A Additional details on quasi-refreshment

In this section, we extend marginal quasi-refreshment beyond what is discussed in the main text and introduce *conditional* quasi-refreshment, which tries to match some conditional distribution of (ρ_t, θ_t) to that corresponding distribution in the target.

Marginal quasi-refreshment We begin by presenting a more general version of Proposition 3.3.

Proposition A.1. *Consider random vectors* $Y, Z, Y', Z' \in \mathbb{R}^d$ *for some* $d \in \mathbb{N}$. *Suppose that* $Y \perp\!\!\!\perp Z$ *and that we have a bijection* $R : \mathbb{R}^d \to \mathbb{R}^d$ *such that* $R(Y') \overset{d}{=} Y$. *Then*

$$\mathrm{D_{KL}}\left(R(Y'), Z' || Y, Z\right) = \mathrm{D_{KL}}\left(Y', Z' || Y, Z\right) - \mathrm{D_{KL}}\left(Y' || Y\right).$$

Proof. Since R is a bijection,

$$\mathrm{D_{KL}}\left(R(Y'), Z' || Y, Z\right) = \mathrm{D_{KL}}\left(Y', Z' || R^{-1}(Y), Z\right).$$

Because $Y \perp\!\!\!\perp Z$, and $R(Y') \overset{d}{=} Y$,

$$\begin{aligned}\mathrm{D_{KL}}\left(Y', Z' || R^{-1}(Y), Z\right) &= \mathrm{D_{KL}}\left(Y' || R^{-1}(Y)\right) + \mathbb{E}\left[\mathrm{D_{KL}}\left((Z'|Y') || Z\right)\right] \\ &= \mathbb{E}\left[\mathrm{D_{KL}}\left((Z'|Y') || Z\right)\right].\end{aligned}$$

Then we add and subtract $\mathrm{D_{KL}}(Y' || Y)$ to obtain the final result,

$$\mathbb{E}\left[\mathrm{D_{KL}}\left((Z'|Y') || Z\right)\right] = \mathrm{D_{KL}}\left(Y', Z' || Y, Z\right) - \mathrm{D_{KL}}\left(Y' || Y\right).$$

□

Then Proposition 3.3 follows immediately from Proposition A.1.

Proof of Proposition 3.3. By setting $Y = \rho, Z = \theta, Y' = \rho_t$, and $Z' = \theta_t$, we arrive at the stated result. □

More generally, in order to apply Proposition A.1, we first need to split the momentum variable into two components $(\rho^{(1)}, \rho^{(2)})$ in such a way that $\rho^{(1)} \perp\!\!\!\perp \rho^{(2)}$ under $(\theta, \rho) \sim \bar{\pi}$, and then set $Y = \rho^{(1)}$ and $Z = (\rho^{(2)}, \theta)$. Since we know that $\rho \sim \mathcal{N}(0, I)$, we have quite a few options. For example:

- Set $Y = \rho$, and $Z = \theta$. Then
$$Y \perp\!\!\!\perp Z \quad \text{and} \quad Y \sim \mathcal{N}(0, I).$$
- Set $Y = \|D\rho\|_2^2$ and $Z = \left(\frac{D\rho}{\|D\rho\|_2}, (I - D)\rho, \theta\right)$ for any binary diagonal matrix $D \in \mathbb{R}^{d \times d}$. Then
$$Y \perp\!\!\!\perp Z \quad \text{and} \quad Y \sim \chi^2(\operatorname{tr} D).$$
- Set $Y = a^T\rho$ for any $a \in \mathbb{R}^d$ such that $\|a\| = 1$, and $Z = \left((I - aa^T)\rho, \theta\right)$. Then
$$Y \perp\!\!\!\perp Z \quad \text{and} \quad Y \sim \mathcal{N}(0, 1).$$

Note again that the first option recovers the marginal quasi-refreshment in the main text. Now we use the above decomposition to design a marginal quasi-refreshment move. Suppose we have run the weighted sparse leapfrog integrator Eq. (7) up to time t, resulting in a current random state θ_t, ρ_t. Let the decomposition of the current state (θ_t, ρ_t) that we select above be denoted Y', Z'. Then the marginal quasi-refreshment move involves finding a map R such that $R(Y') \overset{d}{=} Y$. We can then refresh the state via $(R(Y'), Z')$, and continue the flow. In addition to the parametric approach to designing R presented in the main text, if Y is 1-dimensional—for example, if we are trying to refresh the momentum norm $Y = \|\rho\|_2$—then given that we know the CDF F of Y, we can estimate the CDF $\hat{F}$ of Y' using samples from the flow at timestep t, and then set

$$R(x) = F^{-1}\left(\hat{F}(x)\right).$$

For example, we could use this inverse CDF map technique to refresh the distribution of $\|\rho\|_2^2$ back to $\chi^2(d)$, or to refresh the distribution of $a^T\rho$ back to $\mathcal{N}(0, 1)$.

Conditional quasi-refreshment A conditional quasi-refreshment move is one that tries to make some conditional distribution of (ρ_t, θ_t) match the same conditional in the target. If one is able to accomplish this for *any* conditional distribution (with no requirement of independence as in the marginal case), the KL divergence is guaranteed to reduce.

