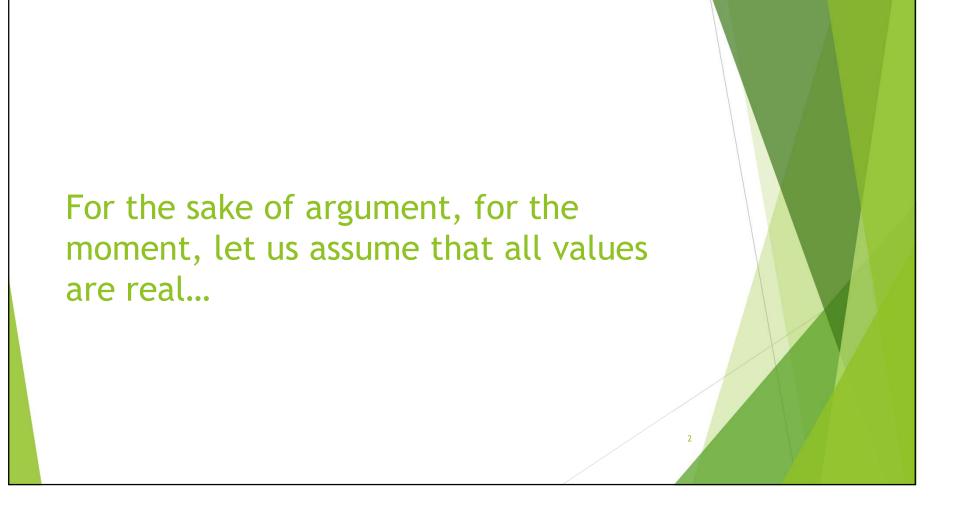


Eric Paquet NRC

28 February 2025



Sampling the prior with a Langevin equation

An energy-based distribution may be sampled with an overdamped Langevin equation:

$$\left| d\mathbf{v}_t = -\nabla_{\mathbf{v}_t} E(\mathbf{v}_t) dt + \mathfrak{D} d\mathbf{w}_t, \quad \mathbf{v}_{t \gg 0} \, \sim \, p\left(\mathbf{v}\right) \right|$$

Drift, diffusion matrices, and Wiener process:

$$d\mathbf{v}_{t} = \underbrace{\left(\mathbf{b} + \mathbf{W}\sigma(\mathbf{W}^{\top}\mathbf{v}_{t} + \mathbf{c})\right)}_{\mathbf{\mu}_{\mathbf{\theta}, t}} dt + \mathbf{\mathfrak{D}}d\mathbf{w}_{t}$$

Euler - Maruyama integration: $\mathbf{v}_{n+1} = \mathbf{v}_n - \nabla_{\mathbf{v}} E(\mathbf{v}_n) \Delta t + \sqrt{\Delta t} \mathfrak{D} \boldsymbol{\xi}_n, \quad \boldsymbol{\xi}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad n \gg 0$

Overdamped Langevin equation and Fokker-Planck equation

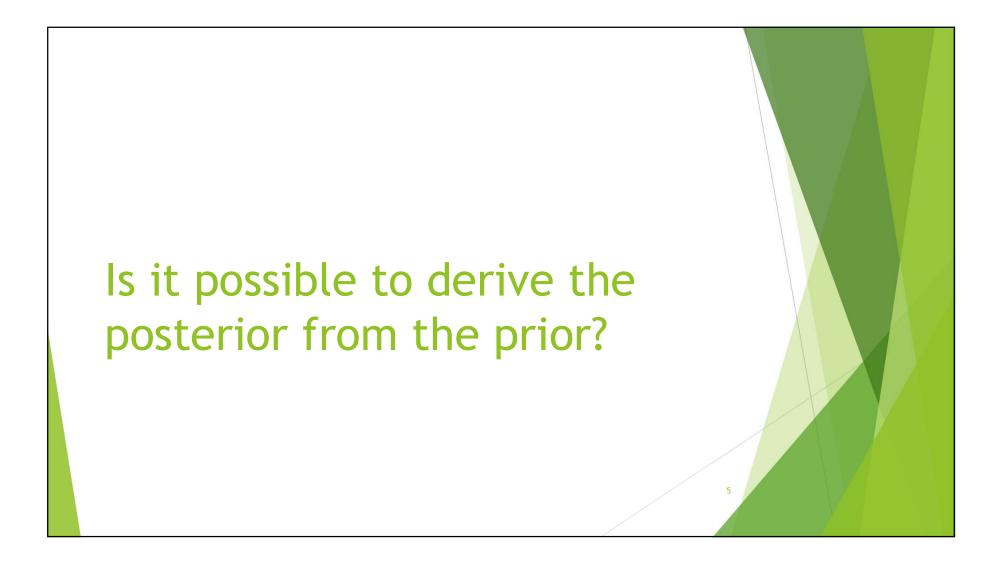
Overdamped Langevin equation:

$$d\mathbf{v}_t = -\nabla E(\mathbf{v}_t)dt + \mathbf{\mathfrak{D}}d\mathbf{w}_t$$

