4th Symposium on Advances in Approximate Bayesian Inference, 2022 1–9

Metropolis Augmented Hamiltonian Monte Carlo

Guangyao Zhou STANNIS@VICARIOUS.COM
Vicarious AI

Abstract

Hamiltonian Monte Carlo (HMC) is a powerful Markov Chain Monte Carlo (MCMC) method for sampling from complex high-dimensional continuous distributions. However, in many situations it is necessary or desirable to combine HMC with other Metropolis-Hastings (MH) samplers. The common HMC-within-Gibbs strategy implies a trade-off between long HMC trajectories and more frequent other MH updates. Addressing this trade-off has been the focus of several recent works. In this paper we propose Metropolis Augmented Hamiltonian Monte Carlo (MAHMC), an HMC variant that allows MH updates within HMC and eliminates this trade-off. Experiments on two representative examples demonstrate MAHMC's efficiency and ease of use when compared with within-Gibbs alternatives.

1. Introduction

Hamiltonian Monte Carlo (HMC) is a popular Markov Chain Monte Carlo (MCMC) method. It samples from complex high-dimensional distributions $\pi(q) \propto \exp(-U(q))$ on continuous variables $q \in \mathcal{R}^n$, where $U(q) : \mathcal{R}^n \to \mathcal{R}$ is commonly referred to as the potential energy. HMC has enjoyed remarkable empirical success, due to the use of powerful symplectic integrators (Leimkuhler and Reich, 2004) (e.g. the leapfrog integrator) to maintain high acceptance probabilities for long-range gradients-guided proposals (Neal, 2012b; Betancourt, 2017). Common HMC implementations involve simulating the Hamiltonian dynamics for multiple leapfrog steps, followed by a Metropolis Hastings (MH) correction. Chen et al. (2020) recently rigorously establishes the importance of using multiple leapfrog steps/long trajectories for HMC's efficiency, especially when compared with the Metropolis Adjusted Langevin Algorithm (MALA) (Dwivedi et al., 2019), a special case of HMC using only one leapfrog step and a widely used algorithm in Bayesian statistics and machine learning.

However, we often face distributions of the form $\pi(q^H, q^O) \propto \exp(-U(q^H, q^O))$, where we can only use HMC for the continuous variables $q^H \in \mathcal{R}^n$, and it is necessary or desirable to use some other MH samplers for the variables q^O (Sec. 4.3 of Neal (2012b)). Some common situations include (1) when q^O are discrete, (2) when we have specialized MH samplers for q^O that are efficient/easy to use, or (3) when q^O are continuous but $\nabla_{q^O} U(q^H, q^O)$ is expensive or impossible to compute. In such cases, we typically adopt an HMC-within-Gibbs strategy (Neal, 2012a; Dang et al., 2019; Kelly et al., 2021), where we alternate between HMC updates and other MH updates. However, this implies a trade-off between long HMC trajectories and more frequent other MH updates. Longer HMC trajectories can help suppress random walk behavior, but might hurt overall sampling since other MH updates can only be done infrequently (between HMC updates). Shorter HMC trajectories allow more frequent other MH updates, but lead to increased random walk behavior.

© G. Zhou.

Algorithm 1: Basic components for sampling from $\pi(q) \propto \exp(-U(q)), q \in \mathcal{R}^n$

```
def leapfrog (q, p, ϵ|U)
    p ← p − (ϵ/2)∇U(q); q ← q + ϵp; p ← p − (ϵ/2)∇U(q);
    return q, p;
end
def MH_correction (q_0, p_0, q, p, v|U, K)
    E_0 ← U(q_0) + K(p_0); E ← U(q) + K(q);
    if |v| ≤ exp(−E + E_0) then
        v ← v exp(−E_0 + E)
    else q, p ← q_0, p_0;
    return q, p, v;
end
def multiple_leapfrogs (q, p, ϵ, L|U)
    for 1 to n do  q, p ← leapfrog (q, p, ϵ|U) ;
    p ← −p;
    return q, p;
end
```

Several recent works have focused on the above trade-off. Neal (2020) combines MALA with partial momentum refreshment (Horowitz, 1991) and non-reversible Metropolis accept/reject decisions to allow more frequent other MH updates while suppressing random walk behavior. Zhou (2020, 2021) proposes mixed HMC (M-HMC) for distributions with mixed discrete and continuous variables to allow making discrete MH updates as part of an HMC trajectory. In this paper, we propose Metropolis Augmented Hamiltonian Monte Carlo (MAHMC) to completely eliminate this trade-off. MAHMC interprets HMC as a deterministic MH proposal, and augments HMC by introducing MH updates as part of the HMC trajectory. MAHMC is generally applicable, can be trivially implemented on top of HMC with no additional overhead, and includes M-HMC with Laplace momentum as a special case. We demonstrate MAHMC's advantage over within-Gibbs alternatives (Neal, 2020) and ease of use on two representative examples in Sec. 4.

