Supplementary Material

A. Background on Fokker-Planck Equation

The Fokker-Planck equation (FPE) associated with a given stochastic differential equation (SDE) describes the time evolution of the distribution on the random variables under the specified stochastic dynamics. For example, consider the SDE:

$$dz = g(z)dt + \mathcal{N}(0, 2D(z)dt), \tag{16}$$

where $z \in \mathbb{R}^n, g(z) \in \mathbb{R}^n, D(z) \in \mathbb{R}^{n\times n}$. The distribution of z governed by Eq. (16) (denoted by $p_t(z)$), evolves under the following equation

$$\partial_t p_t(z) = -\sum_{i=1}^{n} \partial_{z_i}[g_i(z)p_t(z)] + \sum_{i=1}^{n}\sum_{j=1}^{n} \partial_{z_i}\partial_{z_j}[D_{ij}(z)p_t(z)].$$

Here $g_i(z)$ is the i-th entry of vector $g(z)$ and $D_{ij}(z)$ is the (i, j) entry of the matrix D. In the dynamics considered in this paper, $z = (\theta, r)$ and

$$D = \begin{bmatrix} 0 & 0 \\ 0 & B(\theta) \end{bmatrix}. \tag{17}$$

That is, the random variables are momentum r and position θ, with noise only added to r (though dependent upon θ). The FPE can be written in the following compact form:

$$\partial_t p_t(z) = -\nabla^T[g(z)p_t(z)] + \nabla^T[D(z)\nabla p_t(z)], \tag{18}$$

where $\nabla^T[g(z)p_t(z)] = \sum_{i=1}^{n} \partial_{z_i}[g_i(z)p_t(z)]$, and

$$\begin{aligned}
&\nabla^T[D\nabla p_t(\theta, r)] \\
&= \sum_{ij} \partial_{z_i}[D_{ij}(z)\partial_{z_j} p_t(z)] \\
&= \sum_{ij} \partial_{z_i}[D_{ij}(z)\partial_{z_j} p_t(z)] + \sum_{ij} \partial_{z_i}[(\partial_{z_j} D_{ij}(z))p_t(z)] \\
&= \sum_{ij} \partial_{z_i}\partial_{z_j}[D_{ij}(z)p_t(z)].
\end{aligned}$$

Note that $\partial_{z_j} D_{ij}(z) = 0$ for all i, j, since $\partial_{r_j} B_{ij}(\theta) = 0$ (the noise is only added to r and only depends on parameter θ).

B. Proof of Theorem 3.1

Let $G = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$ and $D = \begin{bmatrix} 0 & 0 \\ 0 & B(\theta) \end{bmatrix}$. The noisy Hamiltonian dynamics of Eq. (7) can be written as

$$\begin{aligned}
d\begin{bmatrix} \theta \\ r \end{bmatrix} &= -\begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}\begin{bmatrix} \nabla U(\theta) \\ M^{-1}r \end{bmatrix} dt + \mathcal{N}(0, 2Ddt) \\
&= -G\nabla H(\theta, r)dt + \mathcal{N}(0, 2Ddt).
\end{aligned}$$

Applying Eq. (18), defining $g(z) = -G\nabla H$), the corresponding FPE is given by

$$\partial_t p_t(\theta, r) = \nabla^T[G\nabla H(\theta, r)p_t(\theta, r)] + \nabla^T[D\nabla p_t(\theta, r)]. \tag{19}$$

We use $z = (\theta, r)$ to denote the joint variable of position and momentum. The entropy is defined by $h(p_t(\theta, r)) = -\int_{\theta,r} f(p_t(\theta, r))d\theta dr$. Here $f(x) = x \ln x$ is a strictly convex function defined on $(0, +\infty)$. The evolution of the entropy is governed by

$$\begin{aligned}
\partial_t h(p_t(z)) =& \partial_t \int_z f(p_t(z))dz \\
=& -\int_z f'(p_t(z))\partial_t p_t(z)dz \\
=& -\int_z f'(p_t(z))\nabla^T[G\nabla H(z)p_t(z)]dz \\
&-\int_z f'(p)\nabla^T[D(z)\nabla p_t(z)]dz.
\end{aligned}$$

