Appendix

A Background of SGHMC

SGHMC, Algorithm 1, reduces the computational cost of full-batch methods by using a stochastic gradient in lieu of ∇U. SGHMC estimates $U(\theta)$ by minibatch $\tilde{\mathcal{D}}$. However, naïvely replacing U by $\tilde{U}$ in HMC will lead to divergence from the target distribution (Chen et al., 2014). Therefore, SGHMC adds the additional friction term of L2MC to offset the noise introduced by using minibatch. That is, SGHMC is simulating the dynamics

$$d\theta = \sigma^{-2} r dt,$$

$$dr = -\nabla \tilde{U}(\theta) dt - 2\beta r dt + \mathcal{N}(0, 4(\beta - \hat{\beta})\sigma^2 dt) \tag{7}$$

where β controls the friction term and $\hat{\beta}$ is an estimate of the stochastic gradient noise, which is often set to be zero in practice.

B Proof of Results in Section 4

In this section, we will provide proofs of the results that we asserted in Section 4 about reversibility and skew-reversibility of our algorithms. First, for completeness we re-prove the fact that a skew-reversible chain has stationary distribution π, which is known, but not as well-known as the corresponding result for reversible chains.

Lemma 1. *If G is a skew-reversible chain, that is one that satisfies (5), then π is its stationary distribution.*

Proof. Since G is skew-reversibile, by definition it satisfies for any states x and y the conditions that $\pi(x) = \pi(x^\perp)$ and

$$\pi(x)G(x, y) = \pi(y^\perp)G(y^\perp, x^\perp).$$

By combining these two we can easily get

$$\pi(x)G(x, y) = \pi(y)G(y^\perp, x^\perp).$$

Next, summing up over all x in the while state space Ω,

$$\sum_{x \in \Omega} \pi(x)G(x, y) = \sum_{x \in \Omega} \pi(y)G(y^\perp, x^\perp) = \pi(y) \sum_{x \in \Omega} G(y^\perp, x^\perp).$$

Since $\perp$ denotes an involution, it follows that summing up over x for all x in the state space is equal to summing up over all $x^\perp$, so

$$\sum_{x \in \Omega} \pi(x)G(x, y) = \pi(y) \sum_{x \in \Omega} G(y^\perp, x) = \pi(y),$$

where the last equality follows from the fact that for any Markov chain, the sum of the probabilities of transitioning into all states is always 1. So, we've shown that

$$\sum_{x \in \Omega} \pi(x)G(x, y) = \pi(y)$$

which can be written in matrix form as $\pi G = \pi$; this proves the lemma. □

Lemma 2. *The amortized Metropolis-Hastings procedure described in Section 4 using acceptance probability (3) results in a chain that is reversible with stationary distribution π.*

Proof. According to the algorithm described in Section 4, the probability density of transitioning from state x to state $y \neq x$ via intermediate states $x = x_0, x_1, x_2, \ldots, x_{T-1}, x_T = y$ is

$$\mathbf{E}\left[\tau \cdot \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t)\right],$$

where the expected value here is taken over the randomness used to select the stochastic samples ζ_t. This follows from the law of total expectation. This means that the total probability of transitioning from x to $y \neq x$ is just the integral of this over the intermediate states

$$G(x, y) = \int \mathbf{E}\left[\tau \cdot \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1},$$

Now substituting in the value of τ from (3) gives us

$$\begin{aligned} G(x, y) &= \int \mathbf{E}\left[\min\left(1, \frac{\pi(y)}{\pi(x)} \prod_{t=0}^{T-1} \frac{P(x_{t+1}, x_t; \zeta_t)}{P(x_t, x_{t+1}; \zeta_t)}\right) \cdot \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1} \\ &= \int \mathbf{E}\left[\min\left(\prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t), \frac{\pi(y)}{\pi(x)} \prod_{t=0}^{T-1} P(x_{t+1}, x_t; \zeta_t)\right)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1}. \end{aligned}$$

Multiplying both sides by $\pi(x)$,

$$\pi(x) G(x, y) = \int \mathbf{E}\left[\min\left(\pi(x) \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t), \pi(y) \prod_{t=0}^{T-1} P(x_{t+1}, x_t; \zeta_t)\right)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1}.$$

From here, the fact that G is reversible follows directly from a substitution of $x_t \mapsto x_{T-t}$ in the integral, combined with the observation that the ζ_t are i.i.d. and so exchangeable. □

Lemma 3. *The amortized Metropolis-Hastings procedure for skew-reversible chains described in Section 4 using acceptance probability (6) results in a chain that is skew-reversible with stationary distribution π, as long as π satisfies $\pi(x) = \pi(x^\perp)$.*

Proof. As above, the probability density of transitioning from state x to state $y \neq x$ via intermediate states $x = x_0, x_1, x_2, \ldots, x_{T-1}, x_T = y$ is

$$\mathbf{E}\left[\tau \cdot \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t)\right],$$

where the expected value here is taken over the randomness used to select the stochastic samples ζ_t. This follows from the law of total expectation. This means that the total probability of transitioning from x to $y \neq x$ is just the integral of this over the intermediate states

$$G(x, y) = \int \mathbf{E}\left[\tau \cdot \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1},$$

Now substituting in the value of τ from (3) gives us

$$\begin{aligned} G(x, y) &= \int \mathbf{E}\left[\min\left(1, \frac{\pi(y)}{\pi(x)} \prod_{t=0}^{T-1} \frac{P(x_{t+1}^\perp, x_t^\perp; \zeta_t)}{P(x_t, x_{t+1}; \zeta_t)}\right) \cdot \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1} \\ &= \int \mathbf{E}\left[\min\left(\prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t), \frac{\pi(y)}{\pi(x)} \prod_{t=0}^{T-1} P(x_{t+1}^\perp, x_t^\perp; \zeta_t)\right)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1}. \end{aligned}$$

Multiplying both sides by $\pi(x)$, and leveraging the fact that $\pi(x) = \pi(x^\perp)$,

$$\pi(x) G(x, y) = \int \mathbf{E}\left[\min\left(\pi(x) \prod_{t=0}^{T-1} P(x_t, x_{t+1}; \zeta_t), \pi(y^\perp) \prod_{t=0}^{T-1} P(x_{t+1}^\perp, x_t^\perp; \zeta_t)\right)\right] \, dx_1 \cdot dx_2 \cdots dx_{T-1}.$$

From here, the fact that G is skew-reversible follows directly from a substitution of $x_t \mapsto x_{T-t}^\perp$ in the integral (which is a valid substitution without introducing an extra constant term because the involution $\perp$ is measure-preserving by assumption), combined with the observation that the ζ_t are i.i.d. and so exchangeable. □

Lemma 4. *A skew-reversible chain will become reversible by resampling the momentum at the beginning of outer loop.*

Proof. Assume the chain starts at (θ, r) and ends at (θ^*, r^*). By the skew-detailed balance, we have

$$\pi(\theta, r)G((\theta, r), (\theta^*, r^*)) = \pi(\theta^*, -r^*)G((\theta^*, -r^*), (\theta, -r))$$

Since the momentum is resampled and is independent of θ, we can integrate it and describe the chain in terms of θ

$$\pi(\theta)H(\theta, \theta^*) := \int \pi(\theta)\pi(r)G((\theta, r), (\theta^*, r^*))drdr^*$$

Similarly, we have

$$\pi(\theta^*)H(\theta^*, \theta) := \int \pi(\theta^*)\pi(-r^*)G((\theta^*, -r^*), (\theta, -r))dr^*dr$$

By the skew-detailed balance we know

$$\pi(\theta)H(\theta, \theta^*) = \pi(\theta^*)H(\theta^*, \theta)$$

This proves the lemma. □

C Proof of Theorem 1

Proof. First, we consider resampling momentum in Algorithm 2 and will show that the chain is reversible. We consider the probability of starting from θ and going through a particular sequence of $r_{t+\frac{1}{2}}$ and θ_t and arriving at (θ^*, r^*). We have $G(\theta, \theta^*)$, which is the transition probability from θ to θ^*, as the following

$$G(\theta, \theta^*) = \mathbf{E}\int \mathbf{P}\left(\theta_0, \theta_1, \dots, \theta_{T-1}, \theta^* \middle| \theta, \tilde{U}_0, \dots, \tilde{U}_{T-1}\right) \min(1, a(\boldsymbol{\theta}))d\theta_0 \cdots d\theta_{T-1}$$

where $\boldsymbol{\theta} = \{\theta_0, \dots, \theta_{T-1}, \theta^*\}$ and the expectation is taken over the stochastic energy function samples $\tilde{U}_t$.

