Forward diffusion process - improvements

Since $q(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})$ is Gaussian, we can directly compute $q(\mathbf{x}^{(t)} \mid \mathbf{x}^{(0)})$ in closed-form!

$$\mathbf{x}^{(t)} = \prod_{i=1}^{t} \sqrt{1-\beta_i}\mathbf{x}^{(0)} + \sqrt{1 - \prod_{i=1}^{t}(1-\beta_i)}\epsilon$$

Let us define $\alpha_t = 1 - \beta_t$, and $\bar{a}_t = \prod_{i=1}^{t} \alpha_i$.

$$\mathbf{x}^{(t)} = \sqrt{\bar{\alpha}_t}\mathbf{x}^{(0)} + \sqrt{1-\bar{\alpha}_t}\epsilon; \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Reverse diffusion process

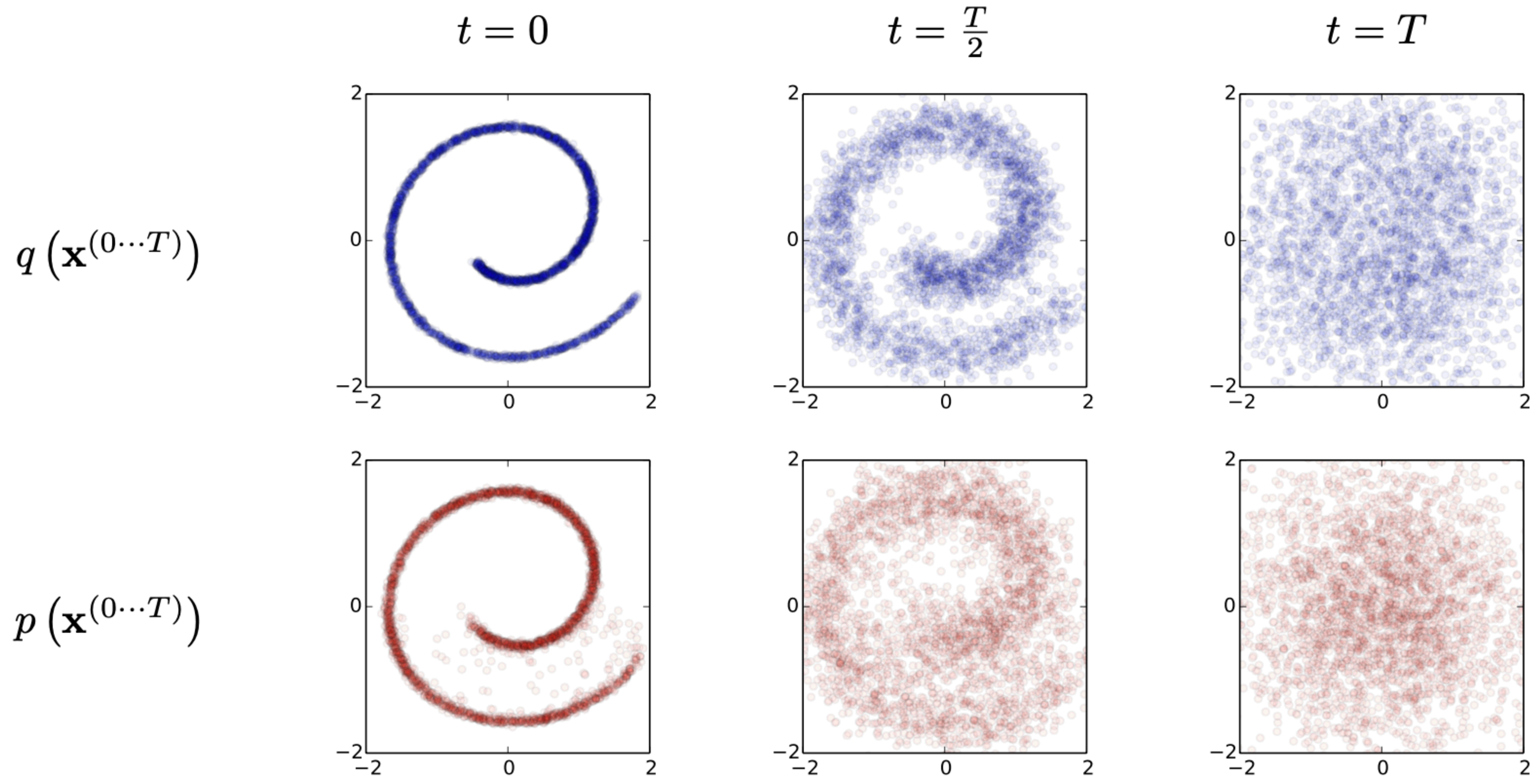
2.2. Reverse Trajectory

The generative distribution will be trained to describe the same trajectory, but in reverse,

$$p\left(\mathbf{x}^{(T)}\right) = \pi\left(\mathbf{x}^{(T)}\right) \tag{4}$$

$$p\left(\mathbf{x}^{(0\cdots T)}\right) = p\left(\mathbf{x}^{(T)}\right)\prod_{t=1}^{T} p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right). \tag{5}$$