2. Background

2.1. HMC

HMC introduces auxiliary momentum $p \in \mathcal{R}^n$ associated with kinetic energy $K(p) = \sum_{i=1}^n p_i^2/2$, and simulates the Hamiltonian dynamics for L steps of size ϵ using the leapfrog integrator (*leapfrog* in Algo. 1), before making the final MH correction (*HMC* in Algo. 2). We observe that we can in fact interpret the multiple leapfrog steps (*multiple_leapfrogs* in Algo. 1) as a deterministic MH proposal, and derive the final MH correction as the usual MH acceptance probability for the deterministic proposal (Sec. 5.2 of Betancourt (2017)).

2.2. MALA and MALA variants

Using $L = 1$ leapfrog step in HMC results in the widely used special case MALA. MALA uses gradients information, and allows more frequent other MH updates, but suffers from random

Algorithm 2: Algorithms for sampling from $\pi(q) \propto \exp(-U(q)), q \in \mathcal{R}^n$

```
def HMC (q_0, ε, L|U, K)                                          // Sec. 2.1
    p_0 ~ N(0, I_n); q, p ← q_0, p_0;
    q, p ← multiple_leapfrogs (q, p, ε, L|U);                     // Defined in Algo. 1
    q, p, v ← MH_correction (q_0, p_0, q, p, Uniform(0,1)|U, K);  // Defined in Algo. 1
    return q;
end
def MALA-P (q_0, p_0, ε, α|U, K)                                  // Sec. 2.2
    n ~ N(0,_n ); p_0 ← αp_0 + √(1 − α^2) n; q, p ← q_0, p_0;
    q, p ← multiple_leapfrogs (q, p, ε, 1|U);                     // Defined in Algo. 1
    q, p, v ← MH_correction (q_0, p_0, q, p, Uniform(0,1)|U, K);  // Defined in Algo. 1
    return q, −p;
end
def MALA-PN (q_0, p_0, v, ε, α, δ|U, K)                           // Sec. 2.2
    n ~ N(0, I_n); p_0 ← αp_0 + √(1 − α^2) n; q, p ← q_0, p_0;
    q, p ← multiple_leapfrogs (q, p, ε, 1|U);                     // Defined in Algo. 1
    q, p, v ← MH_correction (q_0, p_0, q, p, v|U, K);             // Defined in Algo. 1
    v ← (v + 1 + δ) mod 2 − 1;
    return q, −p, v;
end
```

walk behavior due to the use of short trajectories with frequent momentum refreshments. Partial momentum refreshment (Horowitz, 1991) (MALA-P in Algo. 2) was proposed as a possible remedy. However, as Neal (2020) explains, a rejection would lead MALA-P to almost double back on itself, making it less efficient than HMC with long trajectories.

Neal (2020) proposes a non-reversible scheme for Metropolis accept/reject decisions (MALA-PN in Algo. 2) to maintain the ability to make frequent MH updates while further suppressing the random walk behavior that comes from MALA-P doubling back on itself due to rejections. MALA-PN produces long rejection-free runs by encouraging rejections to cluster together, and demonstrates improved performance on multiple problems.

2.3. Within-Gibbs sampling from $\pi(q^H, q^O) \propto \exp(-U(q^H, q^O)), q^H \in \mathcal{R}^n$

For distributions $\pi(q^H, q^O) \propto \exp(-U(q^H, q^O))$, $q^H \in \mathcal{R}^n$ where we want to use MH updates for q^O, we refer to the common strategy of alternating between updates of $q^H \in \mathcal{R}^n$ (using a sampler for continuous distributions) and MH updates of the other variables q^O as within-Gibbs sampling. In our experiments, we use four types of within-Gibbs samplers as baselines: (1) MALA within Gibbs (MALAwG), (2) HMC within Gibbs (HwG), (3) MALA-P within Gibbs (MALA-PwG), and (4) MALA-PN within Gibbs (MALA-PNwG).

