A PRACTICAL GUIDE ON CURRENT AND FUTURE GENERATIVE MODELS FOR CONTINUOUS DATA

Nam Tran
Independent Researcher
namhoangtran1590@gmail.com

ABSTRACT

There is a language barrier that confusticates a broad audience's understanding of generative model. Consequently, specifications given to researchers to design better models at non-academia organisations are also confusing. To overcome this barrier: in section 1, I help beginners casually catchup to experienced readers and clarify what this paper is not about; in section 2, I re-frame all current generative models under 1 'language': the Stochastic Interpolant framework [1]; in section 3, I setup a simple pretraining experiment on 2D toy data distribution that is reproducible on Google Colab with 1 CPU [1]. As this experiment forces all generative models to learn data with little training samples and a very small neural network, it reveals a narrative threading across all current generative models that is aligned with the current research trend to discover a new-better generative model. Finally, I test top-3 performing generative models on a real-world image dataset (i.e. CIFAR-100), the result shows that their performances with respects to each other remain consistent, thus this opens an option to meaningfully stress-test any generative model before making costly pretraining on real-world dataset.

1 Background

Recall the Stable Diffusion where it generates new image without needing user prompt but generates whichever user prompt to a neat extent. A general pipeline of any arbitrary generative model for continuous data (i.e. image data) including Stable Diffusion is illustrated in Figure 1

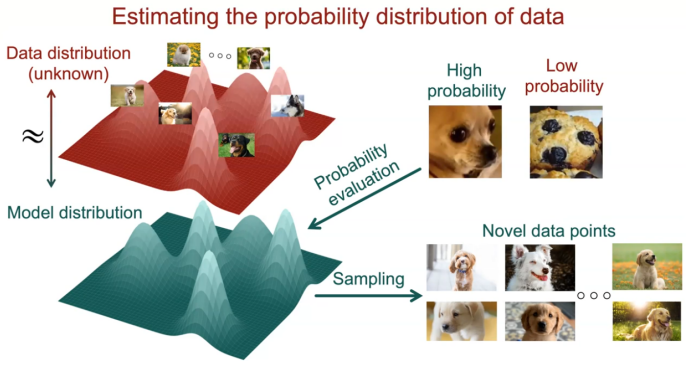

Figure 1: Pipeline of generative model for continuous data, screenshotted from the oral presentation of Yang Song PhD thesis: first, a model learns to approximate the distribution of training dataset; second, a probabilistic evaluation (usually Fréchet Inception Distance) compares how close ground truth mean vs approximated mean and ground truth covariance vs approximated covariance; finally, generate new data is equivalent to sample new data from the model distribution

[1] All experiments and ablation tests can be founded on my github:

1.1 Data distribution

Consider an example 256 $\times$ 128 pixels image in Figure 2, it can be viewed by a density plot per image's channel (i.e. red, green, blue):

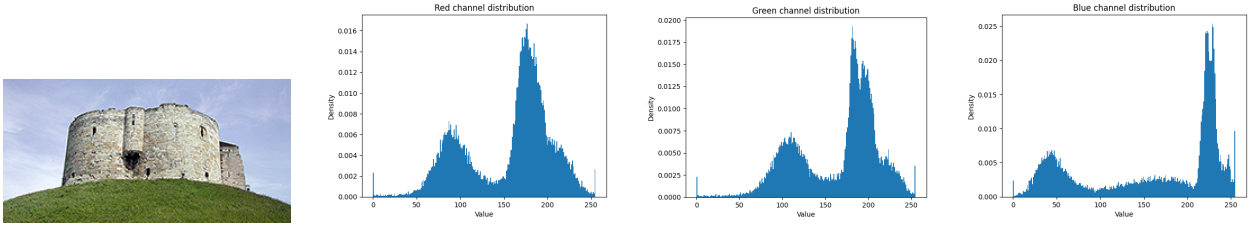

Figure 2: An example image distribution

Figure 2 illustrates a property: although each channel contains 256 $\times$ 128 pixels which means its linear span has 256 $\times$ 128 dimensions, it can still be described by 1 dimensional density function (i.e. image channel distribution). Can 1 density function describe a dataset of images, where each image has 3 channels? Yes, in that case the density function would be referred to as data distribution as illustrated Figure 1.

