α-STABLE DIFFUSSIONS:

Extreme Diffussion.

The main goal of this notebook lies in developing α-stable diffusions. It is based on the following observation: a lot of what makes diffusion models tractable is that the normal distribution is stable.

That is, (from Wikipedia):

Let X_1 and X_2 be independent realizations of a random variable X. Then X is said to be stable if for any constants $a>0$ and $b>0$ the random variable $aX_1 + bX_2$ has the same distribution as $cX+d$ for some constants $c>0$ and d. The distribution is said to be strictly stable if it holds for $d=0$.

The normal, Cauchy and Lévy distributions all have the same property.

The probability density function for a general stable distribution cannot be written analytically, (Why?). A random variable that is stable has the following characteristic function, in general,

$$\psi(t;\alpha,\beta,c,\mu) = \exp\left(it\mu - |ct|^{\alpha}\left(1 - i\beta\,\mathrm{sign}(t)\,\phi\right)\right)$$

$$\phi = \begin{cases} \tan\left(\frac{\pi\alpha}{2}\right) & \alpha \neq 1 \\ -\frac{2}{\pi}\log|t| & \alpha = 1 \end{cases}$$

$\alpha \in (0,2]$ shape — ~~assymmetry~~ concentration.

$\beta \in [-1,1]$ shape — assymetry

$c \in (0,\infty)$ scale

$\mu \in (-\infty,\infty)$ location

Properties

- All stable distributions are infinitely divisible. → if can be expressed as the sum of an arbitrary # of iid r.v.'s.
- With the exception of the normal distribution ($\alpha = 2$), stable distributions are leptokurtic and heavy-tailed distributions
- Closure under convolution.
  - ↳ Can be seen from multiplying chars. function.

α-stable distributions generalize the CLT!

Generalized CLT.

A non-degenerate random variable Z is α-stable for some $0 < \alpha \leq 2$ if and only if there is an independent, identically distributed sequence of random variables $X_1, X_2, \ldots$ and constants $a_n > 0$, $b_n \in \mathbb{R}$ with

$$a_n (X_1 + \cdots + X_n) - b_n \xrightarrow{d.} Z$$

Cauchy Distribution.

A relatively small number of α-stable distributions have a representation in terms of simple functions. Apart from the normal distribution, ($\alpha = 2$), the Cauchy distribution is the most studied. Here are a few properties (Wikipedia):

1. If $X \sim \text{Cauchy}(x_0, \gamma)$ then $kX + l \sim \text{Cauchy}(x_0 k + l, \gamma|k|)$.

2. If $X \sim \text{Cauchy}(x_0, \gamma_0)$ and $Y \sim \text{Cauchy}(x_1, \gamma_1)$ are independent, then $X + Y \sim \text{Cauchy}(x_0 + x_1, \gamma_0 + \gamma_1)$ and
$X - Y \sim \text{Cauchy}(x_0 - x_1, \gamma_0 + \gamma_1)$.

3. If $X \sim \text{Cauchy}(0, \gamma)$ then $\frac{1}{X} \sim \text{Cauchy}(0, \frac{1}{\gamma})$.

4. Has univariate pdf

$$f(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]} \qquad X \sim \text{Cauchy}(x_0, \gamma)$$

Has multivariate pdf:

$$f(x; \mu, \Sigma, k) = \frac{\Gamma\left(\frac{1+k}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\pi^{k/2}|\Sigma|^{\frac{1}{2}}\left[1 + (x-\mu)^T\Sigma^{-1}(x-\mu)\right]^{\frac{1+k}{2}}}$$

Note: $\Sigma = I$ does not imply that the r.v.'s are independent.

5. Has univariate chf:

$$\varphi_X(t) = E\left[e^{iXt}\right] = e^{ix_0 t - \gamma|t|}.$$

and multivariate chf

$$\varphi_X(t) = e^{iX_0(t) - \gamma(t)}$$

$X_0(t), \gamma(t)$ real functions of degree one s.t. $x_0(at) = a x_0(t)$, $\gamma(at) = |a|\gamma(t)$.