2.4. M-HMC for distributions with mixed discrete and continuous variables

Zhou (2020, 2021) proposes M-HMC, an HMC variant that evolves the discrete and continuous variables in tandem for distributions with mixed support. M-HMC naturally supports frequent MH updates within long HMC trajectories, and demonstrates improved perfor-

Algorithm 3: MAHMC. Blue highlights changes on top of naive MH within HMC.

def `MAHMC` $(q_0^H, p_0^H, q_0^O, \epsilon, L | U, K, \mathbb{Q}_i, i = 1, \ldots, N^O, \mathbb{P}^{\mathcal{D}})$ `// Sec. 3.1`
  $q^H, p^H, q^O \leftarrow q_0^H, p_0^H, q_0^O$; $D \sim \mathbb{P}^{\mathcal{D}}(\cdot)$; $\Delta E \leftarrow 0$;
  for $j \leftarrow 1$ **to** L **do**
    if $D_j = 0$ **then**
      $q^H, p^H \leftarrow$ `leapfrog`$(q^H, p^H, \epsilon | U(\cdot, q^O))$; `// Defined in Algo. 1`
    else
      $\tilde{q}^O \sim \mathbb{Q}_{D_j}(\cdot | q^H, q^O)$;
      if $Uniform(0,1) \leq \frac{\exp(-U(q^H, \tilde{q}^O))\mathbb{Q}_{D_j}(q^O | q^H, \tilde{q}^O)}{\exp(-U(q^H, q^O))\mathbb{Q}_{D_j}(\tilde{q}^O | q^H, q^O)}$ **then**
        $q^O \leftarrow \tilde{q}^O$; $\Delta E \leftarrow \Delta E + U(q^H, \tilde{q}^O) - U(q^H, q^O)$;
  end
  $E \leftarrow U(q^H, q^O) + K(p^H)$; $E_0 \leftarrow U(q_0^H, q_0^O) + K(p_0^H)$;
  if $Uniform(0,1) \leq \frac{\exp(-E)}{\exp(-E_0)} \frac{\mathbb{P}^{\mathcal{D}}(D^{-1})\exp(\Delta E)}{\mathbb{P}^{\mathcal{D}}(D)}$ **then** $p^H \leftarrow -p^H$;
  else $q^H, p^H, q^O \leftarrow q_0^H, p_0^H, q_0^O$;
  return q^H, p^H, q^O
end

mance over strong baselines. In Sec. 3.2, we show that the practically useful M-HMC implementation (with Laplace momentum) can be seen as a special case of MAHMC.

3. Metropolis Augmented Hamiltonian Monte Carlo (MAHMC)

3.1. Augmenting HMC with MH updates

Motivated by the interpretation of HMC as a deterministic MH proposal (Sec. 2.1), for a given distribution $\pi(q^H, q^O) \propto \exp(-U(q^H, q^O))$, MAHMC allows more frequent MH updates for q^O by combining leapfrog steps for q^H and MH updates for q^O (including the MH correction) into a single MH proposal, followed by an additional final MH correction.

Formally, for some given step size ϵ and number of steps L, an MAHMC iteration makes use of the leapfrog integrator $leapfrog(q, p, \epsilon | U(\cdot, q^O))$ (Algo. 1) and N^O MH proposals $\mathbb{Q}_i(\tilde{q}^O | q^H, q^O), i = 1, \ldots, N^O$ to construct an MH proposal making L total updates of q^H, q^O. For a sequence $D \in \left\{0, 1, \ldots, N^O\right\}^L$ of L integers, define $D^{-1} = (D_L, D_{L-1}, \ldots, D_1)$. Starting from q_0^H, q_0^O, MAHMC first resamples the momentum p_0^H from $N(0, I_n)$, then samples a sequence of L variable updates represented as a sequence D of L integers from some distribution $\mathbb{P}^{\mathcal{D}}(D)$, applies the variable updates one at a time, before making a final MH correction. See Algo. 3 for a detailed description of an MAHMC iteration.

Critically, the use of MH correction in each MH update serves as a mechanism to prevent MAHMC from deviating too much into low-probability regions. As we empirically verify in Sec. 4, even with the additional MH updates as part of the trajectory, MAHMC can maintain high acceptance probabilities for long-range proposals, similar to HMC. This eliminates the need to balance long HMC trajectories and more frequent other MH updates, and contributes to MAHMC's improved performance over other within-Gibbs alternatives.