► Corresponding Fokker-Planck equation; diffusion process:

$$\frac{\partial p\left(\mathbf{v},t\right)}{\partial t} = \sum_{i} \frac{\partial \left[\nabla E(\mathbf{v})_{i} p\left(\mathbf{v},t\right)\right]}{\partial v_{i}} + \frac{1}{2} \sum_{i,j} \frac{\partial^{2} \left[\left(\mathfrak{D} \mathfrak{D}^{T}\right)_{i,j} p\left(\mathbf{v},t\right)\right]}{\partial v_{i} \partial v_{j}}$$

4



Change of measure and Girsanov's theorem

► Change of measure:

$$\mathbb{P} \to \mathbb{Q} \Rightarrow \mathbf{w}_t^{\mathbb{P}} \to \mathbf{w}_t^{\mathbb{Q}}$$

Adapted Wiener process; measure transformation; adjust the drift of the process, the diffusions matrix remains the same:

$$\mathbf{w}_t^{\mathbb{Q}} = \mathbf{w}_t^{\mathbb{P}} + \int_0^t \mathbf{u}_s \, ds \Rightarrow \boxed{d\mathbf{w}_t^{\mathbb{Q}} = d\mathbf{w}_t^{\mathbb{P}} + \mathbf{u}_t dt}$$

▶ Girsanov's theorem; Radon-Nikodym derivative:

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{q\left(\bullet\right)}{p\left(\bullet\right)} = \exp\left(-\int_{0}^{T} \mathbf{u}_{t}^{\dagger} d\mathbf{w}_{t}^{\mathbb{P}} - \frac{1}{2} \int_{0}^{T} \|\mathbf{u}_{t}\|_{2}^{2} dt\right)$$

6

Transformed stochastic differential equation and Kullback - Leibler divergence

- ► Girsanov's theorem ensures that the new process is Brownian and is a martingale (no bias; fair game; the conditional expectation of the next value in the sequence is equal to the present value; Itō calculus).
- ► Transformed stochastic differential equation (new drift):

$$\left| d\mathbf{v}_t^{\mathbb{Q}} \right| = \mathbf{\mu}_{\mathbf{\theta},t} dt + \mathbf{\mathfrak{D}} d\mathbf{w}_t^{\mathbb{Q}} = \left(\mathbf{\mu}_{\mathbf{\theta},t} + \mathbf{\mathfrak{D}} \mathbf{u}_t \right) dt + \mathbf{\mathfrak{D}} d\mathbf{w}_t^{\mathbb{P}}$$

▶ Non-biased KL divergence (see Radon-Nikodym derivative):

$$\left| D_{KL} \left(\left. p \left(\bullet \right) \right\| q \left(\bullet \right) \right) = \mathbb{E}_p \left[\ln \frac{p}{q} \right] = - \mathbb{E}_{\mathbb{P}} \left[\ln \frac{d\mathbb{Q}}{d\mathbb{P}} \right] = \int_0^T \frac{1}{2} \mid\mid \mathbf{u}_t \mid\mid_2^2 dt \right]$$

Variational autoencoder, prior and posterior

▶ The posterior is obtained by changing the measure of the prior:

$$\mathbf{z}_{t}^{\mathbb{P}} \sim p_{\boldsymbol{\theta}}\left(\mathbf{z}_{t}^{\mathbb{P}}\right) \Rightarrow d\mathbf{z}_{t}^{\mathbb{P}} = \underbrace{\left(\mathbf{b} + \mathbf{W}\sigma(\mathbf{W}^{\top}\mathbf{z}_{t}^{\mathbb{P}} + \mathbf{c})\right)}_{\boldsymbol{\mu}_{\boldsymbol{\theta},t}\left(\mathbf{z}_{t}^{\mathbb{P}}\right)} dt + \boldsymbol{\mathfrak{D}}d\mathbf{w}_{t}^{\mathbb{P}}$$

