Qeios

Peer Review

Review of: "Enhancing Sample Generation of Diffusion Models using Noise Level Correction"

Luke Shaw[1]

1. Departament de Matemàtiques, Universitat Jaume I de Castellón, Spain

I think that the results of the paper are relevant and interesting, and the general idea of the approach is well-motivated. I do not think it is necessary to perform more experiments, as many have been performed, and the proposed methods are probably competitive with state-of-the-art plug-and-play models such as DDRM, although this is still not fully clear from the experiments in section 4. However, I have two big complaints to make.

The first is that the presentation is rather sloppy and needs to be improved, as there is a lot of needless abuse of notation and imprecise wording. A non-exhaustive list:

- The phrase “the noise level no longer approximates distance (2)” does not refer to anything, since it cannot refer to Eq. (2), and there is no other candidate for a referent. The phrase “the randomly sampled noise naturally concentrates around the norm $\sqrt{n}$” is sloppy - it is true that the expected value of the norm of the noise is $\sqrt{n}$, assuming normality of the noise, but that is somewhat different from the given statement.
- The training process does not use the expectation $\mathbb{E}$ - it uses an average over a finite training set. Especially when the authors introduce the training of the noise correction network (NCN), this obfuscates matters, even more so because it is unclear whether the NCN is “trained alongside” the supposedly “fixed, pre-trained” denoiser network or not; presumably in the cases of toy experiments, yes, in the other examples, no. There should also be a citation regarding how the training is carried out.
- In eq.(10), one must put over what set the inf is taken, presumably over $x\in\mathbb{R}^n$.
- The NCN $\widehat{r}_t$ is introduced as “the residual”, but the residual of what? Please put an expression so the reader has some idea.

 doi.org/10.32388/2NORJT

- In section 3.3, when the “initial sample estimate” is introduced as $x_{0\vert t}$, a hat should be used to match Eq. (5).
- My biggest complaint is that the notation relating to the projective method is a mess - for example, it does not make sense to have the projection of $\widehat{x}_{0\vert t}$ as $x_{0\vert t}$, since this suggests that the latter is somehow errorless. The use of $x_{(k)}$ is not necessarily a bad route to go, but the presentation of the algorithm in Eqs. (20), (21) is convoluted, and terms are introduced prior to definition or without definition. Eq.(21) is insufficient to define the process since x_{0} should be defined, which is only the case later, but then it is defined as $\sigma_{max}\bar{\epsilon}$: σ_{max} is not defined until the algorithm appears, and $\bar{\epsilon}$ is never defined. It should be rewritten from scratch. Also, Algorithm 2, line 6 uses Proj rather than proj; certain lines are in blue but are not referred to as special at any point in the text.
- The presentation of Algorithm 1 is marginally better but still leaves a lot to be desired, and while perhaps it is useful to have η as a sliding parameter (as in the DDIM paper), this is not justified or referenced at any point, so I don't see why one cannot just include the DDIM and DDPM cases in a more readable way. In general, it could be much more readable; for example, one has $\widehat{\sigma}_{t-1}=(1+r)\sigma_{t-1}$, but this is written in an unnecessarily different way, which hinders understanding. There is also a typo in line 3 with a subscript bracket _).
- η is redefined three different times, with different meanings each time. I refuse to believe that this is necessary.

My second major complaint is that the relative performance of the NLC networks and the uncorrected networks is not well-clarified in section 4. For example, it is noted that r_{θ} is roughly 10% of the size of the main denoiser network; one would then think that for a fair comparison, the FID values in Tables 1 and 2 should be multiplied by 1.1 for the NLC methods. However, if one reads further on, in Table 9 in section D.2, the inference times for NLC are roughly 20% larger (18% for CIFAR10, 27% for ImageNet), so perhaps even a factor of 1.2 would be more reasonable. I also note in passing that the authors incorrectly state a 10% figure based on the same Table 9, which is eminently false.

Furthermore, the true cost of IterProj-NLC in section 4.3 is unclear, since it is an iterative method. For example, does it have Kmax set to 100, so that it never uses more iterations than the competitor non-noise-corrected methods? And if this is the case, does it ever use less? That would be interesting, and if it were the case, it would be an additional benefit of the algorithm.

Finally, in section 4.1, I don't see why one divides by $\sqrt{n}\sigma_t$, when a more suitable distance estimation bias would, I think, be given by dividing by $dist_{\mathcal{K}}(\widehat{x}_t)$.

Declarations

Potential competing interests: No potential competing interests to declare.