Next, we want to derive the probability density in terms of r and $\boldsymbol{\eta} = \{\eta_0, \dots, \eta_{T-1}\}$. This involves a change of variables in the PDF formula. We notice that $\boldsymbol{\theta}$ is a bijective function of r and $\boldsymbol{\eta}$. By the rule of change of variables, we know that

$$\begin{aligned} &\mathbf{P}\left(\theta_0, \theta_1, \dots, \theta_{T-1}, \theta^* \middle| \theta, \tilde{U}_0, \dots, \tilde{U}_{T-1}\right) \min(1, a(\boldsymbol{\theta})) \\ &= \mathbf{P}\left(r, \eta_0, \dots, \eta_{T-1} \middle| \theta, \tilde{U}_0, \dots, \tilde{U}_{T-1}\right) \min(1, a(\boldsymbol{\eta}, r)))det^{-1}(D_{(\boldsymbol{\eta}, r)}(\boldsymbol{\theta})) \end{aligned}$$

where $D_{(\boldsymbol{\eta}, r)}(\boldsymbol{\theta}, \theta^*)$ is the Jacobian matrix.

To get this Jacobian matrix, we first apply the chain rule,

$$D_{(\boldsymbol{\eta}, r)}(\boldsymbol{\theta}) = D_{\boldsymbol{r}}(\boldsymbol{\theta}) \cdot D_{(\boldsymbol{\eta}, r)}\boldsymbol{r}$$

where $\boldsymbol{r} = \{r, r, \dots, r_{T-\frac{1}{2}}\}$.

Since the derivative of θ_t with respect to any $r_{s-\frac{1}{2}}$ for $s > t$ is zero, it follows that $D_{\boldsymbol{r}}(\boldsymbol{\theta})$ will be triangular, and so the determinant is just the product of the diagonal entries. From our formula for the update rule,

$$\begin{aligned} \theta_t &= \theta + \frac{1}{2}\epsilon\sigma^{-2}r_{t-\frac{1}{2}}, \text{ for } t = 0, T \\ \theta_t &= \theta_{t-1} + \epsilon\sigma^{-2}r_{-\frac{1}{2}}, \text{ for } t = 1, \dots, T-1. \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial\theta_0}{\partial r_{t-\frac{1}{2}}} &= \frac{1}{2}\epsilon\sigma^{-2}I_d, \text{ for } t = 0, T \\ \frac{\partial\theta_t}{\partial r_{t-\frac{1}{2}}} &= \epsilon\sigma^{-2}I_d, \text{ for } t = 1, \dots, T-1. \end{aligned}$$

It follows that

$$\det\left(D_{\boldsymbol{r}}(\boldsymbol{\theta})\right) = \frac{1}{4^d}\left(\epsilon\sigma^{-2}\right)^{(T+1)d}.$$

Similarly, the derivative of η_t with respect to any $r_{s-\frac{1}{2}}$ for $s > t$ is zero, it follows that $D_{(\boldsymbol{\eta},r)}\boldsymbol{r}$ will be triangular, and so the determinant is just the product of the diagonal entries. That is,

$$D_{\boldsymbol{r}}(\boldsymbol{\eta}, r)) = \prod_{t=0}^{T-1} \frac{\partial \eta_t}{\partial r_{t+\frac{1}{2}}}.$$

From our original formula for the update rule,

$$r_{t+\frac{1}{2}} = r_{t-\frac{1}{2}} - \epsilon\nabla\tilde{U}_t(\theta_t) - \epsilon\beta\left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right) + \eta_t$$

we have

$$(1+\epsilon\beta)r_{t+\frac{1}{2}} = r_{t-\frac{1}{2}} - \epsilon\nabla\tilde{U}_t(\theta_t) - \epsilon\beta r_{t-\frac{1}{2}} + \eta_t,$$

and so

$$\frac{\partial \eta_t}{\partial r_{t+\frac{1}{2}}} = (1+\epsilon\beta)I_d.$$

It follows that

$$\det\left(D_{(\boldsymbol{\eta},r)}\boldsymbol{r}\right) = (1+\epsilon\beta)^{-Td}$$

Now we can get that

$$\begin{aligned}\det\left(D_{(\boldsymbol{\eta},r)}(\boldsymbol{\theta})\right) &= \det\left(D_{\boldsymbol{r}}(\boldsymbol{\theta})\right)\cdot\det\left(D_{(\boldsymbol{\eta},r)}\boldsymbol{r}\right)\\ &= (1+\epsilon\beta)^{-Td}\cdot\frac{1}{4^d}\left(\epsilon\sigma^{-2}\right)^{(T+1)d}\end{aligned}$$

Thus,

$$\begin{aligned}G(\theta,\theta^*) &= \mathbf{E}\int \mathbf{P}\left(\theta_0,\theta_1,\ldots,\theta_{T-1},\theta^*\middle|\theta,\tilde{U}_0,\ldots,\tilde{U}_{T-1}\right)\min(1,a(\boldsymbol{\theta}))d\theta_0\cdots d\theta_{T-1}\\ &= (1+\epsilon\beta)^{Td}\cdot 4^d\left(\epsilon\sigma^{-2}\right)^{-(T+1)d}\,\mathbf{E}\int\mathbf{P}\left(r,\eta_0,\ldots,\eta_{T-1}\middle|\theta,\tilde{U}_0,\ldots,\tilde{U}_{T-1}\right)\\ &\quad\min(1,a(\boldsymbol{\eta},r)))d\theta_0\cdots d\theta_{T-1}\end{aligned}$$

By the distribution of r and η_t, we know that

$$\begin{aligned}&\mathbf{P}\left(r,\eta_0,\eta_2,\ldots,\eta_{T-2}\middle|\theta,\tilde{U}_0,\ldots,\tilde{U}_{T-1}\right)\\ &= \left(2\pi\sigma^2\right)^{\frac{-d}{2}}\cdot\exp\left(-\frac{\|r\|^2}{2\sigma^2}\right)\cdot\prod_{t=0}^{T-1}\left(8\pi\epsilon\beta\sigma^2\right)^{\frac{-d}{2}}\cdot\exp\left(-\frac{\|\eta_t\|^2}{8\epsilon\beta\sigma^2}\right)\\ &= \left(2\pi\sigma^2\right)^{\frac{-d}{2}}\cdot\left(8\pi\epsilon\beta\sigma^2\right)^{\frac{-Td}{2}}\cdot\exp\left(-\frac{\|r\|^2}{2\sigma^2}\right)\cdot\exp\left(-\frac{1}{8\epsilon\beta\sigma^2}\sum_{t=0}^{T-1}\|\eta_t\|^2\right).\end{aligned}$$

Notice that

$$\begin{aligned}
\sum_{t=0}^{T-1} \|\eta_t\|^2 &= \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) + \epsilon\beta \left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right) \right\|^2 \\
&= \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 \\
&\qquad + 2\epsilon\beta \left(r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right)^T \left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right) + \epsilon^2\beta^2 \left\| r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right\|^2 \\
&= \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 + 2\epsilon\beta \left(\left\| r_{t+\frac{1}{2}} \right\|^2 - \left\| r_{t-\frac{1}{2}} \right\|^2 \right) \\
&\qquad + 2\epsilon^2\beta \nabla \tilde{U}_t(\theta_t)^T \left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right) + \epsilon^2\beta^2 \left\| r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right\|^2 \\
&= 2\epsilon\beta \left(\left\| r_{T-\frac{1}{2}} \right\|^2 - \|r\|^2 \right) + \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 \\
&\qquad + 4\epsilon\beta\sigma^2 \left(\rho_{t+\frac{1}{2}} - \rho_{t-\frac{1}{2}} \right) + \epsilon^2\beta^2 \left\| r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right\|^2 \\
&= 2\epsilon\beta \left(\|r^*\|^2 - \|r\|^2 \right) + 4\epsilon\beta\sigma^2 \left(\rho_{T-\frac{1}{2}} - \rho_{-\frac{1}{2}} \right) \\
&\qquad + \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 + \epsilon^2\beta^2 \left\| r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right\|^2 .
\end{aligned}$$