6. The KL-divergence between two Cauchy distributions has a closed form

$$KL(p_{x_{0,1},\gamma_1} ; p_{x_{0,2},\gamma_2}) = \log\frac{(\gamma_1 + \gamma_2)^2 + (x_{0,1} - x_{0,2})^2}{4\gamma_1\gamma_2}$$

Motivating Question:

DDPM's algorithm for training is:

1. Repeat
2. $X_0 \sim q(X_0)$ $\longmapsto$ sample from data.
3. $t \sim \text{Unif}(\{1, \dots, T\})$. $\longmapsto$ T is # of timesteps.
4. $\epsilon \sim N(0, I)$.
5. Take gradient step on.

$$\nabla_\theta \| \epsilon - \epsilon_\theta(X_t, t) \|^2 \longmapsto \nabla_\theta \| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \|^2$$

6. Until converged

A crucial step, (5), ~~loss~~ is dependent upon the stability of the normal, because for some schedule of β_t's,

$$X_t = \sqrt{\bar{\alpha}_t} X_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon.$$

Does the Cauchy distribution, with its stability and existing KL divergence determine a viable diffusion process with different or desirable (extremes?) properties

From the review and there are a few things that worry me:

(1) - stability might not be enough? Maybe compyacy too is needed? Maybe compyacy is related to stability, right?

(2) - Will it be enough to exchange μ and σ to X_0 and λ? (2019?)

(3) - Ho's DDPM claims that for β_t small, the reverse process is also Gaussian. He cites Sohl-Dickerstein (2015?) for this claim, who in turn cites Feller. (an old article).
How does this claim hold for the Cauchy case?

With all that in mind, the simplest litmus test for this idea might be assuming univariate distributions (bivariate top), and run with it.

A. Gaussian Diffusion Process gradually transforms data x into random noise by adding increasing amounts of Gaussian noise at each timestep t.

Bishop's New book

has a section on diffusion, and although I find it lacking, it does have one stellar problem, that I hope will help: problem 20.7, where he asks you to show that the reverse $q(z_{t-1} | z_t)$ is approximately Gaussian when β_t is small. If I can reproduce this, then I might be able to reproduce it for the Cauchy case.

Anyway, Bishop's diffusion:

Forward Encoder:

$$z_1 = \sqrt{1-\beta_1}\, x + \sqrt{\beta_1}\, \epsilon_1, \qquad \epsilon_1 \sim N(\epsilon_1 | 0, I).$$

$\beta_1 < 1$ is the variance.

then $$q(z_1 | x) = N(z_1 | \sqrt{1-\beta_1}\, x, \beta_1 I).$$

More generally, $$z_t = \sqrt{1-\beta_t}\, z_{t-1} + \sqrt{\beta_t}\, \epsilon_t, \qquad \epsilon_t \sim N(\epsilon_t | 0; I).$$

$$q(z_t | z_{t-1}) = N(z_t | \sqrt{1-\beta_t}\, z_{t-1}, \beta_t I).$$

$$\beta_1 < \beta_2 < \dots < \beta_T$$

Diffusion Kernel.

$$q(z_1, \dots, z_t | x) = q(z_1 | x) \prod_{\tau=2}^{t} q(z_\tau | z_{\tau-1}).$$

One can marginalize over intermediate variables and find that:

$$q(z_t | x) = \mathcal{N}(z_t | \sqrt{\alpha_t}\, x, (1-\alpha_t) I)$$

$$\alpha_t = \prod_{\tau=1}^{t} (1 - \beta_\tau),$$

or $z_t = \sqrt{\alpha_t}\, x + \sqrt{1-\alpha_t}\, \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\epsilon_t | 0, I).$

and as $T \to \infty$, we have

$$q(z_T | x) = \mathcal{N}(z_T | 0, I).$$

and $\Leftarrow$ $\underbrace{q(z_T)}_{\text{indep. of } x} = \mathcal{N}(z_T | 0, I)$

Conditional distribution.