Proposition A.2. *Consider random vectors $Y, Z, Y', Z' \in \mathbb{R}^d$ for some $d \in \mathbb{N}$. Suppose for each $s \in \mathbb{R}^d$ we have a bijection $R_s : \mathbb{R}^d \to \mathbb{R}^d$ such that $(R_s(Y') \mid Z' = s) \overset{d}{=} (Y \mid Z = s)$. Then*

$$\mathrm{D_{KL}}\left(R_{Z'}(Y'), Z' || Y, Z\right) = \mathrm{D_{KL}}\left(Z' || Z\right).$$

Proof. By assumption the distribution of Y given $Z = s$ is the same as that of $R_s(Y')$ given $Z' = s$; so the result follows directly from the decomposition

$$\begin{aligned}\mathrm{D_{KL}}\left(R_{Z'}(Y'), Z' || Y, Z\right) &= \mathrm{D_{KL}}\left(Z' || Z\right) + \mathbb{E}\left[\mathrm{D_{KL}}\left((R_S(Y') | Z' = S) || (Y | Z = S)\right)\right], \quad S \overset{d}{=} Z' \\ &= \mathrm{D_{KL}}\left(Z' || Z\right).\end{aligned}$$

□

Conditional moves are much harder to design than marginal moves in general. One case of particular utility occurs when one is willing to assume that (θ_t, ρ_t) are roughly jointly normally distributed. In this case,

$$\begin{bmatrix}\theta_t \\ \rho_t\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mu_\theta \\ \mu_\rho\end{bmatrix}, \begin{bmatrix}\Sigma_{\theta\theta} & \Sigma_{\theta\rho} \\ \Sigma_{\theta\rho}^T & \Sigma_{\rho\rho}\end{bmatrix}\right),$$

so one can refresh the momentum ρ_t by updating it to $R_{\theta_t}(\rho_t)$, where

$$R_s(x) = \Sigma^{-1/2}\left(x - \mu_\rho - \Sigma_{\theta\rho}^T \Sigma_{\theta\theta}^{-1}(s - \mu_\theta)\right) \quad \text{and} \quad \Sigma = \Sigma_{\rho\rho} - \Sigma_{\theta\rho}^T \Sigma_{\theta\theta}^{-1} \Sigma_{\theta\rho}.$$

Here Proposition A.2 applies by setting $Y = \rho, Z = \theta, Y' = \rho_t$, and $Z' = \theta_t$. In order to use this quasi-refreshment move, one can either include the covariance matrices and mean vectors as tunable parameters in the optimization, or use samples from the flow at step t to estimate them directly.

B Details of experiments

We begin by describing in detail the differences between `HIS-Full`/`UHA-Full` and `HIS-Coreset`/`UHA-Coreset`. It is worth noting noting that we were unable to train `HIS`/`UHA` with full-dataset flow dynamics; even on a small 2-dimensional Gaussian location model with 100 data points, these methods took over 8 minutes for training. Therefore, as suggested by [21, 23], we train the leapfrog step sizes and annealing parameters by using a random minibatch of the data in each iteration to construct the flow dynamics. To then obtain valid ELBO estimates for comparison, we generate samples from the trained flow with leapfrog transformations based on the full dataset (`HIS-Full`/`UHA-Full`). As a simple heuristic baseline that also provides a valid ELBO estimate, we compare to the trained flow with leapfrog transformations based on a fixed, uniformly sampled coreset (`HIS-Coreset`/`UHA-Coreset`) of the same size as used for `SHF`.

We now describe the detailed settings that apply across all three experiments in the main text. In all experiments, `SHF` uses the quasi-refreshment from Eq. (8) initialized using the warm-start procedure in Section 3 with a batch of 100 samples. We use the same number of leapfrog, tempering, and (quasi-)refreshment steps for all of `SHF`, `HIS` and `UHA` respectively. The leapfrog step sizes are initialized at the same value for all three methods. For all three methods (`SHF`, `HIS`, and `UHA`), the unnormalized log target density used in computing the ELBO objective is estimated using a minibatch of $S = 100$ data points for each optimization iteration. For both `HIS` and `UHA`, the leapfrog transitions themselves are also based on a fresh uniformly sampled minibatch of size 30—the same as the coreset size for `SHF`—at each optimization iteration. `HIS` and `UHA` both also involve tempering procedures, requiring optimization over the tempering schedule $0 \le \beta_1 \le \cdots \le \beta_{R-1} \le \beta_R = 1$, where R denotes the number of tempering steps (equal to the number of quasi-refreshment / refreshment steps in all methods). We consider a reparameterization of $(\beta_1, \ldots, \beta_R)$ to $(\alpha_1, \ldots, \alpha_{R-1})$, where $\alpha_r = \mathrm{Logit}(\beta_{r+1}/\beta_r)$ for $r \in \{1, \ldots, R-1\}$, ensuring a set of unconstrained parameters. The initial value for each r is set to $\alpha_r = 1$. For `UHA`, the initial damping coefficient of its partial momentum refreshment is set to 0.5. We estimate all evaluation metrics using 100 samples. To estimate KSD, we use the IMQ base kernel with its parameters set to the same values as outlined in [45] ($\beta = -\frac{1}{2}$ and $c = 1$).

In addition to `HIS` and `UHA`, we also include adaptive HMC and NUTS in our comparison of density evaluation and sample generation times. We tune adaptive HMC and NUTS with a target acceptance rate of 0.65 during a number of burn-in iterations equal to the number of output samples, and include the burn-in time in the timing results for these two methods. Finally, we compare the quality of coresets constructed by `SHF` against `UNI` and `Hilbert-OMP`. For both `UNI` and `Hilbert-OMP`, we use NUTS to draw samples from the coreset posterior approximation. For `Hilbert-OMP`, we use a random log-likelihood function projection of dimension $100d$, where the true posterior parameter is of dimension d, generated from the Laplace approximation [47].

