Some Notes on Generative Modelling

January 15, 2025

Abstract

I recently had a chat with a colleague of mine, who is an expert in Stochastic Processes in the context of Markov chain Monte Carlo methods. The purpose of the chat was to give them some broad introduction to (some version of) contemporary Generative Modelling, with an emphasis on principles which would be familiar and interesting to them, based on their skillset. I document here some notes which I made as a part of this chat.

The Task

The general setup of generative modelling is something loosely like the following: given data $\mathcal{D} = \{x_i : i \in [\![N]\!]\} \subseteq \mathcal{X}$ (perhaps nominally drawn from some distribution $\pi \in \mathcal{P}(\mathcal{X})$), construct a distribution $\widehat{\pi} \in \mathcal{P}(\mathcal{X})$ which reproduces and or extends π in some sense.

Representing Distributions

This distribution can be specified in various forms, perhaps as a density (normalised or un-normalised), as an algorithm which produces samples, or as some other characterisation of a probability measure. Part of the fun is that there is a bit of flexibility in how to go about this.

Functionalities

Depending on the way in which you intend to use this $\widehat{\pi}$, different functionalities might be of more or less interest. Sometimes it is important to be able to evaluate the density of $\widehat{\pi}$ exactly, whereas in other contexts, it might be sufficient to evaluate it up to a multiplicative constant. Sometimes, it is important to be able to condition on some of the variables and infer the remaining variables. Sometimes, it is important to have access to lower-dimensional marginals of the distribution. Each of these functionalities could turn out to be more or less tractable, exactly or approximately so. Again, trading off different functionalities appears to be part of the fun,

Duality

As a sampling researcher, it bears mentioning that many innovations in Generative Modeling turn out to have some dual interpretation in the context of sampling from an unnormalised probability distribution. One hypothesis for why this duality is so prevalent is that generative modelling is often cast (explicitly or otherwise) as minimising the functional

$$\widehat{\pi} \mapsto \mathsf{KL}\left(p_{\text{data}}, \widehat{\pi}\right),$$

whereas the sampling problem is instead more closely related to minimising the functional

$$\widehat{\pi} \mapsto \mathsf{KL}\left(\widehat{\pi}, p_{\text{target}}\right).$$

This is not always precisely the case, but it hints at the connection a bit. In any case, if you see some new method for Generative Modeling, then it's usually not unreasonable to wonder about whether the same principles can be used for sampling (or indeed, whether they've been borrowed from the world of sampling).

Some Core Ingredients

To begin with, let's observe some elementary building blocks with which to start building interesting probability measures:

1. Univariate, closed-form distributions
   (a) That is, e.g. Uniform, Beta, Gaussian, Gamma, and so on. Some would say that the Uniform is already sufficient, but conceptually, it doesn't hurt to treat some of the other familiar distributions as being equally fundamental.
2. Apply a deterministic transformation to a sample.
   (a) In the context of core Monte Carlo, this would relate to things like Inverse Transform Sampling, which transforms a Uniform random variable into whatever you'd like. The extent to which this is 'allowed' for a given random variable really depends on your definition of a special function. Anyways, it's something which we can do quite well in one dimension.
3. Use a change-of-measure on a simple distribution.
   (a) That is, given access to some tractable distribution q, take some non-negative 'weight' function ϖ, and define a new distribution p by asserting that $p \propto q \cdot \varpi$. In the case that we have access to q and ϖ admits a known upper bound, one can generate samples from p by rejection sampling; more generally, the approach of importance (re-)sampling yields approximate access to p. Markov Random Fields are a type of Undirected Graphical Model which can be thought of in these terms, at least to some extent.
4. Take independent products of simple distributions.
   (a) That is, given simple distributions $q_1, q_2, \cdots, q_d$, define a d-dimensional random variable $X = (X_1, X_2, \cdots, X_d)$ by saying that for each $i \in [\![d]\!]$, X_i is drawn from q_i, so that each of these samples are drawn independently. This is related to the widely-used Mean-Field Approximation of both Statistical Physics and Variational Approximation more broadly, but the idea of preferring to work with independent random variables where possible is anyways somewhat universal.
5. Take mixtures of distributions which are defined over a common base space.
   (a) That is, given simple distributions $q_1, q_2, \cdots, q_K$, fix a probability vector $\alpha \in \Delta^{K-1}$, and define a new distribution p by $p = \sum_{k \in [\![K]\!]} \alpha_k \cdot q_k$. This amounts to randomly picking an index k in $[\![K]\!]$, and then drawing a sample from q_k. Perhaps the most common instance of this would be the well-travelled Gaussian Mixture Model.
6. Construct joint distributions through "conditionally-simple" specifications.
   (a) That is, to specify a d-dimensional distribution, for each $i \in [\![d]\!]$, specify the distribution of x_i given $x_{\prec i} = \{x_j : j < i\}$ as being some simple distribution whose parameters are a simple function of $x_{\prec i}$. This is often instantiated through so-called Directed Acyclic Graphs, or Directed Graphical Models.