We are trying to unlearn to undo the noise process, and so, we want $q(z_{t-1} | z_t)$:

$$q(z_{t-1} | z_t) = \frac{q(z_t | z_{t-1})\, q(z_{t-1})}{q(z_t)}.$$

and

$$q(z_{t-1}) = \int q(z_{t-1} | x)\, p(x)\, dx.$$

where $q(z_{t-1} | x) = \mathcal{N}(z_{t-1} | \sqrt{\alpha_{t-1}}\, x, (1-\alpha_{t-1}) I)$

Hence $q(z_{t-1} | z_t)$ is intractable, so we condition on x.

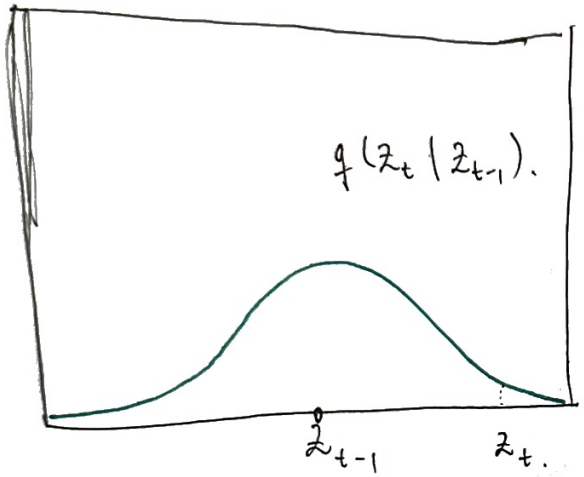
The proof is left to myself in the future ↓

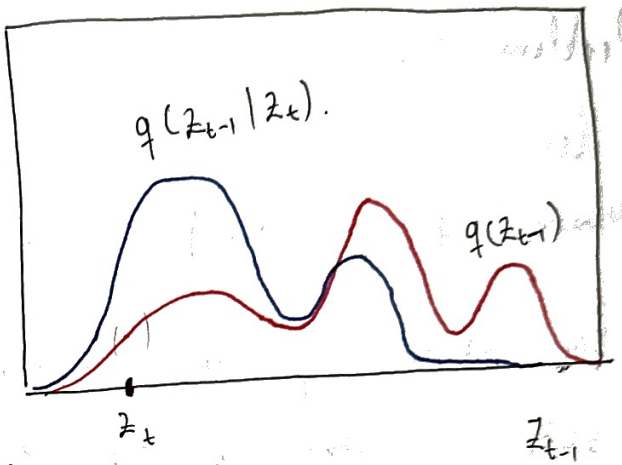
$$q(z_{t-1} | z_t, x) = \frac{q(z_t | z_{t-1}, x)\, q(z_{t-1} | x)}{q(z_t | x)}.$$

$$\hookrightarrow q(z_{t-1} | z_t, x) = \mathcal{N}\left(z_{t-1} \middle| \frac{(1-\alpha_{t-1})\sqrt{1-\beta_t}\, z_t + \sqrt{\alpha_{t-1}}\, \beta_t x}{1-\alpha_t}, \beta_t \frac{(1-\alpha_{t-1})}{1-\alpha_t} I\right)$$

Reverse Decoder.

But we reeeally want $q(z_{t-1}|z_t)$ because if we are doing generative modeling begining from z_T, we can not assume the ~~existence~~ (Knowledge) of X.