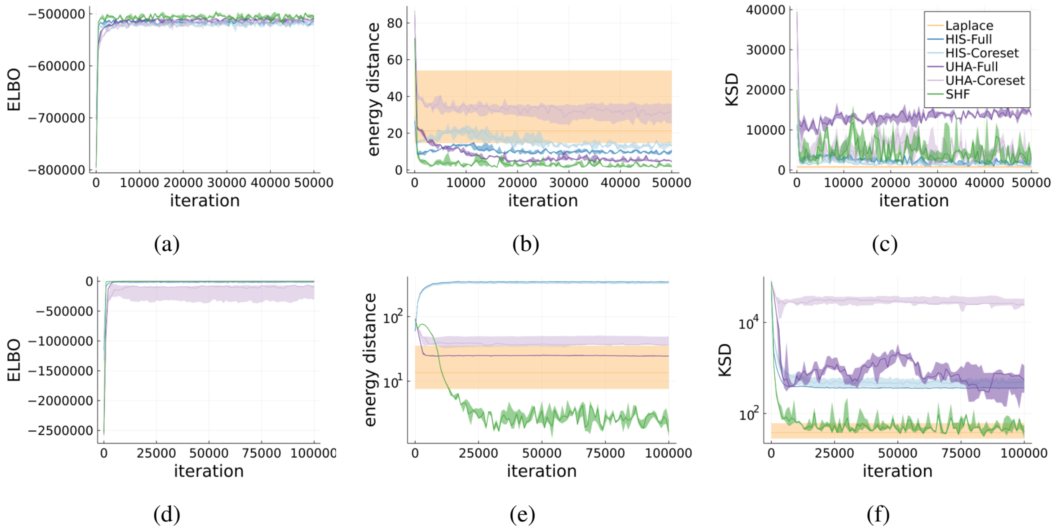


Figure 6: Linear (top row, Figs. 6a to 6c) and logistic (bottom row Figs. 6d to 6f) regression: posterior approximation quality results. The lines indicate the median, and error regions indicate 25th to 75th percentile from 5 runs.

B.1 Synthetic Gaussian

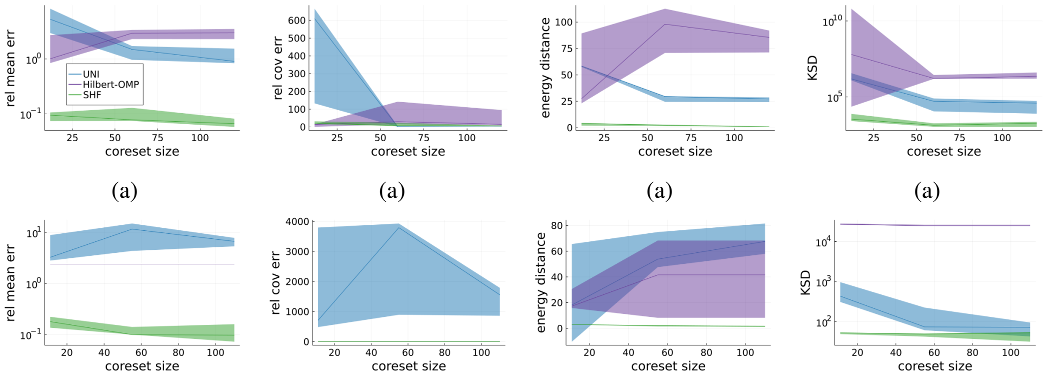
To train `SHF`, a total of 5 quasi-refreshments are used with 10 leapfrog steps in between; a similar schedule is used in `HIS` and `UHA` for momentum tempering and refreshment respectively. The initial distribution is set to $\theta_0 \sim \mathcal{N}(0, I)$ and $\rho_0 \sim \mathcal{N}(0, I)$. For all methods, the number of optimization iterations is set to $20,000$, and the initial leapfrog step size is set to 0.01 across all d dimensions. We train all methods with ADAM with initial learning rate 0.001.

B.2 Bayesian linear regression

To train `SHF`, a total of 8 quasi-refreshments are used with 10 leapfrog steps in between; a similar schedule is used in `HIS` and `UHA` for momentum tempering and refreshment respectively. The initial distribution is set to $\theta_0 \sim \mathcal{N}(15, 0.01I)$ and $\rho_0 \sim \mathcal{N}(0, I)$. For all methods, the number of optimization iterations is set to $50,000$, and the initial leapfrog step size is set to 0.02 across all d dimensions except for the dimension for the $\log \sigma^2$ term, where the step size is set to 0.0002. We train all methods with ADAM with initial learning rate 0.002. We also include the posterior approximation obtained from the Laplace approximation, where we search for the mode of the target density at some location generated from the same distribution as that of θ_0. Figs. 6, 8 and 7 provide additional results for this experiment. Fig. 6a shows that the ELBO obtained from `SHF` is a tighter lower bound of the log normalization constant compared to `HIS` and `UHA`. Fig. 6b shows that `SHF` produces a posterior approximation in terms of energy distance than all other methods; Fig. 6c shows that `SHF` is competitive, if not better, than the other methods in terms of KSD. Figs. 7a to 7d show that the coreset constructed by `SHF` is of better quality than those obtained from `UNI` and `Hilbert-OMP` in terms of all four metrics shown. Finally, Figs. 8a and 8b show the computational gain of using `SHF` to approximate density and generate posterior samples as compared to `HIS` and `UHA`.

