Continuously tempered Hamiltonian Monte Carlo

Matthew M. Graham
University of Edinburgh
m.m.graham@ed.ac.uk

Amos J. Storkey
University of Edinburgh
a.storkey@ed.ac.uk

Abstract

Hamiltonian Monte Carlo (HMC) is a powerful Markov chain Monte Carlo (MCMC) method for performing approximate inference in complex probabilistic models of continuous variables. In common with many MCMC methods however the standard HMC approach performs poorly in distributions with multiple isolated modes. Based on an approach proposed in the statistical physics literature, we present a method for augmenting the Hamiltonian system with an extra continuous *temperature control* variable which allows the dynamic to bridge between sampling a complex target distribution and a simpler uni-modal base distribution. This augmentation both helps increase mode-hopping in multi-modal targets and allows the normalisation constant of the target distribution to be estimated. The method is simple to implement within existing HMC code, requiring only a standard leapfrog integrator. It produces MCMC samples from the target distribution which can be used to directly estimate expectations without any importance re-weighting.

1 Introduction

Hamiltonian Monte Carlo (HMC) [7, 19] has become a workhorse for performing approximate inference in complex high-dimensional probabilistic models of continuous variables. Implementations in probabilistic programming frameworks such as Stan [6] and PyMC3 [21] have allowed increasingly 'black-box' use of HMC methods, leveraging reverse-mode automatic differentiation to efficiently compute the necessary model gradients without manual derivation and using extensions to the original algorithm such as the *No U-Turn Sampler* (NUTS) [12] to adaptively tune the method's free parameters.

In HMC the target density of interest $\mathbb{p}\left[\mathbf{x} = \boldsymbol{x}\right] = \exp\left[-\phi(\boldsymbol{x})\right] / Z$ is defined on a vector variable $\mathbf{x} \in \mathbb{R}^D = \mathcal{X}$, which we will refer to as the *configuration state*. This is augmented with a *momentum state* $\mathbf{p} \in \mathbb{R}^D$. Typically the momentum state is chosen to be independent of the configuration state with marginal $\mathbb{p}\left[\mathbf{p} = \boldsymbol{p}\right] \propto \exp\left[-\tau(\boldsymbol{p})\right]$ such that the joint density factorises as $\mathbb{p}\left[\mathbf{x} = \boldsymbol{x}, \mathbf{p} = \boldsymbol{p}\right] = \mathbb{p}\left[\mathbf{x} = \boldsymbol{x}\right]\mathbb{p}\left[\mathbf{p} = \boldsymbol{p}\right] \propto \exp\left[-\phi(\boldsymbol{x}) - \tau(\boldsymbol{p})\right]$. With analogy to classical dynamics, $\phi(\boldsymbol{x})$ is referred to as the *potential energy* and $\tau(\boldsymbol{p})$ the *kinetic energy*, with $H(\boldsymbol{x}, \boldsymbol{p}) = \phi(\boldsymbol{x}) + \tau(\boldsymbol{p})$ being termed the *Hamiltonian*. By construction, marginalising the joint density over the momenta recovers $\mathbb{p}\left[\mathbf{x} = \boldsymbol{x}\right]$.

Each HMC update involves a discrete-time simulation of the canonical Hamiltonian dynamic which conserves the Hamiltonian and is time-reversible and volume-preserving. Through choice of an appropriate geometric integrator such as the popular Störmer-Verlet or *leapfrog* scheme, the simulated discrete time dynamic remains exactly time-reversible and volume-preserving and also approximately conserves the Hamiltonian even over long simulated trajectories [15].

The discrete time dynamic is used to generate a new proposed state $(\boldsymbol{x}', \boldsymbol{p}')$ given the current state $(\boldsymbol{x}, \boldsymbol{p})$ and this proposal accepted or rejected in a Metropolis step with accept probability $\min\left\{1, \exp\left[H(\boldsymbol{x}, \boldsymbol{p}) - H(\boldsymbol{x}', \boldsymbol{p}')\right]\right\}$. Due to the approximate Hamiltonian conservation the accept probability is typically close to one, and so HMC is able to make long-range moves in high-dimensional state-spaces while still maintaining high acceptance rates, a significant improvement over the behaviour typical of simpler Metropolis-Hastings methods in high-dimensions.

The energy-conservation property which affords this desirable behaviour also however suggests that standard HMC updates are unlikely to move between isolated modes in a target distribution. The Hamiltonian is approximately conserved over a trajectory therefore we have $\phi(\boldsymbol{x}')-\phi(\boldsymbol{x}) \approx \tau(\boldsymbol{p})-\tau(\boldsymbol{p}')$. Typically a quadratic kinetic energy $\tau(\boldsymbol{p}) = \boldsymbol{p}^{\mathrm{T}} \boldsymbol{M}^{-1} \boldsymbol{p} \,/\, 2$ is used corresponding to a Gaussian marginal density on the momentum state. As this kinetic energy is bounded from below by zero, the maximum change in potential energy over a trajectory is approximately equal to the initial kinetic energy.

At equilibrium the momenta will have a Gaussian distribution and so the kinetic energy a χ^2 distribution with mean $D/2$ and variance D [2, 19]. If potential energy barriers significantly larger than $\sim D$ separate regions of the configuration state space the HMC updates are unlikely to move across the barriers meaning impractically long sampling runs will be needed for effective ergodicity.

A common approach in MCMC methods for dealing with multi-modal target distributions is to introduce a concept of *temperature*. In statistical mechanics, the Boltzmann distribution on a configuration $\mathbf{x}$ of a mechanical system with energy function ϕ and in thermal equilibrium with a heat bath at temperature T is defined by a probability density $\exp\left[-\beta\phi(\boldsymbol{x})\right] / \mathcal{Z}(\beta)$ where $\beta = (k_B T)^{-1}$ is the *inverse temperature*, k_B is Boltzmann's constant and $\mathcal{Z}(\beta)$ is the partition function. At high temperatures ($\beta \to 0$) the density function becomes increasingly flat across the configuration state space and correspondingly energy barriers between different regions of the state space become lower.

In the standard statistical mechanics formulation, the distribution in the limit $\beta \to 0$ is an improper flat density across the configuration state space (assuming $\mathbf{x} \in \mathbb{R}^D$). More usefully from a statistical perspective we can use an inverse temperature variable $\beta \in [0, 1]$ to geometrically bridge between a simple distribution with normalised density $\exp\left[-\psi(\boldsymbol{x})\right]$ at $\beta = 0$ and the target distribution at $\beta = 1$

$$\pi\left[\boldsymbol{x} \mid \beta\right] = \frac{1}{\mathcal{Z}(\beta)} \exp\left[-\phi(\boldsymbol{x})\right]^{\beta} \exp\left[-\psi(\boldsymbol{x})\right]^{1-\beta} = \frac{1}{\mathcal{Z}(\beta)} \exp\left[-\beta\phi(\boldsymbol{x}) - (1-\beta)\psi(\boldsymbol{x})\right]. \tag{1}$$

Various approaches for altering the temperature in a simulated system have been proposed. In *simulated tempering* [16] an ordered set of inverse temperature values is chosen

$$\left\{\beta_n\right\}_{n=0}^{N} : 0 = \beta_N < \beta_{N-1} < \cdots < \beta_0 = 1 \tag{2}$$

and a probability distribution $\mathbb{P}\left[\beta = \beta_n\right] \;\forall n \in \{0 \dots N\}$ specified. MCMC updates are performed on the augmented state $(\mathbf{x}, \beta)$ with $\mathbf{x}$ samples for which $\beta = 1$ converging in distribution to the target.