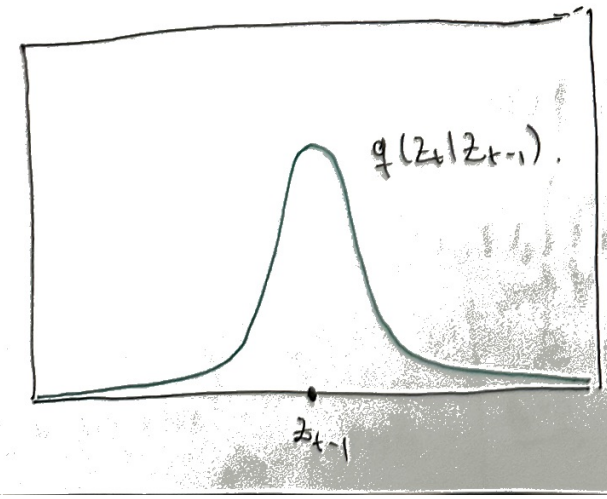


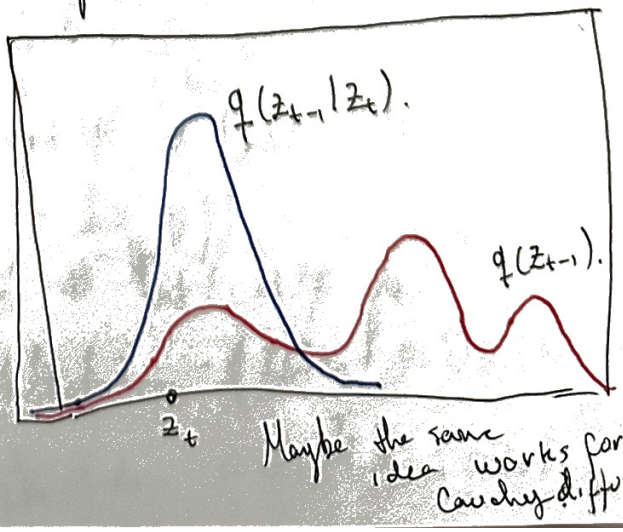


$$q(z_t|z_{t-1}) = \mathcal{N}(z_t|\sqrt{1-\beta_t}z_{t-1}, \beta_t I).$$

$$q(z_{t-1}|z_t) = \frac{q(z_t|z_{t-1})\, q(z_{t-1})}{q(z_t)}.$$

If $q(z_t|z_{t-1})$ is a sufficiently narrow Gaussian then $q(z_{t-1})$ will vary only a small amount over the region in which $q(z_t|z_{t-1})$ has significant mass, and hence $q(z_{t-1}|z_t)$ will be approximately Gaussian.





Maybe the same idea works for Cauchy diffusion?

More formally, $q(z_{t-1}|z_t)$ will be approximately Gaussian by making a Taylor expansion of $\ln q(z_{t-1}|z_t)$ around the point z_t as a function of z_{t-1}. This also shows that for small variance, the reverse distribution $q(z_t|z_{t-1})$ will have a covariance that is close to $\beta_t I$.

Problem 20.7.

Consider the inverse conditional $q(z_{t-1}|z_t) = \frac{q(z_t|z_{t-1})\,q(z_{t-1})}{q(z_t)}$

where $q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}\,z_{t-1}, \beta_t I)$

By taking log and making a Taylor expansion of $q(z_{t-1})$ centred on z_t, show that, for small β_t, $q(z_{t-1}|z_t)$ is approximately Gaussian with mean z_t and covariance $\beta_t I$.

$$\log q(z_{t-1}|z_t) = \log q(z_t|z_{t-1}) + \log q(z_{t-1}) - \log q(z_t)$$

$$\log q(z_{t-1}|z_t) \approx \log q(z_t|z_{t-1}) + \log q(z_t) - \log q(z_t)$$

$$q(z_{t-1}|z_t) \propto q(z_t|z_{t-1})\,q(z_{t-1})$$

$$\log q(z_{t-1}|z_t) \propto \log q(z_t|z_{t-1}) + \log q(z_{t-1})$$

$q(z_{t-1}|z_t)$

$$\log q(z_{t-1}) \approx \log q_{z_{t-1}}(z_t) + (z_{t-1} - z_t)\, D\, q_{z_{t-1}}(z_t)$$

