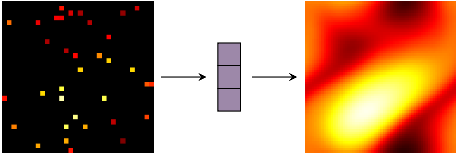
Autoencoders in Function Space

Justin Bunker[1], Mark Girolami[1,3], **Hefin Lambley**[4],
Andrew M. Stuart[2], and T. J. Sullivan[4]



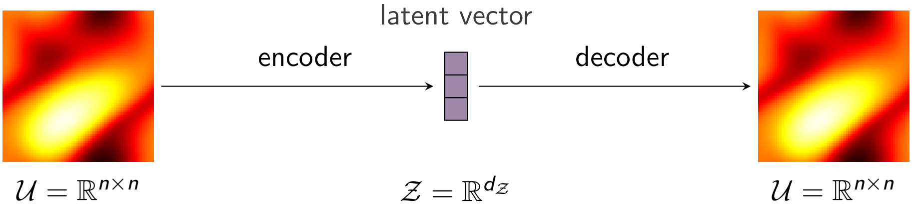
UNIVERSITY OF CAMBRIDGE
1

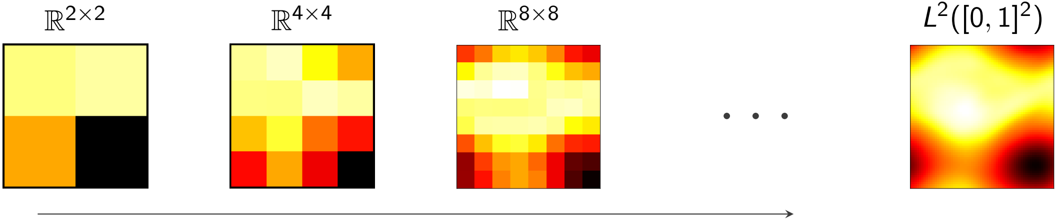
Caltech
2

The Alan Turing Institute
3

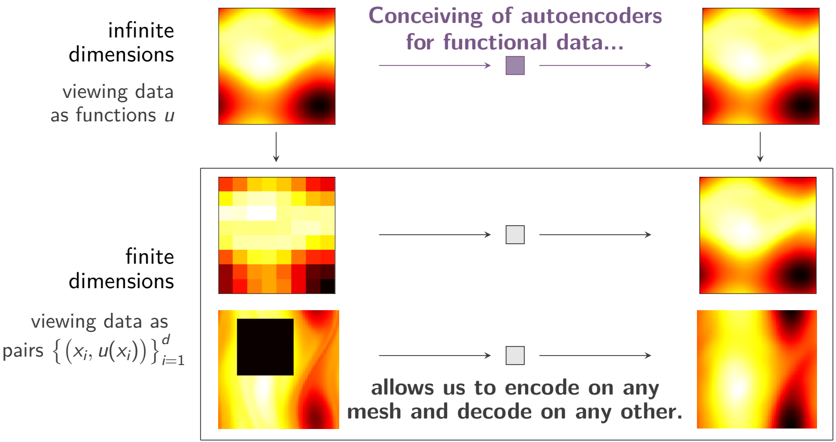
UNIVERSITY OF WARWICK
4

Autoencoders are machine-learning models for
dimension reduction and **generative modelling**





In **scientific applications** and in **image processing**, it is useful to **view discretised data as approximations of the underlying functions**.


infinite dimensions
viewing data as functions u
Conceiving of autoencoders for functional data...
finite dimensions
viewing data as pairs $\{(x_i, u(x_i))\}_{i=1}^{d}$
allows us to encode on any mesh and decode on any other.

The autoencoder problem in the continuum

Given | **Learn** | **Choose**

$\mathcal{U}$

data space
(separable Banach space)

encoder $\rightarrow$

$\leftarrow$ decoder

$\mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}}$

latent space

Υ

data distribution

$\mathbb{P}_z$

latent distribution

The autoencoder problem in the continuum

Given | **Learn** | **Choose**

$\mathcal{U}$ — data space (separable Banach space)

encoder: $\mathcal{U} \to \mathcal{Z}$

decoder: $\mathcal{Z} \to \mathcal{U}$

$\mathcal{Z} = \mathbb{R}^{d_{\mathcal{Z}}}$ — latent space

Υ — data distribution

$\mathbb{P}_z$ — latent distribution

This work in a nutshell:

variational autoencoder (Kingma & Welling, 2014) ⟶ **functional variational autoencoder (FVAE)**
"probabilistic" encoder and decoder

regularised autoencoder ⟶ **functional autoencoder (FAE)**
"deterministic" encoder and decoder.