B.3 Bayesian logistic regression

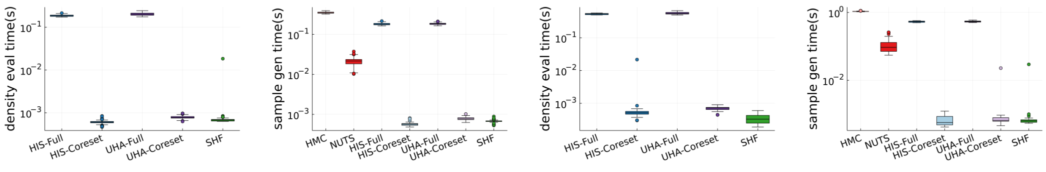
To train `SHF`, a total of 8 quasi-refreshments are used with 10 leapfrog steps in between; a similar schedule is used in `HIS` and `UHA` for momentum tempering and refreshment respectively. The initial distribution is set to $\theta_0 \sim \mathcal{N}(15, 10^{-4}I)$ and $\rho_0 \sim \mathcal{N}(0, I)$. For all methods, the number of optimization iterations is set to $100,000$, and the initial leapfrog step size is set to 0.0005 across all d dimensions. We train all methods with ADAM with initial learning rate 0.001. We also include the posterior approximation obtained from the Laplace approximation, where we search for the mode of the target density at some location generated from the same distribution as that of θ_0. Figs. 6 to 8 provide additional results for this experiment, where similar conclusions as in the Bayesian linear regression experiment can be drawn.



(a) (a) (a) (a)

(a) (a) (a) (a)

Figure 7: Linear (top row, Figs. 7a to 7d) and logistic (bottom row Figs. 7e to 7h) regression: posterior approximation quality results. The lines indicate the median, and error regions indicate 25th to 75th percentile from 5 runs.



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

Figure 8: Linear (Figs. 8a and 8b) and logistic (Figs. 8c and 8d) regression: timing results based on 100 samples.

C Gaussian KL upper bound proof

Proof of Proposition 3.1. Suppose we are given a particular choice $\mathcal{I} \subseteq [N]$ of N indices of size $|\mathcal{I}| = M \in \mathbb{N}$, $M \leq N$. Let $\mathcal{W}_\mathcal{I} = \left\{w \in \mathbb{R}_+^N : \forall n \in [N], n \notin \mathcal{I} \implies w_n = 0\right\}$. In the d-dimensional normal location model, the exact and $w \in \mathcal{W}_\mathcal{I}$-coreset posteriors are multivariate Gaussian distributions, denoted as $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_w, \Sigma_w)$ respectively, with mean and covariance

$$\Sigma_1 = \frac{1}{1+N} I, \quad \mu_1 = \Sigma_1 \left(\sum_{n=1}^N X_n\right) \quad \text{and} \quad \Sigma_w = \frac{I}{1 + \left(\sum_{n\in\mathcal{I}} w_n\right)}, \quad \mu_w = \Sigma_w \left(\sum_{n\in\mathcal{I}} w_n X_n\right).$$

The KL divergence between these two distributions is

$$\mathrm{D_{KL}}\left(\pi_w || \pi\right) = \frac{1}{2}\left[-d \log\left(\frac{1+N}{1+\sum_{n\in\mathcal{I}} w_n}\right) - d + d\left(\frac{1+N}{1+\sum_{n\in\mathcal{I}} w_n}\right) + (\mu_1 - \mu_w)^T \Sigma_1^{-1} (\mu_1 - \mu_w)\right].$$

We can bound this quantity above by adding the constraint $\sum_{n\in\mathcal{I}} w_n = N$, yielding

$$\min_{w \in \mathcal{W}_\mathcal{I}} \mathrm{D_{KL}}\left(\pi_w || \pi\right) \leq \min_{w \in \triangle^{M-1}} \frac{N^2}{2(N+1)} \left\| \bar{X} - \sum_{n\in\mathcal{I}} w_n X_n \right\|^2,$$

where $\bar{X} = \frac{1}{N}\sum_{n=1}^N X_n$. and $\triangle^{M-1}$ is the $M-1$-dimensional simplex $w \geq 0$, $1^T w = 1$. We aim to show that with high probability (over uniform random choice of $\mathcal{I}$ and realizations of $X_n, n \in [N]$), there exists a $w \in \triangle^{M-1}$ such that $\bar{X} = \sum_{n\in\mathcal{I}} w_n X_n$, and hence the optimal KL divergence is 0.

Since $X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I)$, $\sqrt{N}\bar{X} \sim \mathcal{N}(0, I)$, so $\|\sqrt{N}\bar{X}\|^2 \sim \chi^2(d)$ and therefore for any $s > \sqrt{d/N}$,

$$\mathbb{P}\left(\|\bar{X}\| > s\right) \leq \left(\frac{s^2 N}{d} e^{1 - \frac{s^2 N}{d}}\right)^{d/2}. \tag{9}$$

In other words, as N increases, we can expect $\bar{X}$ to concentrate around the origin. Therefore as long as the convex hull of $X_n, n \in \mathcal{I}$ contains a ball of some fixed radius around the origin with high probability, we know that $\bar{X}$ is a convex combination of $X_n, n \in \mathcal{I}$. The radius of the largest origin-centered ball inside the convex hull of $X_n, n \in \mathcal{I}$ can be expressed as

$$r^\star = \min_{a \in \mathbb{R}^d : \|a\|=1, b \geq 0} b \quad \text{s.t.} \quad \forall n \in \mathcal{I}, \quad a^T X_n - b \leq 0 = \min_{a \in \mathbb{R}^d : \|a\|=1} \max_{n \in \mathcal{I}} a^T X_n.$$

