

# **TTIC 31230, Fundamentals of Deep Learning**

David McAllester, Autumn 2023

The Mathematics of Diffusion Models

McAllester, arXiv January 2023

# Diffusion Models

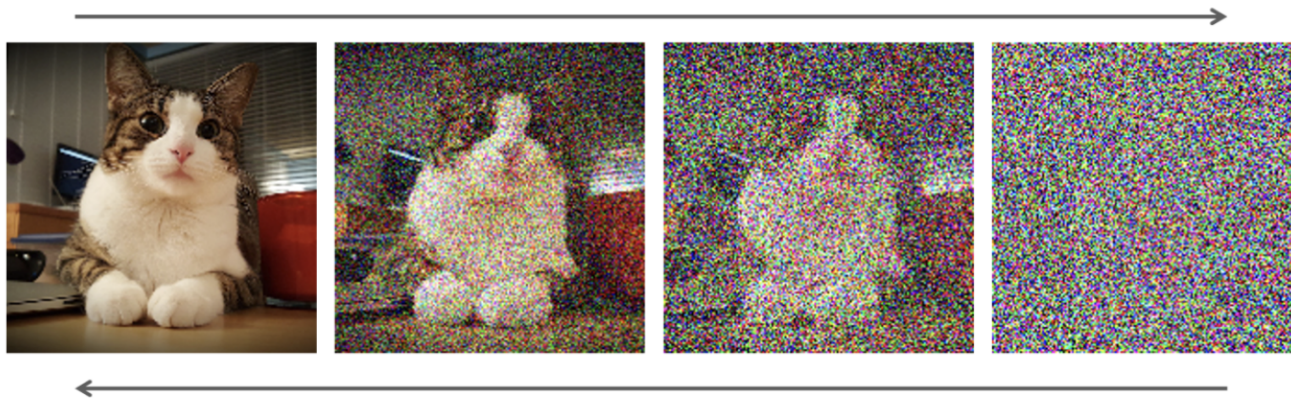
Deep unsupervised learning using nonequilibrium thermodynamics  
Sohl-Dickstein et al., 2015.

Denoising Diffusion Probabilistic Models (DDPM)  
Ho, Jain and Abbeel, June 2020



# Markovian VAEs

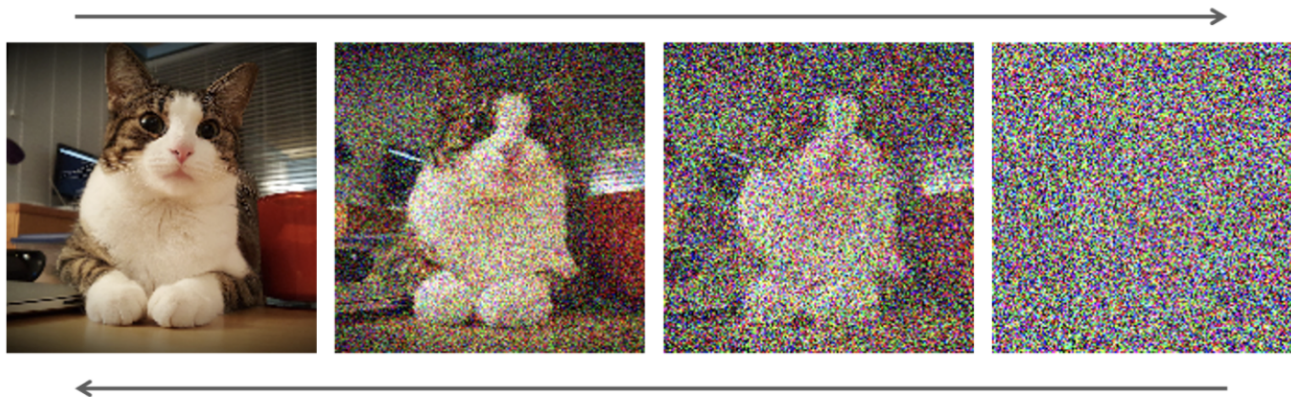
A diffusion models computes and inverts a sequence



