Deep Generative Models

- Modelling **high-dimensional** data is **hard**;
- Deep generation models are **unsupervised** models that **learn** the **distribution** of the **data**;
- Supervised learning: $p(x, y)$, while **unsupervised learning**: $p(x)$;
- **Generative models** use latent (random) variables h such that: $P(x) = \sum_h P(x, h)$;
- Examples of **deep generative models**:
  - **Restricted Boltzmann machines**;
  - **Variational auto-encoders (VAE)**;
  - **Generative adversarial networks (GAN)**;
  - **Denoising diffusion models**;
  - ...

Boltzmann Machines

- **Energy-based model** over binary vectors;
  - The probability distribution via an **energy function** $E(x, h; \theta)$ is $P_\theta(x, h) = \frac{exp(-E(x,h;\theta))}{Z(\theta)}$, where $Z(\theta) = \sum_{x,h} exp(-E(x, h; \theta))$ is the **partition function - maximizing** probability corresponds to **minimizing** energy;
- v are **observed** variables, while h are **hidden/latent** variables:
- $P_\theta(v, h) = \frac{exp(-E(v,h;\theta))}{Z(\theta)}$;
- Energy function: $E(v, h; \theta) = -v^T Rv - v^T Wh - h^T Sh - v^T b - h^T c$;
- Boltzmann machine is an **universal approximator of probability mass functions over discrete variables**;
- Learning a general BM is usually very challenging, so we typically use **restricted Boltzmann machines (RBM)**.

Restricted Boltzmann Machines (RBM)

- Also called **harmoniums**;
- A layer of **observable variables** v and a layer of **hidden variables** h;
- **No intra-layer connections** - $R = 0$ and $S = 0$ - the energy function is: $E(v, h; \theta) = -v^T W h - v^T b - h^T c$;
- The **partition function** $Z(\theta)$ is still intractable, but the **conditionals** $P_\theta(h|v)$ **and** $P_\theta(v|h)$ **are tractable**;
  - Easy to compute and to sample from, because of the **conditional independence**;
  - Without intra-layer connections, $h1, ..., h_N$ are **conditionally independent** given v: $P_\theta(h|v) = \prod_{j=1}^{M} P_\theta(h_j|v)$, where $P_\theta(h_j = 1|v) = \sigma(c_j + W_j v)$;
  - This is **reciprocal** for v: $P_\theta(v|h) = \prod_{i=1}^{N} P_\theta(v_i|h)$, where $P_\theta(v_i = 1|h) = \sigma(b_i + W_i h)$;

- Several recent models are based on **differentiable generator networks**;
- This is a differentiable function $G(h; \theta)$ that maps a **latent variables** h to sample reconstructions x;
- This idea underlies **variational auto-encoders (VAE)** and **generative adversarial networks (GAN)**.

Variational Auto-Encoders (VAE)

- Many latent variable models have **intractable evidence** $P(x)$ and **intractable posterior** $P(h|x)$;
- **Variational inference** is used to approximate these quantities;
- **Auto-encoders** are **unsupervised** models that **learn** a **representation** of the **data**: $x \to h \to \hat{x}$;
- The key idea is to combine **variational inference** with **auto-encoders**;

Assumptions

- **Prior** $P_\theta(h)$ is **tractable**;
- **Conditional likelihood** $P_\theta(x|h)$ is **tractable**;
- **Evidence** $P_\theta(x)$ is **intractable**;

- **Posterior** $P_\theta(h|x)$ is **intractable**.

So we will use **variational inference** to approximate these computations.

Variational Inference and Evidence Lower Bound (ELBO)

- **Variational inference** is a **method** for **approximating intractable posterior distributions**.

  Recap - Shannon Entropy

  - **Shannon entropy** is a measure of **uncertainty** of a **random variable**. Let P be a distribution over X, the **entropy** of P is: $H(P) = -\sum_{x \in X} P(x) log P(x)$;
  - Non-negative: $H(P) \geq 0$;
  - Maximum entropy: $H(P) \leq log|X|$;

  Recap - Kullback-Leibler Divergence

  - **Kullback-Leibler divergence** is a measure of **difference** between two **probability distributions** P and Q over the same **random variable** X. It is defined as: $KL(P||Q) = \sum_{x \in X} P(x) log \frac{P(x)}{Q(x)}$;
  - Non-negative: $KL(P||Q) \geq 0$;
  - $KL(P||Q) = 0$ if and only if $P = Q$;
  - $KL(P||Q) \neq KL(Q||P)$;

Evidence Lower Bound (ELBO)

- ELBO is a central concept in **variational inference**;
- True posterior and **evidence**: $P_\theta(h|x)$ and $P_\theta(x)$;
- For any distribution $Q(h)$:

$$log P_\theta(x) \geq ELBO(Q) \tag{1}$$
$$ELBO(Q) = E_{Q(h)}[log P_\theta(x|h)] - KL(Q(h)||P_\theta(h)) \tag{2}$$

- $ELBO(Q)$ is the **evidence lower bound** of Q;

Variational Inference

- Equality achieved for $Q(h) = P_\theta(h|x)$, but this is **intractable**;
- Key idea:
  - Constraints $Q(h)$ to a **tractable family** of distributions;
    - **Mean-field approximation (MFA)**: $Q(h) = \prod_{i=1}^{M} Q(h_i)$;
  - Look for the **best** $Q(h)$ in this family that **maximizes** $ELBO(Q)$ **- minimizes** $KL(Q(h)||P_\theta(h|x))$;
- Since optimizing $Q(h)$ for every sample is **expensive**, we use **amortized variational inference**;
  - Use an **encoder** with shared parameters ϕ to define $Q_\phi(h|x)$;

Gradients and Reparametrization Trick

- **Reparametrization trick** is a **technique** to **estimate gradients** of **expectations** with respect to **random variables**;
- Let $h = g_\phi(\epsilon, x)$, where $\epsilon \sim p(\epsilon)$ and g_ϕ is a **differentiable function** with parameters ϕ;
- Then, for any **differentiable function** f:

$$E_{Q_\phi(h|x)}[f(h)] = \frac{1}{N}\sum_{n=1}^{N} f(g_\phi(\epsilon_n, x)) \tag{3}$$

Generative Adversarial Networks (GAN)

- **Generative adversarial networks (GAN)** are a **class of generative models** that use **discriminative models** to **learn** the **distribution** of the **data**;
- Key idea:
  - Keep the **generation network** $G = P_\theta(h), P_\theta(x|h)$ fixed - generate data that look like real data;
  - Drop the inference network, use a **discriminator network** $D : X \to 0, 1$ - distinguish between real and generated data;

- **Minimax game** between G and D:
  - **Generator** G tries to **minimize** $logD(x)$;
  - **Discriminator** D tries to **maximize** $logD(x)$ and **minimize** $log(1-D(G(h)))$;
- Trained using **Stochastic Gradient Descent (SGD)**.