By Böröczky and Wintsche [37, Corollary 1.2], S^d can be covered by

$$N_d(\phi) = \frac{C \cdot \cos\phi}{\sin^d \phi} d^{\frac{3}{2}} \log(1 + d\cos^2\phi) \leq \phi^{-d} A_d, \quad A_d = Ce^{\frac{d}{2}} d^{\frac{3}{2}} \log(1+d). \tag{10}$$

balls of radius $0 < \phi \leq \arccos \frac{1}{\sqrt{d+1}}$, where C is a universal constant. Denote the centres of these balls $a_i \in S^d$, $i = 1, \ldots, N_d(\phi)$. Then

$$\begin{aligned} r^\star &\geq \min_{i \in [N_d(\phi)], v \in \mathbb{R}^d : \|v\| \leq \phi} \max_{n \in \mathcal{I}} (a_i + v)^T X_n \\ &\geq \min_{i \in [N_d(\phi)]} \max_{n \in \mathcal{I}} a_i^T X_n - \phi \|X_n\|. \end{aligned}$$

Therefore the probability that the largest origin-centred ball enclosed in the convex hull is small is bounded above by

$$\begin{aligned} \mathbb{P}(r^\star \leq t) &\leq \mathbb{P}\left(\min_{i \in [N_d(\phi)]} \max_{n \in \mathcal{I}} a_i^T X_n - \phi\|X_n\| \leq t \right) \\ &\leq N_d(\phi) \mathbb{P}\left(\max_{n \in \mathcal{I}} a_i^T X_n - \phi\|X_n\| \leq t \right) \\ &= N_d(\phi) \mathbb{P}\left(a^T Z - \phi\|Z\| \leq t \right)^M, \end{aligned}$$

for $a \in S^d$ and $Z \sim \mathcal{N}(0, I)$. Since Z has a spherically symmetric distribution, a is arbitrary, so we can choose $a = 1 \quad 0 \quad \ldots \quad 0^T$. If we let $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(d-1)$ be independent, this yields

$$\begin{aligned} \mathbb{P}\left(a^T Z - \phi\|Z\| \leq t \right) &= \mathbb{P}\left(U - \phi\sqrt{U^2 + V} \leq t \right) \\ &= \mathbb{P}\left(U - t \leq \phi\sqrt{U^2 + V} \right) \\ &\leq \mathbb{P}(U < 2t) + \mathbb{P}\left(V \geq \phi^{-2}(U-t)^2 - U^2, U \geq 2t \right) \\ &\leq \Phi(2t) + \mathbb{P}\left(V' \geq \phi^{-2} t^2 \right), \end{aligned}$$

where $V' \sim \chi^2(d)$ and $\Phi(\cdot)$ is the CDF of the standard normal. Therefore as long as $t > \phi\sqrt{d}$,

$$\begin{aligned} \mathbb{P}\left(V' \geq \phi^{-2} t^2 \right) &= \mathbb{P}\left(V' \geq d \left(\frac{t}{\phi\sqrt{d}} \right)^2 \right) \\ &\leq \left(\left(\frac{t}{\phi\sqrt{d}} \right)^2 e^{1 - \left(\frac{t}{\phi\sqrt{d}} \right)^2} \right)^{d/2}. \end{aligned}$$

We now combine the above results to show that for any $t > \phi\sqrt{d}$,

$$\mathbb{P}(r^\star \leq t) \leq N_d(\phi) \left(\Phi(2t) + d^{-\frac{d}{2}} e^{\frac{d}{2}} \phi^{-d} t^d e^{-\frac{1}{2}\phi^{-2} t^2} \right)^M. \tag{11}$$

Finally we combine the bound on the norm of $\bar{X}$ Eq. (9), the bound on $r^\star$ Eq. (11), and the covering number of S^d Eq. (10) for the final result. For any $t > \max\{\phi\sqrt{d}, \sqrt{d/N}\}$ and $\phi \leq \arccos \frac{1}{\sqrt{d+1}}$,

$$\mathbb{P}\left(\bar{X} \notin \mathrm{conv}(X_n)_{n \in \mathcal{I}} \right) \leq d^{-\frac{d}{2}} e^{\frac{d}{2}} t^d N^{d/2} e^{-\frac{1}{2} t^2 N} + \phi^{-d} A_d \left(\Phi(2t) + d^{-\frac{d}{2}} e^{\frac{d}{2}} \phi^{-d} t^d e^{-\frac{1}{2}\phi^{-2} t^2} \right)^M$$

Set $\phi = \sqrt{\frac{1}{N}}$ and $t = s\sqrt{d/N}$. As long as $N \geq 2$ and $s > 1$, we are guaranteed that $t > \max\{\phi\sqrt{d}, \sqrt{d/N}\}$ and $\phi \leq \arccos \frac{1}{\sqrt{d+1}}$ as required, and so

$$\mathbb{P}\left(\bar{X} \notin \mathrm{conv}(X_n)_{n \in \mathcal{I}} \right) \leq e^{\frac{d}{2}} s^d e^{-\frac{d}{2} s^2} + N^{\frac{d}{2}} A_d \left(\Phi\left(s\sqrt{\frac{4d}{N}} \right) + e^{\frac{d}{2}} s^d e^{-\frac{d}{2} s^2} \right)^M.$$