3.2. Connections to M-HMC

Zhou (2020, 2021) derives M-HMC using auxiliary Hamiltonian dynamics for the discrete variables. We note that the practically useful M-HMC implementation using Laplace momentum is in fact a variant of a special case of MAHMC, where $\mathbb{P}^{\mathcal{D}}$ is implicitly defined using the auxiliary Hamiltonian dynamics and the proposals $\mathbb{Q}_i$ are for single discrete variables (given all other discrete and continuous variables). However, M-HMC differs from MAHMC in its use of a persistent kinetic energy $k^{\mathcal{D}}$ for the MH corrections in all MH updates.

3.3. MAHMC in Algo. 3 satisfies detailed balance with respect to $\pi(q^H, q^O)$

Proof Sketch To establish detailed balance, we make use of the concept of *probabilistic paths*, similar to Sec. 2.3 in Zhou (2020, 2021). Starting from $s = (q_0^H, p_0^H, q_0^O)$, a *probabilistic path* $\boldsymbol{t}$ contains information about all randomness in an MAHMC iteration (Algo. 3):

1. The sequence D of L integers specifying which updates to use at each of the L steps.
2. The actual states $q_j^H, p_j^H, q_j^O, j = 1, \ldots, L$ (before final MH correction) at each step.
3. The sequence of proposed states $\tilde{q}_j^H, \tilde{p}_j^H, \tilde{q}_j^O, j = 1, \ldots, L$ at each step.
   - If $D_j = 0$, $\tilde{q}_j^H, \tilde{p}_j^H = \textit{leapfrog}(q_j^H, p_j^H, \epsilon | U(\cdot, q_j^O)), \tilde{q}_j^O = q_{j-1}^O$.
   - If $D_j > 0$, $\tilde{q}_j^H, \tilde{p}_j^H = q_j^H, p_j^H$, $\tilde{q}_j^O$ is a sample from $\mathbb{Q}_{D_j}(\cdot | q_j^H, q_j^O)$.
4. The sequence of accept/reject decisions $a_j \in \{\texttt{True}, \texttt{False}\}, j = 1, \ldots, L$ at each step. Note that we always accept ($a_j = \texttt{True}$) for leapfrog updates ($D_j = 0$).

We can think of an MAHMC iteration as first sampling a probabilistic path $\boldsymbol{t}$ which brings s_0 to s_L, where $s_j = (q_j^H, p_j^H, q_j^O), j = 1, \ldots, L$, before making an MH correction to either accept s_L or reject and return to s_0. We can *reverse* a probabilistic path $\boldsymbol{t}$ to get probabilistic path $\boldsymbol{t}^{-1}$ which brings s_L back to s_0. $\boldsymbol{t}^{-1}$ uses the sequence of L updates specified by D^{-1}, and reverses the sequence of actual and proposed states (with proper momentum negation) as well as the sequence of accept/rejection decisions. Denote by $\mathbb{P}(\boldsymbol{t}|s_0)$ the probability of sampling $\boldsymbol{t}$ starting from s_0, and $\mathbb{P}(s|\boldsymbol{t}), s \in \{s_0, s_L\}$ the MH correction step in MAHMC, we can derive the transition probability of MAHMC as $\mathbb{P}(s'|s) = \sum_{\boldsymbol{t}:s\to s'} \mathbb{P}(s'|\boldsymbol{t})\mathbb{P}(\boldsymbol{t}|s)$. Define $E(s) = U(q^H, q^O) + K(p^H)$. We can establish the desired detailed balance if we can prove $\exp(-E(s))\mathbb{P}(s'|s) = \exp(-E(s'))\mathbb{P}(s|s')$.

