Bayesian Latent variable model

- Suppose we have N data points $X_i, i = 1, ..., N$ with complex, high-dimensional structure.
- Latent variable models (LVM) aim to explain this complex structure we see in the *observed* $\{X_i\}_{i=1}^N$ by a much simpler structure using some *unobserved* latent variables $\{Z_i\}_{i=1}^N$.
- Bayesian Latent Variable Model (BLVM) assumes that observed X_is are generated by the unobserved latent Z_is via some unknown generative process.

Bayesian latent variable models

For each X_i, $i = 1, ..., N$, BLVM assumes the following

1. Draw Z_i from a (often simple-structured) **prior** distribution $p(Z)$.
2. Conditioned on Z_i, a **likelihood** $p_\theta(\cdot|Z_i)$ links the latent Z_i to the observable data X_i. Here θ represents some global, trainable parameter of the likelihood.

The joint model

Let $\mathbf{X} = \{X_i\}_{i=1}^N$, $\mathbf{Z} = \{Z_i\}_{i=1}^N$ The joint model of the BLVM can therefore be written as

$$p_\theta(\mathbf{X}, \mathbf{Z}) = \prod_{i=1}^{N} p_\theta(X_i|Z_i)p(Z_i) \tag{1}$$

Computation challenges

Fit the model $p_\theta(X_i|Z_i)$ by maximising marginal likelihood is infeasible:

- In general, the marginal likelihood of data $\mathbf{X}$,

$$p_\theta(\mathbf{X}) = \int p_\theta(\mathbf{X}, \mathbf{Z})\, d\mathbf{Z} = \prod_{i=1}^{N} \int p_\theta(X_i|Z_i) p(Z_i) dZ_i = \prod_{i=1}^{N} p_\theta(X_i), \tag{2}$$

  is computationally intractable. As a result, one cannot learn θ directly by maximising the likelihood $p_\theta(\mathbf{X})$ with respect to θ.

Perform posterior inference can be computationally challenging:

- Sampling from the posterior distribution of the latent variable

$$p_\theta(Z_i|X_i) = \frac{p_\theta(X_i|Z_i)p(Z_i)}{p_\theta(X_i)} \propto p_\theta(X_i|Z_i)p(Z_i) \tag{3}$$

  can be challenging due to the intractable normalising constant.

Variational Inference

Variational inference (VI) (Jordan et al., 1999) addresses these two challenges.

- Find probability distribution $q_{\phi_i}(Z|X_i)$, characterised by the variational parameter ϕ_i, that is **close** to the true posterior $p_\theta(Z|X_i)$ in the distribution space.
- Closeness between two distributions P, Q with probability density functions $p(x), q(x)$, respectively, can be measured via KL-divergence

$$KL(P||Q) = \int \log \frac{p(x)}{q(x)} p(x) dx. \tag{4}$$

- The distribution $q_{\phi_i^*}(Z|X_i)$, where

$$\phi_i^* = \arg\min_{\phi_i} KL(q_{\phi_i}(Z|X_i)||p_\theta(Z|X_i)), \tag{5}$$

  is a variational approximation of the true $p_\theta(Z|X_i)$.

Evidence Lower Bound (ELBO)

How does VI link posterior approximation to $p_\theta(\mathbf{X}) = \prod_{i=1}^{N} p_\theta(X_i)$?

Rearranging the VI objective

$$KL(q_{\phi_i}(Z|X_i)||p_\theta(Z|X_i)) = \mathbb{E}_{q_\phi}\left(\log q_{\phi_i}(Z|X_i) - \log p_\theta(Z|X_i)\right) \tag{6}$$

$$= \mathbb{E}_{q_\phi}\left(\log q_{\phi_i}(Z|X_i) - \log p_\theta(Z, X_i) + \log p_\theta(X_i)\right) \tag{7}$$

$$= \log p_\theta(X_i) - \underbrace{\mathbb{E}_{q_{\phi_i}}\left(\log p_\theta(Z, X_i) - \log q_{\phi_i}(Z|X_i)\right)}_{\text{A lower bound of } \log p_\theta(X_i)} \tag{8}$$

Evidence Lower Bound (ELBO)

The log likelihood (i.e. evidence) $\log p_\theta(X_i)$ of X_i is lower bounded by the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta, \phi_i; X_i) = \mathbb{E}_{q_{\phi_i}}\left(\log p_\theta(Z, X_i) - \log q_{\phi_i}(Z|X_i)\right)) \tag{9}$$

$$= \mathbb{E}_{q_{\phi_i}}\left(\log p_\theta(X_i|Z)\right) - KL\left(\log q_{\phi_i}(Z|X_i)||p(Z)\right) \tag{10}$$