Functional variational autoencoder (FVAE)

Idea: view the encoder and decoder as probabilistic.

encoder	u vector in $\mathcal{U}$	$\mapsto$	$\mathbb{Q}_{z\|u}$ distribution over $\mathcal{Z}$
decoder	z vector in $\mathcal{Z}$	$\mapsto$	$\mathbb{P}_{u\|z}$ distribution over $\mathcal{U}$

Choose the following:

family of encoders	$\left(u \mapsto \mathbb{Q}^{\theta}_{z\|u}\right)_{\theta\in\Theta}$
family of decoders	$\left(z \mapsto \mathbb{P}^{\psi}_{u\|z}\right)_{\psi\in\Psi}$
latent distribution	$\mathbb{P}_z$ on $\mathcal{Z}$

Choose the following:

family of encoders	$\left(u \mapsto \mathbb{Q}^{\theta}_{z\|u}\right)_{\theta\in\Theta}$
family of decoders	$\left(z \mapsto \mathbb{P}^{\psi}_{u\|z}\right)_{\psi\in\Psi}$
latent distribution	$\mathbb{P}_z$ on $\mathcal{Z}$

Joint encoder model $\mathbb{Q}^{\theta}_{z,u}$ on (z, u)

$$u \sim \Upsilon,$$
$$z \mid u \sim \mathbb{Q}^{\theta}_{z|u}.$$

Choose the following:

family of encoders $\left(u \mapsto \mathbb{Q}^{\theta}_{z|u}\right)_{\theta\in\Theta}$

family of decoders $\left(z \mapsto \mathbb{P}^{\psi}_{u|z}\right)_{\psi\in\Psi}$

latent distribution $\mathbb{P}_z$ on $\mathcal{Z}$

Joint encoder model $\mathbb{Q}^{\theta}_{z,u}$ on (z,u)

$$u \sim \Upsilon,$$
$$z \mid u \sim \mathbb{Q}^{\theta}_{z|u}.$$

Joint decoder model $\mathbb{P}^{\psi}_{z,u}$ on (z,u)

$$z \sim \mathbb{P}_z,$$
$$u \mid z \sim \mathbb{P}^{\psi}_{u|z}.$$

Choose the following:

family of encoders	$\left(u \mapsto \mathbb{Q}^{\theta}_{z\|u}\right)_{\theta\in\Theta}$
family of decoders	$\left(z \mapsto \mathbb{P}^{\psi}_{u\|z}\right)_{\psi\in\Psi}$
latent distribution	$\mathbb{P}_z$ on $\mathcal{Z}$

Joint encoder model $\mathbb{Q}^{\theta}_{z,u}$ on (z,u)

$$u \sim \Upsilon,$$
$$z \mid u \sim \mathbb{Q}^{\theta}_{z|u}.$$

Joint decoder model $\mathbb{P}^{\psi}_{z,u}$ on (z,u)

$$z \sim \mathbb{P}_z,$$
$$u \mid z \sim \mathbb{P}^{\psi}_{u|z}.$$

Objective Minimise the Kullback–Leibler divergence D_{KL} between joint distributions:

$$\operatorname*{arg\,min}_{\theta\in\Theta,\,\psi\in\Psi} D_{\mathrm{KL}}\left(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\right).$$

When is the FVAE objective valid?

Adopt the standard Gaussian VAE model:

Gaussian encoder family	$u \mapsto \mathbb{Q}^{\theta}_{z\|u} = N\big(f(u;\theta), \alpha I_{\mathcal{Z}}\big)$
Gaussian decoder family	$z \mapsto \mathbb{P}^{\psi}_{u\|z} = N\big(g(z;\theta), \beta I_{\mathcal{U}}\big)$
Gaussian latent distribution	$\mathbb{P}_z = N(0, I_{\mathcal{Z}})$

When is the FVAE objective valid?

