Thanks for all the comments! We answer major comments from each reviewer below; we'll fix the minor ones.

**REVIEWER 1:** **"This paper ranks high in novelty...The experimental results are strong, especially on Text modelling. Moreover, the proposed method significantly computationally more efficient compared to competing approaches...The paper is very well written and easy to understand."** Thanks for the high praise!

**"1. The authors should consider a wider array of distributions to more convincingly demonstrate the capabilities of the proposed flow layers."** We include many synthetic discrete problems in the Experiments section: full rank discrete distributions, addition, and Potts models. We decided against including more continuous/ordinal data problems as our flows don't take advantage of ordinal structure in the Categorical states, and there are many practical problems without this structure.

**"2. Some important details are unclear. E.g. what is the base distribution for sampling? Is it the factorised marginal distribution? If it is how is it estimated in high dimensions (given that the number of data samples needed for an accurate estimate would grow exponentially)?"** We use factorized Categorical base distributions for bipartite flow models and correlated Categorical base distributions parameterized by an autoregressive neural network for autoregressive flow models. Note that the number of parameters for the base distribution is far less than the total number of possible states (e.g., $K \cdot D$ for factorized Categorical instead of $K^D$).

**REVIEWER 2:** **"Originality: This paper is the first demonstration of flow-based models to discrete data. As such, the work is fairly novel....That being said, the main technical contribution amounts to...on top of the existing techniques. I view this simplicity as a benefit of the approach, but some may view this a simple extension of existing techniques."** Thanks for the high praise! We agree about simplicity being a benefit. We explicitly designed discrete flows to be natural extensions of the continuous case.

**"Quality: The technical and experimental aspects of the paper are well-executed. Clarity: The presentation of the approach is incredibly clear."** Thanks!

**"For the most part, the experiments section is also clear. Some details of the models and training set-up are unclear, particularly in the toy examples from sections 4.1 - 4.3. Additional details in the supplementary material would help to clear up confusion."** Great idea. We'll clean this up .

**Adding discrete image datasets.** Demonstrating discrete flows on image data is a good idea. So far, we focused on text applications to show a domain normalizing flows hadn't yet been applied to. It's a ripe area given that both flexible modeling with bidirectionality and nonautoregressive generation are of huge interest for text. We'd love to explore images in future work, in particular pushing on the ordinality of pixel intensities to better handle data quantization. Hoogeboom et al. (2019) provide excellent complementary work in that direction.

**"In Section 3.1, 3.2., or in the supplementary, it would be helpful to have an expanded discussion around when discrete flows are invertible and what difficulties there are in ensuring this aspect. This discussion could also include the invertible discrete functions alluded to in Section 5."** In the revision, we'll add a section about the expressivity of discrete flows, what the set of transformations are (permutation-based), and designing parameterized invertible discrete functions.

**REVIEWER 3:** **"The approach taken at this point is to ignore the problem, and employ the straight-through estimator (which the authors argue work well for problems where K is not too large)."** The straight-through estimator is effective and commonly used in many discrete optimization problems to compute gradients through the non-differentiable argmax operation. In future work we hope to investigate whether other gradient approximations (including non-gradient based optimization) improve performance.

**"The paper is clearly original, and I imagine it will be of great significance (it brings two interests together, namely, flexible flow-based density estimation, and modelling discrete data)."** Thanks!

**"Since you count on backpropagation via straight-through estimator (STE), the derivative of mod K becomes relevant (as it will be necessary for chain rule to update the parameters of the NN that predict flow parameters). The best I could gather is that it's probably 1 everywhere except when (u + sigma x) is exactly divisible by K. Is that correct?"** You're correct that the operations are non-differentiable. The goal with continuous relaxations is not to compute the true gradients, but rather to provide a useful signal for improving the loss. We'll add these nuances to the paper.

**"You mention STE works well if K is not too large, but is that all?"** That's a great question. Depth (number of flows) affects gradients, since the bias explicitly accumulates as each flow uses an approximation. This is also the case for dimensionality (sequence length). We haven't found complexity of the networks parameterizing the flow to make a difference, but this requires further investigate. Note an additional complication follows your above point: "true gradients" are not well-defined. So instead of examining gradient bias, bias may be formalized by comparing, for example, the minima from the discrete loss to the minima from the relaxed loss.