2.00 and 2.10, ..., 19.90, 20.00.

this example.

Although in practice it is rarely the case that an optimization can be carried out beforehand to determine the best value of c for fastest convergence of GLA, we nevertheless use GLA with the optimal friction $c = 0.7$ for comparison purposes. We will show that TuneGLA outperforms even this optimized GLA, demonstrating that c really needs to be tuned locally.

In Figure 4, GLA with $c = 0.7$ (the optimal fixed value), TuneGLA (GLA with friction tuning) which adaptively tunes c but does not anneal (Eq. 6), AnnealGLA (GLA with temperature annealing) which uses an inverse linear cooling schedule ($C = 10T_f$) but does not tune c (Eq. 20), and AnnealTuneGLAs that tune and anneal with respectively $\alpha = 0$ and $\alpha = 0.7$ are compared. We observe that tuning friction and annealing temperature individually accelerates the convergence, and their effects are additive. Therefore, the proposed AnnealTuneGLA has the fastest rate of convergence. In addition, here the choice of $\alpha = 0$ slightly outperforms $\alpha = 0.7$, which is set to be the value of the optimal c. The optimal choice of α has not been investigated.

4   Acknowledgement

References

[1] J. C. Mattingly and A. M. Stuart, Markov Process. Related Fields **8**, 199 (2002), inhomogeneous random systems (Cergy-Pontoise, 2001).
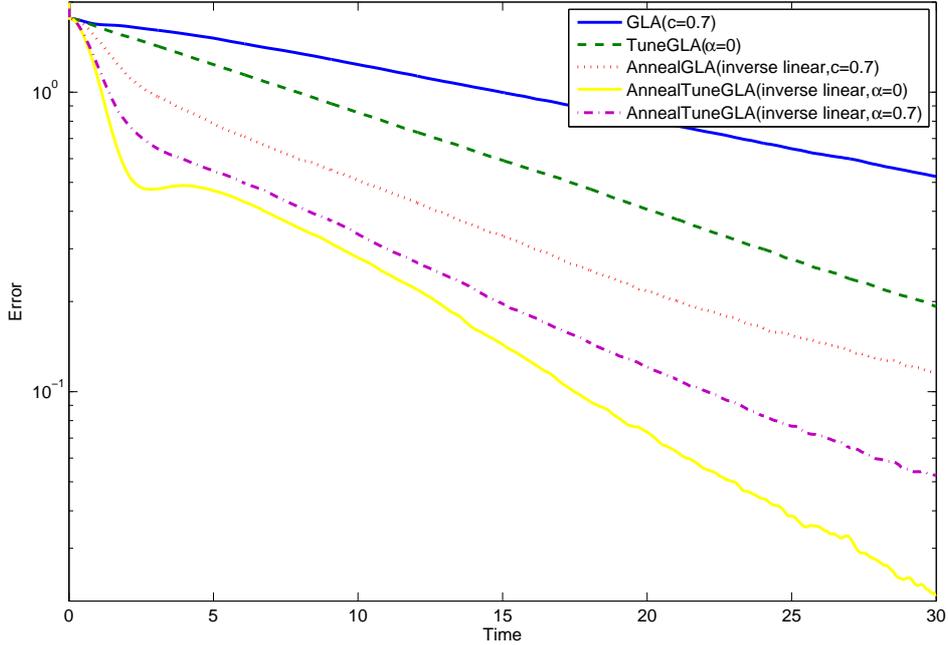
Figure 4: Comparison of errors of GLA, TuneGLA with c adaptively tuned, AnnealGLA with inverse linear cooling schedule, and AnnealTuneGLA with both. $c = 0.7$ that ensures fastest GLA convergence (Figure 3) is used in GLA and AnnealGLA. A comparison between choices of α (which indicates the value of c when curvature of potential is negative for tuning (Eq. 6) is also presented. Total simulation time=30 is fixed, and error at each step throughout the simulation is recorded. Simulation step length is $h = 0.01$. Error at time t is calculated by $|\frac{1}{M} \sum_{i=1}^{M} q^i(t) - \mathbb{E}q(\infty)|$, where $M = 10000$ is the total number of independent trajectories, $q^i(t)$ is the position of the ith trajectory at time t, and $\mathbb{E}q(\infty)$ is well approximated by empirical average of an ensemble of 20000 GLA trajectories at total simulation time of 300. The constant of initial temperature in the inverse linear cooling (Eq. 20) is $C = 10T_f$ and applies to all three AnnealGLAs.