By substituting this above and recalling that $\rho_{-\frac{1}{2}} = 0$,

$$\begin{aligned}
G(\theta, \theta^*) &= (1+\epsilon\beta)^{Td} \cdot 4^d \left(\epsilon\sigma^{-2} \right)^{-(T+1)d} \mathbf{E} \int \mathbf{P} \left(r, \eta_0, \eta_2, \ldots, \eta_{T-2}, \theta^* \middle| \theta, \tilde{U}_0, \ldots, \tilde{U}_{T-1} \right) \\
&\qquad \min(1, a(\boldsymbol{\eta}, r)) d\theta_0 \cdots d\theta_{T-1} \\
&= (1+\epsilon\beta)^{Td} \cdot 4^d \left(\epsilon\sigma^{-2} \right)^{-(T+1)d} \mathbf{E} \int \left(2\pi\sigma^2 \right)^{\frac{-d}{2}} \cdot \left(8\pi\epsilon\beta\sigma^2 \right)^{\frac{-Td}{2}} \cdot \exp\left(-\frac{\|r\|^2}{2\sigma^2} \right) \\
&\qquad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot 2\epsilon\beta \left(\|r^*\|^2 - \|r\|^2 \right) \right) \\
&\qquad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot 4\epsilon\beta\sigma^2 \left(\rho_{T-\frac{1}{2}} - \rho_{-\frac{1}{2}} \right) \right) \\
&\qquad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 + \epsilon^2\beta^2 \left\| r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right\|^2 \right) \\
&\qquad \cdot \min(1, a) d\theta_0 \cdots d\theta_{T-1} \\
&= (1+\epsilon\beta)^{Td} \cdot 4^d \left(\epsilon\sigma^{-2} \right)^{-(T+1)d} \mathbf{E} \int \left(2\pi\sigma^2 \right)^{\frac{-d}{2}} \cdot \left(8\pi\epsilon\beta\sigma^2 \right)^{\frac{-Td}{2}} \\
&\qquad \cdot \exp\left(-\frac{1}{4\sigma^2} \left(\|r^*\|^2 + \|r\|^2 \right) \right) \cdot \exp\left(-\frac{1}{2} \rho_{T-\frac{1}{2}} \right) \\
&\qquad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left\| r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 + \epsilon^2\beta^2 \left\| r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}} \right\|^2 \right) \\
&\qquad \cdot \min(1, a) d\theta_0 \cdots d\theta_{T-1}
\end{aligned}$$

where $\boldsymbol{r}$ are to be understood as functions of the θ_t, and the integral is taken over θ_t.

Substituting the expression of a, then the term inside the integral is

$$\begin{aligned}
&\left(2\pi\sigma^2\right)^{\frac{-d}{2}} \cdot \left(8\pi\epsilon\beta\sigma^2\right)^{\frac{-Td}{2}} \cdot \exp\left(-\frac{1}{4\sigma^2}\left(\left\|r^*\right\|^2 + \left\|r\right\|^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{1}{2}\rho_{T-\frac{1}{2}}\right) \\
&\quad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left\|r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2 + \epsilon^2\beta^2\left\|r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right\|^2\right) \\
&\quad \cdot \min\left(1, \exp\left(U(\theta) - U(\theta^*) + \rho_{T-\frac{1}{2}}\right)\right) \\
&= \left(2\pi\sigma^2\right)^{\frac{-d}{2}} \cdot \left(8\pi\epsilon\beta\sigma^2\right)^{\frac{-Td}{2}} \cdot \exp\left(-\frac{1}{4\sigma^2}\left(\left\|r^*\right\|^2 + \left\|r\right\|^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left\|r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2 + \epsilon^2\beta^2\left\|r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right\|^2\right) \\
&\quad \cdot \exp\left(U(\theta)\right) \cdot \min\left(\exp\left(-U(\theta) - \frac{1}{2}\rho_{T-\frac{1}{2}}\right), \exp\left(-U(\theta^*) + \frac{1}{2}\rho_{T-\frac{1}{2}}\right)\right).
\end{aligned}$$

Finally, this probability multiplied by the probability of θ_0, which is $\frac{1}{Z}\exp\left(-U(\theta)\right)$, is

$$\begin{aligned}
&\pi(\theta)G(\theta, \theta^*) \\
&= \frac{1}{Z} \cdot (1+\epsilon\beta)^{Td} \cdot 4^d\left(\epsilon\sigma^{-2}\right)^{-(T+1)d} \\
&\quad \mathbf{E}\int \left(2\pi\sigma^2\right)^{\frac{-d}{2}} \cdot \left(8\pi\epsilon\beta\sigma^2\right)^{\frac{-Td}{2}} \cdot \exp\left(-\frac{1}{4\sigma^2}\left(\left\|r^*\right\|^2 + \left\|r\right\|^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left\|r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2 + \epsilon^2\beta^2\left\|r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right\|^2\right) \\
&\quad \cdot \exp\left(U(\theta)\right) \cdot \min\left(\exp\left(-U(\theta) - \frac{1}{2}\rho_{T-\frac{1}{2}}\right), \exp\left(-U(\theta^*) + \frac{1}{2}\rho_{T-\frac{1}{2}}\right)\right) d\theta_0 \cdots d\theta_{T-1} \\
&= \frac{1}{Z} \cdot (1+\epsilon\beta)^{Td} \cdot 4^d\left(\epsilon\sigma^{-2}\right)^{-(T+1)d} \cdot \left(2\pi\sigma^2\right)^{\frac{-d}{2}} \cdot \left(8\pi\epsilon\beta\sigma^2\right)^{\frac{-Td}{2}} \cdot \\
&\quad \mathbf{E}\int \exp\left(-\frac{1}{4\sigma^2}\left(\left\|r^*\right\|^2 + \left\|r\right\|^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left(\left\|r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2 + \epsilon^2\beta^2\left\|r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right\|^2\right)\right) \\
&\quad \cdot \min\left(\exp\left(-U(\theta) - \frac{1}{2}\rho_{T-\frac{1}{2}}\right), \exp\left(-U(\theta^*) + \frac{1}{2}\rho_{T-\frac{1}{2}}\right)\right) d\theta_0 \cdots d\theta_{T-1} \\
&= \frac{1}{Z} \cdot (1+\epsilon\beta)^{Td} \cdot \beta^{-\frac{Td}{2}} \cdot 2^{-\frac{3(T-1)d}{2}} \cdot \pi^{-\frac{(T+1)d}{2}} \cdot \epsilon^{-\frac{(3T+2)d}{2}} \cdot \sigma^{(T+1)d} \\
&\quad \cdot \mathbf{E}\int \exp\left(-\frac{1}{4\sigma^2}\left(\left\|r^*\right\|^2 + \left\|r\right\|^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \left(\left\|r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2 + \epsilon^2\beta^2\left\|r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right\|^2\right)\right) \\
&\quad \cdot \exp\left(-\frac{U(\theta) + U(\theta^*)}{2}\right) \cdot \exp\left(-\frac{1}{2}\left|U(\theta) - U(\theta^*) + \rho_{T-\frac{1}{2}}\right|\right) d\theta_0 \cdots d\theta_{T-1}.
\end{aligned}$$

And writing this out explicitly in terms of

$$\rho_{T-\frac{1}{2}} = \frac{1}{2}\epsilon\sigma^{-2}\sum_{t=0}^{T-1}\nabla\tilde{U}_t(\theta_t)^T\left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right),$$

we get

$$\begin{aligned}
&\pi(\theta)G(\theta,\theta^*)\\
&= \frac{1}{Z}\cdot(1+\epsilon\beta)^{Td}\cdot\beta^{-\frac{Td}{2}}\cdot 2^{-\frac{3(T-1)d}{2}}\cdot\pi^{-\frac{(T+1)d}{2}}\cdot\epsilon^{-\frac{(3T+2)d}{2}}\cdot\sigma^{(T+1)d}\\
&\quad\cdot\mathbf{E}\int\exp\left(-\frac{1}{4\sigma^2}\left(\|r^*\|^2+\|r\|^2\right)\right)\\
&\quad\cdot\exp\left(-\frac{1}{8\epsilon\beta\sigma^2}\cdot\sum_{t=0}^{T-1}\left(\left\|r_{t+\frac{1}{2}}-r_{t-\frac{1}{2}}+\epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2+\epsilon^2\beta^2\left\|r_{t-\frac{1}{2}}+r_{t+\frac{1}{2}}\right\|^2\right)\right)\\
&\quad\cdot\exp\left(-\frac{U(\theta)+U(\theta^*)}{2}\right)\\
&\quad\cdot\exp\left(-\frac{1}{2}\left|U(\theta)-U(\theta^*)+\frac{1}{2}\epsilon\sigma^{-2}\sum_{t=0}^{T-1}\nabla\tilde{U}_t(\theta_t)^T\left(r_{t-\frac{1}{2}}+r_{t+\frac{1}{2}}\right)\right|\right)d\theta_0\cdots d\theta_{T-1}.
\end{aligned}$$

