Non-Volume Preserving Hamiltonian Monte Carlo and No-U-Turn Samplers

In section 1 of the supplementary material, we formalise baseline NUTS algorithm (Hoffman and Gelman, 2014) with more details to facilitate the proofs presented in the subsequent sections. In section 2, we prove Proposition 1 of the main text. That is, we show that if the parameter $\Delta_{\max}$ is set to any value less than ∞, the termination criterion that is based on this parameter interferes with detailed balance. In Section 3, we rigorously prove Theorem 6.1 in the main text. That is, the draws from NoVoP NUTS algorithm converge to the draws from the target distribution. Section 4 provides additional experimental results. Sections 2 to 4 are not a prerequisite to each other and can be read independently.

1 Baseline NUTS

Baseline NUTS algorithm is equivalent to the following steps:

Step 1. Draw a slice variable

Starting from a current state $\mathbf{z}^{(t)}$ with probability $\pi_{\mathbf{Z}}\left(\mathbf{z}^{(t)}\right) := \exp(-H(\mathbf{z}^{(t)}))$, an auxiliary *slice variable*, u, is drawn from:

$$u \sim \text{Uniform}(0, \pi_{\mathbf{Z}}(\mathbf{z}^{(t)})) \tag{1}$$

Step 2. Construct a traced set

An ordered set of states, namely *traced set* (or *traced path*) $\mathcal{B}$, is generated that includes $\mathbf{z}^{(t)}$ and should satisfy the following condition (known as *balance tree property*), required for detailed balance:

$$P(\mathcal{B}|\mathbf{z}) = P(\mathcal{B}|\mathbf{z}') \qquad \forall \mathbf{z}, \mathbf{z}' \in \mathcal{B} \tag{2}$$

That is, the probability of generating $\mathcal{B}$ starting from any of its members should be the same.

$\mathcal{B}$ is generated iteratively, by starting from $\mathcal{B}_1 := \{\mathbf{z}^{(t)}\}$ and either expanding it by adding the same number of new states to it, $\mathcal{B}_{i+1} := \mathcal{B}_i \sqcup \mathcal{B}_{i+1}^{\text{new}}$, or terminating and returning it as the final set $\mathcal{B}$. To make $\mathcal{B}_{i+1}^{\text{new}}$, given $\mathcal{B}_i$ of cardinality $||\mathcal{B}_i|| = 2^{i-1}$, an equiprobable forward or backward direction is picked and 2^{i-1} leapfrog steps are taken in that direction, starting from the leftmost or rightmost state in $\mathcal{B}_i$, respectively. There are two termination conditions under which $\mathcal{B}_{i+1}^{\text{new}}$ is discarded and $\mathcal{B}_i$ is returned as the final $\mathcal{B}$:

(a) When a U-turn happens in the way:

The condition can be verified as follows. $\mathcal{B}_{i+1}^{\text{new}}$ is split into two parts (according to the order defined by leapfrogs) and each part into two parts recursively. Termination deterministically happens if for $\mathcal{B}_{i+1}^{\text{new}}$ or any such subset $\mathcal{S}$, the following condition holds:

$$[(\mathbf{q}^+(\mathcal{S}) - \mathbf{q}^-(\mathcal{S})) \cdot \mathbf{p}^-(\mathcal{S}) < 0] \vee [(\mathbf{q}^+(\mathcal{S}) - \mathbf{q}^-(\mathcal{S})) \cdot \mathbf{p}^+(\mathcal{S}) < 0] \tag{3}$$

where $(\mathbf{q}^-(\mathcal{S}), \mathbf{p}^-(\mathcal{S}))$ and $(\mathbf{q}^+(\mathcal{S}), \mathbf{p}^+(\mathcal{S}))$ are the first and last elements of $\mathcal{S}$ visited via leapfrog mechanism.

(b) When the probability of a newly generated state $\mathbf{z}'$ is extremely low:

$$u \not< \exp(\Delta_{\max} - H(\mathbf{z}')) \tag{4}$$

where $\Delta_{\max}$ is a positive parameter set by the user. This is equivalent to the termination probability,

$$P(\text{TERMINATE-b}) = 1 - \min\left(1, \exp(\Delta_{\max})\frac{\exp(-H(\mathbf{z}'))}{\exp(-H(\mathbf{z}))}\right) \tag{5}$$

Therefore, according to (5), the probability of termination via this criterion is non-zero if $H(\mathbf{z}') > H(\mathbf{z}) + \Delta_{\max}$. That is, the Hamiltonian preservation is violated by an error of magnitude more than $\Delta_{\max}$, where $\Delta_{\max}$ is a positive parameter recommended to be set large (e.g. 1000 according to (Hoffman and Gelman, 2014)).

If neither of the above termination conditions holds then,

1. The set of candidate states is expanded:
$$\mathcal{B}_{i+1} := \mathcal{B}_i \sqcup \mathcal{B}_{i+1}^{\text{new}}$$
2. If U-turn condition given by (3) holds for $\mathcal{S} = \mathcal{B}_{i+1}$, then termination happens and $\mathcal{B}_{i+1}$ is return as the final $\mathcal{B}$. Otherwise, the iterative expansion of $\mathcal{B}$ continuous.

