We thank all the reviewers for the valuable comments.

To Reviewer 1:

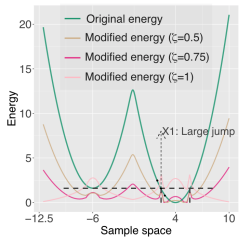
Q1. *Details on the vanishing gradient problem in flat-histogram SGLD:* The original step function in formula (4) leads to $\frac{\partial \log \Psi_\theta(u)}{\partial u} = \frac{1}{\Psi_\theta(u)} \frac{\partial \Psi_\theta(u)}{\partial u} = 0$ almost everywhere. Combining it with (6) leads to $\nabla_x \log \varpi_{\Psi_\theta}(x) = -\left[1 + \zeta\tau \frac{\partial \log \Psi_\theta(u)}{\partial u}\right] \frac{\nabla_x U(x)}{\tau} = -\frac{\nabla_x U(x)}{\tau}$. Therefore, the naive flat-histogram SGLD will behave like SGLD and fail to converge to the flattened density (2).

Q2. *Advantages of CSGLD over M-SGD*: (i) CSGLD belongs to the class of adaptive biasing force algorithms and has the potential to exponentially speed-up the computation [1], i.e. the fine-tuned CSGLD should outperform the fine-tuned M-SGD. (ii) CSGLD also belongs to the class of dynamic importance sampling algorithms and is able to quantify uncertainty for its estimation and prediction. While M-SGD is an optimization algorithm and produces a point estimate only. [1] Long-time convergence of an adaptive biasing force (ABF) method. Nonlinearity. 2008.

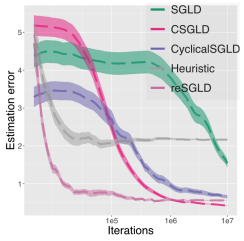
To Reviewer 2:

Q1. *The working distribution is different from the original one*: CSGLD belongs to the class of dynamic importance sampling algorithms, for which each sample is generated with an importance weight and they can be used for inference of the target distribution via a weighted averaging estimator. We have provided theoretical guarantee for the convergence of the weighted averaging estimator. The samples from the target distribution can also be obtained via an importance resampling step from the pool of importance samples generated by CSGLD.

Q2. *Asymmetric modes with a heuristic baseline:* Following section 4.1, we set $\pi(x) = 1/6N(-6, 1) + 5/6N(4, 1)$ and included a baseline called *Heuristic*, which injects a Gaussian noise $N(0, 2^2)$ whenever $|\nabla \widetilde{U}(x)| < 0.1$. As shown by Figure (b) of this rebuttal, *Heuristic* performs quite well in the early period due to the heuristic random walk helping to escape local traps. However, it leads to a very large prediction error in the long run as the sampling equilibrium was broken by *Heuristic*, but the samples were not properly weighted.



(a) Original v.s. trial densities



(b) Estimation errors

Q3. *Discussions on the number of partitions:* A fine partition will lead to a smaller discretization error, but it increases the risk in stability. In particular, it will lead to large bouncy jumps around optima (a large negative learning rate, i.e., $\frac{\log \theta(2) - \log \theta(1)}{\Delta u} \ll 0$ in formula (8) will be caused there). Empirically, we suggest to partition the sample space into a moderate number of subregions, e.g. 10-1000, to balance between stability and discretization error.

To Reviewer 3:

Q1. *Drawbacks of simulated annealing (SA) and replica exchange SGLD (reSGLD)/parallel tempering:* SA can only be used in optimization, and it might get stuck in a poor local minimum if the temperature decreases too fast. The reSGLD requires a large correction to reduce the bias, which may lead to a large bias and insignificant accelerations.

Q2. *Missing baselines*: We further compared CSGLD with CyclicalSGLD and reSGLD on an asymmetric mixture distribution. All algorithms were run 10^7 iterations. For CSGLD, we set $\widetilde{U}(x) \sim \mathcal{N}(U(x), 4^2)$, the default stepsize $\alpha = 0.1$ and temperature $T = 1$. For CyclicalSGLD, we set $\alpha_0 = 1$ and 100 cycles and the threshold $\beta = 0.9$; for reSGLD, we additionally include a high-temperature process with $T = 5$ and a correction of 16. Figure (b) of this rebuttal shows that CSGLD is inferior to the baselines at the beginning, but it eventually outperforms the baselines as the learning of θ_k is mature. We will include the baselines and references in the next version.

Q3. The gradient-vanishing problem in SGLD is not clear: Please refer to our reply to Q1 of Reviewer 1.

To Reviewer 4:

Q1. *Comments on bizarre peaks*: A bizarre peak always indicates that there is a local minimum of the same energy in somewhere of the sample space, as the sample space is partitioned according to the energy function in CSGLD. For example, Figure (a) of this rebuttal shows a bizarre peak at x_1. Although x_1 is not a local minimum, it has the same energy as "-6" which is a local minimum. Note that in CSGLD, x_1 and "-6" belongs to the same subregion.

Q2. *In very low density level with no driving forces.* Within the same subregion, CSGLD is reduced to SGLD. Therefore, in the very low density region, it will not move farther away, but still move toward the high density region.

Q3. *Choices of the partition and ζ*: Regarding the partition, please refer to our response to Q3 of reviewer 2. With respect to ζ, we suggest to tune ζ to obtain the largest barrier reductions such as $\zeta = 0.75$ in Figure 1(a) of the paper.