We appreciate the reviewers for their valuable comments on the improvement of this paper. The reviews are insightful and constructive; we believe that all issues mentioned in the comments can be properly addressed in the final version.

Response to Two Common Concerns:

1. Performance on CIFAR10. It is possible that ResNet-20 on the CIFAR10 in our experiment were undertrained due to the early-stopping we applied. In the initial experiment, we setup the maximum training epochs as 1000; but the procedure would be manually stopped when the performance on the validation set had stopped from increasing for 5 consecutive epochs. For each experimental setting, we conducted 5 independent runs; each time the training of ResNet-20 has been finalized within about 150 epochs, *i.e.* it encountered performance plateau after 150 epochs. As for all hyperparameters for baselines, we have taken the recommended settings and then conducted a simple grid search within a small interval to determine the best fit. As we haven't been over-tuning our method and under-tuning the baselines, we still believe the current result reflects some advantageous tendency of our method. Of course, we will redo the experiment and manage to obtain a more thorough understanding in the final version.

2. Choice of hyperparameters. In fact, we haven't put much more effort in tuning hyperparameters in our experiment and the result seemed satisfactory: the **batchsizes** are set as default values in a typical day-to-day neural network training; the **threshold** σ_* can be readily estimated from the sample variances within the objective function evaluated on a small number of mini-batches from the target dataset. As for the **bandwidth** λ, it influences the accuracy by governing the quality of approximation on compensation distribution: with higher bandwidth, the approximation becomes more accurate and its computation will in return be more time-consuming; with better approximation, the detailed balance will be better preserved, which will lead to more accurate samples. We have found that given relatively small batch variance (as is in our experiment), the accuracy and complexity can be balanced quite well by setting λ to a **moderate value**; also, we observed that the value of λ is not a sensitive factor that needs much tuning for better performance. Indeed, the design of **temperature ladders** is challenging in the context of machine learning due to the absence of physical guidance; nevertheless, some of the approaches developed for physics may still be applied safely to machine learning, *e.g.* the geometric layout as is applied in this paper. In general, the configuration of a ladder will depend on the specific application and also the architecture of the network; it can be optimized through a grid search. According to our observation, ladders of **eight** temperatures allocated by **geometric** factor 0.05 functioned well in all our experiments.

To Reviewer 1. *1.* Please see **Common Concerns**. *2.* We've noticed the Leimkuhler splitting in the very beginning. The omission of this scheme in our paper is primarily due to our focus on replica-exchange protocol. Leimkuhler's scheme is interchangeable with the Euler splitting in our algorithm.

To Reviewer 2. *1.* Please see **Common Concerns**. *2.* Our RE protocol works under the circumstances where the evaluation of energy function is perturbed by Gaussian noise. No matter whether the RE criterion is Barker's test (as in our proposal) or Metropolis' alternative, the it involves Gaussian deconvolution, which in either case has no exact analytical solution; we have to leverage approximation to make it work. Nevertheless, a more detailed analysis on the discrepancy is beneficial, we will provided it as a complementary section.

To Reviewer 3. *1.* Please see **Common Concerns**. *2.* We claim our analytical approximation being more efficient in generating compensation variable z_C, where our proposal enables Gibbs sampler whereas Seita's numerical solution needs lookup tables. The latter is way slower than the former. Detailed comparison will be reported as a complementary section. *3.* Please see **Common Concerns**. *4.* We have conducted the comparison in a different manner, we run each method with the same epochs, which we believe reflects the performance in practice. For those baselines with much slower sampling speed, our advantage lies in the time efficiency in real world. These latest ensemble methods will be compared and discussed. *5.* Actually, since Gaussian distribution decays much faster than the logistic, no matter how large the variance Gaussian noise is, in theory, the correction distribution can be obtained at arbitrary precision. In practice, we predefined the noise threshold σ_*^2 in order to simplify the computation: with a fixed σ_*^2, no need to recompute the correction distribution, we simply compensate the actual variance σ^2 up to σ_*^2 and reuse the recomputed numerics. Multiple mini-batches might be required in the rare case the actual variance σ^2 exceeds the threshold σ_*^2.

To Reviewer 4. Thank you very much for your kind support and endorsement.

To Reviewer 5. *1.* We've examined on several latest architectures with residual connections, namely ResNet, DenseNet, Transformer, and Residual LSTM; empirical findings indicate that the gradient noise, no matter how deep a network will be, resembles Gaussian variables. Hence, albeit found in AlexNet, the heavy-tail phenomenon is not a common situation for all neural architectures; at least for some of the latest models, the conventional assumption of Gaussianity is to some extent still valid. Furthermore, all evidences in our experiment support presuming the constant variance in Gaussian noise. The *i.i.d.* assumption relies on the fact that the dataset is built upon examples collected independently from a certain data distribution. *2.* We've noticed that Metropolis criterion is optimal whereas the Barker's alternative is of 70% efficiency compared with the former. The reason Barker's is leveraged is that his proposal is based on logistic distribution, which resembles Gaussian and is super-smooth. It sacrifices some of the efficiency for much smaller discrepancy and much better analytic characteristics. The traditional RE methods with Metropolis' test either fails to address Gaussian noise or encounters severe problems (*e.g.* delta functions) in deriving correction distributions. Zanella's proposal will be examined carefully. *3.* Please see **Common Concerns**.