Step 3. Decide the next state

When the final $\mathcal{B}$ is made, a set $\mathcal{C} \subset \mathcal{B}$ (of *chosen states*) is generated deterministically, conditioned on u:

$$\forall \mathbf{z} \in \mathcal{B}, \qquad P((\mathbf{z} \in \mathcal{C})|u) = \begin{cases} 1 & \text{if } u \leq P(\mathbf{z}) \\ 0 & \text{otherwise} \end{cases}, \tag{6}$$

and a member of $\mathcal{C}$ is drawn uniformly (i.e. with probability $1/||\mathcal{C}||$) as the next state $\mathbf{z}^{(t+1)}$.

2 Effect of $\Delta_{\max}$ on the convergence of baseline NUTS

Proposition 1. *If $\Delta_{\max}$ is set to any value less than ∞, the termination criterion (5) violates the intended condition (2).*

Proof. This can be shown by a simple pathological example. Consider a 1D potential function U with a discontinuity at point d and of height h in the neighbourhood of which, U is flat. Consider points q and q' on the opposite sides the flat region around d. with potentials $U(q) = c$ and $U(q') = c + h$ where c is a constant. Let leapfrog mechanism $\mathcal{F}$ maps $\mathbf{z} = (q, p)$ to $\mathbf{z}' = (q', p)$ and vice versa. (Note that since $\nabla U(\mathbf{q}) = \nabla U(\mathbf{q}') = 0$, the momentum is not changed via $\mathcal{F}$). If $h > \Delta_{\max}$ it can easily be seen that $P(\mathcal{B}|\mathbf{z}) \neq P(\mathcal{B}|\mathbf{z}')$ for $\mathcal{B} := \{\mathbf{z}, \mathbf{z}'\}$, because by (5),

$$\begin{aligned} P(\mathcal{B}|\mathbf{z}) &= P(\mathcal{B}_2 = \{\mathbf{z}, \mathbf{z}'\}|\mathcal{B}_1 = \{\mathbf{z}\}, \mathcal{B}_2^{\text{new}} = \{\mathbf{z}'\}) \\ &= 1 - P(\text{TERMINATE-b}|\mathcal{B}_1 = \{\mathbf{z}\}, \mathcal{B}_2^{\text{new}} = \{\mathbf{z}'\}), \\ &= \min\left(1, e^{\Delta_{\max}} \cdot \frac{e^{-H(\mathbf{z}')}}{e^{-H(\mathbf{z})}}\right) = \min\left(1, e^{\Delta_{\max}} \cdot \frac{e^{-U(q')}}{e^{-U(q)}}\right) = \min\left(1, e^{\Delta_{\max}} \cdot \frac{e^{-c-h}}{e^{-c}}\right) \\ &= \min\left(1, \frac{e^{\Delta_{\max}}}{e^{h}}\right) = \frac{e^{\Delta_{\max}}}{e^{h}} \qquad \text{since } h > \Delta_{\max} \end{aligned}$$

but with a similar reasoning

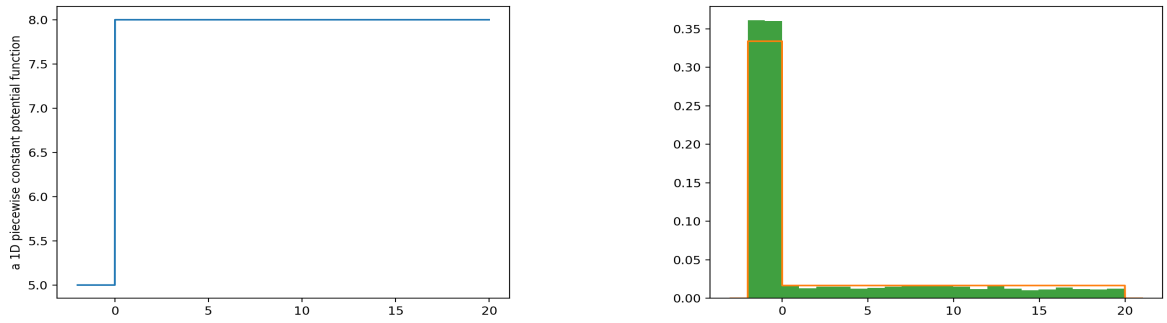
$$P(\mathcal{B}|\mathbf{z}') = \min\left(1, \frac{e^{h}}{e^{\Delta_{\max}}}\right) = 1$$

Which violates condition (2). □

An instance of such a potential function is plotted in Figure 1 (a) where $c = 5$, $h = 3$ and $\Delta_{\max} = 1.0$. Figure 1 (b) depicts its corresponding density function along with the histogram of 5000 samples drawn from it via NUTS algorithm. It can be seen that the samples do not converge to the target distribution.