Evidence Lower Bound (ELBO) and VI

Take a closer look at the identity derived in the last slide

$$\log p_\theta(X_i) = \mathcal{L}(\theta, \phi_i; X_i) + KL(q_{\phi_i}(Z|X_i)||p_\theta(Z|X_i)) \tag{11}$$

- For any fixed θ, minimising $KL(q_{\phi_i}(Z|X_i)||p_\theta(Z|X_i))$ with respect to ϕ_i (performing VI) is equivalent to maximising $\mathcal{L}(\theta, \phi_i; X_i)$ with respect to ϕ_i.

- For any fixed ϕ_i, maximising $\mathcal{L}(\theta, \phi_i; X_i)$ with respect to θ is equivalent to maximising a lower bound of $\log p_\theta(X_i)$. The gap between the two quantities disappears if and only if $q_{\phi_i}(Z|X_i) = p_\theta(Z|X_i)$.

An approximate coordinate ascent perspective

Denote $\mathcal{L}\left(\theta, \{\phi_i\}_{i=1}^N; \mathbf{X}\right) = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(\theta, \phi_i; X_i)$ the ELBO of the log likelihood (evidence) of the full dataset $\log p_\theta(\mathbf{X})$. Let's start with some initial $\theta^{(0)}, \{\phi_i^{(0)}\}_{i=1}^N$.

1. **VI**: Let $\phi_i^{(1)} = \arg\max_{\phi_i} \mathcal{L}\left(\theta^{(0)}, \{\phi_i^{(0)}\}_{i=1}^N; \mathbf{X}\right)$ for $i = 1, ..., N$.
2. **As a result of VI**: Suppose VI went well, i.e. $KL(q_{\phi_i^{(1)}}(Z|X_i)||p_{\theta^{(0)}}(Z|X_i)) \approx 0$, we will have approximately $\log p_{\theta^{(0)}}(\mathbf{X}) \approx \mathcal{L}\left(\theta^{(0)}, \{\phi_i^{(1)}\}_{i=1}^N; \mathbf{X}\right)$.
3. **(Approximate) maximum likelihood**: Let $\theta^{(1)} = \arg\max_\theta \mathcal{L}\left(\theta, \{\phi_i^{(1)}\}_{i=1}^N; \mathbf{X}\right)$.
4. **As a result of approximate ML**:
   $\log p_{\theta^{(1)}}(\mathbf{X}) \geq \mathcal{L}\left(\theta^{(1)}, \{\phi_i^{(1)}\}_{i=1}^N; \mathbf{X}\right) \geq \mathcal{L}\left(\theta^{(0)}, \{\phi_i^{(1)}\}_{i=1}^N; \mathbf{X}\right) \approx \log p_{\theta^{(0)}}(\mathbf{X})$
5. Go back to 1. Repeat.

Fitting BLVM by maximising ELBO

In summary, maximising $\mathcal{L}(\theta, \{\phi_i\}_{i=1}^N; \mathbf{X})$ with respect to θ and ϕ_is does two things simultaneously:

1. It performs variational inference by minimizing the KL divergence between the variational approximation $q_{\phi_i}(Z|X_i)$ and the true posterior, which is equivalent to tightening the bound.
2. It pushes up the value of the marginal log-likelihood $\log p_\theta(X_i)$ by maximising a lower bound of it, thereby fitting the BLVM to the data.

Variational Autoencoder (VAE)

A Variational Autoencoder (Kingma and Welling, 2013) is a deep learning model consisting of two modules

1. **Encoder network** E_ϕ: An *inference* module that takes a data point X as input and outputs a conditional distribution of the corresponding latent variable Z given data X, usually represented by the parameters for the distribution (e.g. mean and covariance matrix of a Gaussian distribution).
2. **Decoder network** D_θ: A *generative* module that takes a latent variable Z as input, and outputs a conditional distribution of data X given Z, again usually represented by the parameters of the distribution.

Variational Autoencoder as a BLVM

VAE can be interpreted as a specific, neural-network based implementation of BLVM:

- Similar to BLVM, a prior $p(Z)$ is put on the latent variable Z.
- The decoder network D_θ, taking Z as its input, plays the role of the likelihood term $p_\theta(X|Z)$ in a BLVM.

Decoder Example

For example, the decoder generative model can define a likelihood

$$p_\theta(X|Z) = \mathcal{N}(X; \mu_\theta(Z), s_\theta(Z)) \tag{12}$$

where $D_\theta(X) = \{\mu_\theta(X), \sigma_\theta(X)\}$ is the output of the decoder.