So does an autoregressive language model

[Sally talked to John]  $\xleftrightarrow{\quad} \xleftarrow{\quad}$  [Sally talked to]  $\xleftrightarrow{\quad} \xleftarrow{\quad}$  [Sally talked]  $\xleftrightarrow{\quad} \xleftarrow{\quad}$  [Sally]

# Markovian VAEs



[Sally talked to John]  $\overset{\rightarrow}{\leftarrow}$  [Sally talked to]  $\overset{\rightarrow}{\leftarrow}$  [Sally talked]  $\overset{\rightarrow}{\leftarrow}$  [Sally]

$$y \overset{\rightarrow}{\leftarrow} z_1 \overset{\rightarrow}{\leftarrow} \dots \overset{\rightarrow}{\leftarrow} z_N$$

## Markovian VAEs

$$y \overset{\rightarrow}{\leftarrow} z_1 \overset{\rightarrow}{\leftarrow} \dots \overset{\rightarrow}{\leftarrow} z_N$$

**Encoder:**  $\text{Pop}(y)$ ,  $P_{\text{enc}}(z_1|y)$ , and  $P_{\text{enc}}(z_{\ell+1}|z_\ell)$ .

**Generator:**  $P_{\text{pri}}(z_N)$ ,  $P_{\text{gen}}(z_{\ell-1}|z_\ell)$ ,  $P_{\text{gen}}(y|z_1)$ .

The encoder and the decoder define distributions  $P_{\text{enc}}(y, \dots, z_N)$  and  $P_{\text{gen}}(y, \dots, z_N)$  respectively.

# The Markovian ELBO

$$\begin{aligned}
H(y) &= E_{\text{enc}} \left[ -\ln \frac{P_{\text{enc}}(y) P_{\text{enc}}(z_1, \dots, z_N | y)}{P_{\text{enc}}(z_1, \dots, z_N | y)} \right] \\
&= E_{\text{enc}} \left[ -\ln \frac{P_{\text{enc}}(y|z_1) P_{\text{enc}}(z_1|z_2) \cdots P_{\text{enc}}(z_{N-1}|z_N) P_{\text{enc}}(z_N)}{P_{\text{enc}}(z_1|z_2, y) \cdots P_{\text{enc}}(z_{N-1}|z_N, y) P_{\text{enc}}(z_N|y)} \right] \\
&\leq E_{\text{enc}} \left[ -\ln \frac{P_{\text{gen}}(y|z_1) P_{\text{gen}}(z_1|z_2) \cdots P_{\text{gen}}(z_{N-1}|z_N) P_{\text{gen}}(z_N)}{P_{\text{enc}}(z_1|z_2, y) \cdots P_{\text{enc}}(z_{N-1}|z_N, y) P_{\text{enc}}(z_N|y)} \right] \\
&= \begin{cases} E_{\text{enc}} [-\ln P_{\text{gen}}(y|z_1)] \\ + \sum_{i=2}^N E_{\text{enc}} KL(P_{\text{enc}}(z_{i-1}|z_i, y), P_{\text{gen}}(z_{i-1}|z_i)) \\ + E_{\text{enc}} KL(P_{\text{enc}}(Z_N|y), p_{\text{gen}}(Z_N)) \end{cases}
\end{aligned}$$

## Markovian VAEs

$$y \overset{\rightarrow}{\leftarrow} z_1 \overset{\rightarrow}{\leftarrow} \cdots \overset{\rightarrow}{\leftarrow} z_N$$

- autoregressive models
- diffusion models
- StyleGan? (layers of resolution)
- U-Nets? (layers of resolution)

A grand unified theory (GUT) of generative AI?

## Diffusion Models



Consider a discrete time process  $z(0), z(\Delta t), z(2\Delta t), z(3\Delta t), \dots$

$$z(0) = y, \quad y \sim \text{Pop}(y)$$

$$z(t + \Delta t) = z(t) + \epsilon \sqrt{\Delta t}, \quad \epsilon \sim \mathcal{N}(0, I)$$

A sum of two Gaussians is a Gaussian whose **variance** is the sum of the two variances.

$$z(t + n\Delta t) = z(t) + \sqrt{n\Delta t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

Here  $\sqrt{n\Delta t}$  is the **standard deviation** of the added noise.



## SDE Notation

In these slides  $\epsilon$  will be a random variable drawn from  $\mathcal{N}(0, I)$ .