$$\mathbf{z}_{t}^{\mathbb{Q}} \sim q_{\boldsymbol{\phi}}\left(\mathbf{z}_{t}^{\mathbb{Q}} \,\middle|\, \mathbf{x}\right) \Rightarrow d\mathbf{z}_{t}^{\mathbb{Q}} = \left(\boldsymbol{\mu}_{\boldsymbol{\theta},t}(\mathbf{z}_{t}^{\mathbb{Q}}) + \boldsymbol{\mathfrak{D}}\mathbf{u}_{\boldsymbol{\phi},t}(\mathbf{z}_{t}^{\mathbb{Q}})\right) dt + \boldsymbol{\mathfrak{D}}d\mathbf{w}_{t}^{\mathbb{P}}$$

▶ Initial conditions; conditioning; Bernoulli distribution; real data:

$$\begin{vmatrix} p_{\mathbf{\theta}} \left(\mathbf{z}_{t}^{\mathbb{P}} \right) \Rightarrow \mathbf{z}_{0}^{\mathbb{P}} \sim \mathcal{B} \left(p \left(x = 1 \right) = 1/2 \right) \\ q_{\mathbf{\phi}} \left(\mathbf{z}_{t}^{\mathbb{Q}} \, \middle| \mathbf{x} \right) \Rightarrow \mathbf{z}_{0}^{\mathbb{Q}} = \mathbf{x} \end{vmatrix}$$

Variational autoencoder and evidence lower bound

Evidence lower bound:

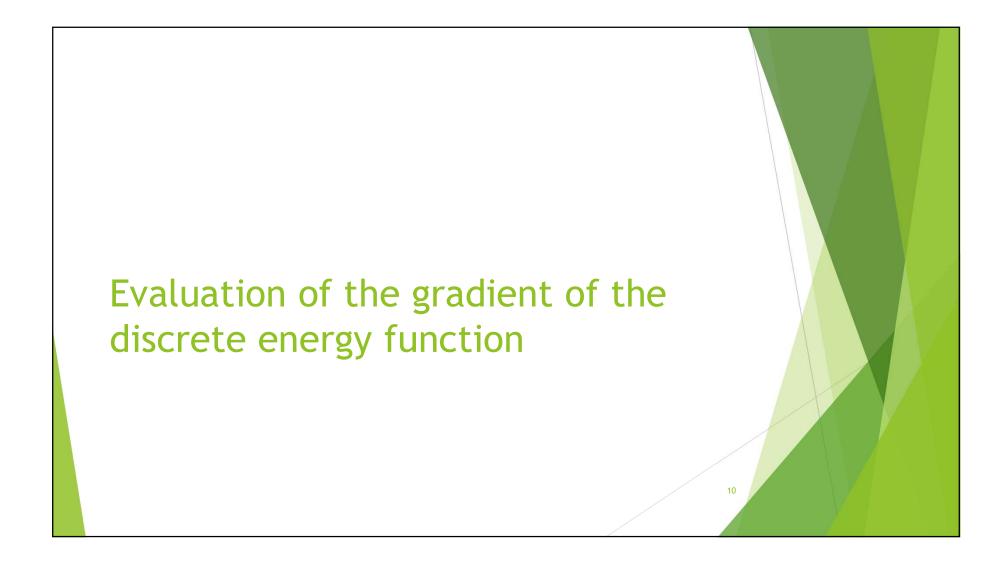
$$\mathcal{L}_{i} = \mathbb{E}_{q_{\phi}\left(\mathbf{z} \middle| \mathbf{x}_{i}\right)} \ln p_{\omega}\left(\mathbf{x}_{i} \middle| \mathbf{z}\right) - D_{KL}\left(p_{\theta}\left(\mathbf{z}\right) \middle\| q_{\phi}\left(\mathbf{z} \middle| \mathbf{x}_{i}\right)\right)$$