The entropy evolution can be described as the sum of two parts: the noise-free Hamiltonian dynamics and the stochastic gradient noise term. The Hamiltonian dynamics part does not change the entropy, since

$$\begin{aligned}
&-\int_z f'(p_t(z))\nabla^T[G\nabla H(z)p_t]dz \\
&= -\int_z f'(p_t(z))\nabla^T[G\nabla H(z)]p_t dz \\
&\quad -\int_z f'(p_t(z))(\nabla p_t(z))^T[G\nabla H(z)]dz \\
&= -\int_z (\nabla f(p_t(z)))^T[G\nabla H(z)]dz \\
&= \int_z f(p_t(z))\nabla^T[G\nabla H(z)]dz = 0.
\end{aligned}$$

In the second equality, we use the fact that $\nabla^T[G\nabla H(z)] = -\partial_\theta\partial_r H + \partial_r\partial_\theta H = 0$. The last equality is given by integration by parts, using the assumption that the probability density vanishes at infinity and $f(x) \to 0$ as $x \to 0$ such that $f(p_t(z))[G\nabla H(z)] \to 0$ as $z \to \infty$.

The contribution due to the stochastic gradient noise can be calculated as

$$\begin{aligned}
&-\int_z f'(p_t(z))\nabla^T[D(z)\nabla p_t(z)]dz \\
&= \int_z (f''(p_t(z))\nabla p_t(z))^T D(z)\nabla p_t(z)dz \\
&= \int_{\theta,r} f''(p_t(z))(\nabla_r p_t(\theta, r))^T B(\theta)\nabla_r p_t(\theta, r)d\theta dr.
\end{aligned}$$

The first equality is again given by integration by parts, assuming that the gradient of p_t vanishes at infinity faster than $\frac{1}{\ln p_t(z)}$. That is, $f'(p_t(z))\nabla p_t(z) = (1 + \ln p_t(z))\nabla p_t(z) \to 0$ such that $f'(p_t(z))[D(z)\nabla p_t(z)] \to 0$ as $z \to \infty$. The statement of Theorem 3.1 immediately follows.

C. Proof of Corollary 3.1

Assume $\pi(\theta, r) = \exp(-H(\theta, r))/Z$ is invariant under Eq. (7) and is a well-behaved distribution such that

$H(\theta, r) \to \infty$ as $\|\theta\|, \|r\| \to \infty$. Then it is straightforward to verify that $\pi(\theta, r)$ and $\ln \pi(\theta, r) \nabla \pi(\theta, r) = \frac{1}{Z} \exp\left(-H(\theta, r)\right) \nabla H^2(\theta, r)$ vanish at infinity, such that π satisfies the conditions of Theorem 3.1. We also have $\nabla_r \pi(\theta, r) = \frac{1}{Z} \exp\left(-H(\theta, r)\right) M^{-1} r$. Using the assumption that the Fisher information matrix $B(\theta)$ has full rank, and noting that $f''(p) > 0$ for $p > 0$, from Eq. (8) of Theorem 3.1 we conclude that entropy increases over time: $\partial_t h(p_t(\theta, r))|_{p_t=\pi} > 0$. This contradicts that π is the invariant distribution.