Now, for this forward path from θ to θ^*, consider the reverse leapfrog trajectory from θ^* to θ. This trajectory will have the same values for $\theta,\theta_0,\ldots,\theta^*$ in the reversed order and will have negated values for $r_{-\frac{1}{2}},r_{1-\frac{1}{2}},\ldots,r_{T-\frac{1}{2}}$ in the reversed order again. Because of this negation, the values of $\rho_{\frac{1}{2}},\rho_{1+\frac{1}{2}},\ldots,\rho_{T-\frac{1}{2}}$ will also be negated. It follows that $\pi(\theta^*)G(\theta^*,\theta)$ will have the same expression.

Therefore,

$$\pi(\theta)G(\theta,\theta^*)=\pi(\theta^*)G(\theta^*,\theta).$$

This shows that Algorithm 2 with resampling momentum is reversible.

Now we show that the chain satisfies *skew detailed balance* and the stationary distribution of θ is $\pi(\theta)$ if not resampling momentum. Skew detailed balance means that the chain satisfies the following condition (Turitsyn et al., 2011)

$$\pi(x)G(x,y)=\pi\left(y^{\perp}\right)G\left(y^{\perp},x^{\perp}\right)$$

where G is the transition probability.

By Section B, we know that a chain that satisfies the above condition will have invariant distribution $\pi(x)$.

In our setting, $x=(\theta,r)$ and $x^{\perp}=(\theta,-r)$. Given this, the skew detailed balance is

$$\pi(\theta,r)G((\theta,r),(\theta^*,r^*))=\pi(\theta^*,-r^*)G((\theta^*,-r^*),(\theta,-r)).$$

Next we will show that Algorithm 2 without resampling momentum satisfies the above condition and it naturally follows that Algorithm 2 without resampling converges to the desired distribution.

We consider the joint distribution of (θ,r). By a similar analysis of resampling case, we can get that

$$\begin{aligned}
&\pi(\theta,r)\cdot G((\theta,r),(\theta^*,r^*))\\
&= \frac{1}{Z}\cdot(1+\epsilon\beta)^{Td}\cdot\beta^{-\frac{Td}{2}}\cdot 2^{-\frac{3(T-1)d}{2}}\cdot\pi^{-\frac{(T+1)d}{2}}\cdot\epsilon^{-\frac{(3T+2)d}{2}}\cdot\sigma^{(T+1)d}\\
&\quad\cdot\mathbf{E}\int\exp\left(-\frac{1}{4\sigma^2}\left(\|r^*\|^2+\|r\|^2\right)\right)\\
&\quad\cdot\exp\left(-\frac{1}{8\epsilon\beta\sigma^2}\cdot\sum_{t=0}^{T-1}\left(\left\|r_{t+\frac{1}{2}}-r_{t-\frac{1}{2}}+\epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2+\epsilon^2\beta^2\left\|r_{t-\frac{1}{2}}+r_{t+\frac{1}{2}}\right\|^2\right)\right)\\
&\quad\cdot\exp\left(-\frac{U(\theta)+U(\theta^*)}{2}\right)\\
&\quad\cdot\exp\left(-\frac{1}{2}\left|U(\theta)-U(\theta^*)+\frac{1}{2}\epsilon\sigma^{-2}\sum_{t=0}^{T-1}\nabla\tilde{U}_t(\theta_t)^T\left(r_{t-\frac{1}{2}}+r_{t+\frac{1}{2}}\right)\right|\right)d\theta_0\cdots d\theta_{T-1}.
\end{aligned}$$

Again, the reverse trajectory will have the same values for $\theta, \theta_0, \ldots, \theta^*$ and will have negated values for $r_{-\frac{1}{2}}, r_{1-\frac{1}{2}}, \ldots, r_{T-\frac{1}{2}}$ in the reversed order. Therefore, $\pi(\theta^*, -r^*)G((\theta^*, -r^*), (\theta, -r))$ will have the same expression.

It follows that

$$\pi(\theta, r)G((\theta, r), (\theta^*, r^*)) = \pi(\theta^*, -r^*)G((\theta^*, -r^*), (\theta, -r))$$

which is what we want. □

D Connection to HMC

When using a full-batch, $\beta = 0$, and resampling, AMAGOLD becomes HMC. To see this, we first notice that with $\beta = 0$ the update rules of θ and r are the same as in HMC. The remaining thing is to show a is also the same as in HMC. We rewrite $\rho_{t+\frac{1}{2}}$ as

$$\rho_{t+\frac{1}{2}} = \rho_{t-\frac{1}{2}} + \frac{1}{2}\sigma^{-2}\left(r_{t-\frac{1}{2}} - r_{t+\frac{1}{2}}\right)^T\left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right) = \rho_{t-\frac{1}{2}} + \frac{1}{2}\sigma^{-2}\left(\left\|r_{t-\frac{1}{2}}\right\|^2 - \left\|r_{t+\frac{1}{2}}\right\|^2\right)$$

As a result,

$$\rho_{T-\frac{1}{2}} = \frac{1}{2}\sigma^{-2}\sum_{t=0}^{T-1}\left(\left\|r_{t-\frac{1}{2}}\right\|^2 - \left\|r_{t+\frac{1}{2}}\right\|^2\right) = \frac{1}{2}\sigma^{-2}\left(\left\|r_{-\frac{1}{2}}\right\|^2 - \left\|r_{T-\frac{1}{2}}\right\|^2\right)$$

It follows that a becomes the same as in HMC.

E Proof of Theorem 2

In this section we prove a bound on the convergence rate of AMAGOLD as compared with second-order Langevin dynamics (L2MC).

Proof. We start with the expression we derived for the transition probability in the proof of reversibility.

$$\begin{aligned}
&\pi(\theta)G(\theta, \theta^*) \\
&= \frac{1}{Z}\cdot(1+\epsilon\beta)^{Td}\cdot\beta^{-\frac{Td}{2}}\cdot 2^{-\frac{(3T-1)d}{2}}\cdot\pi^{-\frac{(T+1)d}{2}}\cdot\epsilon^{-\frac{(3T+2)d}{2}}\cdot\sigma^{(T+1)d} \\
&\quad\cdot\mathbf{E}\int\exp\left(-\frac{1}{4\sigma^2}\left(\|r^*\|^2 + \|r\|^2\right)\right) \\
&\quad\cdot\exp\left(-\frac{1}{8\epsilon\beta\sigma^2}\cdot\sum_{t=0}^{T-1}\left(\left\|r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} + \epsilon\nabla\tilde{U}_t(\theta_t)\right\|^2 + c^2\epsilon^2\beta^2\left\|r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right\|^2\right)\right) \\
&\quad\cdot\exp\left(-\frac{U(\theta)+U(\theta^*)}{2}\right) \\
&\quad\cdot\exp\left(-\frac{1}{2}\left|U(\theta) - U(\theta^*) + \frac{1}{2}\epsilon\sigma^{-2}\sum_{t=0}^{T-1}\nabla\tilde{U}_t(\theta_t)^T\left(r_{t-\frac{1}{2}} + r_{t+\frac{1}{2}}\right)\right|\right)d\theta_0\cdots d\theta_{T-1}.
\end{aligned}$$