Adopt the standard Gaussian VAE model:

Gaussian encoder family	$u \mapsto \mathbb{Q}^{\theta}_{z\|u} = N\big(f(u;\theta), \alpha I_{\mathcal{Z}}\big)$
Gaussian decoder family	$z \mapsto \mathbb{P}^{\psi}_{u\|z} = N\big(g(z;\theta), \beta I_{\mathcal{U}}\big)$
Gaussian latent distribution	$\mathbb{P}_z = N(0, I_{\mathcal{Z}})$

Finite dimensions

$\mathcal{U} = \mathbb{R}^d$

Υ has 'nice' density.

$\longrightarrow$

FVAE is equivalent to a VAE:

$$D_{\mathrm{KL}}\big(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\big) = \text{usual VAE objective} + \text{finite const.}$$

evidence lower bound (ELBO)

When is the FVAE objective valid?

Adopt the standard Gaussian VAE model:

Gaussian encoder family	$u \mapsto \mathbb{Q}^{\theta}_{z\|u} = N\big(f(u;\theta), \alpha I_{\mathcal{Z}}\big)$
Gaussian decoder family	$z \mapsto \mathbb{P}^{\psi}_{u\|z} = N\big(g(z;\theta), \beta I_{\mathcal{U}}\big)$
Gaussian latent distribution	$\mathbb{P}_z = N(0, I_{\mathcal{Z}})$

Finite dimensions

$\mathcal{U} = \mathbb{R}^d$

Υ has 'nice' density.

$\longrightarrow$

FVAE is equivalent to a VAE:

$D_{\mathrm{KL}}\big(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\big) =$ usual VAE objective $+$ finite const.

evidence lower bound (ELBO)

Infinite dimensions

$\mathcal{U} = L^2(0,1)$

Υ is *any* probability distribution on $\mathcal{U}$.

$\longrightarrow$

FVAE's objective is identically infinite:

$D_{\mathrm{KL}}\big(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\big) = +\infty$ for all parameters θ and ψ.

For the FVAE objective to be valid, we must choose the data and decoder to be compatible

Assume $\mathcal{U}$ is a separable Banach space, and take

Gaussian encoder family	$u \mapsto \mathbb{Q}^{\theta}_{z\|u} = N\big(f(u;\theta), \alpha I_{\mathcal{Z}}\big)$
Noise distribution	$\mathbb{P}_\eta$ on $\mathcal{U}$
Shifted decoder family	$z \mapsto \mathbb{P}^{\psi}_{u\|z} = g(z;\theta) + \mathbb{P}_\eta$
Gaussian latent distribution	$\mathbb{P}_z = N(0, I_{\mathcal{Z}})$

Theorem

If $D_{\mathrm{KL}}(\Upsilon \,\|\, \mathbb{P}_\eta) < \infty$, then the objective is well defined:

$$\inf_{\theta\in\Theta,\,\psi\in\Psi} D_{\mathrm{KL}}\big(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\big) < \infty.$$

Examples where FVAE can and cannot be applied

✓ Υ is path distribution of SDE $\quad \mathrm{d}u_t = b(u_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}w_t,\ t \in [0, T]$;
$\mathbb{P}_\eta$ is scaled Brownian motion $\quad \mathrm{d}\eta_t = \sqrt{\varepsilon}\,\mathrm{d}w_t$.

✓ Υ is posterior distribution over function (e.g., from Bayesian inverse problem);
$\mathbb{P}_\eta$ is Gaussian prior distribution.

× Υ is distribution of natural images, viewed as functions (e.g., faces);
very hard to choose $\mathbb{P}_\eta$ such that $D_{\mathrm{KL}}(\Upsilon \,\|\, \mathbb{P}_\eta) < \infty$.

In the cases where FVAE can be applied, we can write

$$D_{\mathrm{KL}}\left(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\right) = \mathop{\mathbb{E}}_{u\sim\Upsilon}\left[\mathcal{L}(u;\theta,\psi)\right] + \text{finite const.}$$

Functional autoencoder (FAE)

Idea: view the encoder and decoder as deterministic.

$$\begin{aligned} &\text{encoder} \quad && \mathcal{U} \ni u && \mapsto && f(u) \in \mathcal{Z} \\ &\text{decoder} \quad && \mathcal{Z} \ni z && \mapsto && g(z) \in \mathcal{U} \end{aligned}$$

Then choose:

$\Big(u \mapsto f(u;\theta)\Big)_{\theta \in \Theta}$ family of encoders

$\Big(z \mapsto g(z;\psi)\Big)_{\psi \in \Psi}$ family of decoders

Objective: Given **regularisation scale** $\beta > 0$, solve

$$\operatorname*{arg\,min}_{\theta \in \Theta,\, \psi \in \Psi} \; \mathbb{E}_{u \sim \Upsilon} \left[\frac{1}{2} \big\| g\big(f(u;\theta);\psi\big) - u \big\|_{\mathcal{U}}^2 + \beta \big\| f(u;\theta) \big\|_2^2 \right].$$

$\rightsquigarrow$ Similar to the VAE objective in finite dimensions with Gaussian model.

