We thank the reviewers for the in-depth reviews. We will fix the typos and add the missing citations.

Comment 1. *Lack of experimental details.*
Response 1. We will add a section in the supplemental material to provide all the experimental details including all the tuning parameters such as learning rate, mini-batch size, number of epochs, etc. The transformation function is parameterized as a DNN with two hidden layers and ReLU non-linearity. The auxiliary distribution is a normal distribution, in which the mean is the output of a DNN and the covariance matrix is a diagonal matrix. The DNN in the auxiliary distribution has the same number of neurons at individual layers as the transformation DNN but with the reverse order, e.g., if the number of neurons from input to output in the transformation DNN are 3-50-50-2, the auxiliary distribution DNN are 2-50-50-3. The results of [Yin and Zhou, 2018] and and [Titsias and Ruiz, 2019] in the binarized MNIST and Fashion-MNIST experiments are taken from [Titsias and Ruiz, 2019]. The results of PBP and dropout in the BNN regression experiments are taken from [Gal and Ghahramani, 2016].

Comment 2. *... claimed that many other distribution can be used as the auxiliary distribution beside normal distribution, but no example of them given ... interesting to see advantage and limitation of using other auxiliary variables ...*
Response 2. The auxiliary distribution is part of the variational posterior [Agakov and Barber, 2004]. The more flexible the auxiliary distribution is, the tighter the lower bound is. We motive the choice of a normal distribution as an approximation to the Dirac delta distribution, which allows us to have a simple formulation and makes theoretical analyses easier. NFW starts with a non-invertible transformation, but optimization of ELBO may result into an invertible function. In that case, both α and β will approach to zero during optimization, so the choice of the auxiliary distribution does not matter much as long as it is able to approach to a Dirac delta distribution. However, when the transformation function is non-invertible or even degenerate, the choice of the auxiliary distribution will impact the tightness of the lower bound. In this case, it will be interesting to see the performance difference due to the different choices of auxiliary distributions, but it is beyond the scope of this paper. It is an interesting topic for future works.

Comment 3. *How does NFW perform with multi-modal and/or overdispresed distributions?*
Response 3. Figure 2 shows the performance of the NFW in terms of modeling three different synthetic distributions including a multi-modal distribution. One challenge that prevents us to do the distribution matching experiments commonly done for flow-based methods is the intractability of the density function after the transformation due to the non-invertibility. Therefore, we can only show the samples in Figure 2.

Comment 4. *... under what condition we can use multiple non-invertible transformation (like NF) and how expensive they are computationally?*
Response 4. It is easy to compose multiple non-invertible transformation like NF, of which the result is still a non-invertible transformation. The auxiliary distribution is only needed at the end of the transformation in order to avoid solving the intractable non-linear system. Therefore, composing multiple non-invertible transformations will not make the lower bound more expensive.

Comment 6. *Lack of comparison with flow-based methods.*
Response 6. There are no published results about applying flow-based methods on BNN regression benchmarks. We will expand the Table 2 with the performance of flow-based methods.

Comment 7. *Are the parameter α and β optimized? How are they optimized?*
Response 7. Both α and β are variational parameters and are optimized together with other model and variational parameters in the ELBO. The optimization of ELBO is not very sensitive to the initial values of α and β. For all the experiments using NFW as variational posteriors (VAE and BNN regression), we use the same initial values of α and β, i.e., $\alpha = \sigma(-5)$ and $\beta = \sigma(-3)$, which are roughly 0.005 and 0.05 respectively. It is also necessary to optimize α and β, because the "regularization" term containing α and β can be viewed as an approximation to the regularization term of normalizing flows. Such an approximation behaves like approximating a logarithm function with a polynomial function (see Figure 1b for an example). Different values of α and β provide good approximations for different values of the parameters of the transformation function. Optimizing α and β together with the transformation parameters leads to a tighter lower bound.

Comment 8. *In Line 157, the authors claim their approach does not require a large sample size. Please clarify if it indicates the data or Monte Carlo ... provide quantitative comparisons regarding to this claim.*
Response 8. The sample size indicates the number of Monte Carlo samples. The resulting gradient estimator can work with only one sample. In the experiments, we set the number of samples to be 10 for faster convergence. In [Yin and Zhou, 2018], the number of samples in the VAE experiment is 100 and in [Titsias and Ruiz, 2019], although only one sample from HMC is used, it takes 5 iterations for burn-in, each of which consists 5 leapfrog gradient steps. Both of the previous works are not able to perform inference with as small as one sample. We will add a plot in the supplemental material to demonstrate the performance changes corresponding to different number of samples.

Comment 9. *How fast the learning is?*
Response 9. Our method is significantly faster than the previous works [Yin and Zhou, 2018] and and [Titsias and Ruiz, 2019], because the sampling process and the auxiliary distribution evaluation are straightforward and our method can work with a small number of Monte Carlo samples (10 vs 100). The exact runtime will be presented in the supplement.