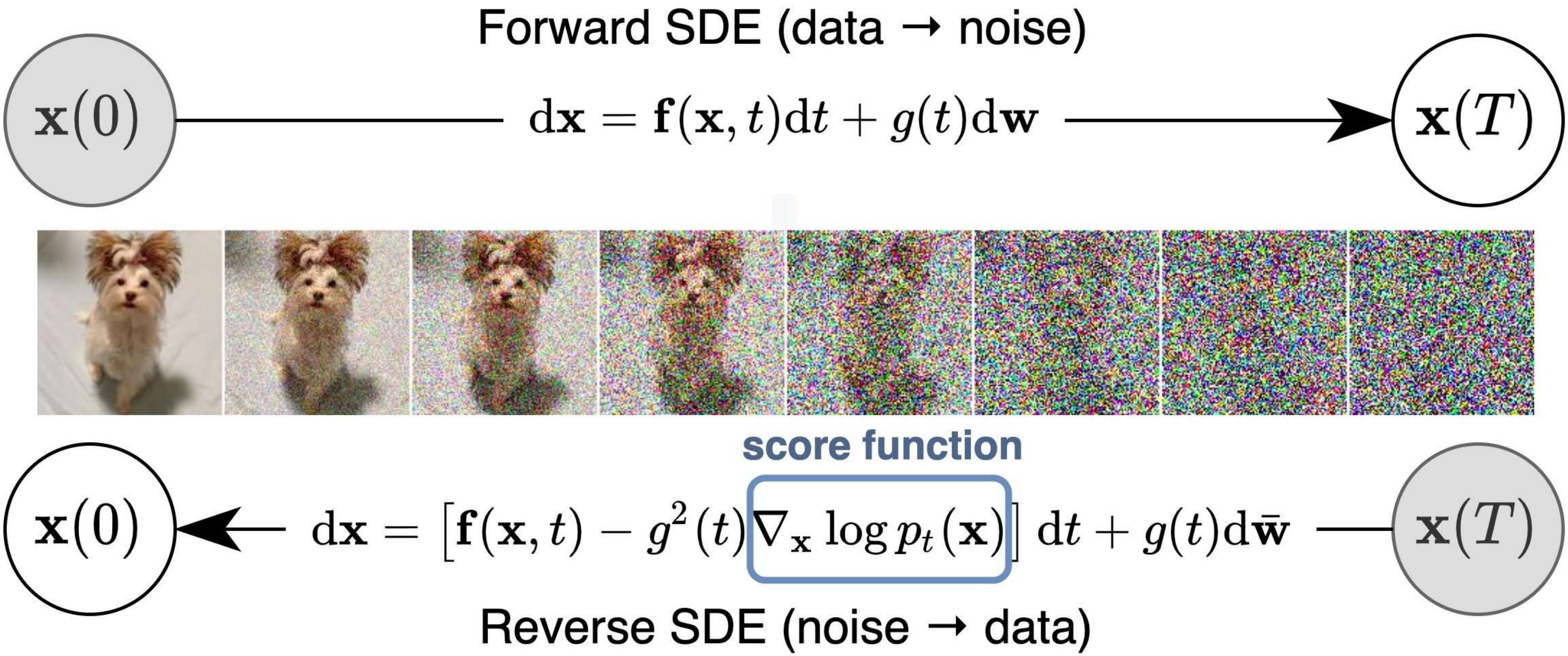
Generative diffusion models, retricted Boltzmann machines and quantum computers

Eric Paquet

NRC

5 December 2024


Forward SDE (data → noise)
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$
$\mathbf{x}(T)$
score function
$\mathbf{x}(0)$
$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}$
$\mathbf{x}(T)$
Reverse SDE (noise → data)

Probability associated with the visible units and free energy

- Energy associated with the RBM:

$$E(\mathbf{v},\mathbf{h}) = -\mathbf{v}^\top \mathbf{W}\mathbf{h} - \mathbf{b}^\top \mathbf{v} - \mathbf{c}^\top \mathbf{h}$$

- Probability associated with the visible units:

$$p(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \equiv \frac{e^{-F(\mathbf{v})}}{Z}$$

- Free energy associated with the visible units:

$$F_{\boldsymbol{\theta}}(\mathbf{v}) = -\mathbf{b}^\top \mathbf{v} - \sum_{j=1}^{H} \ln\left(1 + e^{c_j + \mathbf{W}_j^\top \mathbf{v}}\right), \quad \boldsymbol{\theta} = \{\mathbf{b}, \mathbf{c}, \mathbf{W}\}$$

Score function and mean field

- Score function (no partition function involved):

$$\boxed{\nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) = -\nabla_{\mathbf{v}} F_{\boldsymbol{\theta}}(\mathbf{v})}$$

- Mean field:

$$\boxed{v_i \to \langle v_i \rangle = p\left(v_i = 1\right) \in \left[0{,}1\right]}$$