For both Gaussian and binomial diffusion, for continuous diffusion (limit of small step size β) the reversal of the diffusion process has the identical functional form as the forward process (Feller, 1949). Since $q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)$ is a Gaussian (binomial) distribution, and if β_t is small, then $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)$ will also be a Gaussian (binomial) distribution. The longer the trajectory the smaller the diffusion rate β can be made.



$$q(\mathbf{x}^{(t)} \,|\, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)}) = q(\mathbf{x}^{(t)} \,|\, \mathbf{x}^{(t-1)})$$

$$q(\mathbf{x}^{(t-1)} \,|\, \mathbf{x}^{(t)}, \mathbf{x}^{(t+1)}) \neq q(\mathbf{x}^{(t-1)} \,|\, \mathbf{x}^{(t)})$$

$$p(\mathbf{x}^{(t-1)} \,|\, \mathbf{x}^{(t)}, \mathbf{x}^{(t+1)}) = p(\mathbf{x}^{(t-1)} \,|\, \mathbf{x}^{(t)})$$

$$p(\mathbf{x}^{(t)} \,|\, \mathbf{x}^{(t-1)}, \mathbf{x}^{(t-2)}) \neq p(\mathbf{x}^{(t)} \,|\, \mathbf{x}^{(t-1)})$$