In *parallel tempering*, *Metropolis-coupled MCMC* or *replica exchange* [8, 10, 22], multiple Markov chains on the configuration state are run in parallel, with the i^{th} chain having an associated inverse temperature β_n defined as in (2). MCMC updates are performed on each chain which leave $\pi[\boldsymbol{x} \mid \beta_n]$ invariant. Interleaved with these updates, exchanges of the configuration states of the chains at adjacent inverse temperatures $(\beta_n,\ \beta_{n+1})$ are proposed and accepted or rejected in a Metropolis–Hastings step.

Tempered transitions [18] uses a different approach, deterministically altering the inverse temperature rather than including it as part of the Markov chain state. Again an ordered set of inverse temperatures $\{\beta_n\}_{n=0}^{N}$ are specified as in (2). For each inverse temperature $\beta_n : n > 0$ a pair of transition operators $\hat{T}_n$, $\check{T}_n$ are defined which satisfy a reversibility condition $\hat{T}_n[\boldsymbol{x}' \mid \boldsymbol{x}]\pi[\boldsymbol{x} \mid \beta_n] = \check{T}_n[\boldsymbol{x} \mid \boldsymbol{x}']\pi[\boldsymbol{x}' \mid \beta_n]$ (if $\hat{T}_n = \check{T}_n$ this is just detailed balance). The transition operators associated with lower inverse-temperatures will typically be able to make larger moves in the state space.

These transition operators are applied to the current state in a fixed sequence $\hat{T}_1 \dots \hat{T}_N \check{T}_N \dots \check{T}_1$ to generate a chain of intermediate states, with the idea that in the initial 'heating' stage $\hat{T}_1 \dots \hat{T}_N$ the state will be roughly distributed after each transition according to the target distribution at an increasing series of temperatures before being 'cooled' back down with the transitions $\check{T}_N \dots \check{T}_1$ so that the final state is approximately distributed according to the target (with $\beta_0 = 1$) again. The overall transition from the initial state to the final state in the chain is then accepted or rejected in a Metropolis–Hastings step. A HMC specific variation of this idea in which the tempering is applied within a simulated trajectory by scaling the momenta in a deterministic sequence has also been proposed [19].

In all these methods the number and spacing of the inverse temperatures used, and additional algorithmic choices such as inverse temperature probabilities in simulated tempering or temperature dependence of the transition operators in tempered transitions, are key to getting the methods to perform well in complex high dimensional distributions [1, 3, 18]. To get reasonable performance it

may be necessary to do preliminary pilot runs to guide the choices of these parameters, adding to the computational burden and also difficulty of using these methods in a black-box fashion.

A natural question is whether it is possible to use a continuously varying inverse temperature variable to side-step the need to choose a set of values. *Path sampling* [9] proposes this approach, defining a general *path* as a functional form parametrised by β which continuously maps between the target density $\exp\left[-\phi(\boldsymbol{x})\right] / Z$ at $\beta = 1$ and a base density $\exp\left[-\psi(\boldsymbol{x})\right]$ at $\beta = 0$, with the geometric bridge in (1) a particular example. A joint target $\mathbb{p}\left[\mathbf{x} = \boldsymbol{x},\ \beta = \beta\right] \propto \pi\left[\boldsymbol{x} \mid \beta\right] \rho\left[\beta\right]$ is defined with $\rho\left[\beta\right]$ a user-chosen 'prior' density on the inverse temperature variable. In [9] it is proposed to construct a Markov chain leaving the joint density invariant by alternating updates of $\mathbf{x}$ given β and β given $\mathbf{x}$ using random-walk Metropolis–Hastings or Gibbs sampling transitions for the individual updates.

The main focus of path sampling is estimation of the normalisation constant $Z = \mathcal{Z}(1)$ of the target distribution. To estimate expectations with respect to the target distribution typically an importance re-weighted estimator will be needed to adjust for not having a set of samples for which $\beta = 1$ exactly.

Adiabatic Monte Carlo [3] is another instantiation of the idea of using a continuously varying temperature variable, here specifically in the context of HMC. The original Hamiltonian system $(\mathbf{x}, \mathbf{p})$ is further augmented with a *contact coordinate* $\gamma \in \mathbb{R}$ which is a logit transform of the inverse temperature β. Using the differential geometric theory of contact manifolds, a *contact Hamiltonian* is defined on the augmented system, this defining a contact Hamiltonian flow, which can be considered an instance of the thermodynamical concept of an isentropic or reversible adiabatic process.

Exact simulation of the contact Hamiltonian flow generates a trajectory which conserves the contact Hamiltonian while deterministically traversing the inverse temperature range $[0, 1]$. Simulating the contact Hamiltonian flow is non-trivial in practice however: the contact Hamiltonian includes the log partition function $\log \mathcal{Z}(\beta)$ as a term, the partial derivatives of which require computing expectations with respect to $\pi[\boldsymbol{x} \mid \beta]$ which for most problems is intractable to do exactly. One option is to estimate the expectations with an inner loop running HMC however this adds to the computational burden and makes ensuring overall reversibility of the trajectory difficult. Moreover the contact flow can encounter meta-stabilities whereby the flow halts at an intermediate $\beta \notin \{0, 1\}$ meaning the flow no longer defines a bijection between $\beta = 0$ and $\beta = 1$. This can be ameliorated by regular resampling of the momenta however this potentially increases the random-walk behaviour of the overall dynamic.

An alternative *extended Hamiltonian approach* for simulating a system with a continuously varying temperature was proposed recently in the statistical physics literature [11]. Again the inverse temperature of the system is indirectly set via an auxiliary variable, which we will term a *temperature control variable* $\mathsf{u} \in \mathbb{R}$. This control variable is mapped to an interval $[0, s]$, $0 < s < 1$ via a piecewise defined function f, with the conditions that for a pair of thresholds (θ_1, θ_2) with $0 < \theta_1 < \theta_2$, $f(u) = 0\ \forall\ |u| \leq \theta_1$, $f(u) = s\ \forall\ |u| \geq \theta_2$ and $0 < f(u) < s\ \forall\ \theta_1 < |u| < \theta_2$. In practice we will usually also require that f is continuously differentiable. Appendix C gives some concrete examples.

Unlike Adiabatic Monte Carlo, an additional momenta variable v corresponding to u is also introduced. Although seemingly a minor difference this simplifies the implementation of the approach significantly as the system retains a natural symplectic structure and can continue to be viewed within the usual Hamiltonian dynamics framework. An extended Hamiltonian is then defined on the augmented system

$$H^\star(\boldsymbol{x},\, u,\, \boldsymbol{p},\, v) = H(\boldsymbol{x},\, \boldsymbol{p}) - f(u)G(\boldsymbol{x},\, \boldsymbol{p}) + \omega(u) + \frac{v^2}{2m} \tag{3}$$

where $H(\boldsymbol{x}, \boldsymbol{p})$ is the original Hamiltonian defined previously, $\omega(u)$ is a 'confining potential' on u (analagous to the negative log of $\rho(u)$ in path sampling), and m is the mass (marginal variance) associated with v. $G(\boldsymbol{x}, \boldsymbol{p})$ can be any differentiable function, but the link to tempering approaches comes most obviously from the choice of $G(\boldsymbol{x}, \boldsymbol{p}) = \phi(\boldsymbol{x})$. In this case assuming a separable (no terms coupling $\boldsymbol{x}$ and $\boldsymbol{p}$) Hamiltonian $H(\boldsymbol{x}, \boldsymbol{p}) = \phi(\boldsymbol{x}) + \boldsymbol{p}\boldsymbol{M}^{-1}\boldsymbol{p}/2$ the extended Hamiltonian is

$$H^\star(\boldsymbol{x},\, u,\, \boldsymbol{p},\, v) = \left[1 - f(u)\right]\phi(\boldsymbol{x}) + \omega(u) + \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{p} + \frac{v^2}{2m} \tag{4}$$

with the term $1 - f(u)$ acting analogously to the inverse temperature variable β encountered earlier. Further this extended Hamiltonian remains separable with respect to the extended configuration $(\mathbf{x}, \mathsf{u})$ and extended momentum $(\mathbf{p}, \mathsf{v})$ and so can be efficiently simulated using, for example, a standard leapfrog integrator. The reversible and volume-preserving simulated dynamic can then be used as a proposal generating mechanism for a Metropolis–Hastings step as in the usual HMC approach.