✓ Objective has finite infimum as long as $\mathbb{E}_{u \sim \Upsilon}[\|u\|^2] < \infty$

With access to data distribution Υ in function space

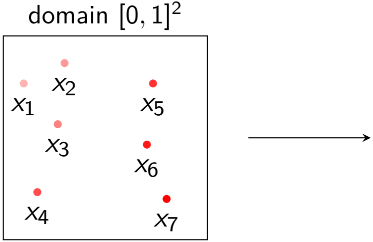
Objective $$\arg\min_{\theta\in\Theta,\,\psi\in\Psi} \mathbb{E}_{u\sim\Upsilon}\left[\mathcal{L}(u;\theta,\psi)\right] + \text{finite const.}$$

With access to training data $\{u_i\}_{i=1}^N \sim \Upsilon$ in function space

Empirical objective $$\arg\min_{\theta\in\Theta,\,\psi\in\Psi} \sum_{i=1}^{N} \mathcal{L}(u_i;\theta,\psi).$$

But we don't have access to the *functions* u_i, just their discrete representations!

We represent discretisations of functions using point clouds

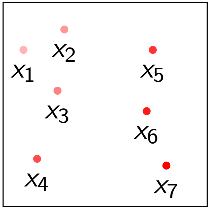


discretised function as point cloud

$$\left\{ \left(x_j, u(x_j)\right) : j = 1, \ldots, M \right\}$$

We represent discretisations of functions using point clouds

domain $[0,1]^2$



discretised function as point cloud

$$\left\{ (x_j, u(x_j)) : j = 1, \ldots, M \right\}$$

Many operations on functions can be discretised on a point cloud—for example:

$$\int_{[0,1]^2} u(x) \, \mathrm{d}x \approx \frac{1}{M} \sum_{j=1}^{M} u(x_j).$$

Since the loss $\mathcal{L}$ from FVAE and FAE consists of function-space norms and inner products (e.g., the L^2-norm), these can be approximated with point-cloud data.

Our proposed architectures

Encoder Define MLPs κ and ρ and let

$$f(u;\theta) = \rho\left(\int_\Omega \kappa\big(x, u(x);\theta\big)\,\mathrm{d}x;\theta\right).$$

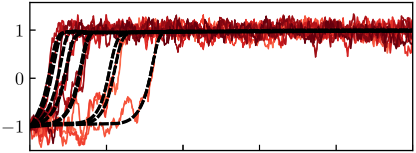
Decoder Parametrise g through coordinate MLP γ:

$$g(z;\psi)(x) = \gamma(z, x;\psi).$$

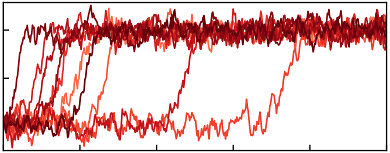
FVAE example problem: Brownian dynamics

Data: Υ distribution on $\mathcal{U} = C([0,5], \mathbb{R})$ of $\mathrm{d}u_t = -\nabla U(u_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}w_t,\ u_0 = -1,$

(a) Data $u \sim \Upsilon$ and reconstructions $g(f(u))$



(b) Samples from generative model



More applications of FVAE in our paper, e.g., motivated by **molecular dynamics** learning a Markov state model from **irregularly sampled transition paths**.

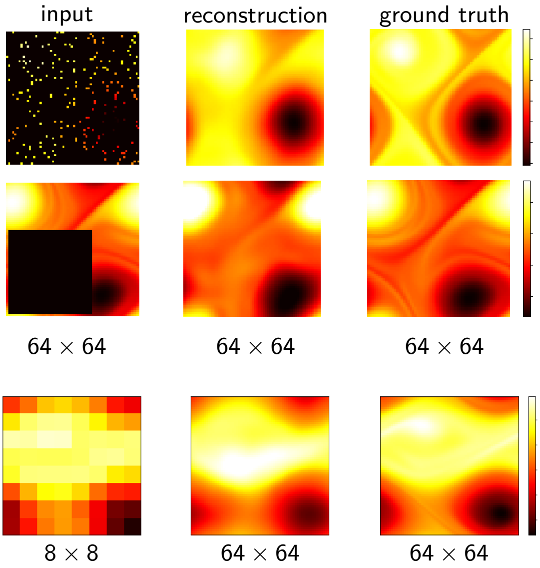
FAE example problem: applications to **inpainting** and **superresolution**