Reverse diffusion process

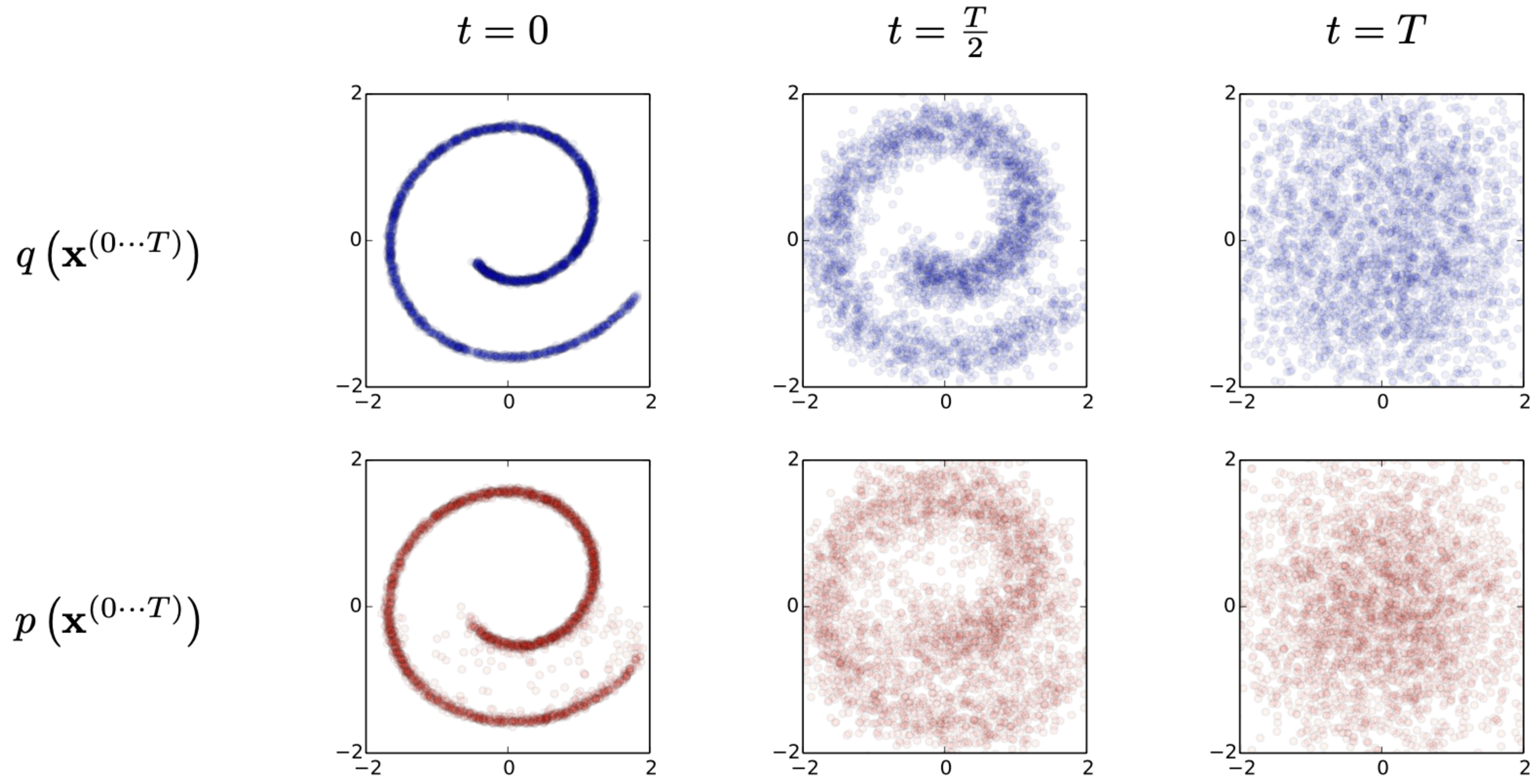
2.2. Reverse Trajectory

The generative distribution will be trained to describe the same trajectory, but in reverse,

$$p\left(\mathbf{x}^{(T)}\right) = \pi\left(\mathbf{x}^{(T)}\right) \tag{4}$$

$$p\left(\mathbf{x}^{(0\cdots T)}\right) = p\left(\mathbf{x}^{(T)}\right)\prod_{t=1}^{T} p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right). \tag{5}$$

For both Gaussian and binomial diffusion, for continuous diffusion (limit of small step size β) the reversal of the diffusion process has the identical functional form as the forward process (Feller, 1949). Since $q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)$ is a Gaussian (binomial) distribution, and if β_t is small, then $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)$ will also be a Gaussian (binomial) distribution. The longer the trajectory the smaller the diffusion rate β can be made.



