Response to the reviewers, ICLR2023

UMass Amherst

November 18, 2022

1 Reviewer 1

1.1 Weaknesses

1. **Not enough motivation for constraining in terms of Lipschitz functions, since neural networks are not a good representative of all Lipschitz functions with a certain Lipschitz constant. Why not gradient-norm penalty?**

   Response: We impose the Lipschitz constraint directly on the neural nets using spectral normalization, [Miyato et al.(2018)Miyato, Kataoka, Koyama, and Yoshida]. But the referee is correct, we could have also used a soft constraint as in GP-WGANs [Gulrajani et al.(2017)Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville] or in Section 4 of [Birrell et al.(2022)Birrell, Dupuis, Katsoulakis, Pantazis, and Rey-Bellet]. We included some relevant details in Appendix E. ”Neural network architectures” paragraph.

2. **The claim that Lipschitz constant makes the particle system ”stable” should be made more precise. Stability in what sense?**

   Response: By that statement we mean: (1) stability in the numerical analysis sense in terms of CFL condition. This is discussed in some detail in the Appendix. Furthermore, (2) stability due to regularization of the KL divergence by the Wasserstein (see Theorem 2.1 part 2) that allows us to compare empirical distributions of data. We have added an explanatory Background subsection in Section 3.

3. **The infinite-speed of diffusions is true, but diffusion is not the only way to implement Fokker-Plank. It can also be implemented deterministically using $\nabla \log$ (density) which is Lipschitz under standard assumptions on density.**

   Response: We included such a discussion on determinstic probabilistic flows in the new version of the paper, both in the main text and in the appendix. However we want to point out that the Lipschitz condition mentioned by the referee is practically uncheckable in generative modeling (but can be true in sampling problems based on maximum principle arguments for parabolic PDE). Furthermore such regularity conditions can be checked only if there is a well-defined density. In our context we have the ability to work directly with the empirical distribution of the data.

4. **Not in depth discussion of the numerical results, conclusions, and limitations.**

   Response: We trimmed, reorganized and summarized our findings in Section 6 and added more discussion of results in the Appendix.

1.2 Clarity, Quality, Novelty And Reproducibility:

The paper is very clear, and has high quality. The idea of using restricted variational forms of f-divergences for gradient flow appears in the following paper. But the combination of using the Lipschitz constrained class and latent space is original.

J. Fan, Q. Zhang, A. Taghvaei, Y. Chen. "Variational Wasserstein gradient flow", ICML 2022

Response: We would like to thank the referee for pointing out this highly relevant paper. Restricted variational forms of f-divergences have appeared in numerous papers in the literature. However the paper [Birrell et al.(2022)Birrell, Dupuis, Katsoulakis, Pantazis, and Rey-Bellet] provides a unifying approach that also encompasses previous methods, we refer to that paper for both literature discussion and new results. Both the present manuscript and the paper cited by the referee [Fan et al.(2022)Fan, Zhang, Taghvaei, and Chen] focus on a gradient flow perspective for their corresponding divergences. However, the time-discretization of their respective gradient flows is different. We use the forward-Euler discretization using the discriminator while in [Fan et al.(2022)Fan, Zhang, Taghvaei, and Chen] the authors use the implicit JKO scheme. Therefore, in [Fan et al.(2022)Fan, Zhang, Taghvaei, and Chen], just like in GANs, an additional Neural Net needs to be introduced in their algorithm compared to ours. The connections between the two papers remain intriguing because of the similarity between the 2-Wasserstein-based JKO and our 1-Wasserstein infimal convolution formula in Theorem 2.1.

1.3 Summary Of the Review:

The paper is well written in general and contain original ideas. However, it needs more motivation for considering Lipschitz constrained class. Also, it discusses two very nice and independent ideas in one single paper. Gradient flows in the latent space is independent of Lipschitz constrained gradient flows. And it is not clear that which of these two ideas are responsible for the improved numerical results compared to GAN

Response: We have explored GPA with and without latent spaces in the Examples. But definitely a latent space can be of huge help, and in addition we also have related theoretical guarantees in Theorem 5.1

2 Reviewer 2

2.1 Weaknesses

1. **The motivation for the approach could be strengthened. In particular, it is unclear what the benefits are in terms of utilizing this Lipschitz regularized formulation versus the original variational formulation over continuous and bounded functions. More motivation and discussion along these lines would improve the readability of the paper.**

   Response: We have added a new explanatory Background subsection in Section 3 to address these issues. We also added a corresponding numerical experiment.

2. **Some experimental results show that the generative capabilities do not improve with more data. For example, the MNIST generation results do not appear to improve with more samples both visually and quantitatively (as given by the FID score).**

   Response: We agree this is an issue and we are currently working on resolving the engineering aspects of our software for the specific imaging applications and in general on upscaling our methods. Here we have focused primarily on examples that can give a first demonstration of the capabilities of the proposed algorithms, and on what problems they are best suited for, for instance the ability to learn from very few samples.