These approaches are separately both quite useful and also quite limited. Things can get quite a bit more interesting when we let them interact.

Cooking with Probability Distributions

In general, it turns out to be most interesting to work with high-dimensional distributions anyways, so it's useful to throughout take as a given that we have already constructed some simple reference distribution which makes sense natively on our high-dimensional space; in practice, this is often done by starting with a high-dimensional Uniform or Gaussian distribution (which are anyways obtained through some combination of Ingredients 1 and 4).

Now, some more ambitious examples:

1. Mixing together Ingredients 4 and 5 yields so-called Sum-Product Networks, a class of efficiently-parameterised mixture models with strong connections to Graphical Models, and even stronger connections to the world of Probabilistic Circuits. These are perhaps not all that well-known in the Statistics community, but would probably be of some interest there.

2. Mixing together several instances of Ingredient 2 yields various generative models which construct deterministic transformations of random variables which are somehow 'deep'. When all of the transformations are invertible diffeomorphisms (satisfying a couple of other practical structural constraints), then this yields something commonly known as a Normalising Flow. When the transformations are left completely general (but perhaps often mapping from lower-dimensional spaces into higher-dimensional spaces), then this is quite close to what people would call a Generative Adversarial Network.

3. Ingredient 3 can be instantiated in various ways. When the weight function ϖ is left quite general (and perhaps not explicitly forced to have $\int q\left(\mathrm{d}x\right) \cdot \varpi\left(x\right) = 1$), then this is what people would call an Energy-Based Model, drawing on a connection to spin models in Statistical Physics. If ϖ is somehow 'spatially local' with respect to the coordinates of $\mathcal{X}$, then one might moreover call it a Markov Random Field. When $\varpi\left(x\right) = \exp\left(\langle \xi, x \rangle\right)$ for some vector ξ of appropriate length, then this is an Exponential Family. The names tend to denote context much more strongly than they do mathematical details.

4. Ingredient 6 also wears many masks. A reasonably common approach is to let the conditional distributions live in some parametric exponential family, with the natural parameters being taken linear in the preceding variables, i.e.

   $$P\left(x_i \mid x_{\prec i}\right) \propto q\left(x_i\right) \cdot \exp\left(\langle \xi_i, x_{\prec i} \rangle\right)$$

   for some ξ_i. This would be fair to call an Autoregressive (Density) Model, although that term is probably used even when the conditional distributions are not in the exponential family, and the dependence on the preceding variables is nonlinear. This can be further combined with the Normalising Flow methodology alluded to earlier to give rise to the class of Autoregressive Flows.

5. A general approach which combines neatly with many of the above is to construct a joint distribution with more variables than are ultimately observed, and then forget some of them. This is the paradigm of Latent Variable Models, where the joint distribution of unobserved and observed random variables is implicitly quite structured (perhaps even rigid), whereas the distribution of only the observed random variables might be more complex and flexible. The Variational Autoencoder is a 'deep' variant of the same model class.

