We thank the reviewers for their insightful feedback. We are encouraged that they find our motivation and idea to be appealing (**R1**), theoretically insightful (**R2**), new (**R2**) and clearly presented (**R2**, **R4**). We are equally glad they found our results strong (**R1**), sound (**R2**), promising and thorough (**R3**), appropriately positioned w.r.t. prior work (**R1**, **R3**), and outperforming the state-of-the-art (**R2**, **R4**). Moreover, we are pleased that **R2** and **R4** appreciated our mathematical insights on why the energy score is better than the softmax score for out-of-distribution detection. We address reviewer comments below and will incorporate all feedback.

[**R1**, **R2**] **Can you evaluate on additional in-distribution dataset?** We have run experiments using SVHN as in-distribution dataset. Using energy score consistently outperforms softmax score by **8.56**% (FPR at 95% TPR), which suggests its general applicability beyond CIFAR datasets. We will definitely include these results in the final version.

[**R2**, **R3**] **How to distinguish methodological differences w.r.t. prior work?** As discussed in L240–L248, there are several major distinctions. We have moved the expanded discussions to Section 3 for clarity, as **R3** suggested. To recap:

- The idea of using the energy score for OOD detection is novel and theoretically motivated. JEM [10] used likelihood score $\log p(\mathbf{x})$ from **generative** modeling or softmax score; we instead derive the energy score from a **pure discriminative** model. We showed both mathematical insights (L103–L117) and empirical success that energy is better than softmax score—which was not presented in OE [14] or JEM [10].
- JEM's optimization is intractable as it requires estimating the normalized densities w.r.t. all inputs $\mathbf{x}$. Ours is easy to optimize using standard SGD, and doesn't involve proper normalization.
- JEM only uses in-distribution data. Our training objective (6) leverages auxiliary outlier data and explicitly enlarges the energy gap, which is also different from OE.

[**R1**] **Is it realistic to assume we have OOD data to optimize the objective? Does one source of OOD data also help another?** Thank you for the question! In fact, we are not the first to use auxiliary data for OOD detection. Previous work (Outlier Exposure [14]) showed one can leverage a large, diverse auxiliary outlier dataset for better estimation on the decision boundary for OOD detection. Since the auxiliary data (i.e., 80 million TinyImages) is sufficiently diverse, this fine-tuning process helps generalize to different test-time OOD data, as evidenced in Table 1.

[**R1**] **Is there a train/test split of the OOD data for fine-tuning?** Following common practice, we used 80M TinyImages as auxiliary OOD training data; we tested on six OOD datasets (L143) that are different from training.

[**R2**] **Why are the $\log p(x)$ results in JEM [10] much worse?** Very insightful question! While the energy score can be interpreted from the likelihood view, the optimization processes of JEM vs. ours are very different. We believe the difficulty in optimization led to the worse results we see in JEM. Specifically, JEM's objective estimates the joint distribution $\log p_\theta(\mathbf{x}, y) = \log p_\theta(\mathbf{x}) + \log p_\theta(y|\mathbf{x})$. The first term maximizes the likelihood from a generative modeling perspective, which requires estimating normalized densities. As pointed out by Grathwohl et al. [10] in the limitations (Sec. 6), *"Energy based models can be very challenging to work with. Since normalized likelihoods cannot be computed, it can be hard to verify that learning is taking place at all ... Furthermore, the gradient estimators we use to train JEM are quite unstable and are prone to diverging if the sampling and optimization parameters are not tuned correctly"*. In contrast, our training objective is purely discriminative and can be more easily optimized using standard SGD.

[**R2**] **Why having two margin loss parameters?** One problem with the single margin is that the difference of the energy scores between in- and out-of-distribution is relatively fixed; however, the energy scores can change over the training process. This results in lower accuracy of in-distribution classification (-0.57% for CIFAR-10 and -1.11% for CIFAR-100) because the training is less stable due to the drift of energy scores from batch to batch. Instead, using two margin hyper-parameters results in overall good performance on both in-distribution classification and OOD detection.

[**R3**] **Does the method work on more complex (CIFAR/SVHN) $\rightarrow$ simpler dataset (MNIST)?** Yes. On the CIFAR-10-pretrained WideResNet, using the energy score yields **14.50**% FPR when evaluated against MNIST as OOD data, outperforming 51.73% using the softmax score.

[**R3**, **R4**] **Other architectures/toy example/more comparison between energy and JEM:** Yes, consistent improvement was also observed on other architectures such as AllConv and DenseNet. We will include these results and the suggested toy example, more comparisons with JEM in the final version.

[**R3**] **Were these 10 runs done over random hyper-parameter configurations or 10 runs with the best selected hyperparameters' values?** The latter. We will make sure to report variances in the final version.

[**R4**] **What if you use self-normalized output, replacing logits with log score?** We found the suggestion of self-normalization very interesting, and investigated further. Mathematically, the log score is equivalent to logit shifted by the free energy, which is *sample specific* and therefore might change the overall score distribution. This shifting in actuality leads to the free energy score (on self-normalized output) collapsing to 0, i.e., $\mathrm{LogSumExp}(\log\ \mathit{softmax}_i) = 0$, and did not work well. We believe the normalizing factor needs to be *sample independent* to ensure $E(\mathbf{x}, f)$ is proportional to the density. We are excited to explore this as part of future work.

[**R2**, **R3**, **R4**] **Typo/Writing clarity:** All fixed. We thank the reviewers for the careful read and helpful suggestions!