D. FPE for Second-Order Langevin Dynamics

Second-order Langevin dynamics can be described by the following equation

$$\begin{aligned} d\begin{bmatrix} \theta \\ r \end{bmatrix} &= -\begin{bmatrix} 0 & -I \\ I & B \end{bmatrix} \begin{bmatrix} \nabla U(\theta) \\ M^{-1} r \end{bmatrix} dt + \mathcal{N}(0, 2\tau D dt) \\ &= -\left[D + G\right] \nabla H(\theta, r) dt + \mathcal{N}(0, 2\tau D dt), \end{aligned} \tag{20}$$

where τ is a temperature (usually set to 1). In this paper, we use the following compact form of the FPE to calculate the distribution evolution under Eq (20):

$$\partial_t p_t(\theta, r) = \nabla^T \{[D+G]\left[p_t(\theta, r) \nabla H(\theta, r) + \tau \nabla p_t(\theta, r)\right]\}. \tag{21}$$

To derive this FPE, we apply Eq. (18) to Eq (20), defining $g(z) = -(D + G)\nabla H$, which yields

$$\partial_t p_t(\theta, r) = \nabla^T \{[D+G]\left[\nabla H(\theta, r) p_t(\theta, r)\right]\} + \nabla^T \left[\tau D \nabla p_t(\theta, r)\right].$$

Using the fact that $\nabla^T \left[G \nabla p_t(\theta, r)\right] = -\partial_\theta \partial_r p_t(\theta, r) + \partial_r \partial_\theta p_t(\theta, r) = 0$, we get Eq. (21). This form of the FPE allows easy verification that the stationary distribution is given by $\pi(\theta, r) \propto e^{-\frac{1}{\tau} H(\theta, r)}$. In particular, if we substitute the target distribution into Eq. (21), we note that

$$\left[e^{-\frac{1}{\tau} H(\theta, r)} \nabla H(\theta, r) + \tau \nabla e^{-\frac{1}{\tau} H(\theta, r)}\right] = 0$$

such that $\partial_t \pi(\theta, r) = 0$, implying that π is indeed the stationary distribution.

The compact form of Eq. (21) can also be used to construct other stochastic processes with the desired invariant distribution. A generalization of the FPE in Eq. (21) is given by Yin & Ao (2006). The system we have discussed in this paper considers cases where $G = \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix}$ and D only depends on θ. In practice, however, it might be helpful to make G depend on θ as well. For example, to make use of the Riemann geometry of the problem, as in Girolami & Calderhead (2011) and Patterson & Teh (2013), by adapting G according to the local curvature. For us to consider these more general cases, a correction term needs to be added during simulation (Shi et al., 2012). With that correction term, we still maintain the desired target distribution as the stationary distribution.

E. Reversibility of SGHMC Dynamics

The dynamics of SGHMC are not reversible in the conventional definition of reversibility. However, the dynamics satisfy the following property:

Theorem E.1. *Assume $P(\theta_t, r_t|\theta_0, r_0)$ is the distribution governed by dynamics in Eq. (20), i.e. $P(\theta_t, r_t|\theta_0, r_0)$ follows Eq. (21), then for $\pi(\theta, r) \propto \exp(-H(\theta, r))$,*

$$\pi(\theta_0, r_0) P(\theta_t, r_t|\theta_0, r_0) = \pi(\theta_t, -r_t) P(\theta_0, -r_0|\theta_t, -r_t). \tag{22}$$

Proof. Assuming π is the stationary distribution and P^* the reverse-time Markov process associated with P: $\pi(\theta_0, r_0) P(\theta_t, r_t|\theta_0, r_0) = \pi(\theta_t, r_t) P^*(\theta_0, r_0|\theta_t, r_t)$. Let $\mathcal{L}(p) = \nabla^T \{[D + G]\left[p \nabla H(\theta, r) + \tau \nabla p\right]\}$ be the generator of Markov process described by Eq. (21). The generator of the reverse process is given by $\mathcal{L}^*$, which is the adjoint operator of $\mathcal{L}$ in the inner-product space $l^2(\pi)$, with inner-product defined by $\langle p, q \rangle_\pi = E_{x \sim \pi(x)}[p(x) q(x)]$. We can verify that $\mathcal{L}^*(p) = \nabla^T \{[D - G]\left[p \nabla H(\theta, r) + \tau \nabla p\right]\}$. The corresponding SDE of the reverse process is given by

$$d\begin{bmatrix} \theta \\ r \end{bmatrix} = \left[D - G\right] \nabla H(\theta, r) + \mathcal{N}(0, 2\tau D dt),$$

which is equivalent to

$$d\begin{bmatrix} \theta \\ -r \end{bmatrix} = \left[D + G\right] \nabla H(\theta, -r) + \mathcal{N}(0, 2\tau D dt).$$