Statistical Analogues

Already, a number of these examples have antecedents in the Statistics literature. Generative Adversarial Networks are somehow structurally related to Approximate Bayesian Computation (ABC), to the extent that there are few structural assumptions enforced beyond "you know how to generate samples from the model". Autoregressive Models are connected to Directed Acyclic Graphs, Structural

Causal Models, and Graphical Models at large. Variational Autoencoders and their ilk are descendants of the humble Mixture Model and its sibling, the Hidden Markov Model. Energy-Based Models are another angle on Unnormalised Likelihood Models. This is less about credit assignment and more about how to engage with new models as they approach: usually (though not always!), there is some classical, low-dimensional antecedent to even the most striking recent advances.

I do like to mention that there seems to be pretty limited historical antecedents for the Normalising Flows methodology in the world of parametric statistical. This is a bit interesting to me on the grounds that even computationally, fitting a shallow normalising flow to data (when appropriately parameterised) can be cast as a problem of convex minimisation. My speculation is that in the past, perhaps people were thinking of model specification more probabilistically and less algorithmically, and so this type of model wouldn't have been front-of-mind conceptually.

Depth

One shorthand for contemporary Generative Modelling might be 'Old-School Density Estimation, but with Deep Neural Networks'. This does miss the point in some ways, but hits an important point: the use of Deep Neural Networks in these methods is often a major ingredient of their success.

There is also another gentler notion of depth which is pervasive in this context: instead of trying to solve the generation problem in a 'one-shot' fashion, many methodologies instead construct the probability distribution of interest in an incremental manner. This could mean gluing together conditional distributions, constructing transformations as composites of near-identity transformations, going from a single latent variable to a sequence of them, and so on.

More concretely:

1. While a Normalising Flow might be formulated as

$$X \sim \mathcal{N}\left(0_d, \mathbf{I}_d\right), \qquad Y = T\left(X\right)$$

for some diffeomorphism T, one might equally write

$$\begin{aligned} X_0 &\sim \mathcal{N}\left(0_d, \mathbf{I}_d\right) \\ X_\ell &= T_\ell\left(X_{\ell-1}\right) \\ Y &= X_L \end{aligned}$$

for some simpler sequence of diffeomorphisms $\{T_\ell : \ell \in [\![L]\!]\}$.

2. While a Variational Autoencoder might be formulated as

$$Z \sim \mathcal{N}\left(0_d, \mathbf{I}_d\right), \qquad X \mid Z \sim \mathcal{N}\left(G\left(Z\right), \sigma^2 \cdot \mathbf{I}_d\right)$$

for some function 'generator' G, the 'Hierarchical Variational Autoencoder' instead proceeds roughly as

$$\begin{aligned} Z_0 &\sim \mathcal{N}\left(0_d, \mathbf{I}_d\right) \\ Z_\ell \mid Z_{\ell-1} &\sim \mathcal{N}\left(G_\ell\left(Z_{\ell-1}\right), \sigma_\ell^2 \cdot \mathbf{I}_d\right) \\ X \mid Z_L &\sim \mathcal{N}\left(H\left(Z_L\right), \sigma_X^2 \cdot \mathbf{I}_d\right). \end{aligned}$$

3. While an Energy-Based Model might be described through its density as $p \propto q \cdot \varpi$, the procedure for drawing samples from p might instead be iterative in character, i.e. instead of $X \sim p$, one calls some sampling routine $X \sim \mathsf{MCMC}\left(\text{target} = p\right)$, which might implicitly involve a sequence of many steps. In practice, this was often even an inhomogeneous MCMC routine, in the spirit of Simulated Annealing.

Finally, when adding in these intermediate steps can already make life easier in certain ways, there is an additional operation which can sometimes make life even easier: invoke an infinite amount of intermediate steps - each of which is infinitesimal - and describe the generative process directly in continuous time.