Due to the condition $f(u) = 0 \; \forall \; |u| < \theta_1$, the set of sampled configuration states $\mathbf{x}$ which have associated $|\mathsf{u}| < \theta_1$ will (assuming the chain is ergodic) asymptotically converge in distribution to the target, and so can be used to estimate expectations without any importance re-weighting.

In contrast to contact Hamiltonian flow dynamic in Adiabatic Monte Carlo, the effective inverse temperature $1 - f(u)$ will not necessarily consistently increase or decrease when simulating the extended Hamiltonian dynamic. Due to the associated momentum v the inverse temperature progression will not stall due to meta-stabilities, however if there are large barriers in the 'extended potential energy' $[1 - f(u)]\phi(\boldsymbol{x}) + \omega(u)$ as u is varied then the dynamic will tend not explore the full distribution of u values well, limiting the gains from the augmentation.

To counter this issue, it is proposed in [11] to use an adaptive history-dependent biasing potential on u to try to achieve a flat density across a bounded interval $|u| < \theta_2$, using for example metadynamics [14]. The resulting non-Markovian updates bias the invariant distribution of the configuration state however this can be accounted for either by a re-weighting scheme [5], or using a vanishing adaptation in an initial 'warm-up' stage as for example used to tuning the mass matrix and step-size in Stan [6].

Although use of adaptive methods like meta-dynamics can help substantially in using the method in a black-box fashion, it is also instructive to consider how we might flatten the marginal density on u using non-adaptive methods. In some simpler cases this can remove the need for adaptive methods altogether, and in more complex cases should still help improve robustness.

2 Method

We use a variation of the original extended Hamiltonian approach, by geometrically bridging between the target distribution and a simple base distribution with *normalised* density $\exp[-\psi(\boldsymbol{x})]$ as in (1). Unlike the cases considered in [11] this allows the $\beta(u) = 1 - f(u)$ to span the full $[0, 1]$ interval (i.e. setting $s = 1$ in definition of $f(u)$ above) as the conditional distribution on the configuration state given $\beta(u) = 0$ remains well-defined rather than an improper flat density.

As another small alteration, we define the temperature control variable u to have a circular topology, wrapping at the boundaries of the interval $[-1, 1]$. This removes the requirement to choose a 'confining potential' to ensure u remains bounded. Finally we introduce a term $\beta(u) \log \zeta$ into the Hamiltonian, with $\log \zeta$ chosen as some deterministic approximation to $\log Z$. This gives an extended Hamiltonian

$$H^\star(\boldsymbol{x}, u, \boldsymbol{p}, v) = \beta(u)\phi(\boldsymbol{x}) + [1 - \beta(u)]\,\psi(\boldsymbol{x}) + \beta(u) \log \zeta + \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{p} + \frac{v^2}{2m}. \tag{5}$$

We can trivially marginalise out the momenta from the joint distribution defined by this Hamiltonian, and further marginalising over the configuration state $\mathbf{x}$ we have that

$$\mathbb{p}\left[\mathsf{u} = u\right] \propto \mathscr{Z}\left[\beta(u)\right] = \zeta^{-\beta(u)} \int_{\mathcal{X}} \exp\left\{-\beta(u)\phi(\boldsymbol{x}) - [1 - \beta(u)]\,\psi(\boldsymbol{x})\right\} \mathrm{d}\boldsymbol{x}. \tag{6}$$

By applying Hölder's and Jensen's inequalities we can bound $\mathscr{Z}[\beta(u)]$ (see Appendix A for details)

$$\beta(u)\left\{\log Z - \log \zeta - D^{b \to t}\right\} \leq \log \mathscr{Z}\left[\beta(u)\right] \leq \beta(u)\left\{\log Z - \log \zeta\right\} \tag{7}$$

where $D^{b \to t}$ indicates the *Kullback–Leibler* (KL) divergence from the base to target distribution. If $\log \zeta = \log Z$ the upper-bound is zero, implying a flat upper bound on the marginal density on u. If additionally $D^{b \to t} = 0$, the bound becomes tight and we will have a flat marginal density on u.

In reality we do not know $\log Z$ and cannot choose a base distribution such that the KL divergence is zero as we wish to use a simple density amenable to exploration. However we can see that under the constraint of the base distribution allowing exploration, we wish to minimise the KL divergence to the target distribution. Further we want to find a ζ as close to Z as possible.

Variational inference is an obvious route for tackling both problems, allowing us to fit a base density from some simple parametric family (e.g. Gaussian) by directly minimising the KL divergence term in (7) and simultaneously giving a lower bound on $\log Z$. In some cases we can make use of variational inference methods specifically aimed at the target distribution family, however more generally *Automatic Differentiation Variational Inference* (ADVI) [13] provides a black-box framework for fitting variational approximations to differentiable target densities.

A potential problem is that the classes of target distribution that we are most interested in applying our approach to — those with multiple isolated modes — are precisely the same distributions that simple variational approximations will tend to fit poorly, the KL divergence being minimised favouring 'mode-seeking' solutions [4], which fit only one mode well. This both limits how small the divergence from the base to target can be made, but also crucially is undesirable for tempering approaches as we wish the base distribution to allow the dynamic to move between modes in the target.

One option would be to instead to minimise the reversed form of the KL divergence from the target to base distribution, this tending to produce 'mode-covering' solutions that match global moments. Methods such as *expectation propagation* (EP) [17] do allow moment-matching approximations to be found (and also give an approximation for Z) and may be a good option in some cases. EP is applicable to a more limited class of distributions than ADVI however and can have convergence issues so is less amenable to a general black-box approach.

An alternative is to fit multiple local variational approximations $\left\{q_i(\boldsymbol{x})\right\}_{i=1}^{L}$ by minimising the variational objective from multiple random parameter initialisations (discarding any duplicate solutions), each approximating a single mode well. We can then combine these local approximations into a global approximation. One option is to use a mixture of the local approximations

$$q(\boldsymbol{x}) = \sum_{i=1}^{L} \left[w_i q_i(\boldsymbol{x})\right] \qquad w_i = \frac{\exp(-K_i)}{\sum_{j=0}^{L}\left[\exp(-K_j)\right]} \tag{8}$$

with K_i the KL divergence from q_i to target distribution (minus the unknown log Z) which corresponds to the variational objective being minimised for each q_i. If the local approximations have non-overlapping support this will lead to a global approximation which is guaranteed to be closer in KL divergence than any of the local approximations [24]. Often we may wish to use local approximations with overlapping support (e.g. Gaussian) where the guarantee does not apply. However for the cases of target distributions with multiple isolated modes the overlap between local Gaussian approximations to each mode will often be minimal and so the method is still a useful heuristic.

