Table 1: FED, LM score & perplexity on EMNLP 2017 News. Lower is better. Addition to metrics in the submitted paper, we evaluate perplexity on the self-generated samples to verify the correctness. The discriminator constraints and the entropy maximization are used in all experiments.

Estimator	FED		Perplexity		
	Train	Val.	Self	Train	Val.
Gumbel-Softmax	0.0141	0.0218	13	5267	6369
REINFORCE	0.0117	0.0182	25	68	72
Taylor	0.0105	0.0149	26	67	72

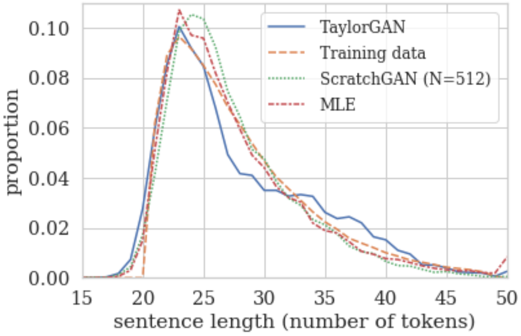

Figure 1: Sentence length statistics.

Reviewer #1

- **Response to Weakness 1, 2**: Regarding traditional estimators for discrete random variables, there are comparisons with REINFORCE and straight-through in Table 2†, discussions in line 133† and Appendix C†. Gumbel-Softmax was reported with bad performance in Zhu *et al.* [3], which is confirmed in our experiments. As shown in Table 1, Taylor estimator outperforms Gumbel-Softmax on FED and perplexity. We argue that the inferior performance is the consequence of biased and spiky distribution explained in both line 47-51† and the unusually high perplexity on real data, even with temperature annealing during the training phase [2].

Reviewer #2

- **Response to Question 1**: If the estimation is performed with Monte-Carlo sampling on $\mathbf{y}$ by a bad proposal Γ, the variance can be high. However, in equation (11)†, the gradient is calculated by *summing* up all $\mathbf{y}$s instead of sampling. Therefore, the variance is reduced by incorporating more samples if Γ is uniform enough. The choice of distribution is *arbitrary* (if computed efficiently), so there is no such “target” distribution to be fit or accuracy concern in equation (13)†. We have experimented kernels such as Epanechnikov and tricube, but none of them outperforms Gaussian. Which kernel function achieves the best balance between bias and variance is an interesting question, and we leave the theoretical investigation to future study.
- **Response to Question 2**: We evaluate the results with perplexity as well as language model scores. The perplexity in Table 2† refers to the inverse of the per-word probability of the model generating the *validation data*, which is independent of the generated samples. We guess the metric you mentioned may be the “LM score” defined in line 199†, for which we indeed trained another model following the settings of de Masson d’Autume *et al.* [1], mentioned in line 198†.

Reviewer #3

- **Response to Weakness**: The quality-diversity curve given temperature sweep is plotted in Figure 3†, which is mentioned in line 219†. We will emphasize this in the caption of Figure 3† in the final submission.

Reviewer #4

- **Response to Question**: The shorter samples in Table 4† are not a expected behavior but a coincidence. TaylorGAN does not tend to generate shorter sentences, as shown in Figure 1. de Masson d’Autume *et al.* [1] has found some correlation between the sentence length and the FED score and then designed their model accordingly. We, on the other hand, do not utilize this correlation. Among our metrics, the BLEU score penalizes short sentences by an exponentially decaying term.

References

[1] C. de Masson dAutume, S. Mohamed, M. Rosca, and J. Rae. Training language gans from scratch. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 4300–4311. Curran Associates, Inc., 2019.

[2] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax, 2016.

[3] Y. Zhu, S. Lu, L. Zheng, J. Guo, W. Zhang, J. Wang, and Y. Yu. Texygen: A benchmarking platform for text generation models. In *SIGIR on Research & Development in Information Retrieval*, pages 1097–1100, 2018.

† refers to the submitted paper and the supplementary material.