From Energy-Based Models to Score-Based Models

As a quick aside, in the context of Energy-Based Models, one might note that for 'natural' strategies for sampling from p, one might only ever actually need to work with quantities of the form $p\left(y\right)/p\left(x\right)$ or $\nabla \log p\left(x\right)$. This suggests that one could perhaps focus directly on these operational quantities, and suggests the paradigms of 'Ratio-Based Models' and 'Score-Based Models' (although since the unit 'score' comes with some statistical baggage, the more physically-motivated 'Force-Based Models' might have been a more fortunate choice).

Edging Towards Recent Developments

Recent developments are overwhelmingly focused on the very exciting 'Denoising Diffusion' approach to Generative Modeling, which overlaps substantially with a number of the preceding strategies. I present here a few versions of the story (geared towards an audience comfortable with stochastic processes), so that one can skim the menu, pick their favourite dish, and then persist with it.

Continuous-Time Variational Autoencoders

Consider the following continuous-time formulation of a Hierarchical Variational Autoencoder:

$$\mathbf{Z} = \{Z_t : 0 \leq t \leq T\} \sim \mathsf{Markov\_Process}, \qquad X \mid Z_T \sim \mathsf{Whatever}.$$

The usual difficulties in fitting such a model are inherited from those of classical latent variable models, i.e. inferring the distribution of the latent states $\mathbf{Z}$ given the observed data X.

In the context of statistical modeling, this can be a frustration, since one tends to parameterise the dynamics of the process in some (hopefully) scientifically-meaningful way. In the context of Generative Modeling, one is not really tied down by the same concerns. Bearing this in mind, one can make the following (ambitious) leap:

1. Specify the distributions of $X \mid Z_T$ and $\mathbf{Z} \mid X$ as something very nice and explicit, and then
2. Fit the dynamics by the method of Maximum Likelihood (or, equivalently, by the EM Algorithm with an 'automatic' choice of E-step).

One instantiation of this would be to specify the model as

$$\begin{aligned} Z_0 &\sim \mathcal{N}\left(0_d, \mathbf{I}_d\right) \\ \mathrm{d}Z_t &= b_t\left(Z_t\right)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t \\ X \mid Z_T &= Z_T, \end{aligned}$$

and then assert the conditional dynamics

$$\begin{aligned} Y_0 &= X \\ \mathrm{d}Y_s &= -Y_s\,\mathrm{d}s + \sqrt{2}\,\mathrm{d}W_s \end{aligned}$$

with $Y_s = Z_{T-s}$ for $0 \leq s \leq T$. Under this formulation, one can fit the path of drifts $\mathbf{b} = \{b_t : 0 \leq t \leq T\}$ by a neat Maximum Likelihood formulation which reduces to a 'denoising'-type least-squares problem.

Smoothed Score-Based Models

In the context of Energy-Based and Score-Based Models, a challenge appeared: in regions of $\mathcal{X}$ for which data is scarce, it would (empirically) often be the case that the estimated value of $\nabla \log p$ would be very close to 0, i.e. the implied energy landscape would be quite flat, and typical MCMC-based sampling schemes would slow down badly.

The practical fix was, in effect, to construct some smoothed distribution p_t which would necessarily put a bit of mass everywhere. The natural candidate was to take $p_t = p \star \mathcal{N}\left(0_d, t \cdot \mathbf{I}_d\right)$, i.e. the law of $X + \sqrt{t} \cdot G$, where $X_0 \sim p$, $G \sim \mathcal{N}\left(0_d, \mathbf{I}_d\right)$, and then estimate $s_t = \nabla \log p_t$. This sequence of time-varying scores $\mathbf{s} = \{s_t : 0 \leq t \leq T\}$ can then be used in e.g. a simulated annealing scheme, at least heuristically.