1.2 Linear span and ambient space

Consider an example 2D data distribution in Figure 3, where its linear span has 2 dimensions, but its ambient space is 3 dimensions. Here, there's a simple observation: a distribution's linear span has less number of dimensions than the ambient space that contains the distribution. The Manifold hypothesis is practically a continuation of this observation with 1 difference: the linear span of an arbitrary real-world data distribution has **far less** number of dimensions than the number of dimension that the ambient space has. Note, deep neural network's width (i.e. "hidden_dim" in model configuration) indicates the number of dimensions that an ambient space has, it's usually 128 or 768 or 4096 or 7168 depending on the model.

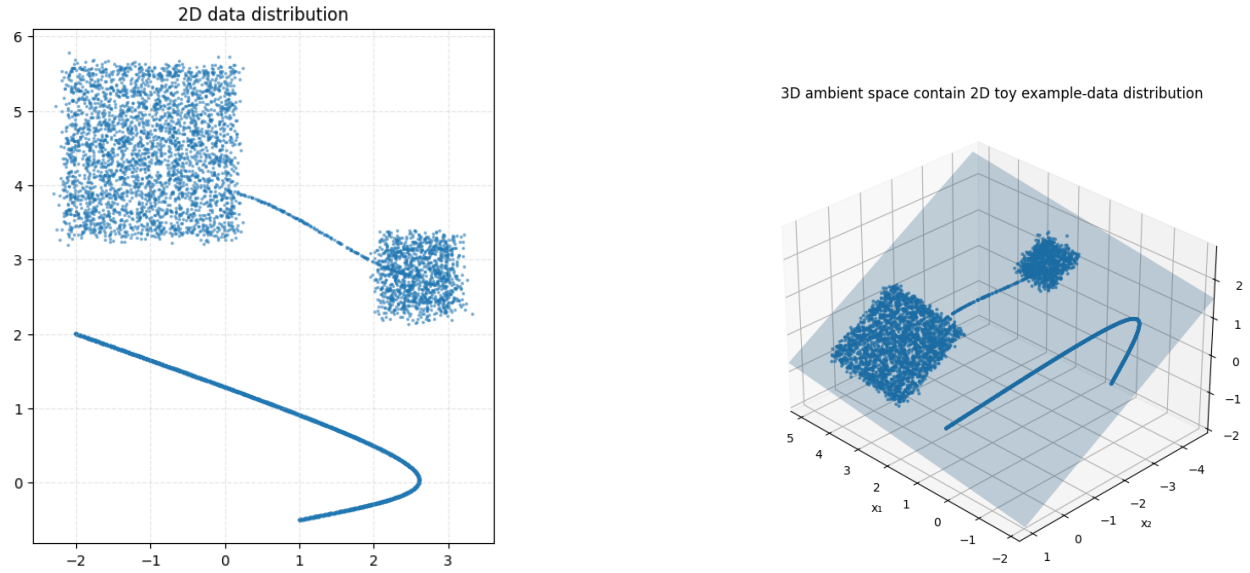

Figure 3: A data distribution and its 2D linear span, contained in 3D ambient space. Note, I intentionally phrase "data distribution and its 2D linear span" rather than "2D data distribution" to prepare beginner reader for the next subsection

For example, per figure 9 of [2], the data distribution of daisy images **likely** had 10D linear span, and was definitely contained in 128D ambient space; and according tables 1 and 2 of [3] the data distribution for word tokens **likely** had 5-6D linear span for Mistral7B language model, but was definitely contained in 4096D ambient space. Here, I bold

the word **likely** to emphasis a detail introduced in Figure 2: no one knows the exact number of dimension that a data distribution's linear span has. Still, people do know that all arbitrary data distribution obey properties 1, 2, 3 of the Curse of Dimensionality.

