Score-based generative diffusion models, restricted Boltzmann machines and the Gumbel trick (The equations festival!)

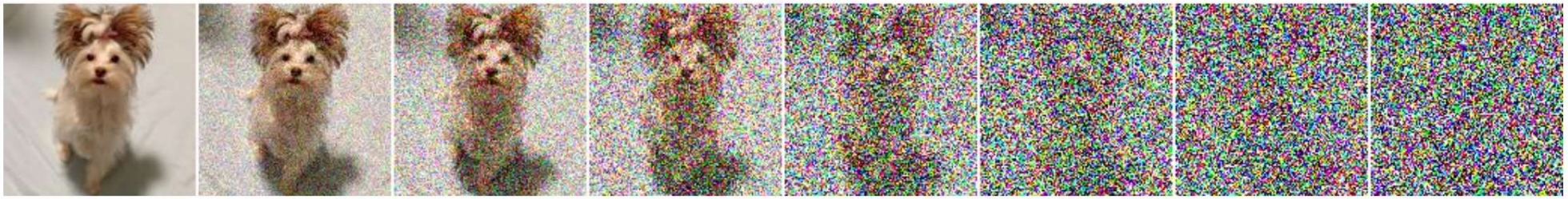
Eric Paquet

NRC

12 December 2024 and 24 January 2025

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 (evidence)

- Energy associated with the RBM:

$$E_{\boldsymbol{\theta}}(\mathbf{v}, \mathbf{h}) = -\mathbf{v}^{\top}\mathbf{W}\mathbf{h} - \mathbf{v}^{T}\mathbf{b} - \mathbf{h}^{T}\mathbf{c}$$

- Probability associated with the visible units; evidence:

$$p_{\boldsymbol{\theta}}(\mathbf{v}) = \frac{1}{Z}\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$$

$$\boldsymbol{\theta} = \left\{\mathbf{b}, \mathbf{c}, \mathbf{W}\right\}$$

Stochastic differential equations and Fokker-Planck equation

- Forward or noising equation; drift and diffusion matrices; Weiner process:

$$d\mathbf{v} = \mathbf{F}_t \mathbf{v} dt + \mathbf{G}_t d\mathbf{w}, \quad \mathbf{v} \in \mathbb{R}^D, \mathbf{F} \in \mathbb{R}^{D \times D}, \mathbf{G} \in \mathbb{R}^{D \times D}, \mathbf{w} \in \mathbb{R}^D$$

- Backward of denoising process; score function (https://arxiv.org/abs/2011.13456); multiple steps (the more, the better (bad news!):

$$d\mathbf{v} = \left[\mathbf{F}_t \mathbf{v} - \mathbf{G}_t \mathbf{G}_t^T \nabla_{\mathbf{v}} \ln p_t \left(\mathbf{v} \right) \right] dt + \mathbf{G}_t d\overline{\mathbf{w}}$$

- Corresponding Fokker-Planck equation:

$$\frac{\partial p_t \left(\mathbf{v} \right)}{\partial t} = -\nabla \cdot \left\{ \left[\mathbf{F}_t \mathbf{v} - \frac{1}{2} \mathbf{G}_t \mathbf{G}_t^T \nabla_{\mathbf{v}} p_t \left(\mathbf{v} \right) \right] p_t \left(\mathbf{v} \right) \right\}$$

The missing link: the score function

- Score function:

$$\boxed{\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{v}) = \nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v})}$$

- Gradient; let's assume a continuous relaxation:

$$\nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) = \nabla_{\mathbf{v}} \ln \left(\sum_{\mathbf{h}} e^{-E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h})} \right) - \cancel{\nabla_{\mathbf{v}} \ln Z}$$

$$\nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) = \frac{\sum_{\mathbf{h}} e^{-E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h})} \nabla_{\mathbf{v}} \left(-E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h}) \right)}{\sum_{\mathbf{h}} e^{-E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h})}}$$