[2] J. Mattingly, A. Stuart, and D. Higham, Stoch. Proc. Appl. **101**, 185 (2002).

[3] E. J. Barth, B. B. Laird, and B. J. Leimkuhler, J. Chem. Phys. **118**, 5759 (2003).

[4] D. Hamelberg, J. Mongan, and J. A. McCammon, J. Chem. Phys. **120**, 11919 (2004).

[5] F. Karsch and E. Laermann, Rep. Prog. Phys. **56**, 1347 (1993).

[6] F. Bassetti and P. Diaconis, Illinois J. of Math. **50**, 67 (2005).

[7] S. Nosé, Mol. Phys. **52**, 255 (1984).

[8] W. G. Hoover, Phys. Rev. A **31**, 1695 (1985).

[9] H. J. C. Berendsen *et al.*, J. Chem. Phys. **81**, 3684 (1984).

[10] H. C. Andersen, J. Chem. Phys. **72**, 2384 (1980).

[11] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth, Phys. Lett. B **195**, 216 (1987).

[12] G. O. Roberts and R. L. Tweedie, Bernoulli **2**, 341 (1996).

[13] P. Turq, F. Lantelme, and H. L. Friedman, J. Chem. Phys. **66**, 3039 (1977).

[14] T. Schlick, *Molecular Modeling and Simulation* (Springer, New York, 2002), pp. 435–438.

[15] T. Schneider and E. Stoll, Phys. Rev. B **17**, 1302 (1978).

[16] E. Cancès, F. Legoll, and G. Stoltz, M2AN **41**, 351 (2007).

[17] S. H. Cheung and J. L. Beck, J. Eng. Mech. **135**, 243 (2009).

[18] M. Tao, H. Owhadi, and J. E. Marsden, arXiv:1006.4657. Submitted (unpublished).

[19] M. Tao, H. Owhadi, and J. E. Marsden, (2010), arXiv:0908.1241v2. Accepted by SIAM MMS.

[20] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).

[21] M. R. Sørensen and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000).

[22] L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett. **426**, 168 (2006).

[23] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).

[24] Y. Sugita and Y. Okamoto, Chem. Phys. Lett. **314**, 141 (1999).

[25] G. M. Torrie and J. P. Valleau, J. Comput. Phys. **23**, 187 (1977).

[26] X. Wu and S. Wang, J. Chem. Phys. **110**, 9401 (1999).

[27] A. F. Voter, Phys. Rev. Lett. **78**, 3908 (1997).

[28] J. Goodman and J. Weare, Comm. App. Math. Comp. Sci. **5**, 65 (2010).

[29] B. J. Berne and J. E. Straub, Curr. Opin. Struct. Biol. **7**, 181 (1997).

[30] R. M. Neal, Statist. Comput. **6**, 353 (1996).

[31] N. Bou-Rabee and H. Owhadi, SIAM J. Numer. Anal. **48**, 278 (2010).

[32] N. Bou-Rabee and E. Vanden-Eijnden (unpublished).

[33] O. Häggström, *Finite Markov Chains and Algorithmic Applications*, 1st ed. (Cambridge University Press, Cambridge UK, 2002).

[34] H. Cohn and M. Fielding, SIAM J. Optimiz. **9**, 779 (1999).

[35] E. Triki, Y. Collette, and P. Siarry, Eur. J. Oper. Res. **166**, 77 (2005).

[36] S. Geman and D. Geman, IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 (1984).

[37] B. Hajek, Math. Oper. Res. **13**, 311 (1988).

[38] P. Diaconis and D. Stroock, Ann. Appl. Probab. **1**, 36 (1991).

[39] B. Hassibi, in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE Computer Society, Washington DC, 2000), pp. 737–740.

[40] N. J. Higham, Numer. Algorithms **15**, 227 (1997).

[41] M. Freidlin and A. Wentzell, *Random Perturbations of Dynamical Systems*, 2nd ed. (Springer, New York USA, 1998).

[42] C. Schütte and W. Huisinga, in *EQUADIFF 99-International Conference on Differential Equations (B. Fiedler, K. Griger and J. Sprekels, eds.)* (World Scientific, Singapore, 2000), Vol. 2, pp. 1247–1262.

[43] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, 2nd ed. (Springer, New York, 1998).

[44] H. Szu and R. Hartley, Phys. Lett. B **122**, 157 (1987).