Since

$$\theta_t = \theta_{t-1} + \epsilon\sigma^{-2}r_{t-\frac{1}{2}},$$

if we define θ_{-1} and θ_T by convention such that

$$\frac{\theta_{-1}+\theta_0}{2} = \theta \qquad \text{and} \qquad \frac{\theta_{T-1}+\theta_T}{2} = \theta^*,$$

it follows that for all $t \in \{0, \ldots, T-1\}$

$$r_{t+\frac{1}{2}} - r_{t-\frac{1}{2}} = \epsilon^{-1}\sigma^2\left(\theta_{t+1} - 2\theta_t + \theta_{t-1}\right)$$

and

$$r_{t+\frac{1}{2}} + r_{t-\frac{1}{2}} \epsilon^{-1} \sigma^2 \left(\theta_{t+1} - \theta_{t-1} \right)$$

so we can write the above transition probability explicitly in terms of the θ_t as

$$\begin{aligned}
&\pi(\theta) G(\theta, \theta^*) \\
&= \frac{1}{Z} \cdot (1 + \epsilon\beta)^{Td} \cdot \beta^{-\frac{Td}{2}} \cdot 2^{-\frac{(3T-1)d}{2}} \cdot \pi^{-\frac{(T+1)d}{2}} \cdot \epsilon^{-\frac{(3T+2)d}{2}} \cdot \sigma^{(T+1)d} \\
&\quad \cdot \mathbf{E} \int \exp\left(-\frac{c}{4\sigma^2} \left(\|r^*\|^2 + \|r\|^2 \right) \right) \\
&\quad \cdot \exp\Bigg(-\frac{1}{8\epsilon\beta\sigma^2} \cdot \sum_{t=0}^{T-1} \Bigg(\left\| \epsilon^{-1} \sigma^2 \left(\theta_{t+1} - 2\theta_t + \theta_{t-1} \right) + \epsilon \nabla \tilde{U}_t(\theta_t) \right\|^2 \\
&\quad + c^2 \epsilon^2 \beta^2 \left\| \epsilon^{-1} \sigma^2 \left(\theta_{t+1} - \theta_{t-1} \right) \right\|^2 \Bigg) \Bigg) \cdot \exp\left(-\frac{U(\theta) + U(\theta^*)}{2} \right) \\
&\quad \cdot \exp\left(-\frac{1}{2} \left| U(\theta) - U(\theta^*) + \frac{1}{2} \epsilon \sigma^{-2} \sum_{t=0}^{T-1} \nabla \tilde{U}(\theta_t)^T \left(\epsilon^{-1} \sigma^2 \left(\theta_{t+1} - \theta_{t-1} \right) \right) \right| \right) \\
&\quad \cdot d\theta_0 \cdots d\theta_{T-1}.
\end{aligned}$$

Simplifying this a bit, we get

$$\begin{aligned}
&\pi(\theta) G(\theta, \theta^*) \\
&= \frac{1}{Z} \cdot (1 + \epsilon\beta)^{Td} \cdot \beta^{-\frac{Td}{2}} \cdot 2^{-\frac{(3T-1)d}{2}} \cdot \pi^{-\frac{(T+1)d}{2}} \cdot \epsilon^{-\frac{(3T+2)d}{2}} \cdot \sigma^{(T+1)d} \\
&\quad \cdot \mathbf{E} \int \exp\left(-\frac{\sigma^2}{\epsilon^2} \left(\|\theta^* - \theta_{T-1}\|^2 + \|\theta_0 - \theta\|^2 \right) \right) \\
&\quad \cdot \exp\left(-\frac{\sigma^2}{8\epsilon^3\beta} \cdot \sum_{t=0}^{T-1} \left\| \theta_{t+1} - 2\theta_t + \theta_{t-1} + \epsilon^2 \sigma^{-2} \nabla \tilde{U}_t(\theta_t) \right\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{\beta\sigma^2}{8\epsilon} \cdot \sum_{t=0}^{T-1} \|\theta_{t+1} - \theta_{t-1}\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{U(\theta) + U(\theta^*)}{2} \right) \\
&\quad \cdot \exp\left(-\frac{1}{2} \left| U(\theta) - U(\theta^*) + \frac{1}{2} \sum_{t=0}^{T-1} \nabla \tilde{U}_t(\theta_t)^T \left(\theta_{t+1} - \theta_{t-1} \right) \right| \right) \, d\theta_0 \cdots d\theta_{T-1}.
\end{aligned}$$

Next, let

$$\begin{aligned}
N_t &= \nabla \tilde{U}_t(\theta_t) - \nabla U_t(\theta_t), \\
A_t &= \theta_{t+1} - 2\theta_t + \theta_{t-1} + \epsilon^2 \sigma^{-2} \nabla U_t(\theta_t), \\
B_t &= \theta_{t+1} - \theta_{t-1}, \\
C_t &= U(\theta) - U(\theta^*) + \frac{1}{2} \sum_{t=0}^{T-1} \nabla U_t(\theta_t)^T \left(\theta_{t+1} - \theta_{t-1} \right).
\end{aligned}$$

Notice that only N_t depends on the randomness of the stochastic gradient samples. Then,

$$\begin{aligned}
&\pi(\theta)G(\theta,\theta^*) \\
&= \frac{1}{Z}\cdot(1+\epsilon\beta)^{Td}\cdot\beta^{-\frac{Td}{2}}\cdot 2^{-\frac{(3T-1)d}{2}}\cdot\pi^{-\frac{(T+1)d}{2}}\cdot\epsilon^{-\frac{(3T+2)d}{2}}\cdot\sigma^{(T+1)d} \\
&\quad\cdot\int\mathbf{E}\Bigg[\exp\left(-\frac{\sigma^2}{\epsilon^2}\left(\left\|\theta^*-\theta_{T-1}\right\|^2+\left\|\theta_0-\theta\right\|^2\right)\right) \\
&\quad\cdot\exp\left(-\frac{\sigma^2}{8\epsilon^3\beta}\cdot\sum_{t=0}^{T-1}\left(\left\|A_t\right\|^2+2\epsilon^2\sigma^{-2}A_t^TN_t+\epsilon^4\sigma^{-4}\left\|N_t\right\|^2\right)\right) \\
&\quad\cdot\exp\left(-\frac{\beta\sigma^2}{8\epsilon}\cdot\sum_{t=0}^{T-1}\left\|B_t\right\|^2\right) \\
&\quad\cdot\exp\left(-\frac{U(\theta)+U(\theta^*)}{2}\right) \\
&\quad\cdot\exp\left(-\frac{1}{2}\left|C_t+\frac{1}{2}\sum_{t=0}^{T-1}N_t^TB_t\right|\right)\Bigg]\,d\theta_0\cdots d\theta_{T-1}.
\end{aligned}$$

Now, for any constant $c > 1$, we can bound

$$\begin{aligned}
\mathbf{E}\left[\left|\sum_{t=0}^{T-1}N_t^TB_t\right|\right] &\le \sqrt{\mathbf{E}\left[\left(\sum_{t=0}^{T-1}N_t^TB_t\right)^2\right]} \\
&= \sqrt{\sum_{t=0}^{T-1}B_t^T\mathbf{E}\left[N_tN_t^T\right]B_t} \\
&\le \sqrt{\sum_{t=0}^{T-1}\frac{V^2}{d}\left\|B_t\right\|^2} \\
&\le \frac{V^2}{d}\frac{\epsilon}{2(c-1)\beta\sigma^2}+(c-1)\frac{\beta\sigma^2}{2\epsilon}\sum_{t=0}^{T-1}\left\|B_t\right\|^2.
\end{aligned}$$

Additionally, we know that $\mathbf{E}\left[N_t\right]=0$ and

$$\mathbf{E}\left[\left\|N_t\right\|^2\right]=\mathbf{E}\left[\mathbf{tr}\left(N_tN_t^T\right)\right]\le\mathbf{tr}\left(\frac{V^2}{d}I\right)=V^2.$$