▶ Applying Girsanov's theorem and using the Radon - Nikodym derivative:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_t^{\mathbb{Q}}} \left[\left(\sum_{i=1}^{N} \ln p_{\boldsymbol{\omega}} \left(\mathbf{x}_i \mid \mathbf{z}_{t,i}^{\mathbb{Q}} \right) - \int_0^T \frac{1}{2} \mid\mid \mathbf{u}_{\boldsymbol{\theta},t} \left(\mathbf{z}_t^{\mathbb{Q}} \right) \mid\mid_2^2 dt \right) \right]$$

$$\mathbf{z}_t^{\mathbb{Q}} \sim d\mathbf{z}_t^{\mathbb{Q}} = \left(\left(\mathbf{b} + \mathbf{W} \sigma (\mathbf{W}^{\top} \mathbf{z}_t^{\mathbb{Q}} + \mathbf{c}) \right) + \mathfrak{D} \mathbf{u}_{\phi, t} (\mathbf{z}_t^{\mathbb{Q}}) \right) dt + \mathfrak{D} \frac{d\mathbf{w}_t^{\mathbb{P}}}{t}, \quad t \gg 0$$

$$\left|\mathbf{z}_{n+1}^{\mathbb{Q}} = \mathbf{z}_{n}^{\mathbb{Q}} + \left(\mathbf{\mu}_{\mathbf{\theta}}(\mathbf{z}_{n}^{\mathbb{Q}}) + \mathbf{\mathfrak{D}}\mathbf{u}_{\mathbf{\phi}}(\mathbf{z}_{n}^{\mathbb{Q}})\right)\Delta t + \mathbf{\mathfrak{D}}\sqrt{\Delta t}\,\boldsymbol{\xi}_{n}, \quad \boldsymbol{\xi}_{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad n \gg 1$$



Binary and continuous energy functions

Binary energy function (RBM):

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

► Continuous relaxation with Gumbel noise; learnable logit variables; temperature annealing:

$$\tilde{v}_{j} = \sigma \left(\frac{\ell_{v_{i}} + g_{v_{i}}}{\tau} \right), \quad \tilde{h}_{j} = \sigma \left(\frac{\ell_{h_{i}} + g_{h_{i}}}{\tau} \right), \quad g_{v_{i}}, g_{h_{i}} \sim \mathcal{G} \left(0, 1 \right)$$

▶ Approximate energy in the continuous space:

$$\tilde{E}(\ell_{\mathbf{v}}, \mathbf{g}_{\mathbf{v}}, \ell_{\mathbf{h}}, \mathbf{g}_{\mathbf{h}}) = \tilde{E}(\tilde{\mathbf{h}}, \tilde{\mathbf{v}}) = -\mathbf{b}^T \tilde{\mathbf{v}} - \mathbf{c}^T \tilde{\mathbf{h}} - \tilde{\mathbf{v}}^T \mathbf{W} \tilde{\mathbf{h}}$$

Metropolis-adjusted Langevin algorithm with constant Gumbel noise

► Energy gradient with respect to the logit variables:

$$\nabla_{\boldsymbol{\ell}_{\mathbf{v}}} \tilde{E}_{\tau} \left(\tilde{\mathbf{h}}, \tilde{\mathbf{v}} \right) = \sum_{j} \frac{\partial \tilde{E}_{\tau} \left(\tilde{\mathbf{h}}, \tilde{\mathbf{v}} \right)}{\partial \tilde{v}_{j}} \frac{\partial \tilde{v}_{j}}{\partial \ell_{v_{j}}}, \dots.$$

► Metropolis-adjusted Langevin algorithm (MALA):

$$\boldsymbol{\ell}^* = \boldsymbol{\ell}^{(t)} - \frac{\alpha}{2} \nabla_{\boldsymbol{\ell}} \tilde{E}_{\tau} \left(\boldsymbol{\ell}^{(t)} \right) + \sqrt{\alpha} \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \mathcal{N} \left(\boldsymbol{0}, \boldsymbol{\mathbf{I}} \right)$$

► Metropolis-Hasting acceptance; Gaussian proposal kernel:

$$A = \min \left[1, \frac{\exp\left[-\tilde{E}_{\tau}\left(\boldsymbol{\ell}^{*}\right)\right] q\left(\boldsymbol{\ell}^{(t)} \middle| \boldsymbol{\ell}^{*}\right)}{\exp\left[-\tilde{E}_{\tau}\left(\boldsymbol{\ell}^{(t)}\right)\right] q\left(\boldsymbol{\ell}^{*} \middle| \boldsymbol{\ell}^{(t)}\right)} \right] \quad \Rightarrow \quad u \sim \mathcal{U}\left(0, 1\right) \Rightarrow \begin{array}{c} u \leq A \Rightarrow \boldsymbol{\ell}^{(t+1)} = \boldsymbol{\ell}^{*} \\ u > A \Rightarrow \boldsymbol{\ell}^{(t+1)} = \boldsymbol{\ell}^{(t)} \end{array}$$