5 Appendix

5.1 Friction accelerated sampling: analysis of linear systems

In this section, we will show that with β fixed, the choice of $c = 2\sqrt{k}$ will enable the fastest convergence of the following system:

$$\begin{cases} dq &= pdt \\ dp &= -kqdt - cpdt + \sigma dW \end{cases} \tag{7}$$

where $\sigma = \sqrt{2c/\beta}$. Assume k is a scalar for the moment. For our purpose, consider positive k, because if k is 0 the system decouples, and if k is negative the system is not ergodic and does not admit an invariant distribution.

The solution to the above linear system can be explicitly written as

$$\begin{cases} q(t) &= B_{11}(t)q(0) + B_{12}(t)p(0) + \int_0^t B_{12}(t-s)\sigma dW_s \\ p(t) &= B_{21}(t)q(0) + B_{22}(t)p(0) + \int_0^t B_{22}(t-s)\sigma dW_s \end{cases} \tag{8}$$

9

where $B(t)$ is the fundamental matrix defined through the following autonomous ODE $\frac{dB}{dt} = \begin{bmatrix} 0 & 1 \\ -k & -c \end{bmatrix} B$, and written in block form to be

$$B(t) = \begin{bmatrix} B_{11}(t) & B_{12}(t) \\ B_{21}(t) & B_{22}(t) \end{bmatrix} = \exp\left(\begin{bmatrix} 0 & 1 \\ -k & -c \end{bmatrix} t\right) \tag{9}$$

After calculating out the matrix exponential, the expectation of position writes as follows

$$
\begin{aligned}
\mathbb{E}q(t) &= B_{11}(t)q(0) + B_{12}(t)p(0) \\
&= \frac{e^{\frac{1}{2}\left(-c+\sqrt{c^2-4k}\right)t}\left(c + \sqrt{c^2-4k}\right) - e^{\frac{1}{2}\left(-c-\sqrt{c^2-4k}\right)t}\left(c - \sqrt{c^2-4k}\right)}{2\sqrt{c^2-4k}}q(0) \\
&\quad + \frac{e^{\frac{1}{2}\left(-c+\sqrt{c^2-4k}\right)t} - e^{\frac{1}{2}\left(-c-\sqrt{c^2-4k}\right)t}}{\sqrt{c^2-4k}}p(0)
\end{aligned}
\tag{10}
$$

Naturally, the expectation approaches 0 as $t \to +\infty$. Recall that c and k are non-negative reals. We will show in the following discussion that the maximum speed of convergence toward 0 will be achieved when $c = 2\sqrt{k}$:

1. When $c^2 - 4k > 0$, $-c - \sqrt{c^2-4k} < -c + \sqrt{c^2-4k} < 0$ and none of the coefficients are zero. Therefore the bottleneck for convergence of $B_{11}(t)$ and $B_{12}(t)$ will be $e^{\frac{1}{2}\left(-c+\sqrt{c^2-4k}\right)t}$, which will be minimized as $c^2 \downarrow 4k$.

2. When $c^2 - 4k = 0$, $B_{11} = \frac{1}{2}e^{-ct/2}(2 + ct)$ and $B_{12} = e^{-ct/2}t$.

3. When $c^2 - 4k < 0$, define a real number $\omega = \sqrt{4k - c^2}$. $B_{11} = e^{-ct/2}(c\sin(\omega t/2)/\omega + \cos(\omega t/2))$ and $B_{12} = e^{-ct/2}2\sin(\omega t/2)/\omega$. Notice $\cos(\omega t/2)$ and $\sin(\omega t/2)$ can not be simultaneously zero, and therefore the convergence rate is controlled by $e^{-ct/2}$, which will be minimized when $c^2 \uparrow 4k$.

Hence when $c = 2\sqrt{k}$ this linear system (7) converges the fastest. Notice that this choice corresponds to a critically damped system (as opposed to overdamped or underdamped).

When the system is linear but multi-dimensional, k can be assumed without loss of generality to be a symmetric matrix, and it can be immediately seen that there is no theoretical difficulty because one can diagonalize k and choose c diagonal wisely. Therefore, any numerical method that calculates the square root of a matrix could work here for getting c. There are many possible numerical approaches on square rooting matrices, for instance by preconditioning if the matrix has some special structure (which is usually the case in molecular systems), or as in [39] or [40], but for consideration of conciseness the authors will not discuss this numerical topic.

5.2 Temperature accelerated sampling: error bound

Denote by μ_N the distribution of (q_N, p_N) using a cooling schedule $T(\cdot)$, by $\pi_{T(N)}$ the B-G distribution at temperature $T(N)$, and by h the integration time-step.