This means $P^*(\theta_0, r_0|\theta_t, r_t) = P(\theta_0, -r_0|\theta_t, -r_t)$. Recalling that we assume Gaussian momentum, r, centered about 0, we also have $\pi(\theta, r) = \pi(\theta, -r)$. Together, we then have

$$\begin{aligned} \pi(\theta_0, r_0) P(\theta_t, r_t|\theta_0, r_0) &= \pi(\theta_t, r_t) P^*(\theta_0, r_0|\theta_t, r_t) \\ &= \pi(\theta_t, -r_t) P(\theta_0, -r_0|\theta_t, -r_t). \end{aligned}$$

□

Theorem E.1 is not strictly detailed balance by the conventional definition since $\mathcal{L}^* \neq \mathcal{L}$ and $P^* \neq P$. However, it can be viewed as a kind of time reversibility. When we reverse time, the sign of speed needs to be reversed to allow backward travel. This property is shared by the noise-free HMC dynamics of (Neal, 2010). Detailed balance can be enforced by the symmetry of r during the re-sampling step. However, we note that we do not rely on detailed balance to have π be the stationary distribution of our noisy Hamiltonian with friction (see Eq. (9)).

F. Convergence Analysis

In the paper, we have discussed that the efficiency of SGHMC decreases as the step size ϵ decreases. In practice,

we usually want to trade a small amount of error for efficiency. In the case of SGHMC, we are interested in a small, nonzero ϵ and fast approximation of B given by $\hat{B}$. In this case, even under the continuous dynamics, the sampling procedure contains error that relates to ϵ due to inaccurate estimation of B with $\hat{B}$. In this section, we investigate how the choice of ϵ can be related to the error in the final stationary distribution. The sampling procedure with inaccurate estimation of B can be described with the following dynamics

$$\begin{cases} d\theta = M^{-1}r\ dt \\ dr = -\nabla U(\theta)\ dt - CM^{-1}r dt + \mathcal{N}(0, 2(C+\delta S)dt). \end{cases}$$

Here, $\delta S = B - \hat{B}$ is the error term that is not considered by the sampling algorithm. Assume the setting where $\hat{B} = 0$, then we can let $\delta = \epsilon$ and $S = \frac{1}{2}V$. Let $\tilde{\pi}$ be the stationary distribution of the dynamics. In the special case when $V = C$, we can calculate $\tilde{\pi}$ exactly by

$$\tilde{\pi}(\theta, r) \propto \exp\left(-\frac{1}{1+\delta}H(\theta, r)\right). \tag{23}$$

This indicates that for small ϵ, our stationary distribution is indeed close to the true stationary distribution. In general case, we consider the FPE of the distribution of this SDE, given by

$$\partial_t \tilde{p}_t(\theta, r) = [\mathcal{L} + \delta\mathcal{S}]\tilde{p}_t(\theta, r). \tag{24}$$

Here, $\mathcal{L}(p) = \nabla^T\{[D+G]\,[p\nabla H(\theta, r) + \nabla p]\}$ is the operator corresponds to correct sampling process. Let the operator $\mathcal{S}(p) = \nabla_r[S\nabla_r p]$ correspond to the error term introduced by inaccurate $\hat{B}$. Let us consider the χ^2-divergence defined by

$$\chi^2(p, \pi) = E_{x\sim\pi}\left[\frac{(p(x)-\pi(x))^2}{\pi^2(x)}\right] = E_{x\sim\pi}\left[\frac{p^2(x)}{\pi^2(x)}\right] - 1,$$

which provides a measure of distance between the distribution p and the true distribution π. Theorem F.1 shows that the χ^2-divergence decreases as δ becomes smaller.

