## Appendix A. Proof of Lemma 1

To illustrate that Equation (6) indeed results in the maximum sum of KL divergence, indicating a systematic endeavour to minimise information loss between the mixture and target distributions, and consequently, leading to a more precise representation of the desired data distribution, consider the situation when $J = 2$ and $S = S^{(1)} \cup S^{(2)}$.

$$\begin{aligned}\Delta KL(S^{(1)}, S^{(2)}) &:= \int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx - \int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(1)},\sigma^{(1)})})\pi(x)dx \\ &+ \int_{S(2)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(1)},\sigma^{(1)})})\pi(x)dx - \int_{S(2)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx \ .\end{aligned} \tag{13}$$

We consider that $\pi(x)$ is calculated by Equation (4) therefore $\mathcal{N}(x;\mu^{(i)},\sigma^{(i)}) \leq \pi \leq \mathcal{N}(x;\mu^{(j)},\sigma^{(j)})$ on $S^{(j)}$ when $i \neq j$ therefore:

$$\begin{aligned}\int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(1)},\sigma^{(1)})})\pi(x)dx &\leq \int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx \\ \int_{S(2)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx &\leq \int_{S(2)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(1)},\sigma^{(1)})})\pi(x)dx\end{aligned} \tag{14}$$

To argue that $S$ is the applicable definition for regions, we define new partitions instead of $S$, say $S = S^{\hat{(1)}} \cup S^{\hat{(2)}}$ such that $S^{\hat{(1)}} = A \cup B$ with $A \subset S^{(1)}$ and $B \subset S^{(2)}$

$$\begin{aligned}&\Delta KL(S^{\hat{(1)}}, S^{\hat{(2)}}) = \Delta KL(S^{(1)}, S^{(2)}) \\ &+ 2\Big[\int_A\int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(1)},\sigma^{(1)})})\pi(x)dx - \int_A\int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx\Big] \\ &+ 2\Big[\int_B\int_{S(2)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx - \int_B\int_{S(1)} \log(\frac{\pi(x)}{\mathcal{N}(x;\mu^{(2)},\sigma^{(2)})})\pi(x)dx\Big] \\ &\leq \Delta KL(S^{(1)}, S^{(2)}) \ .\end{aligned} \tag{15}$$

We can expand the proof by increasing the $J$ number. The outcome will be the same. Therefore the definition $S$ as stated in Equation (6) has maximum KL divergence differences.

## Appendix B. Derivation of Equation (9)

Haario et al. (2001) presented an adaptive Metropolis algorithm, in which the Gaussian proposal distribution is changed throughout the process utilising all available data. That is, the parameters are updated in real-time. The following is a summary of the algorithm:

$$\begin{aligned}\mu_k &= \mu_{k-1} + \lambda_k(x_k - \mu_{k-1}) \\ \sigma_k &= \sigma_{k-1} + \lambda_k((x_k - \mu_{k-1})(x_k - \mu_{k-1})^T - \sigma_{k-1}) \ .\end{aligned} \tag{16}$$

Where:

- $x_{k+1}$ is derived from $P_{\theta_k}(x_k,.)$ where $\theta = (\mu, \sigma)$ is the kernel of a symmetric random walk MH with a gaussian increment distribution $\mathcal{N}(0, \gamma\sigma)$ $\gamma$ is a constant scaling factor that depends only on the dimension of the state space $n_x$, which is maintained constant between iterations;
- $\{\lambda_k\}$ is a nonincreasing series of positive stepsizes with $\sum_{k=1}^{\infty} \lambda_k = \infty$ and$\sum_{k=1}^{\infty} \lambda_k^{1+\delta} < \infty$ for some $\delta > 0$.

In F-HMC we need to update the parameters $\mu_n^{(j)}$ and $\sigma_n^{(j)}$ on the fly. We built our model for parameter updates on the equation (16) . We have defined our parameters with the conditions described here to ensure it fits in the equation. To do so, we defined

$$\omega_n^{(j)} = \frac{\mathcal{N}(x; \mu_n^{(j)}, \sigma_n^{(j)})}{\sum_{j=1}^{J} \mathcal{N}(x; \mu_n^{(j)}, \sigma_n^{(j)})} \ . \tag{17}$$

as $\lambda_k$ in the equation (16). Since $\sigma_n^{(j)}$ is a nonincreasing positive sequence that $\sum_{n=1}^{\infty} \omega_n^{(j)} < \infty$, it fits the constraints given in Harrio's algorithm.