Interestingly, when scheduled appropriately, this time-inhomogeneous dynamics - that is, the process

$$Y_0 \sim p_T, \qquad \mathrm{d}Y_s = s_{T-s}\left(Y_s\right)\, \mathrm{d}s + \sqrt{2}\, \mathrm{d}W_s$$

- could turn out to preserve equilibrium dynamically, i.e. satisfying $X_t \overset{\mathrm{d}}{=} p_t$ for all $0 \leq t \leq T$. This is certainly not true for generic simulated annealing schemes, so it is (even with hindsight) somewhat remarkable that it might do so here.

Stochastic Differential Equation Perspective

The prevailing perspective these days is a little different, and starts with the following construction:

$$\begin{aligned} X_0 &\sim p_0 = p_{\text{data}} \\ \mathrm{d}X_t &= -X_t\, \mathrm{d}t + \sqrt{2}\, \mathrm{d}W_t, \end{aligned}$$

and taking $T \gg 0$ large enough (though actually usually not even that large) that $X_T \overset{\mathrm{d}}{\approx} \gamma = \mathcal{N}\left(0_d, \mathbf{I}_d\right)$. This process is a priori not so interesting for generative purposes, but it might be interesting if reversed, i.e. draw $Y_0 \sim \gamma \overset{\mathrm{d}}{\approx} p_T$, run the same dynamics 'backwards in time', and interpret Y_T as a sample from p_0. With this in mind, it becomes interesting to understand what these dynamics ought to be

Writing p_t for the density of X_t, it will satisfy the parabolic PDE

$$\begin{aligned} \partial_t p_t &= -\nabla_x \cdot \left(p_t \nabla_x \log \gamma\right) + \Delta_x p_t \\ &= \nabla_x \cdot \left(p_t \nabla_x \log \frac{p_t}{\gamma}\right). \end{aligned}$$

Writing $q_s = p_{T-s}$, basic formal calculations show that

$$\begin{aligned} \partial_s q_s &= -\nabla_y \cdot \left(q_s \nabla_y \log \frac{q_s}{\gamma}\right) \\ &= \nabla_y \cdot \left(q_s \nabla_y \log \frac{q_s}{\gamma}\right) - 2 \cdot \nabla_y \cdot \left(q_s \nabla_y \log \frac{q_s}{\gamma}\right) \\ &= \nabla_y \cdot \left(q_s \left\{\nabla_y \log \frac{q_s}{\gamma} - 2 \cdot b_s\right\}\right) \end{aligned}$$

where $b_s\left(y_s\right) = \nabla_y \log\left(\frac{q_s}{\gamma}\right)\left(y_s\right)$. This PDE would equally track the marginals of the time-inhomogeneous process

$$\begin{aligned} Y_0 &\sim p_T \\ \mathrm{d}Y_s &= \left\{-Y_s + 2 \cdot b_s\left(Y_s\right)\right\}\, \mathrm{d}s + \sqrt{2}\, \mathrm{d}W_s, \end{aligned}$$

which shares the same noise structure as the forward process, and is hence a credible candidate for describing the same process, albeit ‘backwards in time’, i.e. this process not only tracks the correct marginals, but recovers the correct path measure in its entirety. One can eventually show that this is indeed the case, though the above is certainly not a full and rigorous argument of this claim.

The term b_s can be interpreted as a sort of { biasing / control / forcing }, and is eventually expressible in terms of another denoising-type problem.

Ordinary Differential Equation Perspective

Forgetting about aiming to reproduce the path measure (which is anyways typically surplus to requirements for the task of Generative Modelling), the same PDE

$$\begin{aligned} \partial_s q_s &= -\nabla_y \cdot \left(q_s \nabla_y \log \frac{q_s}{\gamma} \right) \\ &= -\nabla_y \cdot \left(q_s b_s \right) \end{aligned}$$

is also associated to the time-inhomogeneous deterministic evolution

$$\begin{aligned} Y_0 &\sim p_T \\ \mathrm{d}Y_s &= b_s \left(Y_s \right) \, \mathrm{d}s. \end{aligned}$$

Actually, (essentially) the same process shows up in the literature on high-dimensional convex geometry under various names.