Theorem F.1. *Assume p_t evolves according to $\partial_t p_t = \mathcal{L}p_t$, and satisfies the following mixing rate λ with respect to χ^2 divergence at $\tilde{\pi}$: $\partial_t\chi^2(p_t, \pi)|_{p_t=\tilde{\pi}} \le -\lambda\chi^2(\tilde{\pi}, \pi)$. Further assume the process governed by $\mathcal{S}$ ($\partial_t q_t = \mathcal{S}q_t$) has bounded divergence change $|\partial_t\chi^2(q_t, \pi)| < c$. Then $\tilde{\pi}$ satisfies*

$$\chi^2(\tilde{\pi}, \pi) < \frac{\delta c}{\lambda}. \tag{25}$$

Proof. Consider the divergence change of $\tilde{p}$ governed by Eq.(24). It can be decomposed into two components, the change of divergence due to $\mathcal{L}$, and the change of divergence due to $\delta\mathcal{S}$

$$\begin{aligned} \partial_t\chi^2(\tilde{p}_t, \pi) =& E_{x\sim\pi}\left[\frac{\tilde{p}(x)}{\pi^2(x)}[\mathcal{L}+\delta\mathcal{S}]\tilde{p}_t(x)\right] \\ =& E_{x\sim\pi}\left[\frac{\tilde{p}_t(x)}{\pi^2(x)}\mathcal{L}\tilde{p}_t(x)\right] + \delta E_{x\sim\pi}\left[\frac{\tilde{p}(x)}{\pi^2(x)}\mathcal{S}\tilde{p}_t(x)\right] \\ =& \partial_t\chi^2(p_t, \pi)|_{p_t=\tilde{p}_t} + \delta\partial_t\chi^2(q_t, \pi)|_{q_t=\tilde{p}_t}. \end{aligned}$$

We then evaluate the above equation at the stationary distribution of the inaccurate dynamics $\tilde{\pi}$. Since $\partial_t\chi^2(\tilde{p}_t, \pi)|_{\tilde{p}=\tilde{\pi}} = 0$, we have

$$\lambda\chi^2(\tilde{\pi}, \pi) = \delta\left|(\partial_t\chi^2(q_t, \pi)|_{q_t=\tilde{\pi}})\right| < \delta c.$$

□

This theorem can also be used to measure the error in SGLD, and justifies the use of small finite step sizes in SGLD. We should note that the mixing rate bound λ at $\tilde{\pi}$ exists for SGLD and can be obtained using spectral analysis (Levin et al., 2008), but the corresponding bounds for SGHMC are unclear due to the irreversibility of the process. We leave this for future work.

Our proof relies on a contraction bound relating the error in the transition distribution to the error in the final stationary distribution. Although our argument is based on a continuous-time Markov process, we should note that a similar guarantee can also be proven in terms of a discrete-time Markov transition kernel. We refer the reader to (Korattikara et al., 2014) and (Bardenet et al., 2014) for further details.

G. Setting SGHMC Parameters

As we discussed in Sec. 3.3, we can connect SGHMC with SGD with momentum by rewriting the dynamics as (see Eq.(15))

$$\begin{cases} \Delta\theta = v \\ \Delta v = -\eta\nabla\tilde{U}(x) - \alpha v + \mathcal{N}(0, 2(\alpha - \hat{\beta})\eta). \end{cases}$$

In analogy to SGD with momentum, we call η the learning rate and $1-\alpha$ the momentum term. This equivalent update rule is cleaner and we recommend parameterizing SGHMC in this form.

The $\hat{\beta}$ term corresponds to the estimation of noise that comes from the gradient. One simple choice is to ignore the gradient noise by setting $\hat{\beta} = 0$ and relying on small ϵ. We can also set $\hat{\beta} = \eta\hat{V}/2$, where $\hat{V}$ is estimated using empirical Fisher information as in (Ahn et al., 2012).