In our case however a mixture distribution is unlikely to be a good choice of base distribution as it will tend to itself be multi-modal. We therefore instead propose here to use a base distribution with moments matched to the fitted mixture distribution, e.g. a single Gaussian with mean and covariance matched to the mean and covariance of the mixture.

The relation between the marginal density on u and partition-function $\mathcal{Z}[\beta(u)]$ expressed in (6) also suggests that we can use the u samples from a Markov chain leaving the joint density on $(\mathbf{x}, \mathsf{u})$ invariant to form an estimate of the normalising constant Z. If $\{u^{(s)}\}_{s=1}^{S}$ are a set of MCMC samples of u then, as shown in Appendix B a consistent estimator of Z is defined by

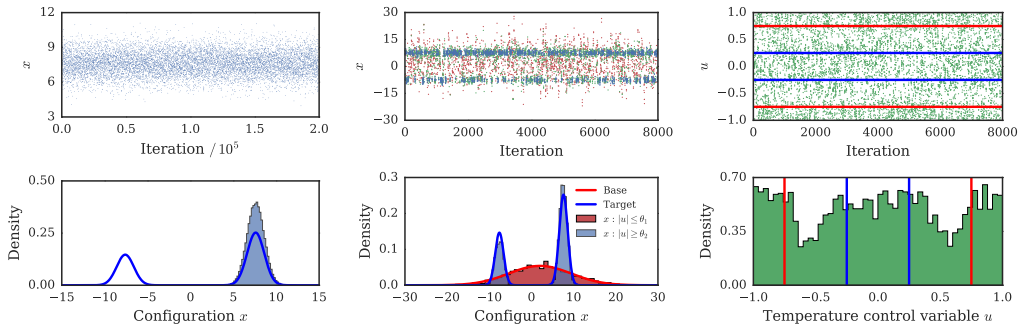
$$Z = \frac{1-\theta_2}{\theta_1}\frac{\mathbb{P}\left[0 \leq |\mathsf{u}| \leq \theta_1\right]}{\mathbb{P}\left[\theta_2 \leq |\mathsf{u}| \leq 1\right]}\zeta = \lim_{S\to\infty}\frac{1-\theta_2}{\theta_1}\frac{\sum_{s=1}^{S}\left\{\mathbb{1}\left[0 \leq |u^{(s)}| \leq \theta_1\right]\right\}}{\sum_{s=1}^{S}\left\{\mathbb{1}\left[\theta_2 \leq |u^{(s)}| \leq 1\right]\right\}}\zeta. \tag{9}$$

We can also use the samples for intermediate $\theta_1 \leq |u| \leq \theta_2$ to form a distinct estimator for Z by using the thermodynamic integration identity used in path sampling [9]. In both cases the quality of the estimator will be dependent on how well the chain explores the marginal density on u - if the chain rarely moves from the target to base distribution or vice versa, the estimates are likely to be poor.

3 Experiments

To validate the method we performed a series of experiments comparing running HMC in the extended and original Hamiltonian systems. For the test model we used a Gaussian mixture relaxation of a Boltzmann machine distribution [23]. In certain parameter regimes the relaxations become highly multi-modal making it a challenging distribution for MCMC methods to explore well. The moments of the relaxation distribution can be calculated from the moments of the original discrete distribution, which for models with a small number of discrete states D can be computed exactly by exhaustive iteration across the 2^D discrete states. This allows ground truth moments to be calculated against which convergence can be checked. The parametrisation used is described in in Appendix D.

All experiments were implemented in Stan and run using `pystan`. Except where otherwise noted we used the adaptive HMC variant NUTS which automatically adjusts number of integrator steps



(a) Original system (x) (b) Extended system (x) (c) Extended system (u)

Figure 1: MCMC samples from running NUTS on univariate Gaussian mixture target density. (a) shows results for standard Hamiltonian system (x only) and (b) and (c) are results for extended Hamiltonian system (x and u). The top row shows trace plots, showing successive samples in the Markov chain. In (b) top the x samples are colour-coded according to the corresponding u value - red: $|u| \leq \theta_1$, blue $|u| \geq \theta_2$ and green otherwise. The red lines in (c) correspond to $\pm\theta_2$ and the blue lines $\pm\theta_1$. Bottom row shows normalised sample histograms (shaded regions). In (a) and (b) bottom plots target density is shown as a thick blue line. In (b) bottom plot base density is additionally shown by a thick red line and histograms are shown for both x samples for which $|u| \leq \theta_1$ (blue region, converging to target density) and $|u| \geq \theta_2$ (red region, converging to base density).

used. Stan also includes a facility to adapt the step size and mass matrix during warm-up iterations however we found this performed poorly in the extended system cases we tested (potentially due to the very differing appropriate scales in the base and target density) so we used a fixed step size of 0.5 when working in the extended system. The Stan models for the original and extended Hamiltonian approaches are provided in Appendix F. We used thresholds $\theta_1 = 0.25$ and $\theta_2 = 0.75$ for the control function $f(u) = 1 - \beta(u)$ and used the p_5 quintic polynomial interpolant from Appendix C.

As an initial toy example we performed inference in a univariate Gaussian mixture target with two well-separated mixture components. Due to the trivial number of modes, for the base density we used a Gaussian with moments exactly matched to the target and used $\log \zeta = \log Z$ rather than the matching moments to a mixture of variational approximations. Figure 1 shows results for single chains in both the original non-augmented system (1a) and extended system (1b and 1c). The trace plot in Figure 1a shows that the dynamic in the non-augmented system is unable to move between the two modes, with the chain remaining confined to the same mode over all 2×10^5 updates, leading to an inaccurate estimate of the density on x in the bottom histogram.

In contrast in the extended system the dynamic is able to regularly jump between the modes in x, with the x samples for which $|u| < \theta_1$ (blue points in 1b trace plot and blue region in 1b density plot) converging quickly in distribution to the multi-modal target density, and correspondingly the x samples for which $|u| \geq \theta_2$ converging to the base density. The plots in 1c show that the u chain is exploring its full marginal distribution well, with minimal autocorrelation evident in the trace plot and the marginal showing an approximately equal flat density for $|u| \leq \theta_1$ and $|u| \geq \theta_2$ as expected due to using the exact relationship $\log \zeta = \log Z$. The density for $\theta_1 < |u| < \theta_2$ shows a pronounced dip, this a result of the non-zero KL divergence between base and target densities. Although the dynamic is still able to easily move across the moderate potential barrier in this toy problem, in more complex systems for which the divergence between base and target will generally be larger it can become increasingly difficult for the dynamic to explore the full range of u values.

As a more complex test case we considered a 19 dimensional Gaussian mixture relaxation corresponding to a Boltzmann machine distribution on a 20 dimensional binary state. The Boltzmann machine weights and biases were randomly generated so as to encourage multi-modality: the eigenspectrum of the weight matrix was shaped to favour multiple large negative and positive eigenvalues and small bias values sampled to discourage a small number of modes dominating. The Gaussian base density was specified by fitting a series of mean field variational approximations to the Boltzmann machine distribution, and matching the first and second moments of a mixture of Gaussian components located at the mean-field solutions. All chains were initialised at at a random sample from the fitted mixture.