From discrete to continuous variables: the Gumbel trick

- Gradient of the evidence (https://doi.org/10.3389/fphy.2021.589626); Gumbel noise:

$$\nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) = \mathbb{E}_{\mathbf{h}\sim p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{v})}\left[-\nabla_{\mathbf{v}} E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h})\right]$$

$$-\nabla_{\mathbf{v}} E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h}) = \mathbf{W}\mathbf{h} + \mathbf{b} \qquad \Rightarrow \nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) = \mathbb{E}_{\mathbf{h}\sim p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{v})}\left[\mathbf{W}\mathbf{h} + \mathbf{b}\right]$$

$$p_{\boldsymbol{\theta}}(h_j = 1 \mid \mathbf{v}) = \sigma((\mathbf{W}^T\mathbf{v} + \mathbf{c})_j) \equiv \sigma\left(\ell_{\boldsymbol{\theta},j}(\mathbf{v})\right)$$

- Continuous approximation of the hidden units, sigmoid; Gumbel trick (https://arxiv.org/abs/2410.22870); logit; annealing parameter (rounds); Gumbel distribution; stochastic:

$$\mathfrak{h}_j(\tau) = \sigma\left(\frac{\overbrace{\left(\mathbf{W}^T\mathbf{v} + \mathbf{c}\right)_j}^{\ell_{\boldsymbol{\theta},j}(\mathbf{v})} + \sigma^{-1}\left(u_j\right)}{\tau}\right) \leftarrow \mathbf{h}_j$$

$$u_j \sim \mathcal{U}(0,1), \quad \sigma^{-1}\left(u_j\right) \sim g_0 - g_1$$

$$\tau = \tau\left(\mathcal{R}\right) \Rightarrow \tau \to^{6} 0 \Leftrightarrow \mathcal{R} = \mathcal{R}_{\max}$$

In a nutshell

$$\nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) = \mathbb{E}_{\mathbf{h}\sim p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{v})}\left[-\nabla_{\mathbf{v}} E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h})\right]$$

$$\mathfrak{h}_j(\tau) = \sigma\left[\frac{\overbrace{\left(\mathbf{W}^T\mathbf{v}+\mathbf{c}\right)_j}^{\ell_{\boldsymbol{\theta},j}(\mathbf{v})} + \sigma^{-1}\left(u_j\right)}{\tau}\right] \leftarrow \mathbf{h}_j$$

- Gradient of the evidence with the Gumbel trick; stochastic; “reparametrisation trick”; the visible units may be binary (no differentiation):

$$\mathbb{E}_{\mathbf{h}\sim p_{\boldsymbol{\theta}}(\mathbf{h}|\mathbf{v})}\left[\cdot\right] \approx \mathbb{E}_{\mathbf{u}}\left[\cdot\middle|\mathfrak{h}(\tau)\right]$$

$$\boxed{\nabla_{\mathbf{v}} \ln p_{\boldsymbol{\theta}}(\mathbf{v}) \approx \mathbb{E}_{\mathbf{u}}[\mathbf{W}\mathfrak{h}(\tau)+\mathbf{b}], \quad \mathbf{u}\sim\mathcal{U}(0,1)}$$

The pros and cons

Mean field	Gumbel trick
Deterministic	Stochastic (as the visible units)
First moment (mean)	All moments
Predict visible units' probability	Predict visible units
Relatively simple	Complex
One step	Annealing

Variance preserving parametrisation

- Closed-form (and optimal) solution for the noising equation:

$\mathbf{F}_t$	$\mathbf{G}_t$
$\frac{1}{2}\frac{d \ln \beta_t}{dt}\mathbf{I}$	$\sqrt{-\frac{d \ln \beta_t}{dt}}\mathbf{I}$

- Noise scheduling: linear, sinusoidal, signal-to-noise ratio.