If we set $s = \sqrt{\log N + 1}$,

$$e^{\frac{d}{2}} s^d e^{-\frac{d}{2} s^2} = N^{-\frac{d}{2}} (\log N + 1)^{\frac{d}{2}} \quad \Phi\left(s\sqrt{\frac{4d}{N}} \right) = \frac{1}{2} + O\left(\sqrt{\frac{\log N}{N}} \right).$$

So then setting $M = d\log_2(N) - \frac{d}{2}\log_2(\log(N)) + \log_2 A_d$ yields the claimed result. □

D Gaussian KL lower bound proof

Proof of Proposition 3.2. Let $z(t) = (\theta_t, \rho_t)$. The evolution of $z(t)$ is determined by a time-inhomogeneous linear ordinary differential equation,

$$\frac{\mathrm{d}z(t)}{\mathrm{d}t} = A(t)z(t) \qquad A(t) = \begin{bmatrix} 0 & 1 \\ -\sigma^{-2} & -\gamma(t) \end{bmatrix},$$

with solution

$$z(t) = e^{B(t)}z(0) \qquad B(t) = \begin{bmatrix} 0 & t \\ -\sigma^{-2}t & g(t) \end{bmatrix}, \qquad g(t) = -\int_0^t \gamma(t)\mathrm{d}t.$$

Therefore by writing $\bar{\pi}_0 = \mathcal{N}\left(m(0), \Sigma(0)\right)$, $z(t) \sim q_t = \mathcal{N}\left(m(t), \Sigma(t)\right)$, where

$$m(t) = e^{B(t)}m(0) \qquad \Sigma(t) = e^{B(t)}\Sigma(0)e^{B(t)^T}. \tag{12}$$

The second result follows because tempered Hamiltonian dynamics where $\gamma(t) = 0$ identically is just standard Hamiltonian dynamics. For the first result, suppose $q_0 = \mathcal{N}\left(\begin{bmatrix}\mu \\ 0\end{bmatrix}, \begin{bmatrix}1 & 0 \\ 0 & \beta^2\end{bmatrix}\right)$ for some $\mu \in \mathbb{R}, \beta \in \mathbb{R}_+$. Note that in this case, it suffices to consider

$$m(0) = \begin{bmatrix}\mu \\ 0\end{bmatrix} \qquad \Sigma(0) = I. \tag{13}$$

This is because the function $\gamma(t)$ is arbitrary; one can, for example, set $\gamma(t) = -\epsilon^{-1}\log\beta$ for $t \in [0, \epsilon)$ for an arbitrarily small $\epsilon > 0$, such that the state at time ϵ is arbitrarily close to the desired initial condition. The KL divergence from q_t to $\bar{\pi}$ is

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) &= \frac{1}{2}\left[\log\left(\frac{\det\Sigma}{\det\Sigma(t)}\right) - 2 + \operatorname{tr}\left(\Sigma^{-1}\Sigma(t)\right) + m(t)^T\Sigma^{-1}m(t)\right] \\ &= \frac{1}{2}\left[\log\left(\frac{\det\Sigma}{\det\Sigma(t)}\right) - 2 + \operatorname{tr}\left(\Sigma^{-1}\left(\Sigma(t) + m(t)m(t)^T\right)\right)\right]. \end{aligned}$$

By Eqs. (12) and (13) and the identity $\operatorname{tr} A^TA = \|A\|_F^2$,

$$\begin{aligned} \mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) &= \frac{1}{2}\left[\log\left(\frac{\det\Sigma}{\det\Sigma(t)}\right) - 2 + \operatorname{tr}\left(\Sigma^{-1}e^{B(t)}\left(I + m(0)m(0)^T\right)e^{B(t)^T}\right)\right] \\ &= \frac{1}{2}\left[\log\left(\frac{\det\Sigma}{\det\Sigma(t)}\right) - 2 + \left\|\Sigma^{-\frac{1}{2}}e^{B(t)}\left(I + m(0)m(0)^T\right)^{\frac{1}{2}}\right\|_F^2\right]. \end{aligned} \tag{14}$$

From this point onward we will drop the explicit dependence on t for notational brevity. Note that B has eigendecomposition $B = VDV^{-1}$ where

$$\begin{aligned} h &= \frac{g}{2t\sigma^{-1}} \qquad \lambda_+ = h + \sqrt{h^2 - 1} \qquad \lambda_- = h - \sqrt{h^2 - 1} \\ V &= t\begin{bmatrix}1 & 0 \\ 0 & \sigma^{-1}\end{bmatrix}\begin{bmatrix}1 & 1 \\ \lambda_+ & \lambda_-\end{bmatrix} \qquad D = t\sigma^{-1}\begin{bmatrix}\lambda_+ & 0 \\ 0 & \lambda_-\end{bmatrix}. \end{aligned}$$

Also note that if $|h| < 1$, this decomposition has complex entries, but e^B is always a real matrix.