Assume that the Markov process of (q_N, p_N) satisfies a uniform geometric ergodicity condition of the type

$$\|\mu_i - \pi_{T(i)}\|_{TV} \le \rho_i \|\mu_{i-1} - \pi_{T(i)}\|_{TV} \tag{11}$$

where $\pi_{T(i)}$ is the ergodic measure towards which the process converges if one step update from $(i-1)^{th}$ step to i^{th} at temperature $T(i)$ is repeated, ρ_i is the convergence rate, and statistical distance is measured in total variation norm, which is defined to be:

$$\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)| \tag{12}$$

where $\mathcal{B}$ is the σ-algebra of measurable space.

By repetitive applications of triangle inequality, we derive from Equation (11) that:

$$a_N \le a_1 p_2 + \sum_{j=2}^{N} b_j p_j \tag{13}$$

with $a_i = \|\mu_i - \pi_{T(i)}\|_{TV}$, $b_i = \|\pi_{T(i-1)} - \pi_{T(i)}\|_{TV}$, and $p_i = \prod_{k=i}^{N} \rho_k$.

We further assume that $0 < \rho_i \le 1 - \frac{h}{h_0} e^{-\frac{C_V}{T(i)}}$ for some constants h_0 (stable step length limit) and C_V (elevation of potential energy). Beyond transition state theory this assumption is motivated by [41], [42], [43], [38], [31] and [2].

Using the assumption $0 \le T(j-1) - T(j) \ll T(j)$, we deduce a bound (function of the cooling schedule) on the sampling error:

$$\|\mu_N - \pi_N\|_{TV} \le \sum_{j=2}^{N} \left(\alpha_j \frac{T(j-1) - T(j)}{T(j)} \prod_{k=j}^{N} \left(1 - \frac{h}{h_0} e^{-\frac{C_V}{T(k)}} \right) + o\left(T(j-1) - T(j) \right) \right)$$
$$+ \prod_{k=2}^{N} \left(1 - \frac{h}{h_0} e^{-\frac{C_V}{T(k)}} \right) \tag{14}$$

where $\alpha_j = \mathbb{E}_{T(j)}[H]/T(j)$ ($\alpha_j = 1$ for harmonic oscillators).

5.3 Temperature accelerated sampling: cooling schedules

Naturally, one would like to minimize the error bound (14) with respect to $T(n)$'s. This is however difficult because of nonlinearity. Instead, we consider the following subsets of cooling schedules (denote by T_f the final temperature at which we want to sample the B-G, and by N the number of steps we can afford to employ):

Inverse logarithmic cooling:

$$T(n) = T_f \frac{\log(N+1)}{\log(n+1)} \tag{15}$$

This is the most popular schedule for optimization ([36, 37], for instance, have been frequently cited), but truncated at T_f before $T \to 0$. Recall inverse logarithmic cooling is $T(n) = \frac{C}{\log(n+1)}$, and C is fixed by requiring $T(N) = T_f$. When N is fixed, there is no need to choose any parameter. This schedule will serve as our benchmark.

Shifted inverse logarithmic cooling:

$$T(n) = T_f + \frac{C}{\log(n+1)} \tag{16}$$

where $C > 0$ is the free parameter to be optimized. $T(N)$ is set to be T_f.

Exponential cooling:

$$T(n) = T_f e^{\tilde{C}(N-n)} = T_f C^{N-n} \tag{17}$$

where $C = e^{\tilde{C}} > 1$ is the free parameter to be optimized.

Shifted exponential cooling:

$$T(n) = T_f + \tilde{C} \cdot C^{-n} \tag{18}$$

where $\tilde{C} > 0$ and $C > 1$ are free parameters. For ease on optimization, we chose $\tilde{C} = 10^{-4} T_f C^N$ so that temperatures 'smoothly' cool to T_f, and are left to optimize only one free parameter.

Linear cooling:

$$T(n) = \frac{n}{N} T_f + (1 - \frac{n}{N}) T_i \tag{19}$$

where $T_i > T_f$ is the free parameter. This is used in [44] for optimization purposes. This seemingly too fast cooling schedule does give a small error bound in typical cases (see below).

Inverse linear cooling:

$$T(n) = 1 / \left(\frac{n}{N} \frac{1}{T_f} + (1 - \frac{n}{N}) \frac{1}{T_i} \right) \tag{20}$$

where $T_i > T_f$ is the free parameter. Instead of linearly interpolating the temperature, this linearly interpolates β which is the inverse of temperature to ensure more steps at low temperatures.

