Denoising Diffusion Models

STATS305B: Applied Statistics II

Scott Linderman

March 5, 2025

Last Time...

Outline:

- Recurrent Neural Networks
- Backpropagation Through Time
- Vanishing Gradients and Gated RNNs
- Other Variations and Uses of RNNs
- Revisiting HMMs
- Linear RNNs and Parallel Inference

Today...

Outline:

- Denoising Diffusion Models
- Noising and Generative Processes
- Evidence Lower Bound
- Continuous Time Limit

Key Ideas

Diffusion models work by

1. Using a fixed, user-defined **noising process** to convert data into noise.
2. Learning to invert this process so that starting from noise, we can generate samples that approximate the data distribution.

We can think of the DDPM as a giant latent variable model, where the latent variables are noisy versions of the data.

As with VAEs, given the latent variables, learning the mapping from latents to observed data is a supervised regression problem.

Noising process

Let $x \equiv x_0 \in \mathbb{R}$ be our observed data (assume scalar for now).

The noising process is a joint distribution over a sequence of latent variables $x_{0:T}$,

$$q(x_{0:T}) = q(x_0) \prod_{t=1}^{T} q(x_t \mid x_{t-1}).$$

where $q(x_0) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_0^{(i)}}(x_0)$ is the empirical measure of the data.

At each step, the latents will become increasingly noisy versions of the original data, until at time T the latent variable x_T is essentially pure noise.

The generative model samples pure noise and attempts to invert the noising process to produce samples that approximate $q(x_0)$.

Gaussian noising process

For continuous data, the standard noising process is a first-order Gaussian autoregressive (AR) process,

$$q(x_t \mid x_{t-1}) = \mathrm{N}(x_t \mid \lambda_t x_{t-1}, \sigma_t^2).$$

The hyperparameters $\{\lambda_t, \sigma_t^2\}_{t=1}^T$ and the number of steps T are fixed (not learned). We restrict $\lambda_t < 1$ so that the process contracts

Conditional Distributions

Since the noising process has linear Gaussian dynamics, we can compute conditional distributions in closed form.

$$
\begin{aligned}
q(x_t \mid x_0) &= \int q(x_t \mid x_{t-1})\, q(x_{t-1} \mid x_0)\, \mathrm{d}x_{t-1} \\
&= \int \mathrm{N}(x_t \mid \lambda_t x_{t-1}, \sigma_t^2)\, \mathrm{N}(x_{t-1} \mid \lambda_{t-1|0} x_0,\, \sigma_{t-1|0}^2)\, \mathrm{d}x_{t-1} \\
&= \mathrm{N}(x_t \mid \lambda_t \lambda_{t-1|0} x_0, \lambda_t^2 \sigma_{t-1|0}^2 + \sigma_t^2) \\
&= \mathrm{N}\left(x_t \mid \lambda_{t|0} x_0, \sigma_{t|0}^2\right),
\end{aligned}
$$

Conditional Distributions

where the parameters are defined recursively,

$$\lambda_{t|0} = \lambda_t \lambda_{t-1|0} = \prod_{s=1}^{t} \lambda_s$$

$$\sigma^2_{t|0} = \lambda_t^2 \sigma^2_{t-1|0} + \sigma^2_t$$

with base case $\lambda_{1|0} = \lambda_1$ and $\sigma^2_{1|0} = \sigma^2_1$.

Variance preserving diffusions

It is common to set,

$$\sigma_t^2 = 1 - \lambda_t^2,$$

in which case the conditional variance simplifies to,

$$\sigma_{t|0}^2 = 1 - \prod_{s=1}^{t} \lambda_s^2 = 1 - \lambda_{t|0}^2.$$

Under this setting, the noising process preserves the variance of the marginal distributions.

If $\mathbb{E}[x_0] = 0$ and $\mathrm{Var}[x_0] = 1$, then the marginal distribution of x_t will be zero mean and unit variance as well.

Limiting Distribution

Consider the following two limits:

1. As $T \to \infty$, the conditional distribution goes to a standard normal, $q(x_T \mid x_0) \to \mathrm{N}(0, 1)$, which makes the marginal distribution $q(x_T)$ easy to sample from.
2. When $\lambda_t \to 1$, the noising process adds infinitesimal noise so that $x_t \approx x_{t-1}$, which makes the inverse process easier to learn.

These two limits are in conflict with one another! If we add a small amount of noise at each time step, the inverse process is easier to learn, but we need to take many time steps to converge to a Gaussian stationary distribution.