To prevent such pathological examples, the designers of NUTS suggest setting $\Delta_{\max}$ to a very large value (like 1000). Therefore, as long as the normalising constant of the target distribution is not large, the interference with detailed balance should be minimal or negligible. Nonetheless, this means that in NUTS, the expansion of $\mathcal{B}$ may continue despite very large errors in the approximation of Hamiltonian.

Unlike the baseline NUTS, the proposed NoVoP-NUTS, does not rely on termination criterion based on $\Delta_{\max}$ and is able to terminate on much lower Hamiltonian approximation errors without violating detailed balance.



(a) (b)

Figure 1: (a) A 1D piecewise constant potential function, (b) its corresponding density function (orange curve) and the histogram of 5000 samples drawn from it with baseline NUTS where $\Delta_{\max}$ is set to 1.

3 Proof of correctness of NoVoP HMC

Theorem 6.1 of the main paper can be restated as follows:

Theorem 3.1. *Draws from NUTS sampler with an arbitrary (not necessarily volume-preserving) bijective transition step converges to draws from a correct stationary distribution if any state $\mathbf{z}'$ in the traced set $\mathcal{B}$ is added to the set of chosen states $\mathcal{C}$, if:*

$$u \leq \pi_{\mathbf{Z}}(\mathbf{z}') \left| \frac{\partial \mathbf{z}'}{\partial \mathbf{z}} \right|$$

where u is NUTS slice variable. That is, the probability of adding a state, $\mathbf{z}'$, to $\mathcal{C}$, given the traced set, $\mathcal{B}$, and slice variable, u, is:

$$P(\mathbf{z}' \in \mathcal{C} | \mathcal{B}, u) := \begin{cases} \mathbb{I}\left[u \leq \pi_{\mathbf{Z}}(\mathbf{z}') \cdot \left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right|\right] & \text{if } \mathbf{z}' \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Proof. Since the probability of choosing a state $\mathbf{z}' \in \mathcal{C}$ as the next state is $1/\|\mathcal{C}\|$, with respect to (7), given $\mathcal{B}$ and u, for any $\mathbf{z}$ and $\mathbf{z}'$ in $\mathcal{B}$, the probability of transition $\mathbf{z} \to \mathbf{z}'$ is,

$$Q(\mathbf{z} \to \mathbf{z}' | \mathcal{B}, u, (\mathbf{z}' \in \mathcal{B})) = \frac{P(\mathbf{z}' \in \mathcal{C} | \mathcal{B}, u, (\mathbf{z}' \in \mathcal{B}))}{\text{size of } C \text{ given } \mathcal{B}, u} = \frac{\mathbb{I}\left[u \leq \pi_{\mathbf{Z}}(\mathbf{z}') \cdot \left|\frac{\partial \mathbf{z}'}{\partial \mathbf{z}}\right|\right]}{\|\{\mathbf{z}^* \in \mathcal{B} \text{ s.t. } u \leq \pi_{\mathbf{Z}}(\mathbf{z}^*) \left|\frac{\partial \mathbf{z}^*}{\partial \mathbf{z}}\right|\}\|} \tag{8}$$

Therefore, to compute $Q(\mathbf{z} \to \mathbf{z}')$, the traced set $\mathcal{B}$ and slice variable u should be marginalised:

$$Q(\mathbf{z} \to \mathbf{z}') = \sum_{\mathcal{B} \in \mathfrak{B}} P(\mathcal{B}|\mathbf{z}) \sum_{\mathbf{z}_b \in \mathcal{B}} \delta(\mathbf{z}' - \mathbf{z}_b) \int_u P_U(u|\mathbf{z}) Q(\mathbf{z} \to \mathbf{z}' | \mathcal{B}, u, (\mathbf{z}' \in \mathcal{B})) du \tag{9}$$

where $\mathfrak{B}$ is the set of all possible traced sets, $\sum_{\mathbf{z}_b \in \mathcal{B}} \delta(\mathbf{z}' - \mathbf{z}_b)$ is the probability that $\mathbf{z}'$ is in $\mathcal{B}$. Meanwhile since u is uniformly drawn from the interval $[0, \pi_{\mathbf{Z}}(z)]$,

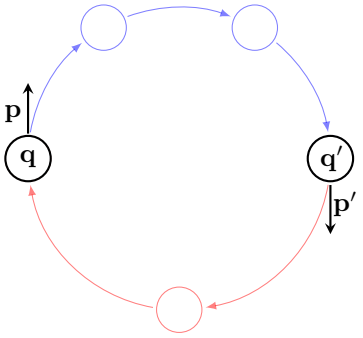
$$P_U(u|\mathbf{z}) := \frac{\mathbb{I}[0 \leq u \leq \pi_{\mathbf{Z}}(\mathbf{z})]}{\pi_{\mathbf{Z}}(\mathbf{z})} \tag{10}$$

In NUTS algorithm, the acceptance probability of the proposal, $\alpha(\mathbf{z} \to \mathbf{z}')$, is always 1. Therefore, non-volume preserving NUTS satisfies the generalised detailed balance condition,