1. $\mathbf{x}^{(T)} \sim \pi(\mathbf{x}^{(T)})$
2. $\mathbf{x}^{(T-1)} \sim p(\mathbf{x}^{(T-1)}\,|\,\mathbf{x}^{(T)})$
3. $\mathbf{x}^{(T-2)} \sim p(\mathbf{x}^{(T-2)}\,|\,\mathbf{x}^{(T-1)})$
4. $\ldots$
5. $\mathbf{x}^{(0)} \sim p(\mathbf{x}^{(0)}\,|\,\mathbf{x}^{(1)})$

Reverse diffusion process

2.2. Reverse Trajectory

The generative distribution will be trained to describe the same trajectory, but in reverse,

$$p\left(\mathbf{x}^{(T)}\right) = \pi\left(\mathbf{x}^{(T)}\right) \tag{4}$$

$$p\left(\mathbf{x}^{(0\cdots T)}\right) = p\left(\mathbf{x}^{(T)}\right)\prod_{t=1}^{T} p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right). \tag{5}$$

For both Gaussian and binomial diffusion, for continuous diffusion (limit of small step size β) the reversal of the diffusion process has the identical functional form as the forward process (Feller, 1949). Since $q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)$ is a Gaussian (binomial) distribution, and if β_t is small, then $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)$ will also be a Gaussian (binomial) distribution. The longer the trajectory the smaller the diffusion rate β can be made.

During learning only the mean and covariance for a Gaussian diffusion kernel, or the bit flip probability for a binomial kernel, need be estimated. As shown in Table App.1, $\mathbf{f}_\mu\left(\mathbf{x}^{(t)}, t\right)$ and $\mathbf{f}_\Sigma\left(\mathbf{x}^{(t)}, t\right)$ are functions defining the mean and covariance of the reverse Markov transitions for a Gaussian, and $\mathbf{f}_b\left(\mathbf{x}^{(t)}, t\right)$ is a function providing the bit flip probability for a binomial distribution. The computational cost of running this algorithm is the cost of these functions, times the number of time-steps. For all results in this paper, multi-layer perceptrons are used to define these functions. A wide range of regression or function fitting techniques would be applicable however, including nonparameteric methods.

1. $\mathbf{x}^{(T)} \sim \pi(\mathbf{x}^{(T)})$
2. $\mathbf{x}^{(T-1)} \sim p(\mathbf{x}^{(T-1)} \,|\, \mathbf{x}^{(T)})$
3. $\mathbf{x}^{(T-2)} \sim p(\mathbf{x}^{(T-2)} \,|\, \mathbf{x}^{(T-1)})$
4. $\ldots$
5. $\mathbf{x}^{(0)} \sim p(\mathbf{x}^{(0)} \,|\, \mathbf{x}^{(1)})$

		Gaussian
Well behaved (analytically tractable) distribution	$\pi\left(\mathbf{x}^{(T)}\right) =$	$\mathcal{N}\left(\mathbf{x}^{(T)}; \mathbf{0}, \mathbf{I}\right)$
Forward diffusion kernel	$q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right) =$	$\mathcal{N}\left(\mathbf{x}^{(t)}; \mathbf{x}^{(t-1)}\sqrt{1-\beta_t}, \mathbf{I}\beta_t\right)$
Reverse diffusion kernel	$p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right) =$	$\mathcal{N}\left(\mathbf{x}^{(t-1)}; \mathbf{f}_\mu\left(\mathbf{x}^{(t)}, t\right), \mathbf{f}_\Sigma\left(\mathbf{x}^{(t)}, t\right)\right)$

Training

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{x}^{(0)}\sim\mathscr{D}_{train}} \log p(\mathbf{x}^{(0)}) \\
&= \mathbb{E}_{\mathbf{x}^{(0)}\sim q(\mathbf{x}^{(0)})} \log p(\mathbf{x}^{(0)}) \\
&= \int q(\mathbf{x}^{(0)}) \log p(\mathbf{x}^{(0)}) d\mathbf{x}^{(0)}
\end{aligned}
$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$
L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}
$$