We show that $\mathbb{P}(s_L|\boldsymbol{t}) = \min\left\{1, \frac{\exp(-E(s_L))\mathbb{P}(\boldsymbol{t}^{-1}|s_L)}{\exp(-E(s_0))\mathbb{P}(\boldsymbol{t}|s_0)}\right\}$ is our desired MH acceptance probability. Note that $\exp(-E(s))\mathbb{P}(s|s) = \exp(-E(s))\mathbb{P}(s|s)$ is trivially true. For $s' \neq s$,

$$
\begin{aligned}
\exp(-E(s))\mathbb{P}(s'|s) &= \exp(-E(s)) \sum_{\boldsymbol{t}:s\to s'} \mathbb{P}(s'|\boldsymbol{t})\mathbb{P}(\boldsymbol{t}|s) \\
&= \exp(-E(s)) \sum_{\boldsymbol{t}:s\to s'} \min\left\{1, \frac{\exp(-E(s'))\mathbb{P}(\boldsymbol{t}^{-1}|s')}{\exp(-E(s))\mathbb{P}(\boldsymbol{t}|s)}\right\} \mathbb{P}(\boldsymbol{t}|s) \\
&= \sum_{\boldsymbol{t}:s\to s'} \min\left\{\exp(-E(s))\mathbb{P}(\boldsymbol{t}|s), \exp(-E(s'))\mathbb{P}(\boldsymbol{t}^{-1}|s')\right\} \\
&= \sum_{\boldsymbol{t}:s'\to s} \min\left\{\exp(-E(s))\mathbb{P}(\boldsymbol{t}^{-1}|s), \exp(-E(s'))\mathbb{P}(\boldsymbol{t}|s')\right\} \\
&= \exp(-E(s'))\mathbb{P}(s|s')
\end{aligned}
$$

For $\frac{\mathbb{P}(\boldsymbol{t}^{-1}|s_L)}{\mathbb{P}(\boldsymbol{t}|s_0)}$, sampling D contributes $\frac{\mathbb{P}^{\mathcal{D}}(D^{-1})}{\mathbb{P}^{\mathcal{D}}(D)}$. Leapfrog updates have no randomness and contribute nothing. For $D_j > 0$, define $p_j = \frac{\exp(-U(\tilde{q}_j^H,\tilde{q}_j^O))\mathbb{Q}_{D_j}(q_j^O|q_j^H,\tilde{q}_j^O)}{\exp(-U(q_j^H,q_j^O))\mathbb{Q}_{D_j}(\tilde{q}_j^O|q_j^H,q_j^O)}$. If $a_j = \texttt{True}$, the MH update contributes $\frac{\mathbb{Q}_{D_j}(q_j^O|q_j^H,\tilde{q}_j^O)\min\{1,1/p_j\}}{\mathbb{Q}_{D_j}(\tilde{q}_j^O|q_j^H,q_j^O)\min\{1,p_j\}} = \frac{\exp(-U(q_j^H,q_j^O))}{\exp(-U(q_j^H,\tilde{q}_j^O))}$, which is captured in the ΔE updates in Algo. 3. If $a_j = \texttt{False}$, the MH update again contributes nothing (as in both $\boldsymbol{t}$ and $\boldsymbol{t}^{-1}$ we make the same proposal followed by a rejection). This proves Algo. 3 uses the correct MH acceptance probability, and establishes the desired detailed balance. ■

Table 1: Results on MDC (Sec. 4.1) and BLR (Sec. 4.2). We show ESS per sample per gradients evaluation of u for MDC, and of the potential energy of τ, β for BLR.

	MALAwG	HMCwG	MALA-PwG	MALA-PNwG	MAHMCwG
MDC	1.0×10^{-4}	4.62×10^{-3}	1.82×10^{-3}	7.38×10^{-3}	$\mathbf{1.78 \times 10^{-2}}$
BLR	1.73×10^{-3}	$7,94 \times 10^{-3}$	6.66×10^{-3}	8.86×10^{-3}	$\mathbf{9.02 \times 10^{-3}}$

4. Experiments

We use the 4 within-Gibbs samplers in Sec. 2.3 as baselines, and follow the setups of Neal (2020) when possible. For MALA-based samplers, we alternate between N^L leapfrog updates for q^H and 1 Gibbs update for q^O. For HwG, we alternate between HMC iterations and Gibbs updates. Although MAHMC can evolve q^H, q^O in tandem, to make the comparison most informative, we consider an MAHMC within Gibbs (MAHMCwG) sampler, where we make $N^U - 1$ Gibbs updates and $N^U N^L$ leapfrog updates within each MAHMC iteration, followed by a Gibbs update. The MAHMC iteration schedules the Gibbs and leapfrog updates similarly to MALA-based samplers, alternating between N^L leapfrog updates and 1 Gibbs update. This leads to a deterministic $\mathbb{P}^{\mathcal{D}}$ that always proposes the same (symmetric) sequence of updates. Note that the use of a deterministic $\mathbb{P}^{\mathcal{D}}$ makes comparison with Neal (2020) straightforward, but in general we have more flexibilities in picking $\mathbb{P}^{\mathcal{D}}$. A simple choice is to randomly pick the update to make at each step with a fixed distribution over the different kinds of available updates (i.e. leapfrog and other updates). The fixed distribution allows us to control the relative frequency of the different updates, and the contribution of $\mathbb{P}^{\mathcal{D}}$ to the final acceptance probability can be easily calculated.