So, since by Jensen's inequality, $\mathbf{E}\left[\exp(X)\right]\ge\exp(\mathbf{E}\left[X\right])$, we can bound this with

$$\begin{aligned}
&\pi(\theta)G(\theta,\theta^*) \\
&\ge \frac{1}{Z}\cdot(1+\epsilon\beta)^{Td}\cdot\beta^{-\frac{Td}{2}}\cdot 2^{-\frac{(3T-1)d}{2}}\cdot\pi^{-\frac{(T+1)d}{2}}\cdot\epsilon^{-\frac{(3T+2)d}{2}}\cdot\sigma^{(T+1)d} \\
&\quad\cdot\int\exp\left(-\frac{\sigma^2}{\epsilon^2}\left(\left\|\theta^*-\theta_{T-1}\right\|^2+\left\|\theta_0-\theta\right\|^2\right)\right) \\
&\quad\cdot\exp\left(-\frac{\sigma^2}{8\epsilon^3\beta}\cdot\sum_{t=0}^{T-1}\left\|A_t\right\|^2\right)\cdot\exp\left(-\frac{\epsilon TV^2}{8\sigma^2\beta}\right) \\
&\quad\cdot\exp\left(-\frac{c\sigma^2\beta}{8\epsilon}\cdot\sum_{t=0}^{T-1}\left\|B_t\right\|^2\right) \\
&\quad\cdot\exp\left(-\frac{U(\theta)+U(\theta^*)}{2}\right) \\
&\quad\cdot\exp\left(-\frac{1}{2}\left|C_t\right|\right)\cdot\exp\left(-\frac{1}{(c-1)Td}\cdot\frac{\epsilon TV^2}{8\sigma^2\beta}\right)\,d\theta_0\cdots d\theta_{T-1}.
\end{aligned}$$

Now, this is a lower bound on the AMAGOLD chain with parameters $(\epsilon, \sigma, \beta)$. Next, we consider the transition probability of a *rescaled* chain, with slightly different parameters, that will be set as a function of c. Specifically, consider the chain with parameters $(\epsilon, \sigma \cdot c^{-1/4}, \beta \cdot c^{-1/2})$. (We will set the parameter c later; at this point in the proof it is just an arbitrary constant $c > 1$.) If we call this rescaled chain G_r, then by substitution of the parameters into the above expression, we get

$$
\begin{aligned}
&\pi(\theta) G_r(\theta, \theta^*) \\
&\geq \frac{1}{Z} \cdot (1 + c^{-1/2} \epsilon \beta)^{Td} \cdot \beta^{-\frac{Td}{2}} \cdot 2^{-\frac{3(T-1)d}{2}} \cdot \pi^{-\frac{(T+1)d}{2}} \cdot c^{-\frac{d}{4}} \cdot \epsilon^{-\frac{(3T+2)d}{2}} \cdot \sigma^{(T+1)d} \\
&\quad \cdot \int \mathbf{E} \exp\left(-\frac{\sigma^2}{c^{1/2} \epsilon^2} \left(\left\| \theta^* - \theta_{T-1} \right\|^2 + \left\| \theta_0 - \theta \right\|^2 \right) \right) \\
&\quad \cdot \exp\left(-\frac{\sigma^2}{8 \epsilon^3 \beta} \cdot \sum_{t=0}^{T-1} \left\| A_t \right\|^2 \right) \cdot \exp\left(-\frac{c \epsilon T V^2}{8 \sigma^2 \beta} \right) \\
&\quad \cdot \exp\left(-\frac{\sigma^2 \beta}{8 \epsilon} \cdot \sum_{t=0}^{T-1} \left\| B_t \right\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{U(\theta) + U(\theta^*)}{2} \right) \\
&\quad \cdot \exp\left(-\frac{1}{2} \left| C_t \right| \right) \cdot \exp\left(-\frac{1}{(c-1)Td} \cdot \frac{c \epsilon T V^2}{8 \sigma^2 \beta} \right) \, d\theta_0 \cdots d\theta_{T-1} \\
&\geq \frac{1}{Z} \cdot (1 + c^{-1/2} \epsilon \beta)^{Td} \cdot \beta^{-\frac{Td}{2}} \cdot 2^{-\frac{3(T-1)d}{2}} \cdot \pi^{-\frac{(T+1)d}{2}} \cdot c^{-\frac{d}{4}} \cdot \epsilon^{-\frac{(3T+2)d}{2}} \cdot \sigma^{(T+1)d} \\
&\quad \cdot \exp\left(-\frac{c \epsilon T V^2}{8 \sigma^2 \beta} \right) \cdot \exp\left(-\frac{1}{(c-1)Td} \cdot \frac{c \epsilon T V^2}{8 \sigma^2 \beta} \right) \\
&\quad \cdot \int \mathbf{E} \exp\left(-\frac{\sigma^2}{\epsilon^2} \left(\left\| \theta^* - \theta_{T-1} \right\|^2 + \left\| \theta_0 - \theta \right\|^2 \right) \right) \\
&\quad \cdot \exp\left(-\frac{\sigma^2}{8 \epsilon^3 \beta} \cdot \sum_{t=0}^{T-1} \left\| A_t \right\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{\sigma^2 \beta}{8 \epsilon} \cdot \sum_{t=0}^{T-1} \left\| B_t \right\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{U(\theta) + U(\theta^*)}{2} \right) \\
&\quad \cdot \exp\left(-\frac{1}{2} \left| C_t \right| \right) \, d\theta_0 \cdots d\theta_{T-1}.
\end{aligned}
$$

On the other hand, consider the transition probability of the full-gradient L2MC chain with parameters $(\epsilon, \sigma, \beta)$. This chain will be the same as the AMAGOLD chain, except that $N_t = 0$ always. So, if we call this chain's transition probability $\bar{G}$, we will have

$$\begin{aligned}
&\pi(\theta)\bar{G}(\theta,\theta^*) \\
&= \frac{1}{Z} \cdot (1+\epsilon\beta)^{Td} \cdot \beta^{-\frac{Td}{2}} \cdot 2^{-\frac{(3T-1)d}{2}} \cdot \pi^{-\frac{(T+1)d}{2}} \cdot \epsilon^{-\frac{(3T+2)d}{2}} \cdot \sigma^{(T+1)d} \\
&\quad \cdot \int \exp\left(-\frac{\sigma^2}{\epsilon^2} \left(\left\| \theta^* - \theta_{T-1} \right\|^2 + \left\| \theta_0 - \theta \right\|^2 \right) \right) \\
&\quad \cdot \exp\left(-\frac{\sigma^2}{8\epsilon^3\beta} \cdot \sum_{t=0}^{T-1} \left\| A_t \right\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{\beta\sigma^2}{8\epsilon} \cdot \sum_{t=0}^{T-1} \left\| B_t \right\|^2 \right) \\
&\quad \cdot \exp\left(-\frac{U(\theta)+U(\theta^*)}{2} \right) \\
&\quad \cdot \exp\left(-\frac{1}{2} \left| C_t \right| \right) \, d\theta_0 \cdots d\theta_{T-1}.
\end{aligned}$$

Using this, we can simplify our bound on the transition probability of the AMAGOLD chain to

$$\begin{aligned}
\pi(\theta)G_r(\theta,\theta^*) \geq{}& \left(\frac{1 + c^{-1/2}\epsilon\beta}{1+\epsilon\beta} \right)^{Td} \cdot c^{-\frac{d}{4}} \cdot \exp\left(-\frac{c\epsilon T V^2}{8\sigma^2\beta} \right) \cdot \exp\left(-\frac{1}{(c-1)Td} \cdot \frac{c\epsilon T V^2}{8\sigma^2\beta} \right) \\
&\cdot \pi(\theta)\bar{G}(\theta,\theta^*).
\end{aligned}$$

Thus,

$$\frac{\pi(\theta)G_r(\theta,\theta^*)}{\pi(\theta)\bar{G}(\theta,\theta^*)} \geq \left(\frac{1 + c^{-1/2}\epsilon\beta}{1+\epsilon\beta} \right)^{Td} \cdot c^{-\frac{d}{4}} \cdot \exp\left(-\frac{c\epsilon T V^2}{8\sigma^2\beta} \right) \cdot \exp\left(-\frac{1}{(c-1)Td} \cdot \frac{c\epsilon T V^2}{8\sigma^2\beta} \right).$$

All that remains to get a bound is to set c appropriately. Since

$$c^{-\frac{d}{4}} = \exp\left(-\frac{d}{4} \log(c) \right) \geq \exp\left(-\frac{d}{4}(c-1) \right),$$

and

$$c^{-1/2} \geq 1 - \frac{c-1}{2},$$

we can bound this with

$$\begin{aligned}
\frac{\pi(\theta)G_r(\theta,\theta^*)}{\pi(\theta)\bar{G}(\theta,\theta^*)} &\geq \left(\frac{1 + \left(1 - \frac{c-1}{2}\right)\epsilon\beta}{1+\epsilon\beta} \right)^{Td} \cdot \exp\left(-\frac{d}{4}(c-1) - \frac{c\epsilon T V^2}{8\sigma^2\beta} - \frac{1}{(c-1)Td} \cdot \frac{c\epsilon T V^2}{8\sigma^2\beta} \right) \\
&\geq \left(1 - \frac{(c-1)\epsilon\beta}{2(1+\epsilon\beta)} \right)^{Td} \cdot \exp\left(-\frac{d}{4}(c-1) - \frac{c\epsilon T V^2}{8\sigma^2\beta} - \frac{1}{(c-1)Td} \cdot \frac{c\epsilon T V^2}{8\sigma^2\beta} \right).
\end{aligned}$$