Scores and Denoising

A quick aside on the form of the drift term here: revert to the pure additive noise setting, wherein $p_t = p \star \mathcal{N} \left(0_d, t \cdot \mathbf{I}_d \right)$. Now, compute

$$\begin{aligned} p_t \left(x_t \right) &\propto \int p_0 \left(x_0 \right) \cdot \exp \left(- \frac{\left\| x_t - x_0 \right\|^2}{2 \cdot t} \right) \, \mathrm{d}x_0 \\ s_t \left(x_t \right) &:= \nabla_{x_t} \log p_t \left(x_t \right) \\ &= \frac{\int p_0 \left(x_0 \right) \cdot \exp \left(- \frac{\left\| x_t - x_0 \right\|^2}{2 \cdot t} \right) \cdot \left\{ \frac{x_0 - x_t}{t} \right\} \, \mathrm{d}x_0}{\int p_0 \left(x_0 \right) \cdot \exp \left(- \frac{\left\| x_t - x_0 \right\|^2}{2 \cdot t} \right) \, \mathrm{d}x_0} \\ &= \frac{1}{t} \cdot \left\{ \mathbf{E} \left[X_0 \mid X_t = x_t \right] - x_t \right\}, \end{aligned}$$

which points to the connection with the denoising problem. Actually, even the second derivatives of $\log p_t$ have quite a nice form:

$$\nabla_{x_t}^2 \log p_t \left(x_t \right) = \frac{1}{t} \cdot \left\{ \mathbf{Cov} \left[X_0 \mid X_t = x_t \right] - \mathbf{I}_d \right\},$$

and obtaining good estimates on the operator norm of $\mathbf{Cov} \left[X_0 \mid X_t = x_t \right]$ can be quite useful in some interesting situations.

One take-away here is that whether you want to simulate the SDE or the ODE version of the time-reversal, it is sufficient to learn how to denoise well.

Conditionally Deterministic Evolutions

A cousin of the Denoising Diffusion is the approach known as 'Flow Matching' or 'Rectified Flow', among other noames. In this world, one posits a stochastic process as follows: draw $X_0 \sim \gamma$, draw $X_1 \sim p_{\text{data}}$, and for $t \in (0,1)$, set

$$X_t = (1-t) \cdot X_0 + t \cdot X_1.$$

Note that $\{X_t : t \in [0,1]\}$ is generally not Markov. This seems like an interesting process. If we could generate X_1 without ever looking at p_{data}, then that would be even more interesting. Flow Matching describes a way to go about this.

Fix a test function ϕ, and compute

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[\phi\left(X_t\right)\right] &= \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[\phi\left((1-t)\cdot X_0 + t \cdot X_1\right)\right] \\ &= \mathbf{E}\left[\frac{\mathrm{d}}{\mathrm{d}t}\phi\left((1-t)\cdot X_0 + t\cdot X_1\right)\right] \\ &= \mathbf{E}\left[\left\langle \nabla\phi\left((1-t)\cdot X_0 + t\cdot X_1\right), X_1 - X_0\right\rangle\right] \\ &= \mathbf{E}\left[\left\langle \nabla\phi\left(X_t\right), X_1 - X_0\right\rangle\right] \\ &= \mathbf{E}\left[\left\langle \nabla\phi\left(X_t\right), \mathbf{E}\left[X_1 - X_0 \mid X_t\right]\right\rangle\right], \end{aligned}$$

where the last step uses the tower property of conditional expectation. Define then

$$u\left(t, x_t\right) = \mathbf{E}\left[X_1 - X_0 \mid X_t = x_t\right]$$

to see that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[\phi\left(X_t\right)\right] = \mathbf{E}\left[\left\langle \nabla\phi\left(X_t\right), u\left(t, X_t\right)\right\rangle\right].$$

Consider for a moment a new stochastic process, defined by

$$\bar{X}_0 \sim \gamma, \qquad \mathrm{d}\bar{X}_t = u\left(t, \bar{X}_t\right)\,\mathrm{d}t.$$