This corresponds to “ $dB$ ” in standard notation for SDEs.

$$z(t + \Delta t) = z(t) + \mu(z, t)\Delta t + \sigma(z, t)\epsilon\sqrt{\Delta t}$$

$$dz = \mu(z, t)dt + \sigma(z, t)dB$$

The first expression is longer but seems clearer to me.

The SDE denotes the limit as  $\Delta t$  in the first equation goes to zero.

## The Diffusion SDE

For the diffusion process (Brownian motion) we have

$$z(0) = y, \quad y \sim \text{Pop}(y)$$

$$z(t + \Delta t) = z(t) + \epsilon \sqrt{\Delta t} \tag{1}$$

$$dz = dB$$

For diffusion we get that (1) holds for all  $t$  and  $\Delta t$ .

## Probability Notation

In these slides unsubscripted probability notation, such as

$$P(z(t + \Delta t)|z(t)),$$

or a conditional expectation such as

$$E[f(y)|z(t)] = E_{y \sim P(y|z_t)}[f(y)],$$

refer the joint distribution on  $y$  and  $z(t)$  defined by diffusion.

# Markovian ELBO

For any Markovian VAE we have

$$-\ln \text{Pop}(y) = -\ln \frac{P(z_N)P(z_{N-1}|z_N) \cdots P(z_1|z_2)P(y|z_1)}{P(z_N|y)P(z_{N-1}|z_N, y) \cdots P(z_1|z_2, y)}$$

$$H(y) = \begin{cases} E[KL(P(z_N|y), P(z_N))] \\ + \sum_{i=2}^N E[KL(P(z_{i-1}|z_i, y), P(z_{i-1}|z_i))] \\ + E[\ln -P(y|z_1)] \end{cases} \quad (2)$$

$$\leq \begin{cases} E[KL(P(z_N|y), P_{\text{gen}}(z_N))] \\ + \sum_{i=2}^N E[KL(P(z_{i-1}|z_i, y), P_{\text{gen}}(z_{i-1}|z_i))] \\ E[-\ln P_{\text{gen}}(y|z_1)] \end{cases} \quad (3)$$

## Reverse-Time Probabilities

In the limit of small  $\Delta t$  it is possible to derive the following.

$$P(z(t - \Delta t)|z(t), y) = \mathcal{N} \left( z(t) + \frac{\Delta t(y - z(t))}{t}, \Delta t I \right)$$

$$P(z(t - \Delta t)|z(t)) = \mathcal{N} \left( z(t) + \frac{\Delta t(E[y|t, z(t)] - z(t))}{t}, \Delta t I \right)$$

## The Reverse-Diffusion SDE

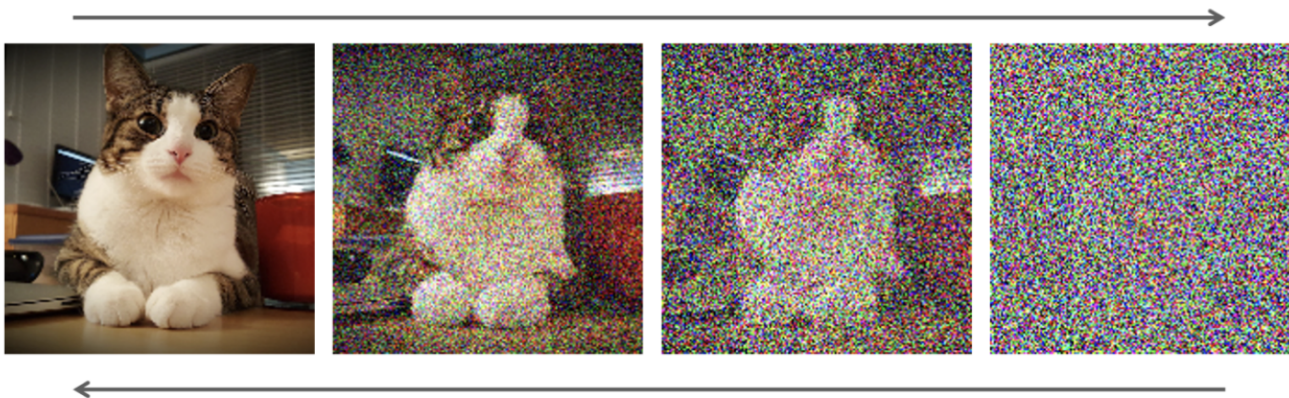
$$P(z(t - \Delta t)|z(t)) = \mathcal{N} \left( z(t) + \frac{\Delta t(E[y|t, z(t)] - z(t))}{t}, \Delta t I \right)$$