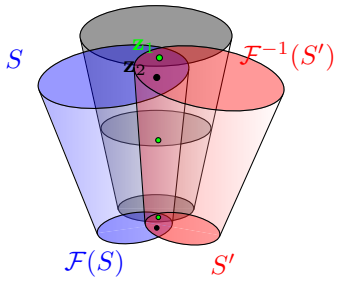
$$\int_{\mathbf{z} \in S} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}' \in S'} Q(\mathbf{z} \to \mathbf{z}') \alpha(\mathbf{z} \to \mathbf{z}') d\mathbf{z}' d\mathbf{z} = \int_{\mathbf{z}' \in S'} \pi_{\mathbf{Z}}(\mathbf{z}') \int_{\mathbf{z} \in S} Q(\mathbf{z}' \to \mathbf{z}) \alpha(\mathbf{z}' \to \mathbf{z}) d\mathbf{z} d\mathbf{z}' \tag{11}$$

iff for arbitrary volumes S and S' in the phase space:

$$\int_{\mathbf{z}\in S} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}'\in S'} Q(\mathbf{z}\to\mathbf{z}')d\mathbf{z}'d\mathbf{z} = \int_{\mathbf{z}'\in S'} \pi_{\mathbf{Z}}(\mathbf{z}') \int_{\mathbf{z}\in S} Q(\mathbf{z}'\to\mathbf{z})d\mathbf{z}d\mathbf{z}' \tag{12}$$



(a) Transition from $(\mathbf{p},\mathbf{q})$ to $(\mathbf{p}',\mathbf{q}')$ via transition steps forward in time (blue path) and transition steps backward in time (red path).



(b) Transition from $\mathbf{z}_1 \in S$ to S' is via two transition steps while transition from $\mathbf{z}_2 \in S$ to S' is via one transition step.

Figure 2: Diagram (a) shows that in NUTS, the transition from one state to another may not be unique since by following forward and backward leapfrogs, different intermediate states may be traced. Diagram (b) shows that different subsets of a volume S may be mapped to S' with different number of transition steps.

A major difference between HMC and NUTS is that in HMC, the transition, $\mathcal{F}$, is unique i.e. it is always via a fixed number, L, of transition steps that are forward in time. The transition in NUTS is more complicated. To start with, as illustrated in Figure 2a, in exceptional cases where a couple of states $\mathbf{z} := (\mathbf{q},\mathbf{p})$ and $\mathbf{z}' := (\mathbf{q}',\mathbf{p}')$ are both on U-turn positions, $\mathbf{z}$ can evolve to $\mathbf{z}'$ via both forward and backward transition steps. Ending in paths that have different intermediate states.

For notational ease, we only consider traced sets $\mathcal{B}$ where the direction of transition steps is forward in time from states S to S', because by symmetry, the discussion holds for the opposite direction as well. This allows us to notate $\mathbf{z}'$ as a function of $\mathbf{z}$ since in NUTS algorithm, for all trace sets $\mathcal{B}$ where $\mathbf{z}$ evolves to $\mathbf{z}'$ via transitions steps that are forward in time, the same set of intermediate states are traced.

Note that the bijective mapping $\mathcal{F}$ may involve different number of transition steps for different subsets of $\mathcal{S}$. Figure 2b, depicts an example where states $\mathbf{z}_1$ and $\mathbf{z}_2$ in S are mapped to S' via 2 and 1 transition steps, respectively. Also if the target region S' is large enough, a traced set $\mathcal{B}$ (that evolves $\mathbf{z}$ forward in time) may contain several states that are in S'.

Again, for brevity, throughout we assume that volumes S and S' are small enough so that any trace set $\mathcal{B}$ contains at most one state from S and one state from S' and that if a trace set contains a pair of states $z\in S$ and $z'\in S'$ then starting from $\mathbf{z}$, by k forward transition steps we end in $\mathbf{z}'$ where k is fixed. That is,

$$\{\mathbf{z},\mathbf{z}'\}\subset\mathcal{B} \iff \mathbf{z}' = \mathcal{F}(\mathbf{z}), \qquad \forall\mathcal{B},\mathbf{z}\in S,\mathbf{z}'\in S' \tag{13}$$

where by definition, $\mathcal{F}$ evolves its input by k forward transition steps. As such, the probability of $\mathbf{z}'\in S'$ being in $\mathcal{B}$ given $\mathbf{z}\in S$ is,

$$\sum_{\mathbf{z}_b\in\mathcal{B}} \delta(\mathbf{z}'-\mathbf{z}_b) := \delta(\mathbf{z}'-\mathcal{F}(\mathbf{z})) \tag{14}$$

These assumptions do not lead to lose of generality since if they do not hold, we can partition S and S' to sufficiently smaller volumes where the assumptions are satisfied. Since if detailed balance holds for the partitions of S and S', then it holds for the original volumes too.

