We thank the reviewers for their detailed and thoughtful comments and provide point-by-point responses below.

=== Reviewer 1 ===

For Table 2, on the SARCOS dataset, the G-GLN is trained with 1200 epochs. Are all models trained with the same number of epochs? The G-GLN was trained for 1200 epochs and all other models were trained for 54,000, as per original paper. We have since run the G-GLN for 2,000 epochs and achieved a state-of-the-art test MSE of 0.11.

Similarly for Table 1, are all models trained for 40 epochs? Yes, including the G-GLN.

I am curious if there is an advantage for the G-GLN model in terms of training epochs. G-GLNs are designed (e.g. convex local loss) to be more data efficient than contemporary methods and truncating results at 1 epoch still yields strong performance. These 1-epoch results will be added to the revision: Boston: 3.37 | Concrete: 7.33 | Energy: 2.33 | Kin8nm: 0.11 | Naval: 0.00 | Power: 3.92 | Protein: 4.38 | Wine: 0.63 | Yacht: 6.21

=== Reviewer 2 ===

Why "side information" is defined as the input features for an input example? The G-GLN architecture is very different from a standard MLP. In the case of a function $y = f(x)$, the input features "x" are input to the GLN as the side information. The side information is used for selecting active weights. The neuron uses the active weights to reweight the distributional predictions coming from the neurons in the previous layer (or "base predictions" in layer 0). This difference gives rise to the advantages of GLNs w.r.t. contemporary methods.

The introduction of related work is not sufficient, and more work on GLN should be given to reflect the advantages or difference of the proposed method, such as the difference from B-GLN: We agree that a more detailed treatment of B-GLNs is valuable for unfamiliar readers. This was pruned so that we could focus the 8 pages on our specific contributions, but we will use the additional page available to accepted papers to expand on connections to related work.

More detail experimental analysis should be connected with your objective function. We thank the reviewer for this comment but are unsure what they intend by experimental analysis. Our objective function is justified theoretically and underpins the success of the whole set of experimental results. In particular, using log-loss for density estimation is standard, but in our work it is foundational because it yields a closed-form representation for a product of exponential-family experts trainable via a convex loss function. This allows us to leverage the no-regret properties of online gradient descent to find a set of weights that gets increasingly close to the maximum likelihood solution as more data is seen.

=== Reviewer 3 ===

One could argue that the proposed approach is a straightforward extension of [2] and [3]. We agree that, once the connection to a weighted product of experts and exponential family members is observed, it is straightforward to obtain an online learning formulation for a weighted product of Gaussians. However, it is certainly not obvious that this gives rise to a natural GLN formulation that works well in practice. This result should be of broad interest to the community, and we believe that the simplicity of the resulting algorithm is actually one of its key strengths given its strong performance on a broad number of tasks.

Some important details of the proposed approach are relegated to the supplementary material. This point is well-taken. Keeping a paper accessible yet interesting to the broadest cross-section of the community is a difficult trade-off, and we will certainly use the additional page of space available to accepted papers to upstream this material.

The authors do not address ... interpretability and robustness to catastrophic forgetting. Thank you for flagging this. These are properties that follow directly from the architecture and inductive biases of GLNs more generally (as has been shown in previous work on the B-GLN), but we could not find commonly accepted catastrophic forgetting / continual benchmarks for regression problems (beyond toy tasks like sine waves). We will revise the manuscript to clarify that the transfer of these properties to the G-GLN has not been shown in this work.

=== Reviewer 4 ===

No concerns were raised by Reviewer 4 other than to indicate that they would appreciate a more detailed assessment of broader impact. We thank the reviewer for their general comments and will include a more comprehensive impact statement in the revised submission, particularly with respect to the (positive) environmental and (potentially negative) privacy implications of improved data efficiency and online modeling.