Since for any $0 \leq x < 1/2$, it holds that $1 - x \geq \exp(-2x)$, as long as

$$\frac{(c-1)\epsilon\beta}{1+\epsilon\beta} \leq 1,$$

it holds that

$$1 - \frac{(c-1)\epsilon\beta}{2(1+\epsilon\beta)} \geq \exp\left(-\frac{(c-1)\epsilon\beta}{1+\epsilon\beta} \right).$$

So, under this assumption,

$$\begin{aligned}\frac{\pi(\theta)G_r(\theta,\theta^*)}{\pi(\theta)\bar{G}(\theta,\theta^*)} &\geq \exp\left(-\frac{(c-1)\epsilon\beta Td}{1+\epsilon\beta} - \frac{d}{4}(c-1) - \frac{c\epsilon TV^2}{8\sigma^2\beta} - \frac{1}{(c-1)Td}\cdot\frac{c\epsilon TV^2}{8\sigma^2\beta}\right) \\ &= \exp\left(-\frac{\epsilon TV^2}{8\sigma^2\beta} - \frac{\epsilon V^2}{8\sigma^2\beta d}\right) \\ &\quad \cdot \exp\left(-(c-1)\left(\frac{(1+\epsilon\beta(1+4T))d}{4(1+\epsilon\beta)} + \frac{\epsilon TV^2}{8\sigma^2\beta}\right) - \frac{\epsilon V^2}{8(c-1)\sigma^2\beta d}\right) \\ &= \exp\left(-\frac{\epsilon TV^2}{8\sigma^2\beta} - \frac{\epsilon V^2}{8\sigma^2\beta d}\right) \\ &\quad \cdot \exp\left(-(c-1)\left(\frac{3}{2}Td + \frac{\epsilon TV^2}{8\sigma^2\beta}\right) - \frac{\epsilon V^2}{8(c-1)\sigma^2\beta d}\right).\end{aligned}$$

If we also assume that

$$\frac{\epsilon V^2}{4\sigma^2\beta d} \leq 1,$$

then

$$\frac{\epsilon TV^2}{8\sigma^2\beta} \leq \frac{Td}{2},$$

and so

$$\begin{aligned}\frac{\pi(\theta)G_r(\theta,\theta^*)}{\pi(\theta)\bar{G}(\theta,\theta^*)} &\geq \exp\left(-\frac{\epsilon TV^2}{8\sigma^2\beta} - \frac{\epsilon V^2}{8\sigma^2\beta d}\right) \\ &\quad \cdot \exp\left(-(c-1)2Td - \frac{\epsilon V^2}{8(c-1)\sigma^2\beta d}\right).\end{aligned}$$

Next, set

$$c-1 = \sqrt{\frac{\epsilon V^2}{16\sigma^2\beta Td^2}}.$$

From this, we will get

$$\begin{aligned}\frac{\pi(\theta)G_r(\theta,\theta^*)}{\pi(\theta)\bar{G}(\theta,\theta^*)} &\geq \exp\left(-\frac{\epsilon TV^2}{8\sigma^2\beta} - \frac{\epsilon V^2}{8\sigma^2\beta d}\right)\cdot\exp\left(-\sqrt{\frac{\epsilon TV^2}{\sigma^2\beta}}\right) \\ &\geq \exp\left(-\frac{\epsilon TV^2}{4\sigma^2\beta} - \sqrt{\frac{\epsilon TV^2}{\sigma^2\beta}}\right).\end{aligned}$$

Now, in order for this to hold, we needed

$$\frac{(c-1)\epsilon\beta}{1+\epsilon\beta} \leq 1.$$

With our setting of c, and our other assumption,

$$c-1 = \sqrt{\frac{\epsilon V^2}{16\sigma^2\beta Td^2}} = \sqrt{\frac{\epsilon V^2}{4\sigma^2\beta d}\cdot\frac{1}{4Td}} \leq \sqrt{\frac{1}{4Td}} \leq 1,$$

so the bound will trivially hold. Thus the only added assumption we needed is the one stated in the Theorem statement, that

$$\frac{\epsilon V^2}{4\sigma^2\beta d} \leq 1.$$

Now we apply the standard Dirichlet form argument. The spectral gap of a Markov chain can be written as (Aida, 1998)

$$\gamma = \inf_{f\in L_0^2(\pi):Var_\pi[f]=1} \mathcal{E}(f)$$

where $L_0^2(\pi)$ denotes the Hilbert space of all functions that are square integrable with respect to probability measure π and have mean zero. $\mathcal{E}(f)$ is the Dirichlet form of a Markov chain associated with transition operator T (Fukushima et al., 2010):

$$\mathcal{E}(f) = \frac{1}{2}\int\int \left[(f(\theta) - f(\theta^*))^2\right] G(\theta, \theta^*)\pi(\theta) d\theta d\theta^*$$

By the expression of the spectral gap, it follows that

$$\begin{aligned}
\gamma &= \inf_{f\in L_0^2(\pi): Var_\pi[f]=1} \left[\frac{1}{2}\int\int \left[(f(\theta) - f(\theta^*))^2\right] G(\theta, \theta^*)\pi(\theta) d\theta d\theta^*\right] \\
&\geq \exp\left(-\frac{\epsilon T V^2}{4\sigma^2\beta} - \sqrt{\frac{\epsilon T V^2}{\sigma^2\beta}}\right) \cdot \inf_{f\in L_0^2(\pi): Var_\pi[f]=1} \left[\frac{1}{2}\int\int \left[(f(\theta) - f(\theta^*))^2\right] \bar{G}(\theta, \theta)\pi(\theta) d\theta d\theta^*\right] \\
&= \exp\left(-\frac{\epsilon T V^2}{4\sigma^2\beta} - \sqrt{\frac{\epsilon T V^2}{\sigma^2\beta}}\right) \cdot \bar{\gamma}
\end{aligned}$$

This finishes the proof. □

F Reformulation of AMAGOLD Algorithm

We reformulate our algorithm by setting $v = \epsilon\sigma^{-2}r$,$b = \epsilon\beta$, $h = \epsilon^2\sigma^{-2}$ and outline the algorithm after reformulation in Algorithm 3.