To understand properties 1 and 3, consider a simple thought experiment: given an euclidean distance between an arbitrary point A and arbitrary point B in a 1D linear span. Let's make a rule that the maximum distance between A and B is 1 and minimum distance between them is 0. Denote A's coordinate is A_1, B's coordinate is B_1: their euclidean distance is $|A_1 - B_1|$ can be any random value between 0 and 1. So I can consider the act of measuring $|A_1 - B_1|$ as sampling a random number of an arbitrary density function, denote as $g \sim p(g)$. In real life, when a person rolls a dice for a lot of time, they would notice that the mean of all observed results converge to the mean of all dice value, which is 3.5. The law of large numbers is a continuation of this empirical observation. Here, when g is sampled for a lot of time, the mean of all observed samples converge to the mean of $p(g)$. Note, this is valid because euclidean distance between arbitrary point is computed for a lot of times in practice.

So, the next step is to approximate the mean of $p(g)$, denote as $\mu = \mathbb{E}[g]$: any sampling-based method listed in [4] is fine enough for this task, I implement the Monte Carlo method because it's easy, which yields Figure 4

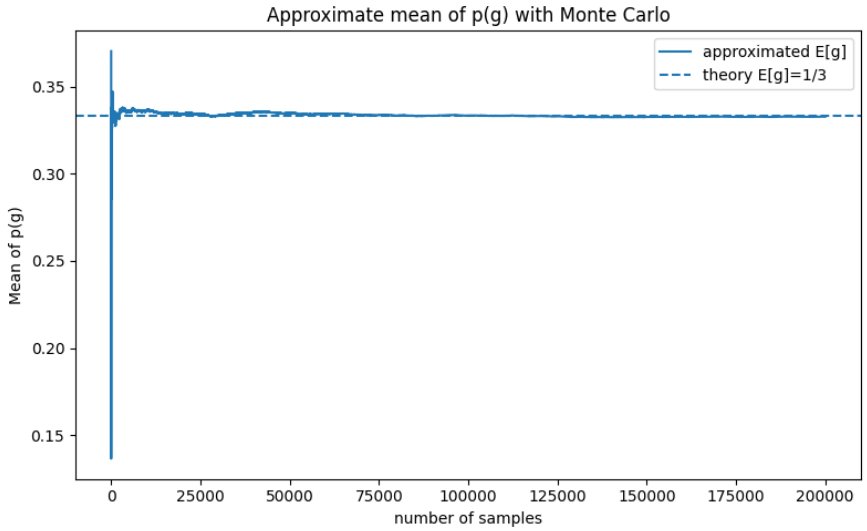

Figure 4: Approximated mean of $|A_1 - B_1|$ using Monte Carlo

Scaling up to d dimensional linear span: the euclidean distance between A and B is now $\sqrt{\sum_{i=1}^{d}(A_i - B_i)^2}$ so its approximate mean now has the form $\mathbb{E}[\sqrt{\sum_{i=1}^{d} g_i^2}]$, where i is the dimensionality index. For example A_1 is a coordinate of A along dimensionality 1. I can reuse Monte Carlo or any sampling-based method in [4], but it would defeat the main objective. That is, to understand the scaling in properties 1 and 3 of the Curse of Dimensionality, so this needs to be solved analytically. Σ can be taken outside of $\mathbb{E}$. But to take $\sqrt{}$ outside of $\mathbb{E}$, I need to use some inequality's trick. The idea is if $\mathbb{E}[\sqrt{}]$ is $\geq$ or $\leq$ than some simpler term with simpler form, I can continue on working with "some simpler term". I choose to use the Cauchy-Schwarz inequality because Equation 1 below can be explained visually to beginner using Pythagoras from high school math:

$$E[\|\boldsymbol{g}\|_2] := \mathbb{E}\left[\sqrt{\sum_{i=1}^{d} g_i^2}\right] \geq \frac{1}{\sqrt{d}} \sum_{i=1}^{d} \mathbb{E}[g_i] \tag{1}$$