By substituting (9) in (12) and under the aforementioned assumptions, the LHS of (12) becomes,

$$\int_{\mathbf{z}\in S} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}'\in S'} \sum_{\mathcal{B}} P(\mathcal{B}|\mathbf{z}) \sum_{\mathbf{z}_b\in\mathcal{B}} \delta(\mathbf{z}'-\mathbf{z}_b) \int_u P_U(u|\mathbf{z}) Q(\mathbf{z}\to\mathbf{z}'|\mathcal{B},u,(\mathbf{z}'\in\mathcal{B}))du d\mathbf{z}' d\mathbf{z} \tag{15}$$

The integral bounds of this expression can be tightened since as illustrated in Figure 2b, transition between $\mathbf{z} \in S$ and $\mathbf{z}' \in S'$ is possible iff,

$$\mathbf{z} \in S \cap \mathcal{F}^{-1}(S') \quad \wedge \quad \mathbf{z}' \in S' \cap \mathcal{F}(S)$$

Furthermore, by substituting relation (14) in (15), the LHS of (12) becomes,

$$\begin{aligned}
&\int_{\mathbf{z} \in S \cap \mathcal{F}^{-1}(S')} \pi_{\mathbf{Z}}(\mathbf{z}) \int_{\mathbf{z}' \in S'} \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathbf{z}'\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \delta(\mathbf{z}' - \mathcal{F}(\mathbf{z})) \int_u P_U(u|\mathbf{z}) Q(\mathbf{z} \to \mathbf{z}'|u, \mathcal{B}) d\mathbf{z}' du d\mathbf{z} \\
&= \int_{\mathbf{z} \in S \cap \mathcal{F}^{-1}(S')} \pi_{\mathbf{Z}}(\mathbf{z}) \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathcal{F}(\mathbf{z})\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \int_u P_U(u|\mathbf{z}) Q(\mathbf{z} \to \mathcal{F}(\mathbf{z})|u, \mathcal{B}) du d\mathbf{z} \qquad (16)
\end{aligned}$$

where in the second line, by using the sifting property of Dirac delta, the integration over S' is removed and all $\mathbf{z}'$ are replaced by $\mathcal{F}(\mathbf{z})$.

Finally, by substituting the definition of P_U (given by (10)) and Q (given by (8)) , in (16), the LHS of (12) becomes:

$$\begin{aligned}
&\int_{\mathbf{z} \in S \cap \mathcal{F}^{-1}(S')} \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathcal{F}(\mathbf{z})\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \pi_{\mathbf{Z}}(\mathbf{z}) \int_u \frac{\mathbb{I}[0 \leq u \leq \pi_{\mathbf{Z}}(\mathbf{z})]}{\pi_{\mathbf{Z}}(\mathbf{z})} \cdot \frac{\mathbb{I}\left[u \leq \pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z})) \left|\frac{\partial \mathcal{F}(\mathbf{z})}{\partial \mathbf{z}}\right|\right]}{\|\{\mathbf{z}^* \in \mathcal{B} \text{ s.t. } u \leq \pi_{\mathbf{Z}}(\mathbf{z}^*) \left|\frac{\partial \mathbf{z}^*}{\partial \mathbf{z}}\right|\}\|} du d\mathbf{z} \\
&= \int_{\mathbf{z} \in S \cap \mathcal{F}^{-1}(S')} \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathcal{F}(\mathbf{z})\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \int_{u=0}^{\min\left\{\pi_{\mathbf{Z}}(\mathbf{z}), \pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))\left|\frac{\partial \mathcal{F}(\mathbf{z})}{\partial \mathbf{z}}\right|\right\}} \frac{1}{\|\{\mathbf{z}^* \in \mathcal{B} \text{ s.t. } u \leq \pi_{\mathbf{Z}}(\mathbf{z}^*) \left|\frac{\partial \mathbf{z}^*}{\partial \mathbf{z}}\right|\}\|} du d\mathbf{z} \qquad (17)
\end{aligned}$$

By similar calculations, the RHS of (12) becomes:

$$\begin{aligned}
&\int_{\mathbf{z}' \in S' \cap \mathcal{F}(S)} \pi_{\mathbf{Z}}(\mathbf{z}') \int_{\mathbf{z} \in S} \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathbf{z}'\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}') \delta(\mathbf{z} - \mathcal{F}^{-1}(\mathbf{z}')) \int_u P_U(u|\mathbf{z}') Q(\mathbf{z}' \to \mathbf{z}|u, \mathcal{B}) d\mathbf{z} du d\mathbf{z}' \\
&= \int_{\mathbf{z}' \in S' \cap \mathcal{F}(S)} \sum_{\mathcal{B} \text{ s.t. } \{\mathcal{F}^{-1}(\mathbf{z}'), \mathbf{z}'\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \int_{u=0}^{\min\left\{\pi_{\mathbf{Z}}(\mathbf{z}'), \pi_{\mathbf{Z}}(\mathcal{F}^{-1}(\mathbf{z}'))\left|\frac{\partial \mathcal{F}^{-1}(\mathbf{z}')}{\partial \mathbf{z}'}\right|\right\}} \frac{1}{\|\{\mathbf{z}^* \in \mathcal{B} \text{ s.t. } u \leq \pi_{\mathbf{Z}}(\mathbf{z}^*) \left|\frac{\partial \mathbf{z}^*}{\partial \mathbf{z}'}\right|\}\|} du d\mathbf{z}' \qquad (18)
\end{aligned}$$

where in the second line, by relying on NUTS balance tree property (given by (2)), $P(\mathcal{B}|\mathbf{z}')$ is substituted by $P(\mathcal{B}|\mathbf{z})$.