Amortised Variational Inference

- The encoder network E_ϕ ,taking X as its input, plays the role of the variational posterior $q_\phi(Z|X)$ of the true posterior $p_\theta(Z|X)$ of Z.
- In original BLVM, each $q_{\phi_i}(Z|X_i)$ has a "private" variational parameter ϕ_i.
- VAE uses a single encoder network E_ϕ, parameterised by a "global" ϕ, as an inference machine that maps any input X to the corresponding variational posterior $q_\phi(Z|X)$.

Encoder Example

For example, the VAE posterior can can define a variational posterior

$$q_\phi(Z|X) = \mathcal{N}(Z; \mu_\phi(X), s_\phi(X)) \tag{13}$$

where $E_\phi(X) = \{\mu_\phi(X), \sigma_\phi(X)\}$ is the output of the encoder.

Reparameterization trick

VAE objective

Using Armortised VI, the ELBO of a VAE can be written as

$$\mathcal{L}_{VAE}(\theta, \phi; \mathbf{X}) = \frac{1}{N} \sum_{i=1}^{N} \Big(\underbrace{\mathbb{E}_{q_\phi(Z|X_i)} \left(\log p_\theta(X_i|Z) \right)}_{\text{Reconstruction}} - \underbrace{KL \left(q_\phi(Z|X_i) || p(Z) \right)}_{\text{Regularisation}} \Big) \tag{14}$$

- Expectation is intractable. But also cannot directly do back propagation on Monte Carlo estimate of $\mathcal{L}(\theta, \phi; \mathbf{X})$ as sampling from $q_\phi(Z|X_i)$ is **non-differentiable**.
- Instead of sampling $Z \sim q_\phi(Z|X_i)$, generate Z via a differentiable transformation $Z = f(\phi, X_i, \epsilon)$ where ϵ is random noise drawn from an independent distribution.
- E.g. sampling Z from a VAE posterior $q_\phi(Z|X) = \mathcal{N}\left(Z; \mu_\phi(X), \sigma_\phi(X)\right)$ is equivalent to $Z = \mu_\phi(X) + \sigma_\phi(X) \odot \epsilon, \ \epsilon \sim \mathcal{N}(0, I_d)$

A schematic illustration

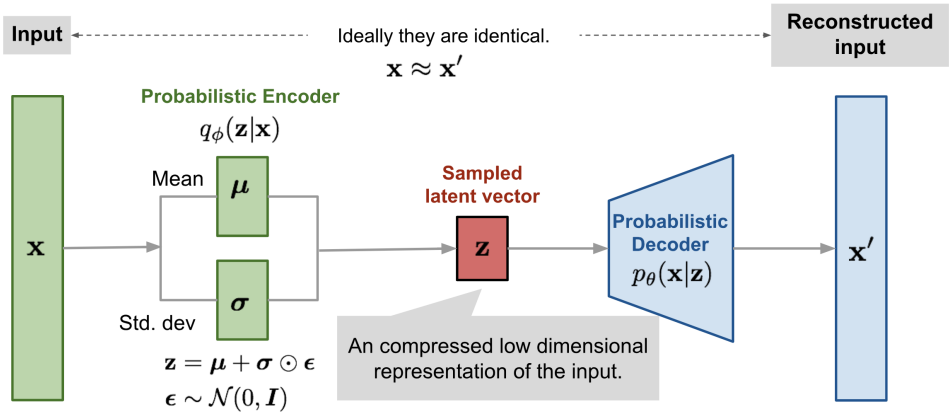

Figure: A schematic illustration of VAE (Image source: From Autoencoder to Beta-VAE, https://lilianweng.github.io/posts/2018-08-12-vae/, Lilian Weng)

β-VAE

β-VAE (Higgins et al., 2017) modifies the original VAE objective by reweighting the regularisation term

β-VAE objective

$$\mathcal{L}_{\beta-VAE}(\theta,\phi;\mathbf{X},\beta)=\frac{1}{N}\sum_{i=1}^{N}\Big(\mathbb{E}_{q_\phi(Z|X_i)}\left(\log p_\theta(X_i|Z)\right)-\beta KL\left(q_\phi(Z|X_i)||p(Z)\right)\Big), \quad (15)$$

where $\beta > 0$ is a hyperparameter.