This equation defines a reverse-diffusion SDE which we can write as

$$z(t - \Delta t) = z(t) + \left( \frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t + \epsilon \sqrt{\Delta t}$$

# Understanding Reverse Diffusion

$$z(t - \Delta t) = z(t) + \left( \frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t + \epsilon \sqrt{\Delta t}$$



$E[y|t, z]$  is averaging over many possible source images  $y$ .

## Estimating $E[y|t, z(t)]$

$$z(t - \Delta t) = z(t) + \left( \frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t + \epsilon \sqrt{\Delta t}$$

We can train a denoising network  $\hat{y}(t, z)$  to estimate  $E[y|t, z(t)]$  using

$$\hat{y}^*(t, z) = \underset{\hat{y}}{\operatorname{argmin}} E (\hat{y}(t, z(t)) - y)^2$$

Assuming universality  $\hat{y}^*(t, z) = E[y|t, z]$ .



## Estimating $E[y|t, z(t)]$

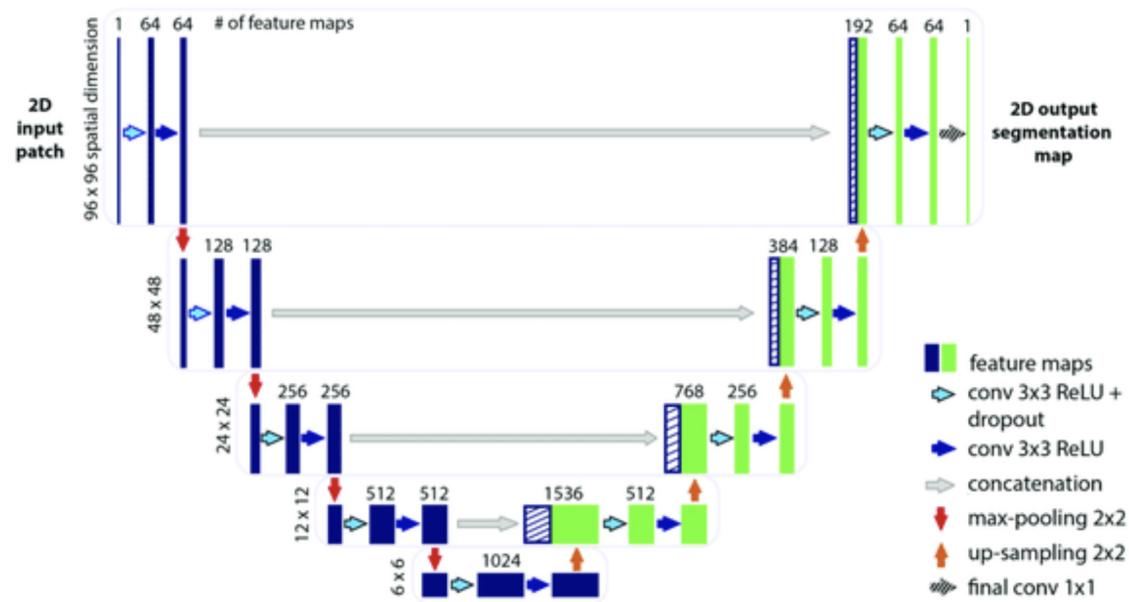
If the population values are scaled so as to have scale 1, then the scale of  $z(t)$  is  $\sqrt{1+t}$ .

$$\hat{y}^* = \underset{\hat{y}}{\operatorname{argmin}} E_{t,z(t)} (\hat{y}(t, z/\sqrt{1+t}) - y)^2$$

$$\hat{E}[y|t, z(t)] = \hat{y}^*(t, z/\sqrt{1+t})$$

$\hat{y}(t, z)$  is a U-Net

In practice  $\hat{y}(t, z)$  is computed with a U-Net.