Optimal error bound: We optimize error bounds (14) for different total numbers of steps (N's) with respect to the cooling schedules described above. As indicated by Table 1, the optimal schedule depends on the size on the total simulation time via N (to be precise, the ratio between Nh and the mixing time of the original system). Unless N is too small or too large, optimal inverse linear cooling produces a small error bound, optimal linear and exponential coolings have close performances as well, and all three optimal cooling schedules are similar. If the number of steps is too small, B-G will not be approximated well by any cooling schedule, and it is better to use the trivial schedule of constant temperature. If the number is instead too large (usually not the case of interest because accelerated sampling is desired), most types of cooling schedules will yield small errors, and surprisingly, shifted exponential cooling outperforms inverse logarithmic cooling, which is a popular cooling schedule for large N.

N	Constant (no cooling)	Inverse log (benchmark)	Shifted inverse log	Exp	Shifted exp	Linear	Inverse linear
200	0.896	1.304	0.950	**0.896**[1]	**0.896**[1]	**0.896**[1]	**0.896**[1]
600	0.718	0.560	0.752	0.372[2]	0.718	**0.365**[2]	0.368[2]
1000	0.575	0.325	0.597	0.266[3]	0.346	0.267[3]	**0.265**[3]
2000	0.331	**0.142**	0.336	0.153[4]	0.161	0.155[4]	0.151[4]
5000	0.063	0.047	0.064	0.046[5]	**0.028**	0.047[5]	0.046[5]

[1]: Achieved by the limiting case of almost constant temperature
[2,3,4,5]: Achieved by almost the same linear-alike optimizers within each row

Table 1: Optimal error bound for different cooling schedules given total steps N. Within each row, bold indicates the minimum error bound. Different values of N are chosen to represent regimes of very small, small, medium, large, very large N's, in the sense of being compared to the total mixing steps which in this case renders the error bound 0.5 with a constant cooling and is $N \approx 1250$.

In these experiments, $T_f = 20$, $C_V = 150$, $h_0/h = 1$, and $\alpha_j = 1$. In this typical setting T_f/C_V is small and the B-G distribution is concentrated in potential wells, h is close to h_0, and $\alpha_j \approx 1$. If the T_f/C_V is large, however, the optimization suggests not to anneal (result not shown). Optimization is done using MATLAB command fmincon.

Numerical validation on choices of cooling schedule: These cooling schedules have been implemented on the example in Section 3. We did not optimize cooling schedules with respect to free parameters but used a heuristic/generic constant instead. Error on the empirical expectation of position has been investigated for each schedule in Figure 5. The ranking of different types of schedules depends on total simulation time and agrees with theoretical prediction (except for large total simulation times which are dominated by numerical error accumulation).

In addition to Figure 5 and the above discussion that compare cooling schedules for different total simulation times, we fix total time and show time dependent errors of different schedules in Figure 6. Here total simulation time is 30 and we are in the medium N regime. Inverse linear cooling indeed has better performances, followed closely by linear cooling, both consistent with the theoretical analysis. Rigorously speaking one
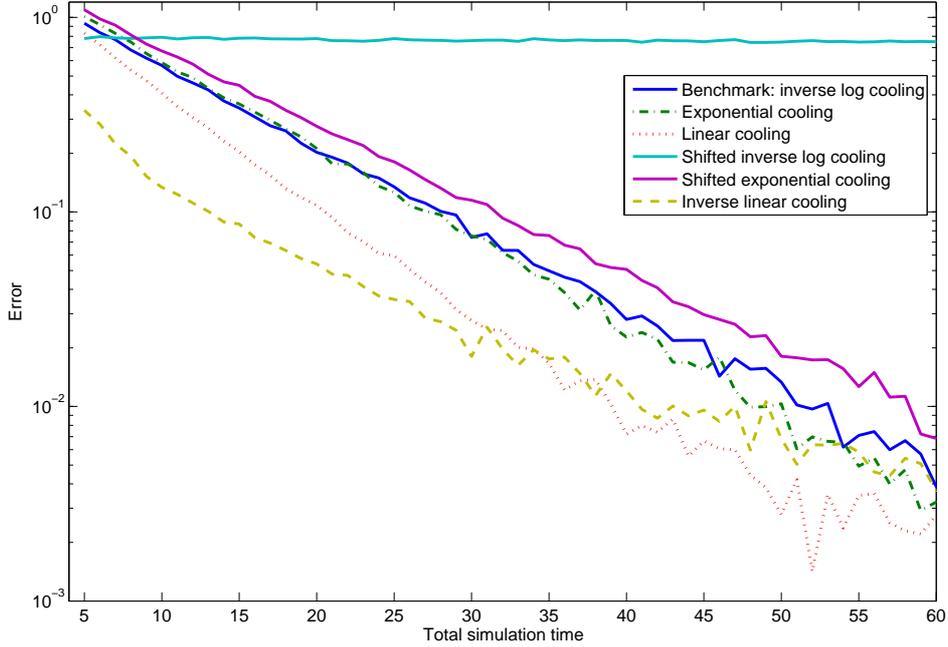
Figure 5: Errors of representative cooling schedules as functions of total simulation time (hence of total simulation step N too). Errors are calculated by $|\frac{1}{M}\sum_{i=1}^{M} q_N^i - \mathbb{E}q(\infty)|$, where $M = 10000$ is the total number of independent trajectories, q_N^i is the Nth step position of the ith trajectory, $N \cdot h$ is the total simulation time and the step length $h = 0.01$. $\mathbb{E}q(\infty)$ is well approximated by empirical average of an ensemble of 20000 TuneGLA trajectories at total simulation time of 300. Constants used in cooling schedules are: Shifted inverse log: $C = 0.01T_f$, Exp: $C = 1.5$, Shifted exp: $T(1) = 2T_f$, Linear: $C = 2T_f$, Inverse linear: $C = 10T_f$. Basically all settings are the same as in Section 3 except for total simulation time and cooling schedule used. Total simulation time is enumerated from 5 to 100 with an increment of 1.