There are then three parameters: the learning rate η, momentum decay α, and minibatch size $|\tilde{\mathcal{D}}|$. Define $\beta = \epsilon M^{-1}B = \frac{1}{2}\eta V(\theta)$ to be the exact term induced by introduction of the stochastic gradient. Then, we have

$$\beta = O\left(\eta\frac{|\mathcal{D}|}{|\tilde{\mathcal{D}}|}\mathcal{I}\right), \tag{26}$$

where $\mathcal{I}$ is fisher information matrix of the gradient, $|\mathcal{D}|$ is size of training data, $|\tilde{\mathcal{D}}|$ is size of minibatch, and η is our learning rate. We want to keep β small so that the resulting dynamics are governed by the user-controlled term and the

sampling algorithm has a stationary distribution close to the target distribution. From Eq. (26), we see that there is no free lunch here: as the training size gets bigger, we can either set a small learning rate $\eta = O(\frac{1}{|\mathcal{D}|})$ or use a bigger minibatch size $|\tilde{\mathcal{D}}|$. In practice, choosing $\eta = O(\frac{1}{|\mathcal{D}|})$ gives better numerical stability, since we also need to multiply η by $\nabla\tilde{U}$, the mean of the stochastic gradient. Large η can cause divergence, especially when we are not close to the mode of distribution. We note that the same discussion holds for SGLD (Welling & Teh, 2011).

In practice, we find that using a minibatch size of hundreds (e.g $|\tilde{\mathcal{D}}| = 500$) and fixing α to a small number (e.g. 0.01 or 0.1) works well. The learning rate can be set as $\eta = \gamma/|\mathcal{D}|$, where γ is the "per-batch learning rate", usually set to 0.1 or 0.01. This method of setting parameters is also commonly used for SGD with momentum (Sutskever et al., 2013).

H. Experimental Setup

H.1. Bayesian Neural Network

The Bayesian neural network model used in Sec. 4.2 can be described by the following equation:

$$P(y = i|x) \propto \exp\left(A_i^T \sigma(B^T x + b) + a_i\right). \tag{27}$$

Here, $y \in \{1, 2, \cdots, 10\}$ is the output label of a digit. $A \in \mathbb{R}^{10\times 100}$ contains the weight for output layers and we use A_i to indicate i-th column of A. $B \in \mathbb{R}^{d\times 100}$ contains the weight for the first layer. We also introduce $a \in \mathbb{R}^{10}$ and $b \in \mathbb{R}^{100}$ as bias terms in the model. In the MNIST dataset, the input dimension $d = 784$. We place a Gaussian prior on the model parameters

$$P(A) \propto \exp(-\lambda_A \|A\|^2), P(B) \propto \exp(-\lambda_B \|B\|^2)$$
$$P(a) \propto \exp(-\lambda_a \|a\|^2), P(b) \propto \exp(-\lambda_b \|a\|^2).$$

We further place gamma priors on each of the precision terms λ:

$$\lambda_A, \lambda_B, \lambda_a, \lambda_b \overset{i.i.d.}{\sim} \Gamma(\alpha, \beta).$$

We simply set α and β to 1 since the results are usually insensitive to these parameters. We generate samples from the posterior distribution

$$P(\Theta|\mathcal{D}) \propto \prod_{y,x\in\mathcal{D}} P(y|x, \Theta)P(\Theta), \tag{28}$$

where parameter set $\Theta = \{A, B, a, b, \lambda_A, \lambda_B, \lambda_a, \lambda_b\}$. The sampling procedure is carried out by alternating the following steps:

- Sample weights from $P(A, B, a, b|\lambda_A, \lambda_B, \lambda_a, \lambda_b, \mathcal{D})$ using SGHMC or SGLD with minibatch of 500 instances. Sample for 100 steps before updating hyper-parameters.
- Sample λ from $P(\lambda_A, \lambda_B, \lambda_a, \lambda_b|A, B, a, b)$ using a Gibbs step. Note that the posterior for λ is a gamma distribution by conditional conjugacy.