Generative process

The generative process is a parameteric model that learns to invert the noise process,

$$p(x_{0:T}; \theta) = p(x_T) \prod_{t=T-1}^{0} p(x_t \mid x_{t+1}; \theta).$$

The initial distribution $p(x_T)$ has no parameters because it is set to the stationary distribution of the noising process, $q(x_\infty)$.

E.g., for the Gaussian noising process above, $p(x_T) = \mathrm{N}(0, 1)$.

Evidence Lower Bound

Like the other latent variable models we studied in this course, we will estimate the parameters by maximizing an **evidence lower bound (ELBO)**,

$$\begin{aligned}\mathscr{L}(\theta) &= \mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_{1:T} \mid x_0)}\left[\log p(x_{0:T};\theta) - \log q(x_{1:T} \mid x_0)\right] \\ &= \mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_{1:T} \mid x_0)}\left[\log p(x_{0:T};\theta)\right] + c,\end{aligned}$$

where $q(x_{1:T} \mid x_0)$ is the conditional distibution of $x_{1:T}$ under the noising process.

Since q is fixed, the objective simplifies to **maximizing the expected log likelihood**.

We can simplify further by expanding the log probability of the generative model,

$$\begin{aligned}\mathscr{L}(\theta) &= \mathbb{E}_{q(x_0)}\sum_{t=0}^{T-1}\mathbb{E}_{q(x_t,x_{t+1} \mid x_0)}\left[\log p(x_t \mid x_{t+1};\theta)\right] \\ &\propto \mathbb{E}_{q(x_0)}\mathbb{E}_{t\sim\mathrm{Unif}(0,T-1)}\mathbb{E}_{q(x_t,x_{t+1} \mid x_0)}\left[\log p(x_t \mid x_{t+1};\theta)\right]\end{aligned}$$

which only depends on pairwise conditionals.

Gaussian generative process

Since the noising process above adds a small amount of Gaussian noise at each step, it is reasonable to model the generative process as Gaussian as well,

$$p(x_t \mid x_{t+1}; \theta) = \mathrm{N}(x_t \mid \mu_\theta(x_{t+1}, t), \widetilde{\sigma}_t^2)$$

where

- $\mu_\theta : \mathbb{R} \times [0, T] \mapsto \mathbb{R}$ is a nonlinear **mean function** that should **denoise** x_{t+1} to obtain the expected value of x_t
- $\widetilde{\sigma}_t^2$ is a fixed variance for the generative process.

Parameter sharing

Rather than learn a separate function for each time point, it is common to parameterize the mean function as a function of both the state x_{t+1} and the time t.

E.g., $\mu_\theta(\cdot, \cdot)$ can be a neural network that takes in the state and a positional embedding of the time t, like the sinusoidal embeddings used in transformers.

Generative Process Variance

You could try to learn the generative process variance as a function of x_{t+1} and t as well, but the literature suggests this is difficult to make work in practice.

Instead, is common to set the variance to either

- $\widetilde{\sigma}_t^2 = \sigma_t^2 = 1 - \lambda_t^2$, the conditional variance in the noising process, which tends to *overestimate* the conditional variance of the true generative process
- $\widetilde{\sigma}_t^2 = \mathrm{Var}_q[x_t \mid x_0, x_{t+1}]$, the conditional variance of the noising process *given* the data x_0 and the next state x_{t+1}. This tends to *underestimate* the conditional variance of the true generative process.

Rao-Blackwellization

Under this Gaussian model for the generative process, we can analytically compute one of the expectations in the ELBO. This is called **Rao-Blackwellization**. It reduces the variance of the objective, which is good for SGD!

Using the chain rule and the Gaussian generative model,

$$\mathbb{E}_{q(x_t, x_{t+1} \mid x_0)}\left[\log p(x_t \mid x_{t+1}; \theta)\right] = \mathbb{E}_{q(x_{t+1} \mid x_0)}\mathbb{E}_{q(x_t \mid x_{t+1}, x_0)}\left[\log \mathrm{N}(x_t \mid \mu_\theta(x_{t+1}, t), \widetilde{\sigma}_t^2)\right]$$

We already computed the conditional distribution $q(x_{t+1} \mid x_0) = \mathrm{N}(x_{t+1} \mid \lambda_{t+1|0} x_0, \sigma^2_{t+1|0})$ above. It turns out the second term is Gaussian as well!