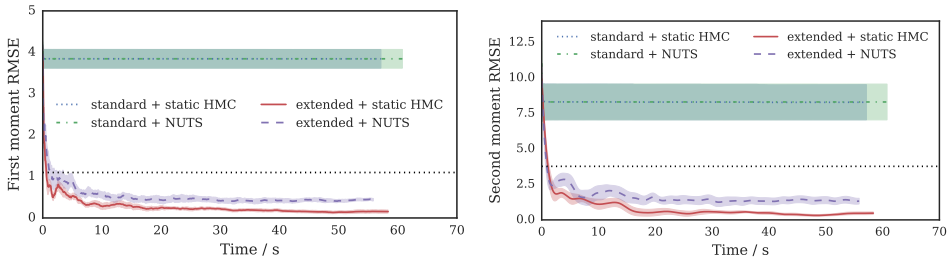


Figure 2: Errors in empirical moments estimated from MCMC samples compared to true values as an increasing number of samples are included in estimator for a 19-dimensional multi-modal Boltzmann machine relaxation target distribution. MCMC dynamics were run both in the standard Hamiltonian system and the proposed extended system, with both non-adaptive 'static HMC' and adaptive NUTS algorithms being tested. For each system / algorithm pair 8 independent chains were run with the curves showing the average RMSE over the chains and the shaded regions ± 1 standard error of the mean. Each curve has been scaled by the average time taken per chain for that setting. The dotted horizontal black line in both plots indicates the corresponding RMSE in the moments of the mixture of variational approximations used to set the base density.

NUTS and static HMC were used to perform MCMC inference in both the original and extended Hamiltonian systems. Plots showing the distributional convergence of the four combinations are shown in figure 2. This is measured by the *root mean squared error* (RMSE) between the empirical and true first and second moments as the number of successive MCMC samples (for which $|u| \leq \theta_1$ in the extended cases) included in the moment estimators is increased. The chains in the extended system converge towards the target distribution, unlike in the original system where trace plots (not shown) suggest the dynamic struggles to move out of an isolated mode. The non-adaptive HMC dynamic (which was not particularly carefully tuned) seems to perform better in the extended system than NUTS — this may be due to the unusual geometry of the extended joint distribution.

Using (9) we can also estimate the normalising constant of the target distribution from the u samples in the extended system chains. The mean absolute error in the $\log Z$ estimate was 0.027 ± 0.007 across the static HMC chains and 0.080 ± 0.017 across the NUTS chains. Both represent a significant improvement over the 0.782 difference between the approximate $\log \zeta$ and the true $\log Z$. Further experimental results for additional random relaxation distributions are shown in Appendix E.

4 Discussion

The method we have presented is a simple augmentation to the standard HMC approach which can both help exploration of distributions with multiple isolated modes and allow estimation of the normalisation constant of the target distribution. Our formulation leverages variational inference to find a simple base distribution approximating the target distribution. It can therefore be seen within the context of class of methods trying to 'bridge the gap' between variational inference and MCMC methods [20], exploiting cheap optimisation based inference methods while still offering the potential of asymptotically exact inference. Initial experimental results are promising however it is likely many of the algorithmic choices we made are far from optimal and so it seems there is much potential to both improve both the computational efficiency and 'black-boxness' of the approach.

Acknowledgments

We thank Ben Leimkuhler for introducing us to the extended Hamiltonian continuous-tempering approach. This work was supported in part by grants EP/F500385/1 and BB/F529254/1 for the University of Edinburgh School of Informatics Doctoral Training Centre in Neuroinformatics and Computational Neuroscience (www.anc.ac.uk/dtc) from the UK Engineering and Physical Sciences Research Council (EPSRC), UK Biotechnology and Biological Sciences Research Council (BBSRC), and the UK Medical Research Council (MRC).

References

[1] G. Behrens, N. Friel, and M. Hurn. Tuning tempered transitions. *Statistics and computing*, 2012.

[2] M. Betancourt. A general metric for Riemannian manifold Hamiltonian Monte Carlo. In *Geometric Science of Information*. Springer, 2013.

[3] M. Betancourt. Adiabatic Monte Carlo. *arXiv preprint arXiv:1405.3489*, 2014.

[4] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[5] M. Bonomi, A. Barducci, and M. Parrinello. Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics. *Journal of computational chemistry*, 2009.

[6] B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 2016.

[7] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 1987.

[8] D. J. Earl and M. W. Deem. Parallel tempering: theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, 2005.

[9] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 1998.

[10] C. J. Geyer. Markov chain Monte Carlo maximum likelihood. *Computer Science and Statistics*, 1991.

[11] G. Gobbo and B. J. Leimkuhler. Extended Hamiltonian approach to continuous tempering. *Physical Review E*, 2015.

[12] M. D. Hoffman and A. Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 2014.

[13] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*, 2016.

[14] A. Laio and M. Parrinello. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 2002.

[15] B. Leimkuhler and S. Reich. *Simulating Hamiltonian dynamics*. Cambridge University Press, 2004.

[16] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 1992.

[17] T. P. Minka. Expectation propagation for approximate Bayesian inference. *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, 2001.

[18] R. M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 1996.

[19] R. M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2011.

[20] T. Salimans, D. P. Kingma, and M. Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, 2015.

[21] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2016.

[22] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, 1986.

[23] Y. Zhang, Z. Ghahramani, A. J. Storkey, and C. A. Sutton. Continuous relaxations for discrete Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 3194–3202, 2012.

[24] O. Zobay. Mean field inference for the Dirichlet process mixture model. *Electronic Journal of Statistics*, 2009.

A Bounding the partition function

To derive the upper-bound we use Hölder's inequality

$$\int_{\mathcal{X}} g(\boldsymbol{x})h(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \leq \left\{\int_{\mathcal{X}} |g(\boldsymbol{x})|^{\frac{1}{a}}\,\mathrm{d}\boldsymbol{x}\right\}^{a}\left\{\int_{\mathcal{X}} |h(\boldsymbol{x})|^{\frac{1}{1-a}}\,\mathrm{d}\boldsymbol{x}\right\}^{1-a} \tag{10}$$

where $a \in [0,\,1]$ and g and h are measurable functions, and the definitions

$$\int_{\mathcal{X}} \exp\left[-\phi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x} = Z \quad \text{and} \quad \int_{\mathcal{X}} \exp\left[-\psi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x} = 1. \tag{11}$$

From (6) (dropping the u dependence of β for clarity) we have

$$\mathcal{Z}(\beta) = \zeta^{-\beta}\int_{\mathcal{X}} \left\{\exp\left[-\phi(\boldsymbol{x})\right]^{\beta}\right\}\left\{\exp\left[-\psi(\boldsymbol{x})\right]^{1-\beta}\right\}\,\mathrm{d}\boldsymbol{x}.$$

Applying Hölder's inequality (10) with $g(\boldsymbol{x}) = \exp[-\phi(\boldsymbol{x})]^{\beta}$, $h(\boldsymbol{x}) = \exp[-\psi(\boldsymbol{x})]^{1-\beta}$ and $a = \beta$

$$\begin{aligned}\mathcal{Z}(\beta) &\leq \zeta^{-\beta}\left\{\int_{\mathcal{X}} \left|\exp\left[-\phi(\boldsymbol{x})\right]^{\beta}\right|^{\frac{1}{\beta}}\,\mathrm{d}\boldsymbol{x}\right\}^{\beta}\left\{\int_{\mathcal{X}} \left|\exp\left[-\psi(\boldsymbol{x})\right]^{1-\beta}\right|^{\frac{1}{1-\beta}}\,\mathrm{d}\boldsymbol{x}\right\}^{1-\beta}\\ &= \zeta^{-\beta}\left\{\int_{\mathcal{X}} \exp\left[-\phi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x}\right\}^{\beta}\left\{\int_{\mathcal{X}} \exp\left[-\psi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x}\right\}^{1-\beta}.\end{aligned}$$

Using (11) and taking logarithms of both sides gives

$$\log \mathcal{Z}(\beta) \leq \beta\left(\log Z - \log\zeta\right).$$

To derive the lower-bound we use Jensen's inequality

$$\varphi\left\{\int_{\mathcal{X}} g(\boldsymbol{x})q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right\} \geq \int_{\mathcal{X}} \varphi\left\{g(\boldsymbol{x})\right\}q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \tag{12}$$

for a concave function φ, normalised density $q : \int_{\mathcal{X}} q(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = 1$ and measurable g.