Then one computes that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[\phi\left(\bar{X}_t\right)\right] = \mathbf{E}\left[\left\langle \nabla\phi\left(\bar{X}_t\right), u\left(t, \bar{X}_t\right)\right\rangle\right]$$

as well. In particular, for any test function ϕ, it holds that

$$\mathbf{E}\left[\phi\left(X_t\right)\right] = \mathbf{E}\left[\phi\left(\bar{X}_t\right)\right],$$

and so the processes $\{X_t : t \in [0,1]\}$, $\left\{\bar{X}_t : t \in [0,1]\right\}$ share the same marginal law. Of course, their joint laws as processes are not the same - the latter is Markovian, the former is not. Actually, the latter is even particularly expressible as the 'Markovian projection' of the former, a notion which can be useful in even greater generality.

Anyhow, the joint law is solely an instrument, and all one really cares about is the law at times $t = 0$ (which should be tractable) and $t = 1$ (which should be the data distribution, or something inspired by it).

The upshot of this is that if u were available, then one could simulate $\bar{X}$ and read off $\bar{X}_1$ as a sample from p_{data}. Perhaps remarkably, u is actually reasonably available: recalling the definition of u as a conditional expectation

$$u\left(t, x_t\right) = \mathbf{E}\left[X_1 - X_0 \mid X_t = x_t\right],$$

one sees that u can be learned by regression. In particular, it holds for each $t \in [0,1]$ that

$$u\left(t, \cdot\right) \in \arg\min \mathbf{E}\left[\left\|u\left(t, X_t\right) - \left(X_1 - X_0\right)\right\|^2\right]$$

for any Hilbertian norm $\|\cdot\|$, where the expectation is taken under the joint law of (X_0, X_1, X_t), which can be sampled from easily given data. The above objective is often described as 'simulation-free', though it would perhaps be more accurate to term it 'discretisation-free'. Note that while the training process is discretisation-free in this sense, the process of generating samples from $\bar{X}$ will again require discretisation of the ODE.

An interesting aspect of this approach is that while it is transport-based in character, it is not really optimal transport in the sense of Monge, Kantorovich, and so on. Instead of saying "find the best transport from γ to p_{data}", one says "here is a specific transport from γ to p_{data}; find out how to evaluate it". Indeed, this specific transport also turns out to be 'pretty good' in a variety of ways. It is really quite explicit by the standards of this field, it can be evaluated a priori at a cost which is independent of most features of the problem, and so on.

Another Markovian Projection

Suppose instead that we consider stochastic processes which begin by sampling

$$X_0 \sim p, \quad Z \sim a,$$

where Z represents some 'side-randomness', possibly living in a different space entirely to X_0. Now, conditionally on (X_0, Z), let the process $X = \{X_t : t \in [0,1]\}$ evolve according to a Markov process with (potentially time-inhomogeneous) generator $\mathcal{L}_t^Z$. This is somehow a stochastic generalisation of the previous setting, so we can consider the same approach to its analysis via test functions:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{E}\left[\phi\left(X_t\right)\right] &= \mathbf{E}\left[\mathcal{L}_t^Z \phi\left(X_t\right)\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\mathcal{L}_t^Z \phi\left(X_t\right) \mid X_t\right]\right] \\ &= \mathbf{E}\left[\widehat{\mathcal{L}_t^{X_t}} \phi\left(X_t\right)\right]\end{aligned}$$

for some operator $\widehat{\mathcal{L}_t^{X_t}}$ whose form will depend greatly on the nature of the $\mathcal{L}^Z$. One can then seek to interpret this operator as the generator of an autonomous and Markovian stochastic process $\bar{X}$, which will again reproduce the same marginals as X.