Conditionals of a Gaussian noising process

Show that

$$q(x_t \mid x_{t+1}, x_0) = \mathrm{N}(x_t \mid \mu_{t|t+1,0}, \sigma^2_{t|t+1,0})$$

where

$$\mu_{t|t+1,0} = a_t x_0 + b_t x_{t+1}$$

Conditionals of a Gaussian noising process

is a **linear combination** of x_0 and x_{t+1} with weights,

$$a_t = \frac{\sigma^2_{t|t+1,0}\,\lambda_{t|0}}{\sigma^2_{t|0}}$$

$$b_t = \frac{\sigma^2_{t|t+1,0}\,\lambda_{t+1}}{\sigma^2_{t+1}}$$

$$\sigma^2_{t|t+1,0} = \left(\frac{1}{\sigma^2_{t|0}} + \frac{\lambda^2_{t+1}}{\sigma^2_{t+1}}\right)^{-1}$$

Derivation

By Bayes rule and the Markovian structure of the noising process,

$$\begin{aligned} q(x_t \mid x_{t+1}, x_0) &\propto q(x_t \mid x_0)\, q(x_{t+1} \mid x_t) \\ &= \mathrm{N}(x_t \mid \lambda_{t|0} x_0, \sigma^2_{t|0})\, \mathrm{N}(x_{t+1} \mid \lambda_{t+1} x_t, \sigma^2_{t+1}) \\ &= \mathrm{N}(x_t \mid \mu_{t|t+1,0}, \sigma^2_{t|t+1,0}) \end{aligned}$$

where, by completing the square,

$$\sigma^2_{t|t+1,0} = \left(\frac{1}{\sigma^2_{t|0}} + \frac{\lambda^2_{t+1}}{\sigma^2_{t+1}} \right)^{-1}$$

$$\mu_{t|t+1,0} = \sigma^2_{t|t+1,0} \left(\frac{\lambda_{t|0} x_0}{\sigma^2_{t|0}} + \frac{\lambda_{t+1} x_{t+1}}{\sigma^2_{t+1}} \right).$$

The forms for a_t and b_t can now be read off.

Gaussian cross-entropy

Finally, to simplify the objective we need the Gaussian cross-entropy,

Let $q(x) = \mathrm{N}(x \mid \mu_q, \sigma_q^2)$ and $p(x) = \mathrm{N}(x \mid \mu_p, \sigma_p^2)$.

Show that,

$$\mathbb{E}_{q(x)}[\log p(x)] = \log \mathrm{N}(\mu_q \mid \mu_p, \sigma_p^2) - \frac{1}{2}\frac{\sigma_q^2}{\sigma_p^2}$$

The Simplified ELBO

Putting it all together,

$$\begin{aligned}
\mathscr{L}(\theta) &= \mathbb{E}_{q(x_0)}\mathbb{E}_t\mathbb{E}_{q(x_{t+1} \mid x_0)}\mathbb{E}_{q(x_t \mid x_0, x_{t+1})}\left[\log p(x_t \mid x_{t+1}; \theta)\right] \\
&= \mathbb{E}_{q(x_0)}\mathbb{E}_t\mathbb{E}_{q(x_{t+1} \mid x_0)}\left[\log \mathrm{N}(a_t x_0 + b_t x_{t+1} \mid \mu_\theta(x_{t+1}, t), \widetilde{\sigma}_t^2) - \frac{1}{2}\frac{\sigma_{t|t+1,0}^2}{\widetilde{\sigma}_t^2}\right] \\
&= \frac{1}{2}\mathbb{E}_{q(x_0)}\mathbb{E}_t\mathbb{E}_{q(x_{t+1} \mid x_0)}\left[\frac{1}{\widetilde{\sigma}_t^2}\left(a_t x_0 + b_t x_{t+1} - \mu_\theta(x_{t+1}, t)\right)^2\right] + c
\end{aligned}$$

where we have absorbed terms that are independent of θ into the constant c.

Denoising mean function

The loss function above suggests a particular form of the mean function,

$$\mu_\theta(x_{t+1}, t) = a_t \hat{x}_0(x_{t+1}, t; \theta) + b_t x_{t+1},$$

where the only part that is learned is $\hat{x}_0(x_{t+1}, t; \theta)$, a function that attempts to **denoise** the current state.

Since x_{t+1} is given and a_t and b_t are determined solely by the hyperparameters, we can use them in the mean function.