$$\log q(z_{t-1}|z_t) \approx -\frac{1}{2\beta_t}(z_t - \sqrt{1-\beta_t}\, z_{t-1})^T (z_t - \sqrt{1-\beta_t}\, z_{t-1})$$

$$+ \log q_{z_{t-1}}(z_t) + (z_{t-1} - z_t)\, D\, q_{z_{t-1}}(z_t)$$

$$- \log 2\pi\beta_t - \log q_{z_t}(z_t)$$

constants wrt z_{t-1}

$$= -\frac{1}{2\beta_t}(z_t - \sqrt{1-\beta_t}\, z_{t-1})^T (z_t - \sqrt{1-\beta_t}\, z_{t-1})$$

$$+ (z_{t-1} - z_t)^T D\, q_{z_{t-1}}(z_t) + \text{const}$$

$a^2 - b^2$

$a^2 - 2ab + b^2$

$= (a-b)^2$

$a = \sqrt{1-\beta_t}\, z_{t-1}$

$$= -\frac{1}{2\beta_t}\left((1-\beta_t) z_{t-1}^T z_{t-1} - 2\sqrt{1-\beta_t}\, z_t^T z_{t-1}\right) + D^T z_{t-1} + \text{const}$$

$$= -\frac{1}{2\beta_t}\left((1-\beta_t) z_{t-1}^T z_{t-1} - 2\left(\sqrt{1-\beta_t}\, z_t + \beta_t D\right)^T z_{t-1}\right) + \text{const}$$

$$= -\frac{1}{2\beta_t}\frac{1}{1-\beta_t}\left(z_{t-1}^T z_{t-1} - 2\left(\frac{1}{\sqrt{1-\beta_t}} z_t + \frac{\beta_t}{1-\beta_t} D\right)^T z_{t-1}\right) + \text{const}$$

$$= -\frac{1}{2\beta_t}\frac{1}{1-\beta_t}\left(z_{t-1} - \left(\frac{1}{\sqrt{1-\beta_t}} z_t + \frac{\beta_t}{1-\beta_t} D\right)\right)^T \left(z_{t-1} - \left(\frac{1}{\sqrt{1-\beta_t}} z_t + \frac{\beta_t}{1-\beta_t} D\right)\right)$$

Expansions at $x = 0$.

$\frac{1}{\sqrt{1-x}} \approx 1 + \frac{x}{2} + \frac{3x^2}{8}$

$\frac{x}{1-x} \approx x + x^2 + x^3$

$\frac{1}{1-x} \approx 1 + x + x^2 + \dots$

small β_t

$$\approx -\frac{1}{2\beta_t}(1)\left(z_{t-1} - \left((1)z_t + \beta_t D\right)\right)^T \left(z_{t-1} - (z_t + \beta_t D)\right)$$

$\beta_t \to 0$

$$\longrightarrow -\frac{1}{2\beta_t}(z_{t-1} - z_t)^T (z_{t-1} - z_t) \sim \mathcal{N}(z_t, \beta_t)$$

Can we apply this to a Cauchy diffusion?

Let's define it.

Forward Encoder:

Consider x given, ~~in~~ in the univariate setting.

$$z_1 = \sqrt{1-\beta_1}\, x + \sqrt{\beta_1}\, \epsilon_1, \quad \epsilon_1 \sim \text{Cauchy}(\epsilon_1 \mid 0, 1). \tag{1}$$

$$\Rightarrow q(z_1 \mid x) = \text{Cauchy}\left(z_1 \mid \sqrt{1-\beta_1}\, x, \sqrt{\beta_1}\, I\right)?$$

From property (1) of the previous page on the Cauchy distribution:

1. If $X \sim \text{Cauchy}(x_0, \gamma)$ then $kX + l \sim \text{Cauchy}(x_0 k + l, \gamma|k|)$.

Hence, if $\epsilon_1 \sim \text{Cauchy}(\epsilon_1 \mid 0, 1)$, then:

$$z_1 \mid x \sim \text{Cauchy}\left(\sqrt{1-\beta_1}\, x, \sqrt{\beta_1}\right). \tag{2}$$

Each successive ~~image~~ point is given by:

(3).

$$z_t = \sqrt{1-\beta_t}\, z_{t-1} + \sqrt{\beta_t}\, \epsilon_t \qquad \epsilon_t \sim \text{Cauchy}(\epsilon_t \mid 0, 1)$$

Where again, by the Markov property and 1.

$$q(z_t \mid z_{t-1}) = \text{Cauchy}\left(z_t \mid \sqrt{1-\beta_t}\, z_{t-1}, \sqrt{\beta_t}\right).$$

Select β_t s.t. $\beta_1 < \beta_2 < \ldots < \beta_T$.

Diffusion Kernel:

Does it follow that

$$q(z_t \mid x) = \text{Cauchy}\left(z_t \mid \sqrt{\alpha_t}\, x, (1-\alpha_t) I\right)$$

$$\text{with} \quad \alpha_t = \prod_{\tau=1}^{t} (1-\beta_\tau). \quad ?$$

We proceed via diffusion.

This clearly holds when $t=1$, and $z_0 = x$.

Assume t holds, that is

$$z_t = \sqrt{\alpha_t}\, x + (1-\sqrt{\alpha_t})\,\epsilon_t \quad \text{and} \quad q(z_t \mid x) = \text{Cauchy}(\sqrt{\alpha_t}\, x, (1-\alpha_t))$$

$$\epsilon_t \sim \text{Cauchy}(\epsilon_t \mid 0, 1).$$

Then

$$z_{t+1} = \sqrt{1-\beta_{t+1}}\, z_t + \sqrt{\beta_{t+1}}\, \epsilon_{t+1}$$

$$= \sqrt{1-\beta_{t+1}}\left(\sqrt{\alpha_t}\, x + (1-\alpha_t)\,\epsilon_t\right) + \sqrt{\beta_{t+1}}\, \epsilon_{t+1}$$

$$= \sqrt{\alpha_{t+1}}\, x + \underbrace{\sqrt{1-\beta_{t+1}}\,(1-\alpha_t)\,\epsilon_t + \sqrt{\beta_{t+1}}\,\epsilon_{t+1}}_{\text{Cauchy}(s \mid 0, \sqrt{1-\beta_{t+1}}(1-\alpha_t)) \;\rightsquigarrow \text{ not induction - sum does not work}}$$

It is not true. But we can then define Cauchy diffusion

$$z_1 = (1-\beta_1)\, x + \beta_1 \epsilon_1, \qquad \epsilon_1 \sim \text{Cauchy}(\epsilon_1 \mid 0, 1).$$

$$z_t = (1-\beta_t)\, z_{t-1} + \beta_t \epsilon_t \qquad \epsilon_t \sim \text{Cauchy}(\epsilon_t \mid 0, 1).$$

$$z_t \mid z_{t-1} \sim \text{Cauchy}(z_t \mid (1-\beta_t) z_{t-1}, \beta_t)$$

Now for the diffusion kernel:

$t=1$, does $q(z_t \mid x) = \text{Cauchy}(z_t \mid \alpha_t x, (1-\alpha_t))$

↳ $z_1 = (1-\beta_1)\, x + \beta_1 \epsilon_1 = \alpha_1 x + (1-\alpha_1)\,\epsilon_1$