By change of random variable $\mathbf{z} = \mathcal{F}^{-1}(\mathbf{z}')$, expression (18) becomes,

$$\begin{aligned}
&\int_{\mathbf{z} \in S \cap \mathcal{F}^{-1}(S')} \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathcal{F}(\mathbf{z})\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \int_{u=0}^{\min\left\{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z})), \pi_{\mathbf{Z}}(\mathbf{z})\left|\frac{\partial \mathcal{F}(\mathbf{z})}{\partial \mathbf{z}}\right|^{-1}\right\}} \\
&\qquad \frac{1}{\|\{\mathbf{z}^* \in \mathcal{B} \text{ s.t. } u \leq \pi_{\mathbf{Z}}(\mathbf{z}^*) \left|\frac{\partial \mathbf{z}^*}{\partial \mathbf{z}}\right| \cdot \left|\frac{\partial \mathcal{F}(\mathbf{z})}{\partial \mathbf{z}}\right|^{-1}\}\|} du \left|\frac{\partial \mathcal{F}(\mathbf{z})}{\partial \mathbf{z}}\right| d\mathbf{z} \qquad (19)
\end{aligned}$$

Finally, by another change of random variable, $w = u\,|\partial \mathcal{F}(\mathbf{z})/\partial \mathbf{z}|$, the expression given by (19) becomes,

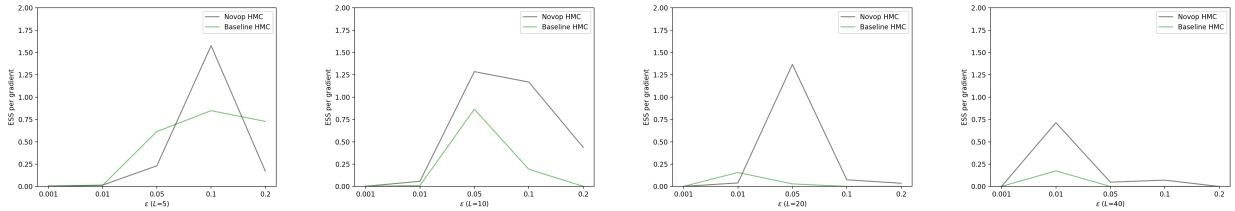
$$\int_{\mathbf{z} \in S \cap \mathcal{F}^{-1}(S')} \sum_{\mathcal{B} \text{ s.t. } \{\mathbf{z}, \mathcal{F}(\mathbf{z})\} \in \mathcal{B}} P(\mathcal{B}|\mathbf{z}) \int_{w=0}^{\min\left\{\pi_{\mathbf{Z}}(\mathcal{F}(\mathbf{z}))\left|\frac{\partial \mathcal{F}(\mathbf{z})}{\partial \mathbf{z}}\right|, \pi_{\mathbf{Z}}(\mathbf{z})\right\}} \frac{1}{\|\{\mathbf{z}^* \in \mathcal{B} \text{ s.t. } w \leq \pi_{\mathbf{Z}}(\mathbf{z}^*) \left|\frac{\partial \mathbf{z}^*}{\partial \mathbf{z}}\right|\}\|} dw d\mathbf{z} \qquad (20)$$

that is equal to the LHS, i.e. (17), which completes the proof. □

4 More Results

4.1 Leapfrog Parameter Configuration

In the experiments presented in the main paper, we used the configurations $L = 10$ and $\epsilon = 0.1$ for Baseline and NoVoP HMC samplers. Nonetheless, it is informative to study the comparative performance of these samplers for different configurations of the leapfrog parameters.



(a) (b) (c) (d)

Figure 3: Effective sample size per gradient versus leapfrog step size ϵ for baseline and NoVoP HMC algorithms where the number of leapfrog steps, L, is set to (a) 5, (b) 10, (c) 20 and (d) 40.

Figure 3 illustrates the effect of tuning the leapfrog parameters for sampling from the experimental model 2 of the main text. The compared samplers are baseline and NoVoP HMC algorithms and the reported performance measure is *Effective sample size* (ESS) per gradient. In sub-figures (a)-(d) the number of leapfrog steps L is set to 5, 10, 20 and 40, respectively and for each configurations of L, the leapfrog step size, ϵ, takes a value in $\{0.001, 0.01, 0.05, 0.1, 0.2\}$. ESS is computed following Hoffman and Gelman (2014) with cutoff threshold 0.05. We choose *summation* as the reduction function f. Due to the symmetry of the target distribution, we know that the true mean of f is 0 and its variance is approximated with high precision from 50,000 samples drawn by a separate run of NoVoP HMC. The results show that in most configuration settings NoVoP HMC outperforms the baseline and the difference is more significant when the parameters are close to their optimal values.