- $\beta = 1$ recovers the VAE objective.
- When $\beta > 1$, the new objective put stronger constraint on the latent, emphasising on being close to the prior (e.g. Independent standard Gaussian).
- Useful for latent factor disentanglement.
- When $\beta \neq 1$, $\mathcal{L}_{\beta-VAE}(\theta,\phi;\mathbf{X})$ doesn't have the ELBO interpretation of a BLVM.

β-VAE

- Intuition: Increasing prior importance $\approx$ down-weighting the likelihood.
- Let us consider a "power posterior" $p_{\theta,\eta}(Z|X_i) \propto p_\theta(X_i|Z)^\eta p(Z)$, $\eta \in (0,1)$.
- Just like how VI on Bayesian posterior leads to VAE, we now show

β-VAE does VI on power posteriors

Using the same ELBO argument, minimising $KL(q_\phi(Z|X_i)||p_{\theta,\eta}(Z|X_i))$ for each X_i is equivalent to maximising a generalised ELBO

$$\mathcal{L}_\eta(\theta,\phi;\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\left(\eta\mathbb{E}_{q_\phi(Z|X_i)}\left(\log p_\theta(X_i|Z)\right) - KL\left(q_\phi(Z|X_i)||p(Z)\right)\right) \tag{16}$$

$$= \eta\mathcal{L}_{\beta-VAE}(\theta,\phi;\mathbf{X},\eta^{-1}), \tag{17}$$

a lower bound of $\log p_{\theta,\eta}(\mathbf{X}) = \sum_{i=1}^N \log p_{\theta,\eta}(X_i) = \sum_{i=1}^N \log \int p_\theta(X_i|Z)^\eta p(Z) dZ$

When we do β-VAE, we are actually doing generalised Bayes.

(Oversimplified) Generalised Bayesian inference

Think of posterior inference as a belief updating procedure:

Bayesian inference

1. Suppose $Z \sim p(Z)$, a prior representing the subjective belief before seeing any data.
2. Assuming the observed data X is generated by $X \sim p(X|Z)$.
3. Update one's belief via Bayes theorem: $Z \sim p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)} \propto p(X|Z)p(Z)$.

Do you REALLY believe the data is generated from $p(X) = \int p(X|Z)p(Z)dZ$? If not:

Generalised Bayesian inference (Bissiri et al., 2016)

1. Suppose $Z \sim p(Z)$, a prior representing the subjective belief before seeing any data.
2. Choose a loss function $l(Z, X)$ reflecting how well Z "goes" with the observed X.
3. Update one's belief ~~via Bayes theorem~~: $Z \sim p(Z|X) \propto \exp(-l(Z, X))p(Z)$.

Info-VAE

- A standard VAE is trained to optimise a trade-off between reconstruction and regularisation.
- **Posterior collapse**, which refers to a failure mode in VAE where the model's latent variables become uninformative and are ignored by the decoder during data reconstruction, can occur when e.g. regularisation is too strong ($\beta >> 1$ in β-VAE) or the decoder network becomes very powerful.
- VAE then cannot produce sensible latent representations or generate diverse samples.
- Info-VAE (Zhao et al., 2017) mitigates this problem by modifying the regularisation term.

Info-VAE

Denote $P(X)$ the true data distribution of $\mathbf{X}$.

ELBO objective as $N \to \infty$

As $N \to \infty$, the ELBO objective becomes

$$\mathcal{L}_{VAE}(\theta, \phi, P(X)) = \underbrace{\mathbb{E}_{P(X)}\mathbb{E}_{q_\phi(Z|X)}\left(\log p_\theta(X|Z)\right)}_{\text{Reconstruction}} - \underbrace{\mathbb{E}_{P(X)}KL\left(q_\phi(Z|X)||p(Z)\right)}_{\text{Regularisation}}\Big) \quad (18)$$

Let $q_\phi(Z) = \int q_\phi(Z|X)P(X)dX$. The regularisation term can be rearranged as

Decomposing the regularisation term

$$\begin{aligned}
\mathbb{E}_{P(X)}KL\left(q_\phi(Z|X)||p(Z)\right) &= \mathbb{E}_{q_\phi(Z|X)P(X)}\left(\log q_\phi(Z|X) - \log p(Z)\right) \\
&= \mathbb{E}_{q_\phi(Z|X)P(X)}\left(\log q_\phi(Z|X) - \log q_\phi(Z) + \log q_\phi(Z) - \log p(Z)\right) \\
&= \underbrace{I_q(X, Z)}_{\text{Mutual Information}} + \underbrace{KL(q_\phi(Z)||p(Z))}_{\text{Marginal regularisation}}
\end{aligned}$$

Info-VAE

- The mutual information $I_q(X, Z)$ characterise to what extent X informs Z through $q_\phi(Z|X)$. This is actually a quantity we do not always want to minimise.
- As a result, one only minimises the marginal regularization $KL(q_\phi(Z)||p(Z))$ characterizing how much the **marginal distribution** of Z generated by the decoder deviates from the prior.
- Replace the intractable $KL(q_\phi(Z)||p(Z))$ by a general sample-based statistical distance $D(q_\phi(Z)||p(Z))$, e.g. MMD distance.