Rearranging (6) and taking logarithms we have

$$\log \mathcal{Z}(\beta) + \beta\log\zeta = \log\left\{\int_{\mathcal{X}} \exp\left\{-\beta\left[\phi(\boldsymbol{x}) - \psi(\boldsymbol{x})\right]\right\}\exp\left[-\psi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x}\right\}.$$

Applying Jensen's inequality (12) with $\varphi = \log$, $q = \exp(-\psi)$ and $g = \exp\left\{-\beta\left[\phi - \psi\right]\right\}$

$$\begin{aligned}\log \mathcal{Z}(\beta) + \beta\log\zeta &\geq \beta\int_{\mathcal{X}} \left[\psi(\boldsymbol{x}) - \phi(\boldsymbol{x})\right]\exp\left[-\psi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x}\\ &= \beta\int_{\mathcal{X}} \left\{\log Z - \log Z - \frac{\log\exp\left[-\psi(\boldsymbol{x})\right]}{\log\exp\left[-\phi(\boldsymbol{x})\right]}\right\}\exp\left[-\psi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x}\\ &= \beta\log Z - \beta\int_{\mathcal{X}} \exp\left[-\psi(\boldsymbol{x})\right]\log\frac{\exp\left[-\psi(\boldsymbol{x})\right]}{\frac{1}{Z}\exp\left[-\phi(\boldsymbol{x})\right]}\,\mathrm{d}\boldsymbol{x}.\end{aligned}$$

Recognising the integral in the last line as the *Kullback–Leibler* (KL) divergence from the base distribution to the target distribution and rearranging we have

$$\log \mathcal{Z}(\beta) \geq \beta\left(\log Z - \log\zeta\right) - \beta D_{\mathrm{KL}}\left[\exp(-\psi)\,\|\,\exp(-\phi)/Z\right].$$

By instead noting (6) can be rearranged into the form

$$\log \mathcal{Z}(\beta) + \beta\log\zeta - \log Z = \log\left\{\int_{\mathcal{X}} \exp\left\{-(1-\beta)\left[\psi(\boldsymbol{x}) - \phi(\boldsymbol{x})\right]\right\}\frac{1}{Z}\exp\left[-\phi(\boldsymbol{x})\right]\,\mathrm{d}\boldsymbol{x}\right\},$$

by an equivalent series of steps we can also derive a bound using the reversed form of the KL divergence from the target to the base distribution

$$\log \mathcal{Z}(\beta) \geq \beta\left(\log Z - \log\zeta\right) - (1-\beta)D_{\mathrm{KL}}\left[\exp(-\phi)/Z\,\|\,\exp(-\psi)\right].$$

B Estimating the target distribution normalising constant Z

We have that for some unknown normaliser C the marginal density on u is

$$\mathbb{p}\left[\mathsf{u}=u\right]=\frac{1}{C}\zeta^{-\beta(u)}\int_{\mathcal{X}}\exp\left\{-\beta(u)\phi(\boldsymbol{x})-\left[1-\beta(u)\right]\psi(\boldsymbol{x})\right\}\,\mathrm{d}\boldsymbol{x}.$$

Defining $\mathcal{U}_1=\left\{u:|u|\leq\theta_1\right\}$ we have that by construction $\beta(u)=1\;\forall u\in\mathcal{U}_1$ and so

$$\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_1\right]=\int_{\mathcal{U}_1}\mathbb{p}\left[\mathsf{u}=u\right]\,\mathrm{d}u=\int_{\mathcal{U}_1}\frac{1}{C}\zeta^{-1}\int_{\mathcal{X}}\exp\left\{-\phi(\boldsymbol{x})\right\}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}u=\frac{Z}{C\zeta}\int_{\mathcal{U}_1}\mathrm{d}u=\frac{2\theta_1 Z}{C\zeta}.$$

Likewise defining $\mathcal{U}_2=\left\{u:\theta_2\leq|u|\leq 1\right\}$ we have that $\beta(u)=0\;\forall u\in\mathcal{U}_2$ and so

$$\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_2\right]=\int_{\mathcal{U}_2}\mathbb{p}\left[\mathsf{u}=u\right]\,\mathrm{d}u=\int_{\mathcal{U}_2}\frac{1}{C}\zeta^{-0}\int_{\mathcal{X}}\exp\left\{-\psi(\boldsymbol{x})\right\}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}u=\frac{1}{C}\int_{\mathcal{U}_2}\mathrm{d}u=\frac{2(1-\theta_2)}{C}.$$

Taking a ratio of these two probabilities gives that

$$\frac{\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_1\right]}{\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_2\right]}=\frac{\theta_1 Z}{(1-\theta_2)\zeta}\quad\Rightarrow\quad Z=\frac{1-\theta_2}{\theta_1}\frac{\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_1\right]}{\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_2\right]}\zeta.$$

If we construct a Markov chain on u which leaves $\mathbb{p}\left[\mathsf{u}=u\right]$ invariant, then a set of samples from the chain $\left\{u^{(s)}\right\}_{s=1}^{S}$ can be used to form consistent estimators for $\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_1\right]$ and $\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_2\right]$

$$\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_1\right]=\lim_{S\to\infty}\frac{1}{S}\sum_{s=1}^{S}\left\{\mathbb{1}\left[u^{(s)}\in\mathcal{U}_1\right]\right\}\quad\text{and}\quad\mathbb{P}\left[\mathsf{u}\in\mathcal{U}_2\right]=\lim_{S\to\infty}\frac{1}{S}\sum_{s=1}^{S}\left\{\mathbb{1}\left[u^{(s)}\in\mathcal{U}_2\right]\right\},$$

where $\mathbb{1}\left[\cdot\right]$ is the indicator function on some predicate, from which we can then form a consistent estimator for Z

$$Z=\lim_{S\to\infty}\frac{1-\theta_2}{\theta_1}\frac{\sum_{s=1}^{S}\left\{\mathbb{1}\left[u^{(s)}\in\mathcal{U}_1\right]\right\}}{\sum_{s=1}^{S}\left\{\mathbb{1}\left[u^{(s)}\in\mathcal{U}_2\right]\right\}}\zeta.$$

C Temperature control function

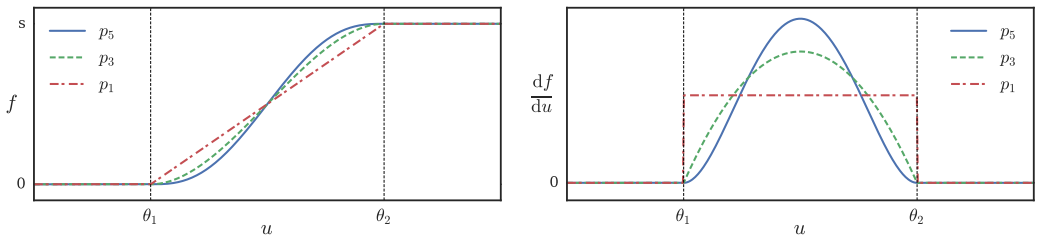


Figure 3: Temperature control functions. Left: Temperature control function $f(u)$ using three different polynomial interpolations for $\theta_1<u<\theta_2$ of order one, three and five. Right: Corresponding gradient of f with respect to u for each of the three polynomial interpolant orders shown in left panel.