Then the log-determinant term in Eq. (14) can be written as

$$\log\left(\frac{\det\Sigma}{\det\Sigma(t)}\right) = \log\frac{\det\Sigma}{\det\Sigma(0)\det e^{2D}} = \log\frac{\sigma^2}{e^{2g}} = \log\frac{\sigma^2}{e^{4ht\sigma^{-1}}} = 2(\log\sigma - 2ht\sigma^{-1}). \tag{15}$$

By $e^B = Ve^DV^{-1}$, the squared Frobenius norm term in Eq. (14) is

$$\begin{aligned} \|\cdot\|_F^2 &:= \left\|\Sigma^{-\frac{1}{2}}e^{B(t)}\left(I + m(0)m(0)^T\right)^{\frac{1}{2}}\right\|_F^2 \\ &= \frac{\sigma^{-2}\left(e_+ - e_-\right)^2}{(\lambda_- - \lambda_+)^2}\left((1+\mu^2) + \sigma^2 + (1+\mu^2)\left(\frac{\lambda_-e_+ - \lambda_+e_-}{e_+ - e_-}\right)^2 + \sigma^2\left(\frac{\lambda_-e_- - \lambda_+e_+}{e_+ - e_-}\right)^2\right). \end{aligned}$$

where $e_+ = e^{t\sigma^{-1}\lambda_+}, e_- = e^{t\sigma^{-1}\lambda_-}$. Let

$$\sinh(t, h) = \sinh\left(t\sigma^{-1}\sqrt{h^2 - 1}\right) \qquad \cosh(t, h) = \cosh\left(t\sigma^{-1}\sqrt{h^2 - 1}\right),$$

we further simplify and get

$$\|\cdot\|_F^2 = \frac{e^{2th\sigma^{-1}}}{\sigma^2}\left(\frac{\sinh(t,h)^2}{h^2-1}(1+\mu^2+\sigma^2)+(1+\mu^2)\left(\frac{h\sinh(t,h)}{\sqrt{h^2-1}}-\cosh(t,h)\right)^2+\right.$$
$$\left.\sigma^2\left(\frac{h\sinh(t,h)}{\sqrt{h^2-1}}+\cosh(t,h)\right)^2\right). \tag{16}$$

Define $a = t\sigma^{-1}\sqrt{h^2-1}$ and $b = \frac{h}{\sqrt{h^2-1}}$ for $h \neq 1$. Then together with Eqs. (15) and (16), Eq. (14) can be written as

$$\mathrm{D}_{\mathrm{KL}}\left(q_t \| \bar{\pi}\right) = \log\sigma - 2ab - 1 +$$
$$\frac{e^{2ab}}{2\sigma^2}\left((b^2-1)\sinh(a)^2(1+\mu^2+\sigma^2)+(1+\mu^2)\left(b\sinh(a)-\cosh(a)\right)^2+\sigma^2\left(b\sinh(a)+\cosh(a)\right)^2\right). \tag{17}$$

Note that if $|h| > 1$, then $a \geq 0$ and $|b| > 1$; if $|h| < 1$, we can write $a = ia'$ for $a' \geq 0$ and $b = ib'$ for $b' \in (-\infty, -1) \cup (1, \infty)$. We now derive lower bounds for Eq. (17) over a and b under three cases: ($|h| > 1, b > 1$), ($|h| > 1, b < -1$), and ($|h| < 1$).

Case ($|h| > 1, b > 1$): Using the identity $\cosh(x)^2 - sinh(x)^2 = 1$ and $\sinh(2x) = 2\sinh(x)\cosh(x)$,

$$\begin{aligned}
&\frac{e^{2ab}}{2\sigma^2}\left((b^2-1)\sinh(a)^2\sigma^2+\sigma^2(b\sinh(a)+\cosh(a))^2\right)\\
=&\frac{e^{2ab}}{2}\left((b^2-1)\sinh(a)^2+(b\sinh(a)+\cosh(a))^2\right)\\
=&\frac{e^{2ab}}{2}\left(2b^2\sinh(a)^2+1+b\sinh(2a)\right)\\
\geq&\frac{e^{2ab}}{2}+\frac{b\sinh(2a)}{2}\\
\geq&2ab,
\end{aligned}$$

where the second last line is obtained by noting that $2b^2\sinh(a)^2 \geq 0$, $e^{2ab} \geq 2ab$, and the last line is obtained by noting that for $|h| > 1$ and $b > 1$, we have $\sinh(2a) \geq 2a$ and $e^{2ab} \geq 1$. Substituting this to Eq. (17),

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(q_t\|\bar{\pi}\right) &\geq \log\sigma - 1 + \frac{e^{2ab}}{2\sigma^2}\left((b^2-1)\sinh(a)^2(1+\mu^2)+(1+\mu^2)(b\sinh(a)-\cosh(a))^2\right)\\
&\geq \log\sigma - 1 + \frac{e^{2ab}(1+\mu^2)}{2\sigma^2}\left(2b^2\sinh(a)^2+1-b\sinh(2a)\right).
\end{aligned} \tag{18}$$

For $|h| > 1, b > 1$, we know $a \geq 0$. When $a \geq 0$ and $b > 1$,

$$\frac{\partial}{\partial a}\left(\log\sigma - 1 + \frac{e^{2ab}(1+\mu^2)}{2\sigma^2}\left(2b^2\sinh(a)^2+1-b\sinh(2a)\right)\right) \geq 0$$

and increases with a. Therefore

$$\operatorname*{arg\,min}_{a\in\mathbb{R}_+}\left(\log\sigma - 1 + \frac{e^{2ab}(1+\mu^2)}{2\sigma^2}\left(2b^2\sinh(a)^2+1-b\sinh(2a)\right)\right) = 0.$$

We also note that when we set $a = 0$, Eq. (18) is constant for all $b > 1$. Then by substituting $a = 0$ to Eq. (18), we get

$$\mathrm{D}_{\mathrm{KL}}\left(q_t\|\bar{\pi}\right) \geq \log\sigma - 1 + \frac{(1+\mu^2)}{2\sigma^2} \geq \log\frac{1+\mu^2}{4\sigma}.$$