We assume gradients evaluation dominates the computation (Neal, 2020), and evaluate performance using effective sample size (ESS) (calculated with Kumar et al. (2019)) per sample per gradients evaluation. Code to reproduce the results is available at https://github.com/StannisZhou/mahmc.

4.1. A distribution with mixed discrete and continuous variables (MDC)

We consider the distribution in Sec. 5 of Neal (2020), where $q^H = (u, v), q^O = (w_1, \ldots, w_{20})$:

$$u \sim N(0,1), v|u \sim N(u, 0.04^2), w_i|u \sim \mathit{Bernoulli}(1/(1+e^u)), i = 1, \ldots, 20$$

Correctness To verify correctness, we compare the histogram of u samples obtained with the 5 samplers, and empirically verify that they match the expected probability density function of $N(0,1)$ and the samplers are sampling from the right distribution.

Efficiency We summarize the results in Tab. 1. From the results, we can see that by combining longer HMC trajectories with more frequent Gibbs updates, MAHMCwG is **3.85x** more efficient than HwG, and **2.4x** more efficient than MALA-PNwG.

For HwG and MALA-PNwG, we use the optimal hyperparameters reported in Sec. 5 of Neal (2020) ($L = 40$ steps and step size $\epsilon = 0.035$ for HwG, and $N^L = 10, \epsilon = 0.03, \alpha = 0.995, \delta = 0.01$ for MALA-PNwG). Our results roughly reproduce the results in Neal (2020) (MALA-PNwG is 1.6x more efficient than HwG, as opposed to 1.83x reporeted in Neal (2020)). The small discrepancy can be due to the use of a different metric (ESS of u instead of ESS of $\mathbb{I}(u \in (-0.5, 1.5))$ as in Neal (2020), where $\mathbb{I}$ represents the indicator function) and a different way to compute ESS (using Kumar et al. (2019)).

To make the comparison informative, we use $N^L = 10$ and $\epsilon = 0.03$ for MALAwG and MALA-PwG, and $\alpha = 0.995$ for MALA-PwG. We observe that partial momentum refreshment and non-reversible Metropolis accept/reject decisions are indeed beneficial in improving the performance of MALAwG/MALA-PwG. However, only MALA-PNwG outperforms HwG.

The optimal performance of MAHMCwG is achieved with $N^U = 10$ and $\epsilon = 0.04$, i.e. in each MAHMC we make $N^U N^L = 100$ leapfrog updates and $N^{U-1} = 9$ Gibbs updates (uniformly spreaded). This is far larger than the optimal number of steps $L = 40$ for HwG, and demonstrates MAHMC's ability to maintain high acceptance probabilities for long-range proposals that includes MH updates. We additionally test MAHMCwG with $N^U = 4$ and $\epsilon = 0.035$, i.e. making $N^U - 1 = 3$ additional Gibbs updates on top of HwG, and observe that the normalized ESS increases to 6.08×10^{-3}. This demonstrates the benefits of more frequent MH updates when we use the same number of leapfrog updates.

4.2. Bayesian Logistic Regression (BLR) with conjugate prior

We apply Bayesian Logistic Regression (BLR) to the breast cancer wisconsin dataset[1], consisting of 569 pairs of 30 dimensional features and targets $y_i \in \{0, 1\}$. We standarize the features and append 1 at the end to get $x_i \in \mathbb{R}^{31}$, and specify BLR with conjugate prior as

$$\tau \sim Gamma(1.0, scale = 100), \beta \sim N(0, \frac{1}{\tau} I_{31}), y_i \sim Bernoulli(sigmoid(x_i^T \beta)), i = 1, \ldots, 569$$

We consider sampling from the posterior $\mathbb{P}(\tau, \beta | X, y)$, where $q^H = \beta$ and $q^O = \tau$, and we make Gibbs updates for τ using $\mathbb{P}(\tau | \beta, X, y)$.