We used the validation set to select parameters for the various methods we compare. Specifically, for SGD and SGLD, we tried step-sizes $\epsilon \in \{0.1, 0.2, 0.4, 0.8\} \times 10^{-4}$, and the best settings were found to be $\epsilon = 0.1 \times 10^{-4}$ for SGD and $\epsilon = 0.2 \times 10^{-4}$ for SGLD. We then further tested $\epsilon = 0.16 \times 10^{-4}$ and $\epsilon = 0.06 \times 10^{-4}$ for SGD, and found $\epsilon = 0.16 \times 10^{-4}$ gave the best result, thus we used this setting for SGD. For SGD with momentum and SGHMC, we fixed $\alpha = 0.01$ and $\hat{\beta} = 0$, and tried $\eta \in \{0.1, 0.2, 0.4, 0.8\} \times 10^{-5}$. The best settings were $\eta = 0.4 \times 10^{-5}$ for SGD with momentum, and $\eta = 0.2 \times 10^{-5}$ for SGHMC. For the optimization-based methods, we use tried regularizer $\lambda \in \{0, 0.1, 1, 10, 100\}$, and $\lambda = 1$ was found to give the best performance.

H.2. Online Bayesian Probabilistic Matrix Factorization

The Bayesian probabilistic matrix factorization (BPMF) model used in Sec. 4.3 can be described as:

$$\begin{aligned} \lambda_U, \lambda_V, \lambda_a, \lambda_b &\overset{i.i.d.}{\sim} \text{Gamma}(1, 1) \\ U_{ki} \sim&\mathcal{N}(0, \lambda_U^{-1}), V_{kj} \sim \mathcal{N}(0, \lambda_V^{-1}), \\ a_i \sim&\mathcal{N}(0, \lambda_a^{-1}), b_i \sim \mathcal{N}(0, \lambda_b^{-1}) \\ Y_{ij}|U, V \sim&\mathcal{N}(U_i^T V_j + a_i + b_j, \tau^{-1}). \end{aligned} \tag{29}$$

The $U_i \in \mathbb{R}^d$ and $V_j \in \mathbb{R}^d$ are latent vectors for user i and movie j, while a_i and b_j are bias terms. We use a slightly simplified model than the BPMF model considered in (Salakhutdinov & Mnih, 2008a), where we only place priors on precision variables $\lambda = \{\lambda_U, \lambda_V, \lambda_a, \lambda_b\}$. However, the model still benefits from Bayesian inference by integrating over the uncertainty in the crucial regularization parameter λ. We generate samples from the posterior distribution

$$P(\Theta|Y) \propto P(Y|\Theta)P(\Theta), \tag{30}$$

with the parameter set $\Theta = \{U, V, a, b, \lambda_U, \lambda_V, \lambda_a, \lambda_b\}$. The sampling procedure is carried out by alternating the followings

- Sample weights from $P(U, V, a, b|\lambda_U, \lambda_V, \lambda_a, \lambda_b, Y)$ using SGHMC or SGLD with a minibatch size of 4,000 ratings. Sample for $2,000$ steps before updating the hyper-parameters.
- Sample λ from $P(\lambda_U, \lambda_V, \lambda_a, \lambda_b|U, V, a, b)$ using a Gibbs step.

The training parameters for this experiment were directly selected using cross-validation. Specifically, for SGD and

SGLD, we tried step-sizes $\epsilon \in \{0.1, 0.2, 0.4, 0.8, 1.6\} \times 10^{-5}$, and the best settings were found to be $\epsilon = 0.4 \times 10^{-5}$ for SGD and $\epsilon = 0.8 \times 10^{-5}$ for SGLD. For SGD with momentum and SGHMC, we fixed $\alpha = 0.05$ and $\hat{\beta} = 0$, and tried $\eta \in \{0.1, 0.2, 0.4, 0.8\} \times 10^{-6}$. The best settings were $\eta = 0.4 \times 10^{-6}$ for SGD with momentum, and $\eta = 0.4 \times 10^{-6}$ for SGHMC.