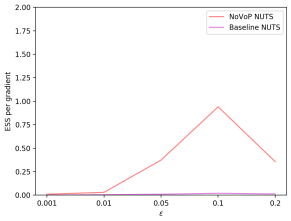


Figure 4: Effective sample size per gradient versus leapfrog step size ϵ for baseline and NoVoP NUTS algorithms.

In Figure 4, the same experiment is repeated for baseline and NoVoP NUTS. Here, the only parameter is ϵ and as the results show, baseline NUTS has difficulty sampling from the studied target distribution and regardless of the parameter tuning, it's performance is extremely poor. The reason is that baseline NUTS does not see the boundaries and is lost in the space. U-turns usually do not happen and the trace set is often expanded to its maximum size. nonetheless, due to the Hamiltonian approximation error, most of the traced states are not added to $\mathcal{C}$ (the ratio of the cardinality of $\mathcal{C}$ to $\mathcal{B}$ is in the order of 0.0001). As such, despite costly computations, baseline NUTS MCMC chains typically stick to a state and have very low ESS. As it can be seen, the performance of NoVoP NUTS is significantly better however, on this model it does not perform as good as tuned NoVoP HMC.

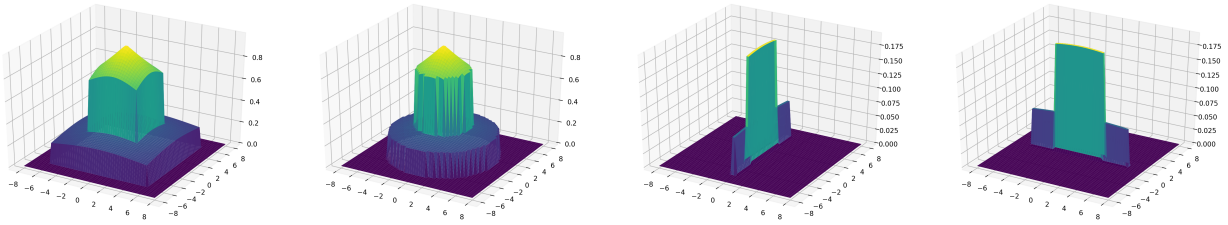
4.2 The geometry of the target experimental distributions

To have a better understanding of the geometry of the experimental distributions (in the main paper), some 2-dimensional instances of such target distributions are plotted in Figure 5. Figure 5-(a) plots a 2D distribution associated with the first experimental model (in the main paper) i.e.

$$\pi_{\mathbf{Q}}(\mathbf{q}) \propto \exp(-U(\mathbf{q})), \qquad U(\mathbf{q}) = \begin{cases} \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } \|\mathbf{q}\|_\infty \leq 3 \\ 1 + \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } 3 \leq \|\mathbf{q}\|_\infty \leq 6 \\ +\infty, & \text{otherwise} \end{cases}$$

where,

$$A := \begin{bmatrix} e^5 & 0 \\ 0 & e^5 \end{bmatrix}$$



(a) (b) (c) (d)

Figure 5: 2D posteriors of Experiment 1 (a) and Experiment 2 (b-d) where in (a) & (b) matrix $A = e^{-5}\mathbf{I}$, in (c) $A = \texttt{diag}([e^5, e^{-5}])$ and in (d) $A = \texttt{diag}([e^{-5}, e^5])$.

Figure 5-(b-d) plots a 2D distribution associated with the second experimental model (in the main paper) i.e.

$$\pi_{\mathbf{Q}}(\mathbf{q}) \propto \exp(-U(\mathbf{q})), \qquad U(\mathbf{q}) = \begin{cases} \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } \|\mathbf{q}\|_2 \leq 3 \\ 1 + \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{if } 3 \leq \|\mathbf{q}\|_2 \leq 6 \\ 50 + \sqrt{\mathbf{q}^\intercal A\, \mathbf{q}}, & \text{otherwise} \end{cases}$$

where in Figure 5-(b), $A := \begin{bmatrix} e^5 & 0 \\ 0 & e^5 \end{bmatrix}$, in Figure 5-(c), $A := \begin{bmatrix} e^5 & 0 \\ 0 & e^{-5} \end{bmatrix}$ and in Figure 5-(d), $A := \begin{bmatrix} e^{-5} & 0 \\ 0 & e^5 \end{bmatrix}$.