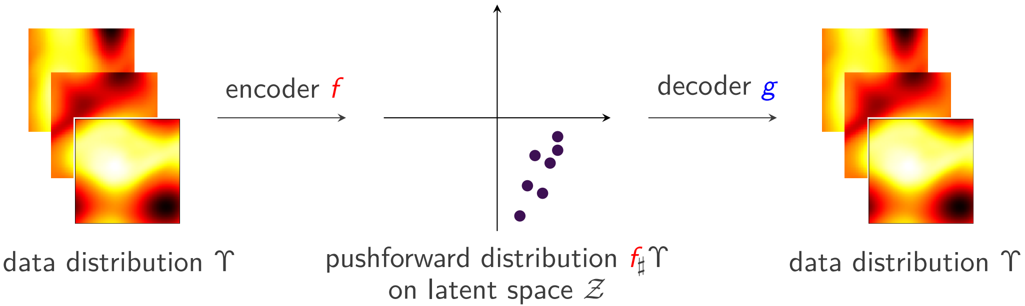
Inpainting

trained at 64×64

Data-driven superresolution

trained at 64×64





Latent generative models While FAE is not inherently a generative model, can learn generative model $\mathbb{P}_z$ to approximate $f_\sharp \Upsilon$ on $\mathcal{Z}$

similar to image generative models such as Stable Diffusion.

Summing things up...

- **Functional variational autoencoder (FVAE)**
  Probabilistic generative model with **built-in uncertainty quantification**.
  Works for **specific classes of data distributions**.

- **Functional autoencoder (FAE)**
  Non-probabilistic autoencoder that can be augmented with a generative model
  Works for **most data distributions** on function space.

Limitations and future work

1. **Barriers to variational inference** in function space;
   can VAEs be extended without the stringent constraints of FVAE?

2. Need for **better architectures** that can be evaluated on any mesh
   e.g., point-cloud architectures such as PointCNN.

3. FVAE and FAE could serve as **building block** for

   supervised learning $\rightsquigarrow$ inspired by PCA-NET
   generative modelling $\rightsquigarrow$ inspired by Stable Diffusion.

More details in our paper:
Justin Bunker, Mark Girolami, **Hefin Lambley**, Andrew M. Stuart, and T. J. Sullivan.
Autoencoders in Function Space. JMLR **26**(165):1–54.

Code package in Python + JAX available at:
`https://github.com/htlambley/functional_autoencoders`

Supplementary slides

Why does absolute continuity fail with the standard Gaussian model?

$$D_{\mathrm{KL}}\left(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\right) < \infty \qquad \Longrightarrow \qquad \mathbb{Q}^{\theta}_{z,u} \ll \mathbb{P}^{\psi}_{z,u}$$

i.e., $\mathbb{P}^{\psi}_{z,u}(A) = 0 \implies \mathbb{Q}^{\theta}_{z,u}(A) = 0.$

Problem: $\mathbb{Q}^{\theta}_{z,u} \not\ll \mathbb{P}^{\psi}_{z,u}$ for the Gaussian model on $\mathcal{U} = L^2(0,1)$.

Why does absolute continuity fail with the standard Gaussian model?

$$D_{\mathrm{KL}}\left(\mathbb{Q}^{\theta}_{z,u} \,\|\, \mathbb{P}^{\psi}_{z,u}\right) < \infty \quad \Longrightarrow \quad \mathbb{Q}^{\theta}_{z,u} \ll \mathbb{P}^{\psi}_{z,u}$$

i.e., $\mathbb{P}^{\psi}_{z,u}(A) = 0 \implies \mathbb{Q}^{\theta}_{z,u}(A) = 0.$

Problem: $\mathbb{Q}^{\theta}_{z,u} \not\ll \mathbb{P}^{\psi}_{z,u}$ for the Gaussian model on $\mathcal{U} = L^2(0,1)$.