Correctness To verify the correctness of the samplers, we first apply the samplers to only the prior distribution

$$\tau \sim Gamma(1.0, scale = 100), \beta \sim N(0, \frac{1}{\tau} I_{31})$$

and similarly verify the marginal distribution of τ is indeed $Gamma(1.0, scale = 100)$ by looking at the histogram of τ samples. We additionally verify that all samplers give rough the same posterior means for β, and we can classify the 569 training data points to an accuracy of 98.77% using posterior samples from the 5 different samplers.

1. Available on the UCI Machine Learning Repository. Accessed through sklearn.

Efficiency We fix $N^L = 5$, and conduct a grid search in

$$N^U \in \{2, 3, 4, 5\}, \epsilon \in \{0.07, 0.08, 0.09, 0.10, 0.11\}, \alpha \in \{0.9, 0.95, 0.98, 0.99\}$$

and $\delta \in \{0.005, 0.01, 0.015, 0.03\}$, and $L \in \{10, 15, 20, 25\}$ for HwG.

Optimal performance is achieved with $\epsilon = 0.11$ for MALAwG, $L = 10, \epsilon = 0.09$ for HwG, $\epsilon = 0.09, \alpha = 0.9$ for MALA-PwG, $\epsilon = 0.1, \alpha = 0.9, \delta = 0.015$ for MALA-PNwG, and $N^U = 2, \epsilon = 0.1$ for MAHMCwG.

For this example, MAHMCwG only slightly outperforms MALA-PNwG. However, we comment that MAHMCwG is easier to tune, as we essentially only need to tune the step size and number of steps, same as in HwG, and using the same setups from HwG usually already gives good performance. However, for MALA-PwG, we need to tune α and δ, which can have significant impacts on performance. For example, for the distribution in 4.1, grid search is done for $\alpha \in \{0.98, 0.99, 0.995, 0.9975, 0.9985, 0.999\}$ in Neal (2020). However, in our experiments we empirically observe that all these values give poor performance, and we have to reduce α to 0.9 to get performance comparable with MAHMCwG, suggesting the potential challenges in tuning α.

4.3. Additional verification of correctness

Since we only used Gibbs updates (which always accept) in our experiments, we additionally verify the correctness of MAHMC by modifying the script used in Zhou (2020, 2021) to make random walk MH updates within HMC for the 1D GMM example using MAHMC. See https://github.com/StannisZhou/mahmc/blob/main/simplified_mixed_hmc.py for the updated script.

References

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Yuansi Chen, Raaz Dwivedi, Martin J Wainwright, and Bin Yu. Fast mixing of Metropolized Hamiltonian Monte Carlo: Benefits of multi-step gradients. *J. Mach. Learn. Res.*, 21: 92–1, 2020.

Khue-Dung Dang, Matias Quiroz, Robert Kohn, Tran Minh-Ngoc, and Mattias Villani. Hamiltonian Monte Carlo with energy conserving subsampling. *Journal of machine learning research*, 20, 2019.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast. *Journal of Machine Learning Research*, 20:1–42, 2019.

Alan M Horowitz. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2): 247–252, 1991.

Jacob Kelly, Richard Zemel, and Will Grathwohl. Directly training joint energy-based models for conditional synthesis and calibrated prediction of multi-attribute data. *arXiv preprint arXiv:2108.04227*, 2021.

Ravin Kumar, Carroll Colin, Ari Hartikainen, and Osvaldo A. Martin. ArviZ a unified library for exploratory analysis of Bayesian models in Python. *The Journal of Open Source Software*, 2019. doi: 10.21105/joss.01143.

Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian dynamics*. Number 14. Cambridge university press, 2004.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012a.

Radford M Neal. MCMC using Hamiltonian dynamics. *arXiv preprint arXiv:1206.1901*, 2012b.

Radford M Neal. Non-reversibly updating a uniform [0, 1] value for Metropolis accept/reject decisions. *arXiv preprint arXiv:2001.11950*, 2020.

Guangyao Zhou. Mixed Hamiltonian Monte Marlo for Mixed Discrete and Continuous Variables. In *Advances in Neural Information Processing Systems*, volume 33, pages 17094–17104. Curran Associates, Inc., 2020.

Guangyao Zhou. Erratum to "Mixed Hamiltonian Monte Carlo for Mixed Discrete and Continuous Variables". https://stanniszhou.github.io/papers/mixed_hmc_erratum.pdf, 2021.