4.3 Generalised Bayesian Belief Update

We consider the *Generalised Bayesian Belief update* framework for binary classification problem (Nishimura et al., 2020), which is based on SECOM data from UCI's machine learning repository.

$$\pi_{\mathbf{Q}}(\mathbf{q}|\mathbf{y}) \propto \pi_{\mathbf{Q}}(\mathbf{q}) \exp\left\{ -\sum_{i=1}^{N} \mathbb{I}[y_i \mathbf{x}_i^\top \mathbf{q} < 0] \right\} \tag{21}$$

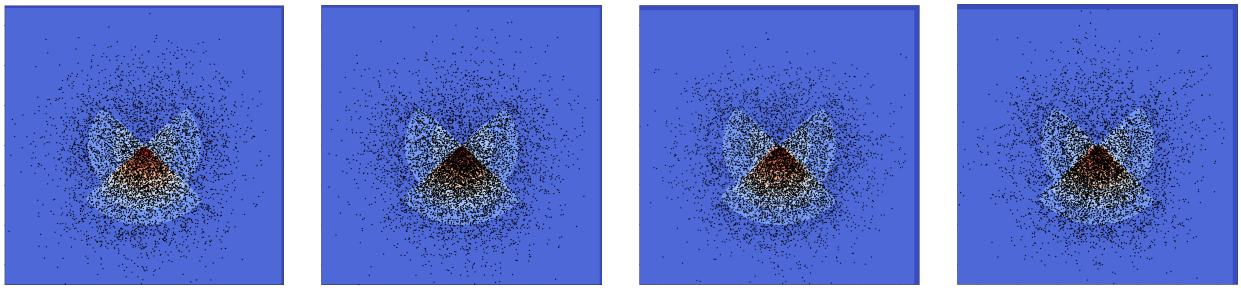
where $\mathbf{x}_i$ are vectors of predictors, $\mathbf{q}$ consists of regression coefficients and $y_i \in \{-1, 1\}$ indicate decisions. The prior distribution of the coefficients $\pi_{\mathbf{Q}}(\mathbf{q})$ is assumed to be standard normal $\mathcal{N}(\mathbf{0}, \mathbf{1})$.

Figure 6 depicts 10,000 samples drawn by baseline and NoVoP HMC and NUTS from this target model. To make a distribution that is suitable for visualisation, we have only consider the first two predictors and the first two data points of SECOM dataset. This leads to a 4-piece, 2 dimensional model. It can be seen that the scatter plots of samples drawn by NoVoP HMC and NUTS look denser than those drawn by their baseline counterparts. The reason is that in the non-volume preserving samplers, the Hamiltonian is preserved better. This directly leads to lower proposal rejection rate in NoVoP HMC in comparison with baseline HMC. In NUTS algorithm, there is no explicit proposal acceptance/rejection step. However, poor Hamiltonian preservation leads to a small set of chosen states $\mathcal{C}$ (even if the traced set $\mathcal{B}$ is very large). Since the current state is always included in $\mathcal{C}$, the probability of choosing the current state as the next state is higher if $\mathcal{C}$ is a small set.

Table 1: Proposal (delegate) acceptance rate versus no. data points (for model given by (21) with dimensionality equal to 5)

	10 data points	20 data points	40 data points	100 data points
Baseline HMC	0.39	0.26	0.13	0.09
NoVoP HMC	**0.54**	**0.50**	**0.44**	**0.53**
Baseline NUTS	0.42	0.35	0.22	0.20
NoVoP NUTS	**0.49**	**0.44**	**0.42**	**0.50**

Note that in models defined by (21), each likelihood function partitions the parameter space to two. As such, the number of partitions can grow up to exponential in the number of data points. On highly piecewise models, the performance of the baseline HMC and NUTS in terms of preserving the Hamiltonian drops while in NoVoP algorithms, the Hamiltonian



(a) Baseline HMC (b) NoVoP HMC (c) Baseline NUTS (d) NoVoP NUTS

Figure 6: 10,000 samples drawn from the experimental model given by (21) with 2 data points and 2 predictors using different sampling methods.

preservation remains high. As such, we expect that in highly piecewise models, the difference between the baseline and non-volume preserving samplers become more evident. To verify this prediction, we fix the dimensionality of the model to 5 and compare the samplers on models with different number of data points. For each sampler, we run 3 MCMC chains of size 1000. All chains start from a same initial state where all position elements are set to 0.1. For baseline and NoVoP HMC samplers, we report the average proposal acceptance rate, whereas in case of baseline and NoVoP NUTS, we report the sum of caridinality of all generated sets $\mathcal{C}$ (of chosen states) divided by the sum of all traced states (even those not added to $\mathcal{B}$) as the delegate of the acceptance rate. All samplers use the same parameters $\epsilon = 0.1$ and $L = 10$ (when applicable). The gradients, $\nabla U(\mathbf{q})$, and all required Jacobians are automatically computed using JAX auto-differentiation framework (Bradbury et al., 2018). The results are presented in Table 1 where it is clear that the non-volume preserving samplers constantly outperform their baseline counterparts and as predicted, the difference becomes more pronounced when the models are highly piecewise.

References

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, and Skye Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018. URL `http://github.com/google/jax`.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Akihiko Nishimura, David B Dunson, and Jianfeng Lu. Discontinuous hamiltonian monte carlo for discrete parameters and discontinuous likelihoods. *Biometrika*, 107(2):365–380, 2020.