Under this parameterization, the loss function reduces to,

$$\mathscr{L}(\theta) = \frac{1}{2} \mathbb{E}_{q(x_0)} \mathbb{E}_t \mathbb{E}_{q(x_{t+1} \mid x_0)} \left[\frac{a_t^2}{\tilde{\sigma}_t^2} \left(x_0 - \hat{x}_0(x_{t+1}, t; \theta) \right)^2 \right] + c$$

One nice thing about this formulation is that the mean function is always outputting *the same thing* — an estimate of the completely denoised data, $\hat{x}_0$, regardless of the time t.

Inverting the noising process

The generative process attempts to invert the noising process, but what is the actual inverse of the process?

Since the noising process is a Markov chain, the reverse of the noising process must be Markovian as well.

$$q(x_{0:T}) = q(x_T) \prod_{t=T-1}^{0} q(x_t \mid x_{t+1})$$

for some sequence of transition distributions $q(x_t \mid x_{t+1})$.

Inverting the noising process

We can obtain those transition distributions by marginalizing and conditioning,

$$\begin{aligned} q(x_t \mid x_{t+1}) &= \int q(x_0, x_t \mid x_{t+1}) \, \mathrm{d}x_0 \\ &= \int q(x_t \mid x_0, x_{t+1}) \, q(x_0 \mid x_{t+1}) \, \mathrm{d}x_0. \end{aligned}$$

Using Bayes' rule,

$$q(x_0 \mid x_{t+1}) = \frac{q(x_0) \, q(x_{t+1} \mid x_0)}{\int q(x_0') q(x_{t+1} \mid x_0') \, \mathrm{d}x_0'}$$

Inverting the noising process

Now recall that $q(x_0) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_0^{(i)}}(x_0)$ is the empirical measure of the data $\{x_0^{(i)}\}_{i=1}^n$. Using this fact, the conditional is,

$$q(x_0^{(i)} \mid x_{t+1}) = \frac{q(x_{t+1} \mid x_0^{(i)})}{\sum_{j=1}^{n} q(x_{t+1} \mid x_0^{(j)})} \triangleq w_i(x_{t+1}),$$

where we have defined the weights $w_i(x_{t+1})$ for each data point $i = 1, \ldots, n$. They are non-negative and sum to one.

Finally, we can give a simpler form for the optimal generative process,

$$\begin{aligned} q(x_t \mid x_{t+1}) &= \sum_{i=1}^{n} w_i(x_{t+1})\, q(x_t \mid x_0^{(i)}, x_{t+1}) \\ &= \sum_{i=1}^{n} w_i(x_{t+1})\, \mathrm{N}(x_t \mid a_t x_0^{(i)} + b_t x_{t+1}, \sigma^2_{t|t+1,0}), \end{aligned}$$

Inverting the noising process

which we recognize as a mixture of Gaussians, all with the same variance, with means biased toward each of the n data points, and weighted by the relative likelihood of $x_0^{(i)}$ having produced x_{t+1}.

For small step sizes, that mixture of Gaussians can be approximated by a single Gaussian with mean equal to the expected value of the mixture,

$$\mathbb{E}[x_t \mid x_{t+1}] = \sum_{i=1}^{n} w_i(x_{t+1}) \left(a_t x_0^{(i)} + b_t x_{t+1} \right)$$

For small steps, this expected value is approximately,

$$\mathbb{E}[x_t \mid x_{t+1}] \approx \frac{x_{t+1}}{\lambda_{t+1}} + \sigma_{t+1}^2 \sum_{i=1}^{n} w_i(x_{t+1}) \left(\frac{\lambda_{t|0} x_0^{(i)} - x_{t+1}}{\sigma_{t|0}^2} \right)$$

(See online notes for derivation.)

Inverting the noising process

Though it's not immediately obvious, the second term in the expectation is related to the **marginal probability**,

$$
\begin{aligned}
q(x_t) &= \frac{1}{n}\sum_{i=1}^{n} q(x_t \mid x_0^{(i)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{N}(x_t \mid \lambda_{t|0} x_0^{(i)}, \sigma_{t|0}^2)
\end{aligned}
$$

Inverting the noising process

Specifically, the second term is the **Stein score function** of the marginal probability,

$$
\begin{aligned}
\nabla_x \log q_t(x_{t+1}) &= \frac{\nabla_x q_t(x_{t+1})}{q_t(x_{t+1})} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} \mathrm{N}(x_{t+1} \mid \lambda_{t|0} x_0^{(i)}, \sigma_{t|0}^2)\left(-\frac{(x_{t+1}-\lambda_{t|0}x_0^{(i)})}{\sigma_{t|0}^2}\right)}{\frac{1}{n}\sum_{j=1}^{n} \mathrm{N}(x_{t+1} \mid \lambda_{t|0} x_0^{(j)}, \sigma_{t|0}^2)} \\
&= \sum_{i=1}^{n} w_i(x_{t+1})\left(\frac{\lambda_{t|0} x_0^{(i)} - x_{t+1}}{\sigma_{t|0}^2}\right)
\end{aligned}
$$