Extended-state Markov chain Monte Carlo algorithm

► Target log-density; logits and Gumbel noise:

$$\boldsymbol{\Phi} \triangleq \left[\left. \boldsymbol{\ell} \right\| \mathbf{g} \right] \Rightarrow \mathcal{E}_{\tau} \left(\boldsymbol{\Phi} \right) = -\tilde{E}_{\tau} \left(\boldsymbol{\Phi} \right) + \sum_{j} \ln \mathcal{G} \left(\boldsymbol{g}_{j} \right), \quad \ln \mathcal{G} \left(\boldsymbol{g}_{j} \right) = -\boldsymbol{g}_{j} - \exp \left(-\boldsymbol{g}_{j} \right)$$

► Metropolis-adjusted Langevin (MALA) update; gradient on both logit variables and Gumbel noise:

$$\mathbf{\Phi}^* = \mathbf{\Phi}^{(t)} + \frac{\alpha}{2} \nabla_{\mathbf{\Phi}} \mathcal{E}_{\tau} \left(\mathbf{\Phi}^{(t)} \right) + \sqrt{\alpha} \mathbf{\varepsilon}_t, \quad \mathbf{\varepsilon}_t \sim \mathcal{N} \left(\mathbf{0}, \mathbf{I} \right)$$

▶ Then proceed as in the previous slide.

Block Gibbs sampling

▶ Free energy; marginalised energy over the hidden variables:

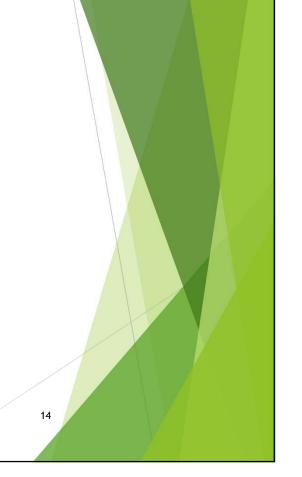
$$F(\mathbf{v}) \equiv E(\mathbf{v}) = -\ln \sum_{\mathbf{h}} \exp[-E(\mathbf{v}, \mathbf{h})]$$

Gradient of the free energy:

$$\nabla_{\mathbf{v}} F(\mathbf{v}) = \mathbb{E}_{\mathbf{h} \sim p(\mathbf{h}|\mathbf{v})} [\mathbf{b} + \mathbf{W}\mathbf{h}]$$

Monte Carlo estimator, Bernoulli distribution:

$$egin{aligned}
abla_{\mathbf{v}} F\left(\mathbf{v}
ight) &pprox rac{1}{K} \sum_{k=1}^{k} \mathbf{b} + \mathbf{W} \mathbf{h}_{\left(k
ight)}^{*}, \quad \mathbf{h}_{\left(k
ight)}^{*} &\sim p\left(\mathbf{h} \,\middle|\, \mathbf{v}
ight) \ p\left(\mathbf{h} \,\middle|\, \mathbf{v}
ight) &= \prod_{i} \mathcal{B} \Big(\sigma\Big(\mathbf{v}^{T} \mathbf{w}_{i} + c_{i}\Big)\Big) \end{aligned}$$



But...

- ▶ The values generated are not binary and, therefore, *cannot* (?) be generated by the quantum computer.
- As a result, a different distribution is sampled.

Conclusions

- ► The prior of the variational autoencoder can be formulated as an overdamped Langevin process.
- By applying Girsanov's theorem, a Brownian and unbiased posterior (martingale) can be derived from the prior by a change of measure (new drift).
- ► The posterior is also a damped Langevin process.
- ► The Fokker-Planck equation governs the time-evolving probability density of this process (diffusion process).
- ▶ Data generation is a one-step process (because of the decoder), but training is not (Langevin equation).
- ▶ It is a variational autoencoder that retains the benefits of a diffusion model in training (prior and posterior), while keeping data generation in a single step.

Bibliography

- Christopher M Bishop and Nasser M Nasrabadi. Pattern recognition and machine learning, volume 4. Springer, 2006.
- Manfred Opper. Variational inference for stochastic differential equations. Annalen der Physik, 531(3):1800233, 2019.
- ➤ Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David K Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In Symposium on Advances in Approximate Bayesian Inference, pp. 1-28. PMLR, 2020.
- Tom Ryder, Andrew Golightly, A Stephen McGough, and Dennis Prangle. Black-box variational inference for stochastic differential equations. In International Conference on Machine Learning, pp. 4423-4432. PMLR, 2018.