Consider $\sqrt{\sum_{i=1}^{d} g_i^2}$: when $d = 2$, it becomes $\sqrt{g_1^2 + g_2^2}$. This is the form of Pythagoras theorem. Now, consider another thought experiment: let g_1 and g_2 forms a right angle triangle as illustrated in Figure 5, let's give a rule where $g_1 + g_2 =$ some fixed number, how does one adjust g_1 and g_2 such that it fits the rule and minimise the hypotenuse of this triangle?

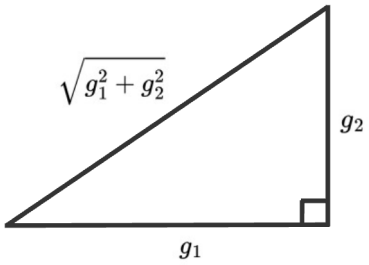

Figure 5: Right angle triangle formed by g_1, g_2 with the hypotenuse as $\sqrt{g_1^2 + g_2^2}$

The answer is assigning $g_1 = g_2 = \frac{\text{some fixed number}}{2}$. Plug this back to the right angle triangle from Figure 5, I arrive an observation that the triangle's hypotenuse is **at least** amounted to $\sqrt{\left(\frac{\text{some fixed number}}{2}\right)^2 + \left(\frac{\text{some fixed number}}{2}\right)^2}$. Let's assign some fixed number as $g_1 + g_2$, which means g_1 and g_2 are now also fixed numbers. In this current setup, the triangle's hypotenuse is **at least** amounted to $\frac{\sqrt{2}}{2}|g_1 + g_2|$ per Equation 2:

$$\sqrt{g_1^2 + g_2^2} \geq \sqrt{\left(\frac{g_1 + g_2}{2}\right)^2 + \left(\frac{g_1 + g_2}{2}\right)^2} = \frac{1}{\sqrt{2}}|g_1 + g_2| \tag{2}$$

For 3D right angle triangle, Equation 2 would have the form: $\sqrt{g_1^2 + g_2^2 + g_3^2} \geq \frac{1}{\sqrt{3}}|g_1 + g_2 + g_3|$. For d dimensional right angle triangle, Equation 2 would have the form: $\sqrt{\sum_{i=1}^{d} g_i^2} \geq \frac{1}{\sqrt{d}}\left|\sum_{i=1}^{d} g_i\right|$, which are the middle term and the right-hand side term in Equation 1.

1.3 Nonlinear span and non-bijective function

1.4 Approximating data distribution

How does a deep neural network parameterise a data distribution? An arbitrary-shape distribution

2 Introduction

In Figure 6 below, I use my research experience and insights from collaborating with senior researchers at non-academia organisations to lay out the priorities when designing a new generative model for practical usage. Note, lower pillar in this illustration indicates greater importance as the whole model implodes if all layers below can't support an arbitrary scaling up operation.

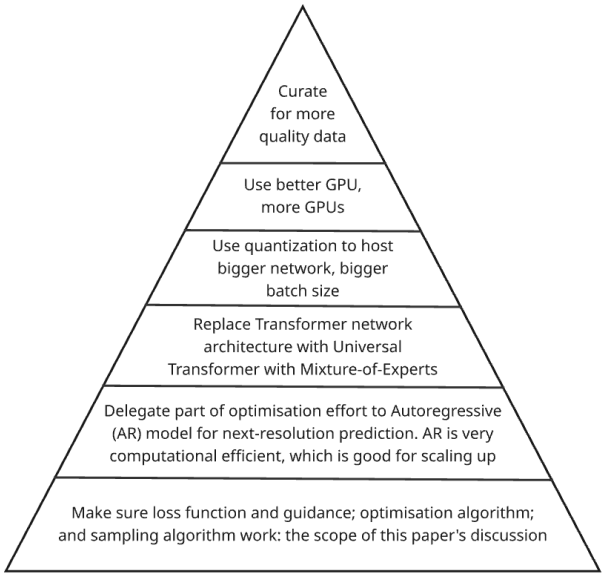

Figure 6: Priorities when designing a new generative model for practical usage. The scope of this paper's discussion is the most fundamental pillar yet least understood

3 Headings: first level

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula. See Section 3.