Final Form

Putting it all together, for small steps, the reverse process is approximately Gaussian with mean and variance,

$$\mathbb{E}[x_t \mid x_{t+1}] \approx \frac{x_{t+1}}{\lambda_{t+1}} + \sigma_{t+1}^2 \nabla_x \log q_t(x_{t+1})$$
$$\mathrm{Var}[x_t \mid x_{t+1}] \approx \sigma_{t+1}^2.$$

This has a nice interpretation: to invert the noise process, **first undo the contraction and then take a step in the direction of the Stein score!**

Continuous time limit

In practice, the best performing diffusion models are based on a continuous-time formulation of the noising process as an SDE (Song et al., 2020).

To motivate this approach, think of the noise process above as a discretization of a continuous process $x(t)$ for $t \in [0, 1]$ with time steps of size $\Delta = \frac{1}{T}$.

That is, map $x_i \mapsto x(i/T)$, $\lambda_i \mapsto \lambda(i/T)$, and $\sigma_i \mapsto \sigma(i/T)$ for $i = 0, 1, \ldots, T$.

Then the discrete model is can be rewritten as,

$$x(t + \Delta) \sim \mathrm{N}(\lambda(t)x(t), \sigma(t)^2),$$

Continuous time limit

or equivalently,

$$x(t+\Delta) - x(t) \sim \mathrm{N}\left(f(x(t),t)\Delta, g(t)^2\Delta\right)$$
$$f(x,t) = \frac{1-\lambda(t)}{\Delta}x$$
$$g(t) = \frac{\sigma(t)}{\Delta}.$$

We can view this as a discretization of the SDE,

$$\mathrm{d}X = f(x,t)\,\mathrm{d}t + g(t)\,\mathrm{d}W$$

where $f(x,t)$ is the **drift** term, $g(t)$ is the **diffusion** term, and $\mathrm{d}W$ is the **Brownian motion**.

The reverse (generative) process can be cast as an SDE as well!

Continuous time limit

Following our derivation of the inverse process above, we can show that the reverse process is,

$$\mathrm{d}X = \left[f(x,t) - g(t)^2 \nabla_x \log q_t(x)\right] \mathrm{d}t + g(t)\,\mathrm{d}W$$

where $\mathrm{d}t$ is a **negative** time increment and $\mathrm{d}W$ is Brownian motion run in reverse time.

Multidimensional models

Very few things need to change in order to apply this idea to multidimensional data $\boldsymbol{x}_0 \in \mathbb{R}^D$.

The standard setup is to apply a Gaussian noising process to each coordinate $x_{0,d}$ independently.

Then, in the generative model,

$$\begin{aligned} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}; \theta) &= \prod_{d=1}^{D} p(x_{t,d} \mid \boldsymbol{x}_{t+1}; \theta) \\ &= \prod_{d=1}^{D} \mathrm{N}(x_{t,d} \mid \mu_\theta(\boldsymbol{x}_{t+1}, t, d), \widetilde{\sigma}_t^2). \end{aligned}$$

The generative process still produces a factored distribution, but we need a separate mean function for each coordinate.

Multidimensional models

Moreover, the mean function needs to consider the entire state $\boldsymbol{x}_{t+1}$. The reason is that $x_{t,d}$ is not conditionally independent of $x_{t+1,d'}$ given $x_{t+1,d}$; the coordinates are coupled in the inverse process since all of $\boldsymbol{x}_{t+1}$ provides information about the $\boldsymbol{x}_0$ that generated it.

Conclusion

There's a lot we didn't cover!

The Stein score function that appeared in the inverse of the noising process allows for connections between denoising score matching {cite:p}`song2019generative` and denoising diffusion models.

Another important topic is **conditional generation**. Suppose we want to take in text and spit out images, like DALL-E 2 or Stable Diffusion. One way to do so is using a diffusion model, but to steer the reverse diffusion based on the text prompt.

Finally, this class was nominally about models for discrete data, but this lecture has focused on continuous diffusions. There has been recent work on discrete denoising diffusion models, which we'll have to cover another time!