For $t+1$, assuming t

$$z_{t+1} = (1-\beta_{t+1})\, z_t + \beta_{t+1}\, \epsilon_{t+1}$$

$$= (1-\beta_{t+1})\left(\alpha_t x + (1-\alpha_t)\,\epsilon_t\right) + \beta_{t+1}\, \epsilon_{t+1}$$

$$= \alpha_{t+1}\, x + \underbrace{(1-\beta_{t+1})(1-\alpha_t)\,\epsilon_t}_{\bullet} + \underbrace{\beta_{t+1}\, \epsilon_{t+1}}_{\bullet\bullet}$$

* $(1-\beta_{t+1})(1-\alpha_t)\epsilon_t \sim \text{Cauchy}(s \mid 0, (1-\beta_{t+1})(1-\alpha_t))$.
** $\beta_{t+1}\epsilon_{t+1} \sim \text{Cauchy}(s' \mid 0, (\beta_{t+1}))$.
* and ** are independent, and so by Cauchy properties

> 2. If $X \sim \text{Cauchy}(x_0, \gamma_0)$ and $Y \sim \text{Cauchy}(x_1, \gamma_1)$ are independent, then
>
> $$X + Y \sim \text{Cauchy}(x_0 + x_1, \gamma_0 + \gamma_1)$$
> $$X - Y \sim \text{Cauchy}(x_0 - x_1, \gamma_0 + \gamma_1).$$

Hence,

$$* + ** \sim \text{Cauchy}(s'' \mid 0, (1-\beta_{t+1})(1-\alpha_t) + \beta_{t+1})$$
$$= \text{Cauchy}(s'' \mid 0, 1 - (1-\beta_{t+1})\alpha_t)$$
$$= \text{Cauchy}(s'' \mid 0, 1 - \alpha_{t+1}).$$

and so, reparameterizing appropriately, we have that

$$q(z_t \mid x) = \text{Cauchy}(z_t \mid \alpha_t x, 1-\alpha_t).$$

Makes sense.

Now, Reverse process. What is, if any, the distribution such that approximates $q(z_{t+1} \mid z_t)$ for small β_t?

$$q(z_{t-1} \mid z_t) \propto q(z_t \mid z_{t-1})\, q(z_{t-1}).$$

$$= \frac{1}{\pi\beta_t\left[1 + \left(\frac{z_t - (1-\beta_t) z_{t-1}}{\beta_t}\right)^2\right]} \cdot q(z_{t-1}).$$

Taylor

$$q(z_{t-1}) \approx q_{z_{t-1}}(z_t) + (z_{t-1} - z_t)\, D\, q_{z_{t-1}}(z_t) + (z_{t-1} - z_t)^2 \frac{\partial^2}{\partial z_{t-1}^2} q_{z_{t-1}}(z_t)\Big|_{z_t}$$

Maybe I need a 2nd order approximation

$$\approx \frac{q_{z_{t-1}}(z_t) + (z_{t-1} - z_t)\, D\, q_{z_{t-1}}(z_t)}{\pi\beta_t\left[1 + \left(\frac{z_t - (1-\beta_t) z_{t-1}}{\beta_t}\right)^2\right]}$$

$=$

$$\log q(z_{t-1} \mid z_t) = \log q(z_t \mid z_{t-1}) + \log q(z_{t-1}) - \log q(z_t)$$

$$= -\log \pi\beta_t - \log\left(1 + \left(\frac{z_t - (1-\beta_t) z_{t-1}}{\beta_t}\right)^2\right) + \log q(z_{t-1}) + \text{const.}$$

$$\approx -\log \beta_t - \log\left(1 + \left(\frac{z_t - (1-\beta_t) z_{t-1}}{\beta_t}\right)^2\right) + \log q(z_t) + (z_{t-1} - z_t)\, D(z_t) + \text{const.}$$

$$= -\log\left(\beta_t^2 + (z_t - (1-\beta_t) z_{t-1})^2\right) + z_{t-1}\, D(z_t) + \text{const}$$

$$= \log\left(\frac{e^{z_{t-1} D(z_t)}}{\beta_t^2 + (z_t - (1-\beta_t) z_{t-1})^2}\right)$$