- Activation function (invertible):

$$a_i = \mathrm{logit}\left(\langle v_i \rangle\right) \in \mathbb{R}, \quad \langle v_i \rangle = \sigma(a_i)$$

Score function associated with the model

- Free energy in terms of the activation function:

$$v_i \rightarrow \left\langle v_i \right\rangle = p\left(v_i = 1\right) = \sigma(a_i) \Rightarrow$$

$$F_{\boldsymbol{\theta}}(\mathbf{a}) = -\mathbf{b}^\top \sigma(\mathbf{a}) - \ln\left(1 + e^{c_j + \mathbf{W}_j^\top \sigma(\mathbf{a})}\right)$$

- Score function associated with the model:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{a}) = -\nabla_{\mathbf{a}} F_{\boldsymbol{\theta}}(\mathbf{a})$$

Loss function

- Score matching:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{a}_0,\mathbf{a}_t}\left[\lambda(t)\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{a}_t,t) - \nabla_{\mathbf{a}_t} \ln p(\mathbf{a}_t \mid \mathbf{a}_0)\right\|_2^2\right]$$

- Activation function associated with the noisy data (forward equation; reparametrisation trick):

$$\mathbf{a}_t \simeq \overbrace{e^{-\frac{1}{2}\int_0^t \beta(s)ds}}^{\mu(\mathbf{a}_0,t)}\mathbf{a}_0 + \sigma_t\boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(0,\mathbf{I}) \Rightarrow$$

$$\mathbf{a}_t \sim p(\mathbf{a}_t \mid \mathbf{a}_0) = \mathcal{N}\left(\mathbf{a}_t; \mu(\mathbf{a}_0,t), \sigma_t\mathbf{I}\right)$$

$$a_{0,i}^{\mathcal{B}} = \operatorname{logit}\left(\frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}} v_{0,i}^{(j)}\right)$$

Training and data generation

- Stochastic optimisation:

$$\hat{\theta} = \left\{\hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\mathbf{W}}\right\} = \underset{\mathbf{b},\mathbf{c},\mathbf{W}}{\arg\min}\, \mathcal{L}(\mathbf{b}, \mathbf{c}, \mathbf{W})$$

- Generation, Euler – Maruyama integration, time consuming:

$$\mathbf{a}_{t-1} = \mathbf{a}_t + \left[f(\mathbf{a}_t, t) - g(t)^2 \mathbf{s}_{\hat{\theta}}(\mathbf{a}_t, t) \right] \Delta t + g(t)\sqrt{\Delta t}\,\zeta, \quad \zeta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

$$\mathbf{a}_T \to \mathbf{a}_{T-1} \to \ldots \to \mathbf{a}_0$$

$$p(v_i = 1) = \langle v_i \rangle = \sigma(a_{0,i}), \quad \mathbb{R} \to]0,1[$$

The RBM as a generative diffusion model sampler I

- Typically, the score function is learned in a direct manner, which implies that the data distribution is not tractable and thus necessitates sampling through the use of the backward (denoising) equation.
- But in our case, the score function is modelled after the free energy of the visible units:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{a}) = -\nabla_{\mathbf{a}} F_{\boldsymbol{\theta}}(\mathbf{a})$$

The RBM as a generative diffusion model sampler II

- The RBM can directly sample the generative diffusion model:

$$\hat{\boldsymbol{\theta}} = \left\{\hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\mathbf{W}}\right\} \Rightarrow F_{\hat{\theta}}(\mathbf{a}) \underbrace{\Rightarrow}_{\langle \mathbf{v} \rangle = \sigma(\mathbf{a})} F_{\hat{\theta}}(\langle \mathbf{v} \rangle) \Rightarrow$$

$$p_{\hat{\theta}}\left(\langle \mathbf{v} \rangle\right) = \frac{e^{-F_{\hat{\theta}}(\langle \mathbf{v} \rangle)}}{Z} \underbrace{\Rightarrow}_{\simeq} p_{\hat{\theta}}\left(\mathbf{v}\right) \simeq \frac{e^{-F_{\hat{\theta}}(\mathbf{v})}}{Z} \Rightarrow \mathbf{v} \sim p_{\hat{\theta}}\left(\mathbf{v}\right)$$

- Either with a Gibbs sampling technique
- **Or directly with D-Wave in one step:**

$$\mathbf{v} \sim \Psi\left(\hat{\boldsymbol{\theta}}\right)$$

Conclusions

- The score function is assimilated to the gradient of the log of the free energy of the visible units
- Visible units, mean field, activation (logit)
- Score matching techniques for learning
- The RBM becomes a one-step sampler for the diffusion process
- As opposed to the reverse stochastic differential equation, which requires hundreds of steps, the generative process can be sampled directly from the RBM either with Gibbs sampling techniques or with D-Wave