3.1 Headings: second level

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j|y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{3}$$

3.1.1 Headings: third level

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Paragraph Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

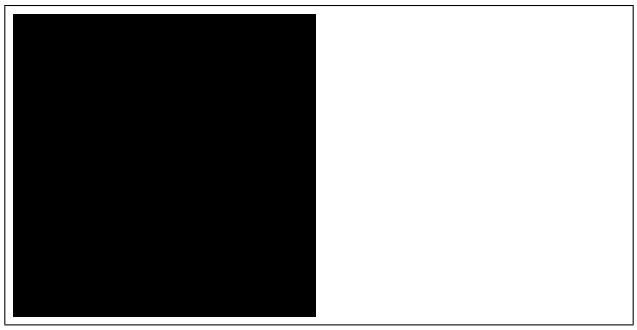

Figure 7: Sample figure caption.

4 Examples of citations, figures, tables, references

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated…

`https://www.ctan.org/pkg/booktabs`

4.1 Figures

Suspendisse vitae elit. Aliquam arcu neque, ornare in, ullamcorper quis, commodo eu, libero. Fusce sagittis erat at erat tristique mollis. Maecenas sapien libero, molestie et, lobortis in, sodales eget, dui. Morbi ultrices rutrum lorem. Nam elementum ullamcorper leo. Morbi dui. Aliquam sagittis. Nunc placerat. Pellentesque tristique sodales est. Maecenas imperdiet lacinia velit. Cras non urna. Morbi eros pede, suscipit ac, varius vel, egestas non, eros. Praesent malesuada, diam id pretium elementum, eros sem dictum tortor, vel consectetuer odio sem sed wisi. See Figure 7. Here is how you add footnotes. [2] Sed feugiat. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut pellentesque augue sed urna. Vestibulum diam eros, fringilla et, consectetuer eu, nonummy id, sapien. Nullam at lectus. In sagittis ultrices mauris. Curabitur malesuada erat sit amet massa. Fusce blandit. Aliquam erat volutpat. Aliquam euismod. Aenean vel lectus. Nunc imperdiet justo nec dolor.

4.2 Tables

Etiam euismod. Fusce facilisis lacinia dui. Suspendisse potenti. In mi erat, cursus id, nonummy sed, ullamcorper eget, sapien. Praesent pretium, magna in eleifend egestas, pede pede pretium lorem, quis consectetuer tortor sapien facilisis magna. Mauris quis magna varius nulla scelerisque imperdiet. Aliquam non quam. Aliquam porttitor quam a lacus. Praesent vel arcu ut tortor cursus volutpat. In vitae pede quis diam bibendum placerat. Fusce elementum convallis neque. Sed dolor orci, scelerisque ac, dapibus nec, ultricies ut, mi. Duis nec dui quis leo sagittis commodo. See awesome Table 1.

4.3 Lists

- Lorem ipsum dolor sit amet
- consectetur adipiscing elit.
- Aliquam dignissim blandit est, in dictum tortor gravida eget. In ac rutrum magna.

[2] Sample of the first footnote.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

5 Conclusion

Your conclusion here

Acknowledgments

This was was supported in part by.....

References

[1] Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions, 2025.

[2] Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning, 2021.

[3] Michael Robinson, Sourya Dey, and Shauna Sweet. The structure of the token space for large language models, 2024.

[4] Michael S. Albergo and Eric Vanden-Eijnden. Learning to sample better, 2023.