The effective inverse temperature $\beta(u)=1-f(u)$ is controlled via a *temperature control function* $f(u)$. Following the same approach as [11] this is piecewise defined as

$$f(u)=\begin{cases}0 & :|u|\leq\theta_1\\ s\times p_i\left(\frac{|u|-\theta_1}{\theta_2-\theta_1}\right) & :\theta_1<|u|<\theta_2\\ s & :|u|\geq\theta_2\end{cases}\tag{13}$$

where $0 < \theta_1 < \theta_2$, $0 \leq s \leq 1$ and p_i is an interpolating polynomial with $p_i(0) = 0$ and $p_i(1) = 1$. One possible choice is simply the linear function $p_1(x) = x$ however this leads to a discontinuous $\frac{\mathrm{d}f}{\mathrm{d}u}$ gradient. A cubic polynomial $p_3(x) = 3x^2 - 2x^3$ as used in [11] leads to continuous f and $\frac{\mathrm{d}f}{\mathrm{d}u}$. If continuity in $\frac{\mathrm{d}^2 f}{\mathrm{d}u^2}$ is also desired a quintic $p_5(x) = 6x^5 - 15x^4 + 10x^3$ can be used. Figure 3 shows f and $\frac{\mathrm{d}f}{\mathrm{d}u}$ for all three of these possibilities. For clarity the functions are plotted only for positive u - in all cases the control function is even in u.

D Gaussian mixture Boltzmann machine relaxation

We define a *Boltzmann machine distribution* on a signed binary state $\mathbf{s} \in \{-1, +1\}^{D_B} = \mathcal{S}$ as

$$\mathbb{P}[\mathbf{s} = \boldsymbol{s}] = \frac{1}{Z_B} \exp\left(\frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{s}^{\mathrm{T}}\boldsymbol{b}\right) \qquad Z_B = \sum_{\boldsymbol{s}\in\mathcal{S}} \left\{\exp\left(\frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{s}^{\mathrm{T}}\boldsymbol{b}\right)\right\}.$$

We introduce an auxiliary real-valued vector random variable $\mathbf{x} \in \mathbb{R}^D$ with conditional distribution

$$\mathbb{p}[\mathbf{x} = \boldsymbol{x} \mid \mathbf{s} = \boldsymbol{s}] = \frac{1}{(2\pi)^{D/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{s}\right)^{\mathrm{T}}\left(\boldsymbol{x} - \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{s}\right)\right]$$

with $\boldsymbol{Q}$ a $D_B \times D$ matrix such that $\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{W} + \boldsymbol{D}$ for some diagonal $\boldsymbol{D}$ which makes $\boldsymbol{W} + \boldsymbol{D}$ positive semi-definite. In our experiments we set $\boldsymbol{D}$ as the solution to the semi-definite programme

$$\min_{\boldsymbol{D}} \left\{\lambda_{\mathrm{MAX}}[\boldsymbol{W} + \boldsymbol{D}]\right\} : \boldsymbol{W} + \boldsymbol{D} \succeq 0 \tag{14}$$

where λ_{MAX} denote the maximal eigenvalue. In general the optimised $\boldsymbol{W} + \boldsymbol{D}$ lies on the semi-definite cone and so has rank less than D_B hence we have $D < D_B$.

The resulting joint distribution on $(\mathbf{x}, \mathbf{s})$ is

$$\begin{aligned}
\mathbb{p}[\mathbf{x} = \boldsymbol{x}, \mathbf{s} = \boldsymbol{s}] &= \frac{1}{(2\pi)^{D/2} Z_B} \exp\left[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{s}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x} - \frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{s} + \frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{s} + \boldsymbol{s}^{\mathrm{T}}\boldsymbol{b}\right] \\
&= \frac{1}{(2\pi)^{D/2} Z_B} \exp\left[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} + \boldsymbol{s}^{\mathrm{T}}(\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}) - \frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{s}\right] \\
&= \frac{1}{(2\pi)^{D/2} Z_B \exp\left(\frac{1}{2}\mathrm{Tr}[\boldsymbol{D}]\right)} \exp\left[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right] \prod_{i=1}^{D_B} \left\{\exp\left[s_i\left(\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{x} + b_i\right)\right]\right\}.
\end{aligned}$$

where $\left\{\boldsymbol{q}_i^{\mathrm{T}}\right\}_{i=1}^{D_B}$ are the D_B rows of $\boldsymbol{Q}$.

We can marginalise over the binary state $\mathbf{s}$ as each $\mathbf{s}_i$ is independent of all the others given $\mathbf{x}$ in the joint distribution. This gives the *Boltzmann machine relaxation* density on $\mathbf{x}$

$$\mathbb{p}[\mathbf{x} = \boldsymbol{x}] = \frac{2^{D_B}}{(2\pi)^{D/2} Z_B \exp\left(\frac{1}{2}\mathrm{Tr}[\boldsymbol{D}]\right)} \exp\left[-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right] \prod_{i=1}^{D_B} \left\{\cosh\left[\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{x} + b_i\right]\right\}$$

which is a specially structured Gaussian mixture density with 2^{D_B} components.

If we define $\mathbb{p}[\mathbf{x} = \boldsymbol{x}] = \frac{1}{Z}\exp[-\phi(\boldsymbol{x})]$ with

$$\phi(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} - \sum_{i=1}^{D_B} \left\{\log\cosh\left[\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{x} + b_i\right]\right\}$$

then the normalisation constant Z of the relaxation density can be related to the normalising constant of the corresponding Boltzmann machine distribution by

$$\log Z = \log Z_B + \frac{1}{2}\mathrm{Tr}[\boldsymbol{D}] + \frac{D}{2}\log(2\pi) - D_B \log 2.$$