Loss function

- Score matching:

$$\mathcal{L}_\tau(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{v}_0,\mathbf{v}_t,\mathbf{u}}\left[\lambda(t)\left\|\mathbb{E}_{\mathbf{u}}[\mathbf{W}\mathfrak{h}(\tau)+\mathbf{b}] - \nabla_{\mathbf{v}_t}\ln p(\mathbf{v}_t \mid \mathbf{v}_0)\right\|_2^2\right]$$

- The conditioned evidence is a solution of the noising equation; reparametrisation trick (https://arxiv.org/abs/2011.13456):

$$\mathbf{u} \sim \mathcal{U}(0,1)$$

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

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

In all its glory!

$$\mathcal{L}_{\tau}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{v}_0,\mathbf{v}_t,\mathbf{u}}\left[\lambda(t)\left\|\mathbb{E}_{\mathbf{u}}[\mathbf{W}\mathfrak{h}(\tau)+\mathbf{b}] - \nabla_{\mathbf{v}_t}\ln p(\mathbf{v}_t \mid \mathbf{v}_0)\right\|_2^2\right]$$

- Score matching:

$$\mathcal{L}_{\tau\tau}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{v}_0,\mathbf{v}_t,\mathbf{u}}\left\|\mathbf{W}\left[\sigma\left(\frac{\left(\mathbf{W}^T\mathbf{v}_t+\mathbf{c}\right)_j + \sigma^{-1}\left(u_j\right)}{\tau}\right)\right] + \mathbf{b} - \nabla_{\mathbf{v}_t}\mathcal{N}\left(\mathbf{v}_t; e^{-\frac{1}{2}\int_0^t \beta(s)ds}\mathbf{v}_0, \sqrt{\int_0^t e^{-\int_s^t \beta(r)dr}\beta(s)ds}\,\mathbf{I}\right)\right\|_2^2, \quad \mathbf{u} \sim \boldsymbol{\mathcal{U}}(0,1)$$

Training and data generation

- Stochastic optimisation; annealing parameter:

$$\hat{\theta} = \left\{\hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\mathbf{W}}\right\} = \arg\min_{\mathbf{b},\mathbf{c},\mathbf{W}} \mathcal{L}_{\tau}(\mathbf{b}, \mathbf{c}, \mathbf{W})\Big|_{\tau \to 0}$$

- Generation; denoising; e.g. Euler – Maruyama integration, time-consuming:

$$\mathbf{v}_{t-1} = \mathbf{v}_t + \left[\mathbf{F}_t \mathbf{v}_t - \mathbf{G}_t \mathbf{G}_t^T \mathbf{s}_{\hat{\theta}}(\mathbf{v}_t, t)\right] \Delta t + \mathbf{G}_t \sqrt{\Delta t}\, \boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0},$$

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

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 score function of the restricted Boltzmann machine:

$$\mathbf{s}_{\hat{\boldsymbol{\theta}}}(\mathbf{v}) = \nabla_{\mathbf{v}} \ln p_{\hat{\mathbf{b}},\hat{\mathbf{c}},\hat{\mathbf{W}}}(\mathbf{v})$$

The RBM as a generative diffusion model sampler II

- The RBM can directly sample the generative diffusion model:

$$\hat{\boldsymbol{\theta}} = \left\{ \hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\mathbf{W}} \right\} \Rightarrow p_{\hat{\boldsymbol{\theta}}}\left(\mathbf{v}\right) \Rightarrow$$

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

- It may be sampled either with a Gibbs sampling technique…
- **Or directly with D-Wave in one step:**

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

Conclusions I

- Missing link: the score function: the latter is assimilated to the score function of the evidence
- Gumbel trick with temperature annealing
- 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

Conclusions II

- The generated data are binary and not real
- The restricted Boltzmann machine may be assimilated to a diffusion model, which may be employed to learn the binary latent space
- Real data could be generated with a Gaussian-Bernoulli RBM, but the latter cannot be simulated on a D-Wave quantum computer

$$E_{\boldsymbol{\theta}}(\mathbf{v},\mathbf{h}) = \sum_{i=1}^{N} \frac{\left(v_i - b_i\right)^2}{2\sigma_i^2} - \sum_{i=1}^{N}\sum_{j=1}^{H} \frac{v_i}{\sigma_i^2} w_{ij} h_j - \sum_{j=1}^{H} c_j h_j, \qquad \mathbf{v} \in \mathbb{R}^N$$