2.2 Clarity, Quality, Novelty And Reproducibility:

Clarity. The paper is fairly clear given the technical nature of the results. There are some areas that can be improved. In particular, there is a lack of motivation for consideration of Lipschitz regularized functionals over continuous and bounded functions. I can see similarities with W-GAN in this regard, but more discussion along these lines would be nice. Also, Figure 1 comparing the differences between GPA and GAN is difficult to read and is quite unclear.

Response: To address these issues, we added more discussion under “related work” in the Introduction, a new Background component in Section 3 and added a new Section in the Appendix on comparisons with other methods. We also refer to the Table in Section 5 using the concept of mobility as means for comparing with GANs.

Novelty. To the reviewer’s knowledge, the use of this general Lipschitz regularized class of divergences for generative modeling appears novel. The technical tools, however, appear heavily inspired by Birrell et al., and it is not clear what are the new contributions on the theoretical side.

Response: The main new elements compared to that paper are the new Gradient flows and assocuiated particle dynamics; the use of latent spaces and “latent” particles; and, the related Data Processing Inequality for auto-encoders in Theorem 5.1 that provide an *a posteriori* error estimate for the approximation of the real space model by the one built on the latent space.

General comments. It is curious that the sample generations do not appear to improve with more examples. In fact, in the MNIST example, the FID score is slightly worse for 2000 samples versus 200 samples. Do the authors have a sense of why this is the case? The theory seems to suggest that having a perfect encoder decoder pairing is important. Is there a connection here to a lack of a perfect encoder decoder pairing? Also, why does it seem to be that the model performs well with few samples?

Response: Indeed these are both good questions. For now GPA seems to perform really well in cases with few available samples, as demonstrated in the experiments. From a theory perspective the performance on low number of samples seems very good due to the use of the (f, Γ)-divergence that allow to transport efficiently arbitrary empirical to target empirical distributions. We are currently working on resolving the issues related to scaling up and optimizing our algorithms.

3 Reviewer 3

3.1 Weaknesses

1. **I found the novelty of the present paper limited. To me, the main novelty is (6), although I think the derivation is more or less identical to that of Dupuis & Mao (2022) from the KL case.**

   Response: The variational derivative is indeed a key tool, but that is only one of the ingredients needed in the proposed methods, which on the theory side also include: gradient descent, interacting particles formulation, dissipation estimates for the gradient descent, and a new DPI for auto-encoders to guarantee performance when using latent space calculations.

2. **The application prospect of the proposed algorithm is questionable. Without using the autoencoder, to me, the proposed algorithm is just to subsample the target empirical distribution. I think there is a lot to be explored here. For instance, if we use a very small number of particles (compared to the number of samples in the target distribution), can we still represent the target distribution faithfully? If so can we use the proposed method as an alternative to k-means?**

   Response: Indeed so far the performance on low number of samples seems very good, largely due to the use of the (f, Γ)-divergence that allows to transport efficiently arbitrary empirical to distributions to target empirical distributions (combining the good mass transport features of Wassesrtein and the mass redistribution features of KL). We have experimented with various number of particles relative to the samples in the target distribution but we have not yet explored carefully the regime suggested by the referee. Indeed this is an intriguing direction, we greatly appreciate the suggestion.

3. **I don't think comparing with GAN is reasonable. GAN is capable of generating endless streams of new samples, whereas for the proposed method you must fix the number of particles ahead of time (it is also unclear to me how many particles were used in the experiments). A good baseline to compare would be, say minimizing the Wasserstein-1 (or W2) distance between particles and the target distribution and comparing the resulting particles. As a result, I do not find the experimental results convincing.**

   Response: The similarities and differences with GANs are discussed in the Table in Section 5. However, we are not sure we completely understand the suggestion by the referee. The referee proposes to minimize Wasserstein distances from the target but we are not sure over what models the optimization should be done.

3.2 Clarity, Quality, Novelty And Reproducibility:

Quite a lot of sentences are not precise and thus confusing. For example, just above Section 4, "This is in sharp contrast with the Fokker Planck equation". How is the finite speed of propagation related to FP equation?

Response: Hopefully these ambiguities are all fixed now: we added more discussion in a new Background component in Section 3 and added a new Section in the Appendix on comparisons with other methods, where the finite speed of propagation is discussed, see also the Appendix Section on numerical analysis aspects of PDEs.

Theorem 2.1 (except the 4th point) seems somewhat irrelevant to the rest of the paper

Response: Theorem 2.1 provides background for the proposed divergences, for instance demonstrates how they interpolate between f-divergences and Wasserstein, while retaining good properties of both due to the infimum convolution representation.

I found the Data Processing Inequality and the argument on the mobility concept in Section 5 confusing and very handwavy. Aside from applying some change of variable in (13)(14) I don't get what the selling point is in this section.