should compare cooling schedules only towards the end of the simulation, because different cooling schedules are at different temperatures in the middle of the simulation; however, the superiority of inverse linear cooling is in fact exhibited throughout the simulation.

These numerical experiments and theoretical bounds indicate that inverse linear cooling is ranked at the top. It is worth pointing out that although annealing accelerates convergence significantly, one has to choose a priori parameters (in most of our cases, total simulation step N and constant C or T_i). This issue usually needs a case-by-case investigation, but C_V (if known) could be used in conjunction with the error bound to determine N and C.
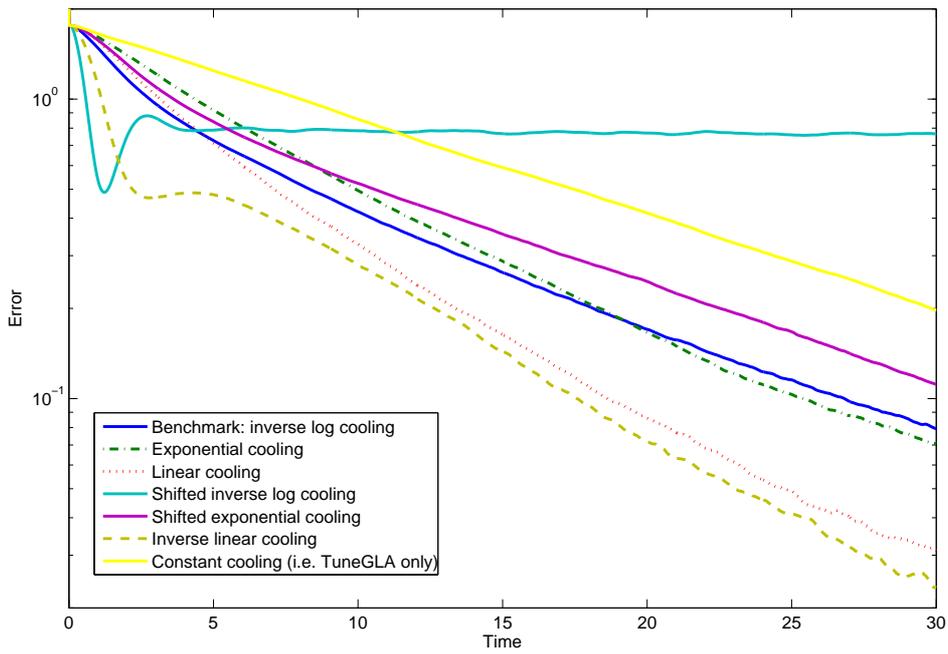
Figure 6: Comparison of errors of TuneGLA with c adaptively tuned and AnnealTuneGLA with different cooling schedules. Again, TuneGLA uses $\alpha = 0.7$, total simulation time=30 is fixed, and all other settings are the same as in Figure 5 and 4 too.