A Report on Calvin Luo's *Understanding Diffusion Models: A Unified Perspective*

Pranav K Nayak

The Paper

The paper presents a comprehensive view at the math behind generative modelling, with the aim of having the reader equipped to understand the rationale behind why Variational Diffusion Models (VDMs) are structured as they are.

It begins by stating that, very broadly, the goal of generative modelling is to learn the prior of the underlying data, $p(\mathbf{x})$. It talks about how this problem is not tractable, and so an approximator to this distribution, the *ELBO*, is maximized instead. The paper then derives the ELBO equation, going into detail about what is actually getting optimized by providing an alternative proof that makes explicit the abstractions adopted during the standard proof that uses Jensen's Inequality.

It then introduces the idea of Variational Autoencoders, and shows how maximizing the ELBO when both encoding and decoding distributions are learned leads to the reconstruction being optimized, as well as to the true prior being matched by the learned one. It builds on this by introducing Markovian Hierarchical VAEs, breaking down the math behind the ELBO when multiple VAEs are stacked on top of each other.

At this point, the paper introduces the reparameterization trick, allowing us to frame an expectation term containing the parameters we'd like to optimize in terms of a linear combination of our optimization parameters and a noise element sampled from a standard normal distribution. This allows us to not have to worry about differentiating an expectation in which the probability density is also parameterized by the variable we'd like to differentiate against.

Once it solidly establishes how the ELBO is used in generative modeling, it starts walking the reader through the derivations underlying VDMs, framing them as MHVAEs with a few assumptions:

- The latent and data dimensions are the same.
- The structure of the encoder distributions is not learned. It is instead set to be Gaussian, centerd around the output of the previous latent layer, and with a diagonal variance, such that the scale of the samples is preserved:

  $$q(x_{t+1}|x_t) = \mathcal{N}(x_{t+1}; \sqrt{\alpha_t}x_t, (1-\alpha_t)\mathbf{I}).$$

- The parameters of the encoders, along with the number of timesteps, is chosen such that the distribution of the latent at the final timestep is a standard Gaussian:

  $$p(x_T) = \mathcal{N}(x_T; \mathbf{0}, \mathbf{I}).$$

Having laid out these assumptions, the paper then walks through multiple derivations of the ELBO, providing intuitions for what is optimized behind the scenes in each of these derivations, before delving into alternative interpretations of what a VDM actually learns by reformulating the loss function at different points of the standard derivation.

Takeaways

Throughout my reading of this paper, I primarily was looking for answers to the following questions:

- Can the diffusion operation (sampling, centering, resampling), be framed in terms of groups? More specifically, can it be understood as an affine transformation, for which group structure is well-defined?
- Does the target of the encoding operation (that the final latent distribution be a standard Gaussian) affect any group structure that arose as an answer to my previous question? Does it preserve it? If not, does it simply change the group structure (i.e. the composition rule while retaining the actual group axioms), or does it break the structure entirely?
- If, in fact, the "directed" sampling operation does have group structure, is the set of these operations also a differentiable manifold?
- If all of the above are true for the noising operation, can anything similar be arrived at for the learned denoising operation?

The painstaking detail that this paper went into when walking its reader through the math has given me some intuition regarding the answers to my questions. This intuition is centered around

- The reparameterization trick, which I believe answers both the first and last of my questions:

  $$x_t \sim \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}) \rightarrow x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon; \quad \epsilon \sim \mathcal{N}(\epsilon; \mathbf{0}, \mathbf{I})$$

- The terms that get tucked away in the ELBO, that are made explicit during the paper's various derivations. Specifically, during many of the derivations, the "prior matching term" is arrived at:

  $$\mathbb{E}_{q(x_{T-1}|x_0)}[D_{KL}(q(x_T|x_{T-1})||p(x_T))].$$

  which has no trainable parameters, and thus does not affect any of the intermediate noising and denoising operations, and is only a function of the number of timesteps T. This looks to be a preliminary answer to my second question.

- During the same derivations, we arrive at the "denoising matching term":

  $$\sum_{t=2}^{T}\mathbb{E}_{q(x_t|x_0)}[D_{KL}(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))]$$

which encourages the learned denoising distribution to be as close as possible to the ground-truth denoising distribution. This means that, if the above term goes to 0, then the properties that are arrived at for $q(x_t|x_{t-1})$ can be extended to $q(x_{t-1}|x_t)$, which then hold for $p_\theta(x_{t-1}|x_t)$, answering my last question.

-