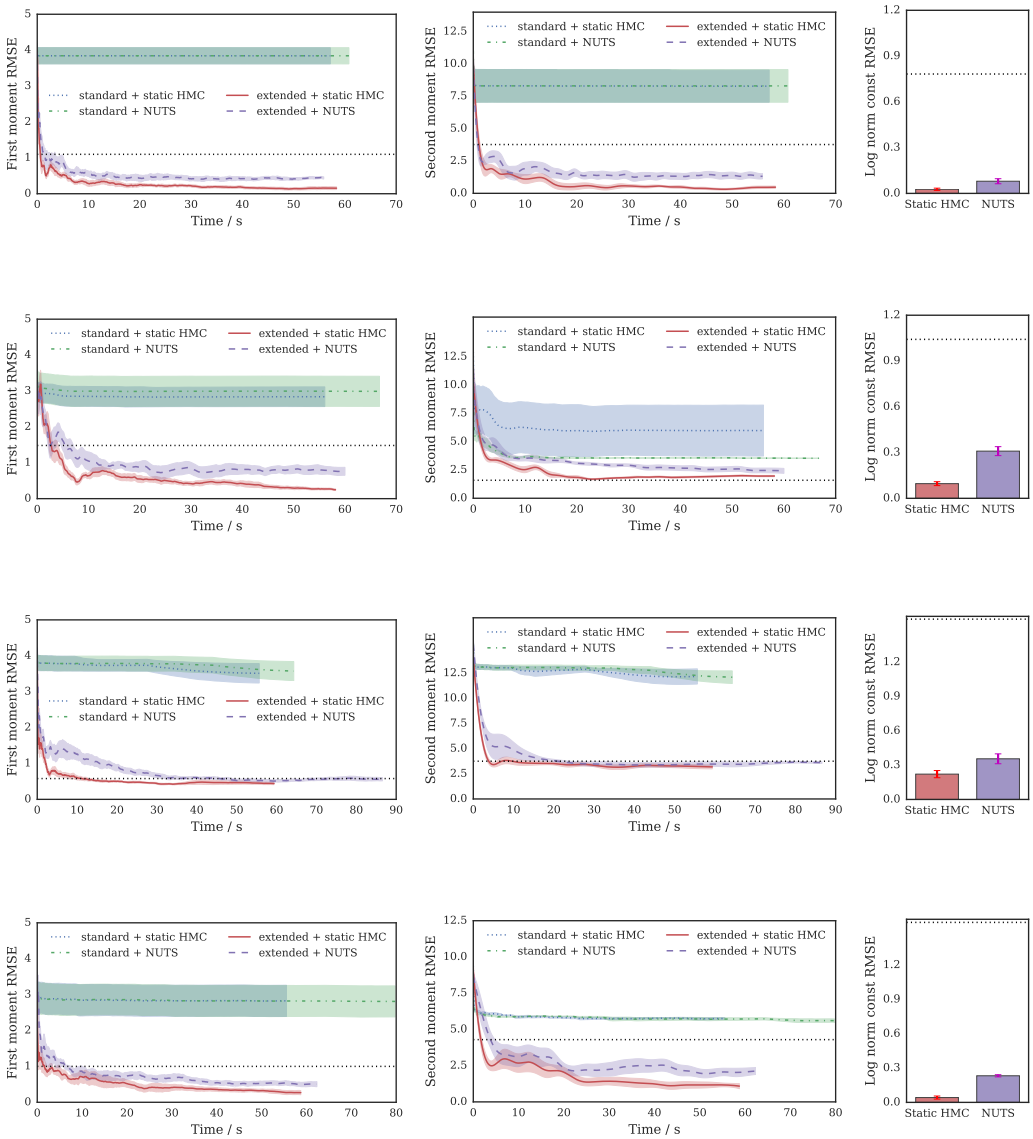
It can also be shown that the first and second moments of the relaxation distribution are related to the first and second moments of the corresponding Boltzmann machine distribution by

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{Q}^{\mathrm{T}}\mathbb{E}[\mathbf{s}] \quad \text{and} \quad \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathrm{T}}\right] = \boldsymbol{Q}^{\mathrm{T}}\mathbb{E}\left[\mathbf{s}\mathbf{s}^{\mathrm{T}}\right]\boldsymbol{Q} + \boldsymbol{I}.$$

E Additional Boltzmann machine relaxation results

The figures below show convergence plots equivalent to those presented in the *Experiments* section of the main paper for additional random Gaussian mixture Boltzmann machine relaxation instances (the first row is a replication of those results for comparison). The weight and bias parameters of the associated Boltzmann machine distribution were generated with exactly the same process as for the test case presented in the paper just with a different random seed.

For every target density 8 independent chains were run for each of: non-adaptive HMC in the original Hamiltonian system, NUTS in the original Hamiltonian system, non-adaptive HMC in the extended Hamiltonian system and NUTS in the extended Hamiltonian system. For each set of parameters a set of local mean field approximations are fitted to the corresponding target density and used to calculate moments for the Gaussian base density. The error between these base density moments (and log normalisation constant) and true values are indicated by dotted black lines in the plots below as a reference for the MCMC convergence.



F Stan model files for Boltzmann machine relaxation experiments

Standard Hamiltonian system with no temperature augmentation:

```
functions {
  // Vectorised log hyperbolic cosine helper.
  vector log_cosh(vector y){
    return y + log(1 + exp(-2 * y)) - log(2);
  }
  // Log probability density of Boltzmann machine relaxation.
  real bm_relaxation_lpdf(vector x, matrix q, vector b){
    return sum(log_cosh(q * x + b)) - 0.5 * x' * x;
  }
}

data {
  // Number of dimension in Boltzmann machine binary state.
  int<lower=0> n_dim_b;
  // Number of dimensions in relaxation configuration state.
  int<lower=0> n_dim_r;
  // Relaxation Q matrix parameters.
  matrix[n_dim_b, n_dim_r] q;
  // Relaxation bias vector parameters.
  vector[n_dim_b] b;
}

parameters {
  // Configuration state.
  vector[n_dim_r] x;
}

model {
  // Set to target to Boltzmann machine relaxation log density.
  x ~ bm_relaxation(q, b);
}
```

Extended Hamiltonian system with temperature control variable:

```
functions {
  // Vectorised log hyperbolic cosine helper.
  vector log_cosh(vector y){
    return y + log(1 + exp(-2 * y)) - log(2);
  }
  // Log probability density of Boltzmann machine relaxation.
  real bm_relaxation_lpdf(vector x, matrix q, vector b){
    return sum(log_cosh(q * x + b)) - 0.5 * x' * x;
  }
  // Circularly wraps unbounded input to [-1, 1].
  real wrap(real u) {
    return fmod(u + 1, 2) + 2 * (u < -1) - 1;
  }
  // Piecewise defined inverse temperature control function.
  real inv_temp(real u, real theta_1, real theta_2) {
    real z;
    z = (fabs(u) - theta_1) / (theta_2 - theta_1);
    if (z <= 0)
      return 1;
    else if (z >= 1)
      return 0;
    else
      return 1 - z^3 * (z * (6 * z - 15) + 10);
  }
}
```

```
data {
  // Number of dimension in Boltzmann machine binary state.
  int<lower=0> n_dim_b;
  // Number of dimensions in relaxation configuration state.
  int<lower=0> n_dim_r;
  // Relaxation Q matrix parameters.
  matrix[n_dim_b, n_dim_r] q;
  // Relaxation bias vector parameters.
  vector[n_dim_b] b;
  // Temperature control function lower threshold.
  real theta_1;
  // Temperature control function upper threshold.
  real theta_2;
  // Target log normalisation constant approximation.
  real log_zeta;
  // Covariance matrix of Gaussian approximation to target.
  matrix[n_dim_r, n_dim_r] sigma;
  // Mean vector of Gaussian approximation to target.
  vector[n_dim_r] mu;
}

transformed data {
  // Approximate covariance Cholesky factor.
  matrix[n_dim_r, n_dim_r] chol_sigma;
  chol_sigma = cholesky_decompose(sigma);
}

parameters {
  // Configuration state.
  vector[n_dim_r] x;
  // Unwrapped temperature control variable.
  real u_unwrapped;
}

transformed parameters {
  // Temperature control variable wrapped to [-1, 1].
  real<lower=-1, upper=1> u;
  // Inverse temperature.
  real<lower=0, upper=1> beta;
  u = wrap(u_unwrapped);
  beta = inv_temp(u, theta_1, theta_2);
}

model {
  // Inverse temperature weighted target density term.
  target += beta * bm_relaxation_lpdf(x | q, b) - beta * log_zeta;
  // Inverse temperature weighted base density term.
  target += (1 - beta) * multi_normal_cholesky_lpdf(x | mu, chol_sigma);
}
```