Algorithm 3 Reformulated AMAGOLD

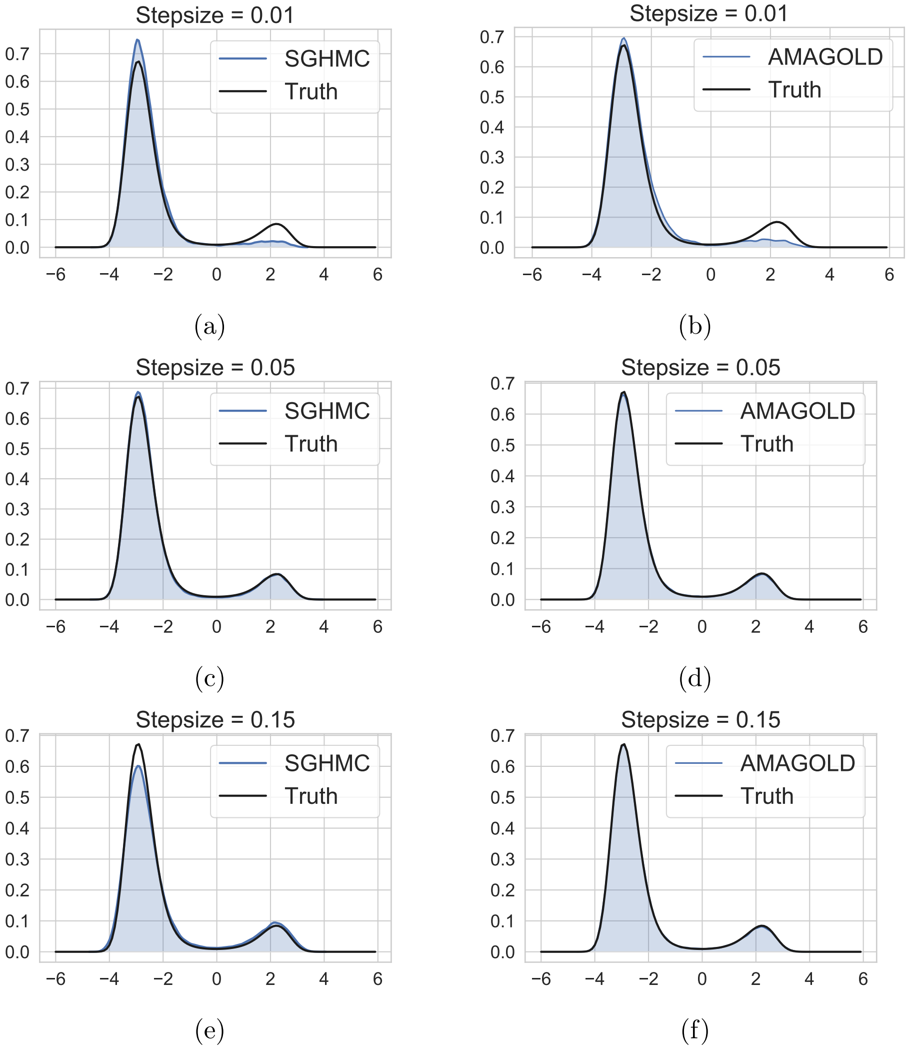
```
given: Energy $U$, initial state $\theta \in \Theta$
loop
  optionally, resample momentum: $v \sim \mathcal{N}(0, h\mathbf{I})$
  initialize momentum and energy acc: $v_{-\frac{1}{2}} \leftarrow v$, $\rho_{-\frac{1}{2}} \leftarrow 0$
  half position update: $\theta_0 \leftarrow \theta + \frac{1}{2}v_{-\frac{1}{2}}$
  for $t = 0$ to $T - 1$ do
    if $t \neq 0$ then
      position update: $\theta_t \leftarrow \theta_{t-1} + v_{t-\frac{1}{2}}$
    end if
    sample noise $\eta_t \sim \mathcal{N}(0, 4hb)$
    sample random energy component $\tilde{U}_t$
    update momentum: $v_{t+\frac{1}{2}} \leftarrow \left((1-b)v_{t-\frac{1}{2}} - h\nabla\tilde{U}_t(\theta_t) + \eta_t\right)/(1+b)$
    update energy acc: $\rho_{t+\frac{1}{2}} \leftarrow \rho_{t-\frac{1}{2}} + \frac{1}{2}\nabla\tilde{U}_t(\theta_t)^T\left(v_{t-\frac{1}{2}} + v_{t+\frac{1}{2}}\right)$
  end for
  half position update: $\theta_T \leftarrow \theta_{T-1} + \frac{1}{2}v_{T-\frac{1}{2}}$
  new values: $\theta^* \leftarrow \theta_T$, $v^* \leftarrow v_{T-\frac{1}{2}}$
  $a \leftarrow \exp\left(U(\theta) - U(\theta^*) + \rho_{T-\frac{1}{2}}\right)$
  with probability $\min(1, a)$ update $\theta \leftarrow \theta^*$, $v \leftarrow v^*$ (as long as $\theta^* \in \Theta$)
  otherwise update $v \leftarrow -v_{-\frac{1}{2}}$
end loop
```

G Additional Experiments Results and Setting Details

G.1 Double Well Potential

We visualize the estimated density on additional step size settings. Consistent with Figure 1d, it is clear here that SGHMC is very sensitive to step size. A small change in step size will cause a big difference in the estimated density. In contrast, AMAGOLD is more robust and can work well with a large range of step sizes.

When the setup of step size is inappropriate, as in Figures 6a and b where it is fixed to be too small, either SGHMC or AMAGOLD converges in the training time. This is because the chain moves too slowly toward the stationary distribution. However, AMAGOLD with step size tuning is able to automatically adjust the step size based on the information provided by M-H step. As shown in Figure 1c, tuned AMAGOLD can determine a step size that causes convergence given the same training time budget. All results are obtained by collecting 10^5 samples with 1000 burn-in samples.



(a) (b) (c) (d) (e) (f)

Figure 6: Estimated densities of SGHMC (1st column) and AMAGOLD (2nd column) on varying step sizes.

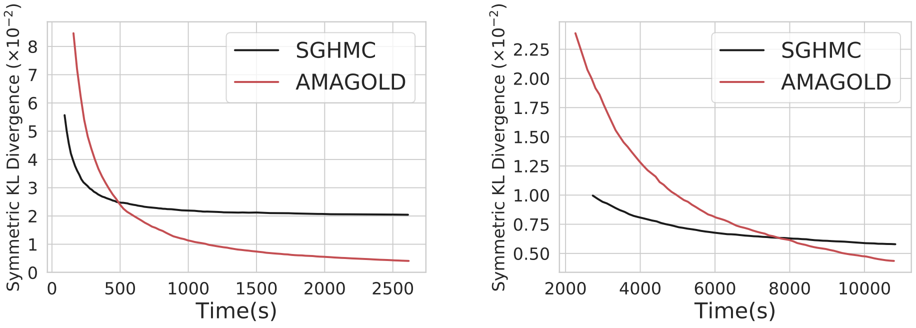
G.2 Two-Dimensional Synthetic Distributions

G.2.1 Analytical Expression

$$
\begin{aligned}
&\text{Dist1: } \mathcal{N}(z_1; z_2^2/4, 1)\mathcal{N}(z_2; 0, 4)\\
&\text{Dist2: } 0.5\mathcal{N}\left(\boldsymbol{z}; 0, \begin{bmatrix} 2 & 1.8 \\ 1.8 & 2 \end{bmatrix}\right) + 0.5\mathcal{N}\left(\boldsymbol{z}; 0, \begin{bmatrix} 2 & -1.8 \\ -1.8 & 2 \end{bmatrix}\right)
\end{aligned}
$$

G.2.2 Runtime Comparisons

We report runtime comparisons between AMAGOLD and SGHMC on Dist1 and Dist2 with step size 0.15 (Figure 7). This experiment uses the analytical energy expression (no data examples), so there is no speed-up of stochastic methods over full-batch methods. At the beginning, SGHMC converges faster due to the lack of M-H step, but eventually it converges to a biased distribution. AMAGOLD is not much slower than SGHMC, which shows that AMA can reduce the amount of computation of adding M-H step while keep the chain unbiased.



(a)

(b)

Figure 7: Runtime comparisons between SGHMC and AMAGOLD on synthetic distributions (a) Dist1 and (b) Dist2.

G.2.3 Additional Note on Figure 2

It is worth noting that, even though it is lower than SGHMC's, AMAGOLD's KL divergence grows when the step size is large compared to full-batch methods. This is because the M-H acceptance probability decreases, causing the chain to converge more slowly. This is expected. It is well-known that stochastic methods are more sensitive to step sizes than full-batch methods (Nemirovski et al., 2009). However, since AMAGOLD's KL divergence grows much slower than SGHMC's, AMAGOLD is more robust to different step sizes

G.3 Bayesian Logistic Regression

We report the acceptance probability of AMAGOLD on *Heart* for varying step sizes in Figure 8. For a large range of step sizes, the acceptance rate is sufficiently high to allow the chain converge fast, demonstrated in Figure 4. The acceptance rate may become very low with a large step size resulting in slow move. But this undesired acceptance probability can be easily detected and avoided in practice.

G.4 Bayesian Neural Networks

The architecture of Bayesian Neural Networks is a two-layer MLP with first hidden layer size 500 and the second hidden layer size 256.

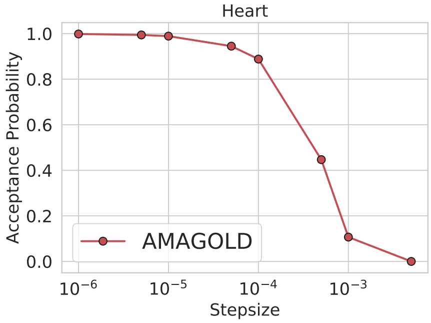


Figure 8: The acceptance probability of the M-H step in AMAGOLD for varying step sizes on the Heart dataset.