Response: The Data Processing Inequality for auto-encoders in Theorem 5.1 provides an error estimate for the approximation of the real space model by the one built on the latent space: in the spirit of numerical analysis of finite elements, e.g. [Ainsworth & Oden(1997)Ainsworth and Oden] we can use the right-hand side of Theorem 5.1 as an *a posteriori* numerical estimate to provide computable performance guarantees, since the upper bound is in the tractable latent space.

The experiments section is very confusing. It feels like the authors try to pack all the numbers and figures in the main text without properly explaining most of them. I think it would be better to focus on two experiments with greater details in the main text and move the rest to the appendix.

Response: We have completely rearranged the Section and followed the suggestions of the referee for a more clear and precise delivery of the findings. Remaining material was moved in the Appendix.

4 Reviewer 4

4.1 Weaknesses

1. **The experiments are somewhat weak. First, the experiments only show results for MNIST and Gene Expression (student-t seems more like a proof of concept than an experiment). These datasets are pretty toy (MNIST is rather low dimensional and easy to solve with simpler methods, and the Gene Expression dataset was just introduced to show merging).**

   Response: Of course this is true regarding MNIST. One of the reasons we considered the gene expression data sets is that they are a fundamental tool for life sciences in general. Here we have focused primarily on examples that can give a first demonstration of the capabilities of the proposed algorithms, and on what problems they are best suited for. We are currently working on upscaling our methods to much larger data sets.

2. **Furthermore, for MNIST in particular, it is very disconcerting that, while the method does perform well in the low-data regime, increasing data from 200 to 2000 samples doesn't improve the method at all (either visually or numerically). This seems like a potentially big drawback for the purported applications. In particular, if the comparison is on the low data regime only, the authors should compare with other methods that also look at this.**

   Response: Indeed we agree this is an issue, we are currently working on resolving the engineering aspects of our software for the specific imaging applications. The very low data regimes give promising results, especially in examples such as the gene expression data set, where unlike in generative modeling for imaging problems, sophisticated data augmentation methods seem harder to develop.

3. **The MNIST results also should be compared with modern methods. In particular, Wasserstein GANs are around 4 years old, and modern methods such as diffusion.**

   Response: Regarding WGAN, our goal here was to compare GPA to another Wasserstein-based generative model. We also compared GPAs to IPM-based methods that were developed more recently, such as MMD gradient flows and KALE, see also Appendix. We have already included a comparison with score-based diffusion models such as SGMs where annealing is necessary. In the context of that example we show that GPAs do not need to be injected with noise. GPAs are also capable of generating distributions with heavy tails where SGMs fail, see the last example in Section 6.

4. **There should be a connection with continuous diffusion models (in particular diffusion also has a similar connection with PDEs like the Fokker-Planck Equation) that the authors should elaborate on and include.**

   Response: We have added related references in the Introduction and a new Section in the Appendix on additional discussion on related work including on score-based diffusion models.

5. **It seems like the method might be pretty computationally expensive (e.g. Table 2 (b)). Could the authors comment more on the training time, as GANs are reasonably fast to train compared with most differential equation methods.**

   Response: In this first paper we focused on the conceptual development of the method and the necessary mathematics, and we are currently working on the full optimization of the algorithms.

4.2 Summary Of The Review:

Overall, I lean (slightly) to accept the paper. This is mostly due to the theoretical niceness of the paper (including it's many connections with gradient flows, f-divergences, and PDEs). **What's stopping me from being fully supportive is the experimental section, which includes relatively mixed results on small toy datasets. However, given the theoretical nature of the work, I do believe that the technical developments outweigh experimental shortcomings.**

I would also ask the authors to include a section on diffusion models and the connections therein if possible.

Response: We appreciate all the suggestions, and indeed we have included a full discussion on continuous-type diffusion models.

References

[Ainsworth & Oden(1997)Ainsworth and Oden] Mark Ainsworth and J.Tinsley Oden. A posteriori error estimation in finite element analysis. *Computer Methods in Applied Mechanics and Engineering*, 142(1):1–88, 1997. ISSN 0045-7825. doi: https://doi.org/10.1016/S0045-7825(96)01107-3. URL `https://www.sciencedirect.com/science/article/pii/S0045782596011073`.

[Birrell et al.(2022)Birrell, Dupuis, Katsoulakis, Pantazis, and Rey-Bellet] Jeremiah Birrell, Paul Dupuis, Markos A Katsoulakis, Yannis Pantazis, and Luc Rey-Bellet. (f-γ)-divergences: Interpolating between f-divergences and integral probability metrics. *Journal of Machine Learning Research*, 23(39):1–70, 2022.

[Fan et al.(2022)Fan, Zhang, Taghvaei, and Chen] Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6185–6215. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/fan22d.html`.

[Gulrajani et al.(2017)Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf`.

[Miyato et al.(2018)Miyato, Kataoka, Koyama, and Yoshida] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. 02 2018.