Objective of Info-VAE (A special case)

$$\mathcal{L}_{InfoVAE}(\theta, \phi, P(X)) = \underbrace{\mathbb{E}_{P(X)}\mathbb{E}_{q_\phi(Z|X)} \log p_\theta(X|Z))}_{\text{Reconstruction}} + \underbrace{\lambda D(q_\phi(Z)||p(Z))}_{\text{Marginal regularisation}} , \quad (19)$$

where $\lambda > 0$ is a hyper parameter.

Diffusion model as a VAE

Can VAE map its inputs into objects other than “distributions of latent vectors"?

- Diffusion model (Ho et al., 2020) can be seen as a VAE that encodes it inputs into **stochastic processes**.
- It uses a **fixed** forward destruction process $\{q(X_t|X_{t-1})\}_{t=1}^{T}$ to corrode real data X_0 down to some unstructured noise X_T.
- Then it trains a **learnable** backward process $\{p_\theta(X_{t-1}|X_t)\}_{t=T}^{1}$ that reverts the forward process and reconstructs data X_0 from noise X_T.

Diffusion model as a VAE

Denote X_0 an observed data point.

	VAE	**Diffusion model**
Latent variable	A (low-dim) random variable Z	A Stochastic process $\{X_{1:T}\}$
Prior	Simple-structured distribution in the latent space	A learnable Markovian denoising process $p_\theta(X_T)\prod_{t=T}^{2} p_\theta(X_{t-1}\|X_t)$
Generation/Decoder	$X_0 \sim p(\cdot\|D_\theta(Z))$ where D_θ is a learnable NN	$X_0 \sim p_\theta(\cdot\|X_1)$ which can either be fixed or learnable
Inference/Encoder	A learnable Neural network E_ϕ taking X_0 as input	A pre-determined Markov chain injecting Gaussian noise into X_0
Training	Maximising ELBO	Maximising ELBO

Table: Diffusion model as a VAE

Note it is just one of many ways to interpret a Diffusion model.

Vector Quantised VAE (VQ-VAE)

- A VQ-VAE (Van Den Oord et al., 2017) embeds data into Discrete latent space instead of a continuous one in vanilla VAE.
- In addition to the encoder and decoder, VQ-VAE additionally maintains a learnable codebook, which is a collection of a fixed number of vectors $\mathcal{E} = \{e_k\}_{k=1}^{K},\ e_k \in \mathbb{R}^D$.
- Suppose we work with image data X of size $H \times W \times C$, VQ-VAE works as:
  1. **Encoder**: Deterministically mapping X to a feature $z_e(X)$ of size $H' \times W' \times D$.
  2. **Vector Quantisation**: For each $h = 1, ..., H', w = 1, ..., W'$, replace the corresponding D dimensional vector in $z_e(X)$ with its closest neighbour $e_{k'}$ from the codebook. *It therefore can be interpreted as a $H' \times W'$ matrix with entries being the corresponding discrete indices.*
  3. **Decoder**: Take the quantised feature as input, reconstruct the input image X.
- Trained by minimising reconstruction without KL-based regularisation.

VQ-VAE is more like an AE instead of a VAE

If VQ-VAE were probabilistic...

- It would treat points in the codebook as random variables and put priors on them.
- It would produce probabilistic output in the VQ step e.g. a Categorical distribution. Instead, VQ-VAE replace such probabilistic output by a deterministic `argmin`.
- It would use ELBO instead of reconstruction loss with ad-hoc regularisations as its objective.

The name can be misleading.

Linear VAE?

Try adding a (linear) armortised VI module to a Bayesian factor analysis model to make it resemble a VAE with linear components!

VAE: marriage of statistics and computation

- VAEs continued to be widely used particularly for unsupervised learning in science.
- Trendy when it first appeared – rapidly increased the base of Bayesian users.
- But people started to become accustomed to the *encoder* and *decoder* analogy from AEs – did not realise it's origins as a BLVM and amortised VI.
- One of the first examples of the modern unification of statistics and computational science.

Bibliography

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Zhao, S., Song, J., and Ermon, S. (2017). Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*.