The U-Nets themselves seem closely related to Markovian VAEs.

## Computing (or Bounding) Entropy

$$H(y) = \begin{cases} E_y[KL(P(z_N|y), P(z_N))] \\ + \sum_{i=2}^N E_{y,z_i}[KL(P(z_{i-1}|z_i, y), P(z_{i-1}|z_i))] \\ + E_{y,z_1}[\ln -P(y|z_1)] \end{cases}$$

For two Gaussian distributions with the same isotropic covariance we have

$$KL \left( \mathcal{N}(\mu_1, \sigma^2 I), \mathcal{N}(\mu_2, \sigma^2 I) \right) = \frac{\|u_1 - \mu_2\|^2}{2\sigma^2}$$

## Computing (or Bounding) Entropy

$$H(y) = \begin{cases} E_y[KL(P(z_N|y), P(z_N))] \\ + \sum_{i=2}^N E_{y,z_i}[KL(P(z_{i-1}|z_i, y), P(z_{i-1}|z_i))] \\ + E_{y,z_1}[\ln - P(y|z_1)] \end{cases}$$

$$P(z(t - \Delta t)|z(t), y) = \mathcal{N} \left( z(t) + \frac{\Delta t(y - z(t))}{t}, \Delta t I \right)$$

$$P(z(t - \Delta t)|z(t)) = \mathcal{N} \left( z(t) + \frac{\Delta t(E_y[y|t, z(t)] - z(t))}{t}, \Delta t I \right)$$

## Computing (or Bounding) Entropy

$$P(z(t - \Delta t)|z(t), y) = \mathcal{N} \left( z(t) + \frac{\Delta t(y - z(t))}{t}, \Delta t I \right)$$

$$P(z(t - \Delta t)|z(t)) = \mathcal{N} \left( z(t) + \frac{\Delta t(E_{y, z_i}[y|t, z(t)] - z(t))}{t}, \Delta t I \right)$$

$$\begin{aligned} KL \left( \frac{P(z(t - \Delta t)|z(t), y)}{P(z(t - \Delta t)|z(t))} \right) &= \left( \frac{\|y - E[y|t, z(t)]\|^2 \Delta t^2}{2t^2 \Delta t} \right) \\ &= \left( \frac{\|y - E[y|t, z(t)]\|^2}{2t^2} \right) \Delta t \end{aligned}$$

## Computing (or Bounding) Entropy

$$\begin{aligned}
 H(y) &= \begin{cases} E_y[KL(P(z_N|y), P(z_N))] \\ + \sum_{i=2}^N E_{y,z_i}[KL(P(z_{i-1}|z_i, y), P(z_{i-1}|z_i))] \\ + E_{y,z_1}[\ln -P(y|z_1)] \end{cases} \\
 &= \sum_{i=2}^N \left( \frac{\|y - E[y|t, z(t)]\|^2}{2t^2} \right)_{t=i\Delta t} \Delta t + E_{y,z_1}[-\ln P(y|z_1)]
 \end{aligned}$$

## Passing to the Integral

$$H(y) = \left\{ \begin{array}{l} \int_{t_0}^{\infty} dt \quad E_{z(t)|y} \left[ \frac{||y - E[y|t, z(t)]||^2}{2t^2} \right] \\ + E_{z(t_0)|y} [-\ln P(y|z(t_0))] \end{array} \right.$$

$$H(y) = \left\{ \begin{array}{l} \int_{t_0}^{\infty} dt \quad E_{y, z(t_0)} \left[ \frac{||y - E[y|t, z(t)]||^2}{2t^2} \right] \\ + H(y|z(t_0)) \end{array} \right.$$

## Mutual Information

$$H(y) = \begin{cases} \int_{t_0}^{\infty} dt \ E_{y,z(t_0)} \left[ \frac{||y - E[y|t, z(t)]||^2}{2t^2} \right] \\ + H(y|z(t_0)) \end{cases}$$

$$H(y) - H(y|z(t_0)) = \int_{t_0}^{\infty} dt \ E_{y,z(t_0)} \left[ \frac{||y - E[y|t, z(t)]||^2}{2t^2} \right]$$

$$I(y, z(t_0)) = \int_{t_0}^{\infty} dt \ E_{y,z(t_0)} \left[ \frac{||y - E[y|t, z(t)]||^2}{2t^2} \right]$$

This is the information minimum mean squared error relation (I-MMSE) relation [Guo et al. 2005].