Case ($|h| > 1, b < -1$): Note in this case, $(b^2-1)\sinh(a)^2\sigma^2 \geq 0$, $\sigma^2\left(b\sinh(a)+\cosh(a)\right)^2 \geq 0$, then we can lower bound Eq. (17) by

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(q_t\|\bar{\pi}\right) &\geq \log\sigma - 2ab - 1 + \frac{e^{2ab}}{2\sigma^2}\left((b^2-1)\sinh(a)^2(1+\mu^2)+(1+\mu^2)\left(b\sinh(a)-\cosh(a)\right)^2\right)\\
&\geq \log\sigma - 2ab - 1 + \frac{e^{2ab}(1+\mu^2)}{2\sigma^2}\left(2b^2\sinh(a)^2+1-b\sinh(2a)\right)\\
&\geq \log\sigma - 2ab - 1 + \frac{e^{2ab}(1+\mu^2)}{2\sigma^2},
\end{aligned} \tag{19}$$

where the last line is obtained by noting $2b^2\sinh(a)^2 - b\sinh(2a) \geq 0$ when $|h| > 1$ and $b < -1$. Also when $|h| > 1$ and $b < -1$, $ab \leq 0$. We then minimize Eq. (19) over $ab \leq 0$, treating ab as a single variable. The stationary point is at $(ab)^\star = \frac{1}{2}\log\frac{2\sigma^2}{1+\mu^2}$. Since Eq. (19) is convex in ab, the optimum is at $(ab)^\star$ if $\frac{2\sigma^2}{1+\mu^2} \leq 1$, and otherwise is at $(ab) = 0$. Therefore

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) &\geq \begin{cases} \log\sigma - 1 + \frac{1+\mu^2}{2\sigma^2} & \frac{2\sigma^2}{1+\mu^2} > 1 \\ \log\frac{1+\mu^2}{2\sigma} & \frac{2\sigma^2}{1+\mu^2} \leq 1 \end{cases} \\
&\geq \log\frac{1+\mu^2}{4\sigma}.
\end{aligned}$$

Case ($|h| < 1$): Write $a = ia'$ and $b = ib'$ where $a' \geq 0$ and $b' \in (-\infty, -1) \cup (1, \infty)$. Using the identities $\sinh(x) = -i\sin(ix)$ and $\cosh(x) = \cos(ix)$, we can write Eq. (17) as

$$\begin{aligned}
&\mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) = \log\sigma + 2a'b' - 1 + \\
&\frac{e^{-2a'b'}}{2\sigma^2}\left((b'^2+1)\sin(a)^2(1+\mu^2+\sigma^2) + (1+\mu^2)\left(b'\sin(a') + \cos(a')\right)^2 + \sigma^2\left(b'\sin(a') - \cos(a')\right)^2\right) \\
&= \log\sigma + 2a'b' - 1 + \frac{e^{-2a'b'}(1+\mu^2+\sigma^2)}{2\sigma^2}\left(b'^2 + 1 + b'f\sin(2a') - b'^2\cos(2a')\right),
\end{aligned}$$

where $f = \frac{1+\mu^2-\sigma^2}{1+\mu^2+\sigma^2}$. We know

$$a'^\star = \operatorname*{arg\,min}_{a' \in \mathbb{R}_+} b'^2 + 1 + b'f\sin(2a') - b'^2\cos(2a') = \frac{1}{2}\tan^{-1}\left(-\frac{f}{b'}\right) + n\pi,$$

where $n \in \mathbb{Z}$ such that $\frac{1}{2}\tan^{-1}\left(-\frac{f}{b'}\right) + n\pi \in \mathbb{R}_+$. Then by $\frac{e^{-2a'b'}(1+\mu^2+\sigma^2)}{2\sigma^2} \geq 0$,

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) &\geq \log\sigma + 2a'b' - 1 + \frac{e^{-2a'b'}(1+\mu^2+\sigma^2)}{2\sigma^2}\left(b'^2 + 1 + b'f\sin(2a'^\star) - b'^2\cos(2a'^\star)\right) \\
&= \log\sigma + 2a'b' - 1 + \frac{e^{-2a'b'}(1+\mu^2+\sigma^2)}{2\sigma^2}\left(b'^2 + 1 - b'^2\sqrt{1+(f/b')^2}\right).
\end{aligned}$$

Since $\sqrt{1+x} \leq 1 + \frac{1}{2}x$,

$$\mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) \geq \log\sigma + 2a'b' - 1 + \frac{e^{-2a'b'}(1+\mu^2+\sigma^2)}{2\sigma^2}\left(1 - \frac{1}{2}f^2\right). \tag{20}$$

The stationary point of Eq. (20) as a function in $a'b'$ is at

$$(a'b')^\star = -\frac{1}{2}\log\frac{2\sigma^2}{(1+\mu^2+\sigma^2)\left(1-\frac{1}{2}f^2\right)}.$$

Since Eq. (20) is convex in $a'b'$ and $a'b' \in \mathbb{R}$, we know the minimum of Eq. (20) is attained at $(a'b')^\star$. Substituting $(a'b')^\star$ back in Eq. (20) and noting $1 - \frac{1}{2}f^2 \geq \frac{1}{2}$, we get

$$\begin{aligned}
\mathrm{D}_{\mathrm{KL}}\left(q_t || \bar{\pi}\right) &\geq \log\frac{(1+\mu^2+\sigma^2)\left(1-\frac{1}{2}f^2\right)}{2\sigma} \\
&\geq \log\frac{1+\mu^2+\sigma^2}{4\sigma} \geq \log\frac{1+\mu^2}{4\sigma}.
\end{aligned}$$

□