We have also defined $ker_{H_n}(x_n,.)$ as $P_{\theta_k}(x_k,.)$ and $H_n^{(j)} = [\mu_n^{(j)}, \sigma_n^{(j)}]$ as parameter updates $\theta = (\mu, \sigma)$ in our algorithm, which conforms to the Harrios algorithm. As a result, the F-HMC updates parameters as follows:

$$\begin{aligned}
\mu_n^{(j)} &= \mu_{n-1}^{(j)} + \frac{\omega_n^{(j)}}{\sum_{i=1}^{n} \omega_i^{(j)}} (x_n - \mu_{n-1}^{(j)}) \\
\sigma_n^{(j)} &= \sigma_{n-1}^{(j)} + \frac{\omega_n^{(j)}}{\sum_{i=1}^{n} \omega_i^{(j)}} ((x_n - \mu_{n-1}^{(j)})(x_n - \mu_{n-1}^{(j)})^T - \sigma_{n-1}^{(j)}) \ .
\end{aligned} \tag{18}$$

## Appendix C. Proof of Lemma 2

### F-HMC satisfies diminishing adaptation condition.

Diminishing condition is defined as $\lim_{n\to\infty} D_n = 0$ where

$$D_n = \sup_{x \in S} |ker_{Hn+1}(x,.) - ker_{Hn}(x,.)| \ . \tag{19}$$

which denotes the distinction between the transition kernels employed throughout iterations $n$ and $n+1$. All we have to do now is prove that $|H_{n+1}^{(j)} - H_n^{(j)}|$ converges to zero with probability 1.

Considering definition of $\mu_n^{(j)}$ and $\sigma_n^{(j)}$ in Equation (10), for all $j = 1, ...J$ we have

$$|\mu_{n+1}^{(j)} - \mu_n^{(j)}| = \frac{\omega_{n+1}^{(j)}}{\sum_{i=1}^{n+1} \omega_i^{(j)}} (x_{n+1} - \mu_n^{(j)}) \ . \tag{20}$$

$$|\sigma_{n+1}^{(j)} - \sigma_n^{(j)}| = \frac{\omega_{n+1}^{(j)}}{\sum_{i=1}^{n+1} \omega_i^{(j)}} ((x_{n+1} - \mu_n^{(j)})(x_{n+1} - \mu_n^{(j)})^T - \sigma_n^{(j)}) \ . \tag{21}$$

Because $S$ is compact, $X_n$, $\mu_n^{(j)}$, and $\sigma_n^{(j)}$ are uniformly bounded, and $0 \leq \omega_i^{(j)} \leq 1$, $\frac{\omega_{n+1}^{(j)}}{\sum_{i=1}^{n+1} \omega_i^{(j)}}$ converges to zero as n approaches $\infty$. Because $\mu_n^{(j)}$ and $X_n$ are both uniformly bounded, $|\mu_{n+1}^{(j)} - \mu_n^{(j)}|$ converges to zero. Similarly, $|\sigma_{n+1}^{(j)} - \sigma_n^{(j)}|$ converges to zero implies that Diminishing adaptation holds.

## Appendix D. Proof of Lemma 3

**F-HMC satisfies Containment condition.**

Containment says that the process's convergence times are bounded in probability. Meaning containment $\epsilon$-convergence time $M_\epsilon(X_n, H_n)$ should be bounded in probability conditional on any $X_0 = x_0$ and $H_0$. In another word for all $\delta > 0$ there is $N$ such that

$$P(M_\epsilon(X_n, H_n) \leq N | X_0 = x_0, H_0) \geq 1 - \delta \ . \tag{22}$$

Where $M_\epsilon(X_n, H_n)$ is $\epsilon$-convergence time and is defined as:

$$M_\epsilon(X_n, H_n) = inf_n\{n \geq 1 : ||ker_{H_n}(x) - \pi(.)|| \leq \epsilon\} \ . \tag{23}$$

According to theorem 21 in Craiu et al. (2015), for each $n \in N$ the mapping $(X_n, H_n) \rightarrow \Psi_n(X_n, H_n) := ||ker_{H_n}(x) - \pi(.)||$ is continuous. Because each $ker_{H_n}(x)$ is Harris ergodic and since $\pi$ is a stationary distribution for $ker_{H_n}(x)$ the mapping $\Psi$ is nonincreasing. Dini's Rudin (1976) theorem states that for each compact subset $C \subset S$:

$$\lim_{n\rightarrow\infty} \sup_{x\in C} \sup_{h\in H_n} \Psi_n(x, h) = 0 \ . \tag{24}$$

Hence, given $C$ and $\epsilon > 0$, there is $n \in N$ with $\sup_{x\in C} \sup_{h\in H_n} \Psi_n(x, h) < \epsilon$. It follows that $\sup_{x\in C} \sup_{h\in H_n} M_\epsilon(x, y) < \infty$ for any fixed $\epsilon > 0$.

Now, if $X_n$ is bounded in probability, then for any $\delta > 0$. we can find a large enough compact subset $C$ such that $P(X_n \notin C) \leq \delta$ for all $n$. Then given $\epsilon > 0$, and if $L := \sup_{x\in C} \sup_{h\in H_n} M_\epsilon(x, h)$, then $L < \infty$, and $P(M_\epsilon(X_n, H_n) > L) \leq \delta$ for all $n$ as well. Since $\delta$ was arbitrary, it follows that $M_\epsilon(X_n, H_n)$ is bounded in probability therefore containment condition holds.