## Bounding Entropy

$$I(y, z(t_0)) = \int_{t_0}^{\infty} dt \ E_{y, z(t_0)} \left[ \frac{||y - \textcolor{red}{E}[y|t, z(t)]||^2}{2t^2} \right]$$

$$\leq \int_{t_0}^{\infty} dt \ E_{y, z(t_0)} \left[ \frac{||y - \textcolor{red}{\hat{E}}[y|t, z(t)]||^2}{2t^2} \right]$$

## The Fokker-Plack Anaylysis (The Score Function)

For  $\epsilon \sim \mathcal{N}(0, I)$  a general SDE can be written as

$$z(t + \Delta t) = z(t) + \mu(z(t), t)\Delta t + \sigma(z(t), t)\epsilon\sqrt{\Delta t}$$

$$dz = \mu(z(t), t)dt + \sigma(z(t), t)dB$$

The diffusion process is the special case of Brownian motion

$$\begin{aligned} z(t + \Delta t) &= z(t) + \epsilon\sqrt{\Delta t} \\ dz &= dB \end{aligned}$$

## The Fokker-Planck Equation

Let  $P_t(z)$  be the probability that  $z(t) = z$ .

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \begin{pmatrix} \mu(z(t), t) P_t(z) \\ -\frac{1}{2} \sigma^2(z(t), t) \nabla_z P_t(z) \end{pmatrix}$$

For the special case of diffusion we have

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left( -\frac{1}{2} \nabla_z P_t(z) \right)$$

## The Score Function

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \begin{pmatrix} \mu(z(t), t) P_t(z) \\ -\frac{1}{2} \sigma^2(z(t), t) \nabla_z P_t(z) \end{pmatrix}$$

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left( -\frac{1}{2} \nabla_z P_t(z) \right)$$

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left[ \left( -\frac{1}{2} \nabla_z \ln P_t(z) \right) P_t(z) \right]$$

## The Score Function

$$\frac{\partial P_t(z)}{\partial t} = -\nabla \cdot \left[ \left( -\frac{1}{2} \nabla_z \ln P_t(z) \right) P_t(z) \right]$$

$\ln P_t(z)$  is the score function.

The time evolution of  $P_t(z)$  can be written as the result of **deterministic** flow given by

$$\frac{dz}{dt} = -\frac{1}{2} \nabla_z \ln p_t(z)$$

## Deterministic Reverse Diffusion

Following the deterministic flow backward in time samples from the population!

$$z(t - \Delta t) = z(t) + \frac{1}{2} \nabla_z \ln p_t(z) \Delta t$$

No reverse diffusion noise!

## Solving for the Score Function

$$P_t(z) = E_y P_t(z|y)$$

$$= E_y \frac{1}{Z(t)} e^{-\frac{\|z-y\|^2}{2t}}$$

$$\nabla_z P_t(z) = E_y P_t(z|y) (y - z)/t$$

$$= E_y \frac{P_t(z) P(y|t, z)}{P(y)} [(y - z)/t]$$

$$= P_t(z) \int dy P(y|t, z) [(y - z)/t]$$

$$= P_t(z) \frac{E[y|t, z] - z}{t}$$

$$\nabla_z \ln P_t(z) = \frac{E[y|t, z] - z}{t}$$

This is Tweedie's formula, Robbins 1956.

## Stochastic vs. Deterministic Reverse Diffusion

$$z(t - \Delta t) = z(t) + \left( \frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t + \epsilon \sqrt{\Delta t}$$

$$z(t - \Delta t) = z(t) + \frac{1}{2} \left( \frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t$$



## Interpolating Stochastic and Deterministic

One can show that for  $\lambda \in [0, 1]$  the following also samples from the population.

$$z(t - \Delta t) = z(t) + \frac{1 + \lambda}{2} \left( \frac{E[y|t, z(t)] - z(t)}{t} \right) \Delta t + \lambda \epsilon \sqrt{\Delta t}$$

**END**