Under encoder model $\mathbb{Q}^{\theta}_{z,u}$:

$$u \sim \Upsilon,$$
$$z \mid u \sim N\big(f(u), \alpha I_{\mathcal{Z}}\big).$$

So $(z,u) \in \mathcal{Z} \times \mathcal{U}$ almost surely

i.e., $\mathbb{Q}^{\theta}_{z,u}(\mathcal{Z} \times \mathcal{U}) = 1.$

Under decoder model $\mathbb{P}^{\psi}_{z,u}$:

$$z \sim N(0, I_{\mathcal{Z}}),$$
$$u \mid z \sim N\big(g(z), \beta I_{\mathcal{U}}\big).$$

So $(z,u) \notin \mathcal{Z} \times \mathcal{U}$ almost surely

i.e, $\mathbb{P}^{\psi}_{z,u}(\mathcal{Z} \times \mathcal{U}) = 0.$

Example: FVAE for stochastic differential equations

On $\mathcal{U} = C\big([0, T], \mathbb{R}^m\big)$:

fix	Υ	distribution of	$\mathrm{d}u_t = b(u_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}w_t,$	$u_0 = 0,$	$t \in [0, T]$
choose	$\mathbb{P}_\eta$	distribution of	$\mathrm{d}\eta_t = \sqrt{\varepsilon}\,\mathrm{d}w_t,$	$\eta_0 = 0,$	$t \in [0, T].$

Example: FVAE for stochastic differential equations

On $\mathcal{U} = C\big([0, T], \mathbb{R}^m\big)$:

fix	Υ	distribution of	$\mathrm{d}u_t = b(u_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}w_t,$	$u_0 = 0,$	$t \in [0, T]$
choose	$\mathbb{P}_\eta$	distribution of	$\mathrm{d}\eta_t = \sqrt{\varepsilon}\,\mathrm{d}w_t,$	$\eta_0 = 0,$	$t \in [0, T]$.

Proposition: finite infimum of FVAE objective

By the Girsanov theorem, assuming b is "nice",

$$D_{\mathrm{KL}}\big(\Upsilon \,\|\, \mathbb{P}_\eta\big) = \mathop{\mathbb{E}}_{u\sim\Upsilon}\left[\frac{1}{2\varepsilon}\int_0^T \big\|b(u_t)\big\|^2\,\mathrm{d}t\right].$$

Example: FVAE for stochastic differential equations

On $\mathcal{U} = C([0, T], \mathbb{R}^m)$:

$$\textbf{fix} \quad \Upsilon \quad \text{distribution of} \quad \mathrm{d}u_t = b(u_t)\,\mathrm{d}t + \sqrt{\varepsilon}\,\mathrm{d}w_t, \quad u_0 = 0, \quad t \in [0, T]$$

$$\textbf{choose} \quad \mathbb{P}_\eta \quad \text{distribution of} \quad \mathrm{d}\eta_t = \sqrt{\varepsilon}\,\mathrm{d}w_t, \quad \eta_0 = 0, \quad t \in [0, T].$$

Proposition: finite infimum of FVAE objective

By the Girsanov theorem, assuming b is "nice",

$$D_{\mathrm{KL}}\left(\Upsilon \,\|\, \mathbb{P}_\eta\right) = \mathop{\mathbb{E}}_{u\sim\Upsilon}\left[\frac{1}{2\varepsilon}\int_0^T \left\|b(u_t)\right\|^2 \mathrm{d}t\right].$$

Theorem: FVAE objective for stochastic differential equations

$$D_{\mathrm{KL}}\left(\mathbb{Q}^\theta_{z,u} \,\|\, \mathbb{P}^\psi_{z,u}\right) = \mathop{\mathbb{E}}_{u\sim\Upsilon}\left[\mathcal{L}(u;\theta,\psi)\right] + D_{\mathrm{KL}}(\Upsilon \,\|\, \mathbb{P}_\eta),$$

$$\mathcal{L}(u;\theta,\psi) = \mathop{\mathbb{E}}_{z\sim\mathbb{Q}^\theta_{z|u}}\left[\frac{1}{\varepsilon}\langle g(z;\psi), u\rangle^\sim_{H^1} - \frac{1}{2\varepsilon}\left\|g(z;\psi)\right\|^2_{H^1}\right] + D_{\mathrm{KL}}\left(\mathbb{Q}^\theta_{z|u} \,\|\, \mathbb{P}_z\right)$$

$\rightsquigarrow$ Similar arguments apply to other noise processes, e.g., OU noise.