A nice example of this is the following 'erasure-based' approach to discrete spaces: draw

$$X_0 \sim \mathsf{Unif}\left(\{\pm 1\}^D\right), \qquad X_1 \sim p_{\mathrm{data}},$$

for $i \in [\![D]\!]$, draw $\tau_i \overset{\mathrm{iid}}{\sim} \mathsf{Unif}\left([0,1]\right)$, and set

$$X_t^i = X_0^i \cdot \mathbf{1}_{[0,\tau^i)}\left(t\right) + X_1^i \cdot \mathbf{1}_{[\tau^i,1]}\left(t\right),$$

i.e. for $t < \tau_i$, the ith coordinate is a fair coin flip, and for $t \geq \tau^i$, it matches a sample from the data distribution. The Markovian projection of this process is then a time-inhomogeneous Markov jump process whereby each coordinate of X_t flips to its negative at some rate $\lambda^i\left(t, X_t\right)$, where the rates $\left\{\lambda^i : i \in [\![D]\!]\right\}$ collectively solve some quite-explicit regression problem.

In this area, a key skill appears to be quite literate in terms of interpreting linear operators as the generator of a suitable Markov process. A recent work titled 'Generator Matching' is much in this spirit.

Martingale Posteriors, Bayesian Flow Networks, etc.

Consider now a model where

$$\begin{aligned} & X_0 \sim p_{\mathrm{data}} \\ \text{for } t \geq 1, \qquad & Y_t \mid X_0 = x_0 \sim \omega\left(\mathrm{d}y; x_0\right),\end{aligned}$$

where ω is some 'observation model' through which to gradually infer X_0. For reasonable ω, theory for the consistency of Bayesian posterior distributions yields that in various natural senses

$$\mathsf{Posterior}\left(X_0 \in A \mid Y_{1:T}\right) \overset{T\to\infty}{\to} \mathbf{1}_A\left(X_0^\star\right).$$

Now, let's rewrite the joint distribution of $(X_0, Y_1, Y_2, \cdots, Y_T)$ in a "predictive" or "y-centric" manner:

$$\begin{aligned}
Y_1 &\sim \mathsf{Predictive}_1\left(\mathrm{d}y_1\right) \\
Y_2 \mid Y_1 = y_1 &\sim \mathsf{Predictive}_{2|1}\left(\mathrm{d}y_2 \mid y_1\right) \\
&\cdots \\
Y_T \mid Y_{1:T-1} = y_{1:T-1} &\sim \mathsf{Predictive}_{T|T-1}\left(\mathrm{d}y_T \mid y_{1:T-1}\right) \\
X_0 \mid Y_{1:T} = y_{1:T} &\sim \mathsf{Posterior}\left(\in \mathrm{d}x_0 \mid y_{1:T}\right),
\end{aligned}$$

i.e. generate all of the data first, and then infer the parameter from it. This suggests a dual view on generative modelling: learn how to predict different parts of the sample from incomplete information, and then infer the sample itself. That is to say, a machine for prediction and a machine for inference can be combined to yield a machine for generation. This is quite related to the perspectives of Fong-Holmes-Walker in their proposal of 'Martingale Posterior Distributions', though contextualised rather differently.

Montanari's note on 'Sampling, diffusions, and stochastic localization' essentially advances this perspective in marginally different language. He notes that the usual Denoising Diffusion models correspond to a continuous-time version of this idea, wherein

$$Y_t = t \cdot X_0 + W_t,$$

for some independent Wiener process $\{W_t : t \geq 0\}$, modulo some elementary rescalings of time and space. He also proposes a cute approach for generating random samples with nonnegative coefficients, proposing to observe them indirectly through a Poisson process with rate X, i.e. $Y_t = \mathsf{VecPP}\left(\text{rate} = t \cdot X_0\right)$. Simulating this process in an 'X_0-free' way then amounts to adding points at rate $\lambda\left(t, y_t\right) = \mathbf{E}\left[X_0 \mid Y_t = y_t\right]$, i.e. another nice conditional expectation. A nice aspect of this is that while X is treated as nonnegative and perhaps taking arbitrary nonnegative values, the observation process Y is count-valued, demonstrating a neat decoupling of the process of interest from the instrumental observation process.