We thank all of the reviewers for their time, effort and engagement with our work. Feedback is addressed below, and all comments will be incorporated in the final manuscript revision.

**=== General Comments ===**

**Extension to non-binary targets.** To deal with different settings, note that gated geometric mixing neurons can be defined for any member of the exponential family, not just the Bernoulli case as considered in this work. This includes both the Categorical distribution and the Gaussian distribution, which enables them to perform multi-class classification and regression respectively. A concrete example of this is the "Gaussian Gated Linear Networks" paper (which can be found on arXiv) that shows SOTA results on many regression problems. In the categorical case, the geometric mixture probability of a class $c$ in $\{1, 2, \ldots, C\}$ given $m$ categorical distributions $p_i(.)$ with weight $w_i$ for $i = 1..m$, is $\Pr(c) := \exp\{\sum_{i=1}^{m} w_i \log p_i(c)\} / \sum_{c'=1}^{C} \exp\{\sum_{i=1}^{m} w_i \log p_i(c')\}$, which requires an additional factor of $C$ work (like softmax) per neuron, and no additional space complexity. NCTL is orthogonal to this specific choice of neuron parameterization, but we elected to focus on classification tasks that are well-established in the continual learning community.

**On scale of datasets.** We agree that the continual learning problem is far more complex than captured by current standard datasets. We elected for the datasets that allowed us to make the most direct comparison to existing SOTA methods. We agree that these datasets are small compared to those considered in large-scale representation learning (e.g. ImageNet), but it's worth noting that (1) the two fields are solving very different problems, and (2) that even MNIST variants are sufficiently complex to clearly stratify the performance of competing methods (the function of a challenge dataset). This largely indicates that online/continual learning is an under-developed field compared to large-scale representation learning (ImageNet) or language modelling, and currently there is little advantage (but much cost) to running experiments at this scale. That said, we can imagine several ways in which GLNs could be augmented e.g. with convolutional inductive biases to scale to larger problems, and agree this is necessary for ImageNet-style problems, but this would be a paper in its own right and is orthogonal to what we were aiming to achieve with this work.

**On clarity of exposition.** Thank you for all the helpful suggestions, they are well taken. If accepted we will endeavor to use the additional page of space to properly address the additional related work mentioned, add further details to the appendix and expand the discussion in the *broader impact* statement.

**Specific Comments (not addressed above):**

**=== Reviewer 1 ===**

**Comparison to EWC.** This is a misunderstanding so we thank the reviewer for raising it. EWC was trained in the same batch (not online) regime described by the original authors, with additional access to information regarding task IDs and boundaries. NCTL has access to none of this information, and is the only method that has been trained subject to these additional constraints. The simple reason is that NCTL is the first method (to our knowledge) that can operate in this strictly more difficult regime. We will clarify and emphasize this critical point in text.

**Why is the method not compared to the original GLN?** The only similar experiment in the original GLN paper is the evaluating of robustness to catastrophic forgetting (negative backward transfer), owing to the simhash-like inductive bias induced by half-space gating. We can add this baseline to the next revision but note that vanilla GLNs have no forward-transfer properties like NCTL, which is the more difficult direction addressed in this work.

**Source code and reproducibiity.** One can find independent reimplementations of GLNs and FMN online. We will open source our implementation in time for NeurIPS. Unfortunately we were unable to obtain the necessary approvals in time for submission/review, but this has since been corrected.

**=== Reviewer 3 ===**

**On Scalability.** There is a slight misunderstanding regarding the asymptotic time complexity of the algorithm. Assuming a fixed network structure and a constant sized memory pool of $k$ items, the per example running time of NCTL is $O(k \log n)$. To process $n$ symbols, the time complexity is $O(nk \log n)$ with a space overhead of $O(k \log n)$. We will clarify this further in the next revision of the paper.

**Why does NCTL outperform the Oracles in Figure 4. Don't the Oracles also have potential for forward transfer?** Consider this training setup: 1000 steps on Task 1, 1000 steps on Task 2, and another 1000 steps on Task 1. The stronger Oracle maintains two networks, one trained on 2000 steps of T1 and the other on 1000 of T2. NCTL (using the exact same underlying network/params) is exposed to all 3000 steps in an unlabelled stream, thus the only way it can outcompete the Oracle on T1 is to leverage information from the 1000 steps on T2. By definition this demonstrates positive transfer. We will clarify our explanation.