Training

2.3. Model Probability

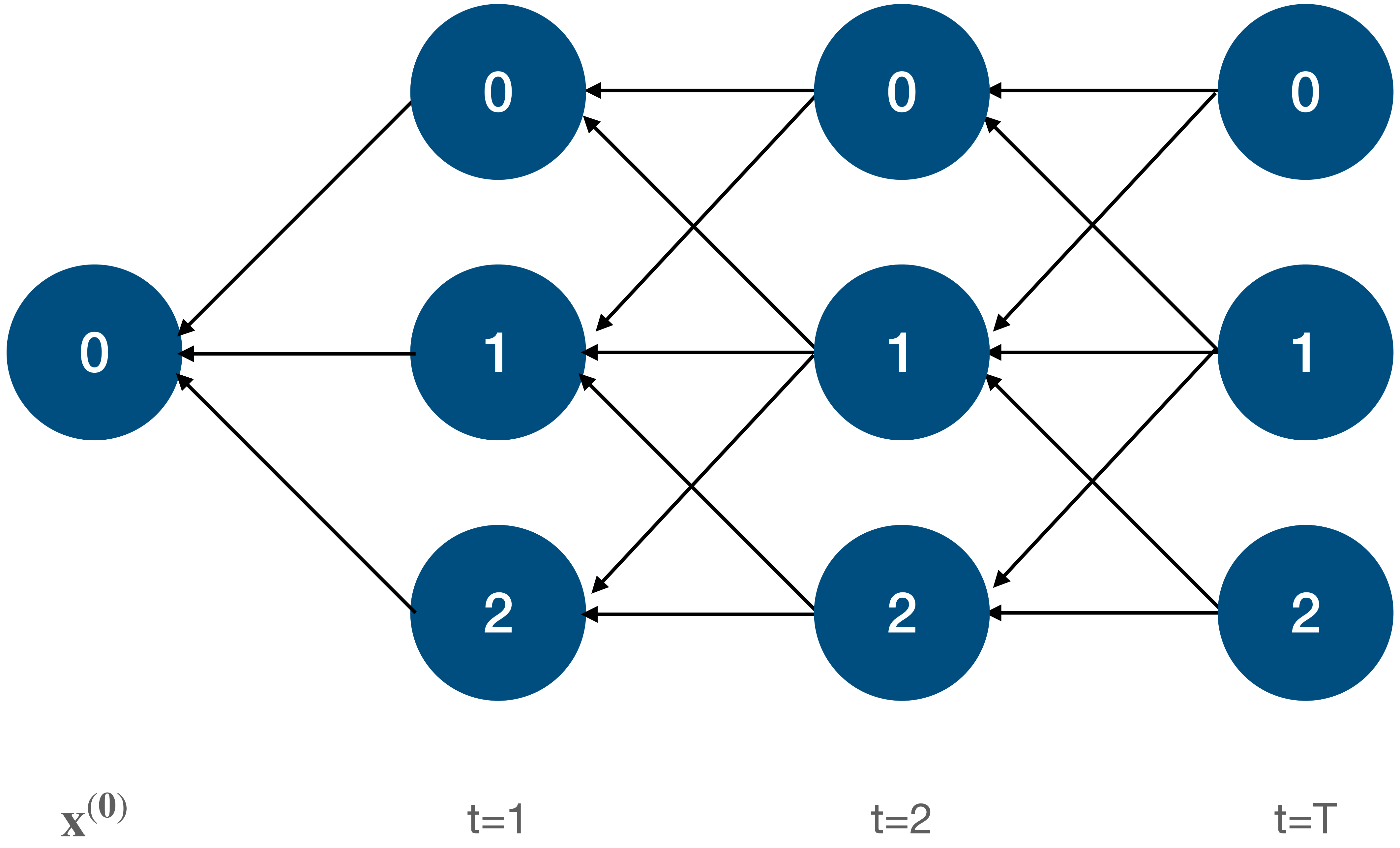
The probability the generative model assigns to the data is

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right). \tag{6}$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$



Training

2.3. Model Probability

The probability the generative model assigns to the data is

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right). \tag{6}$$

Naively this integral is intractable – but taking a cue from annealed importance sampling and the Jarzynski equality, we instead evaluate the relative probability of the forward and reverse trajectories, averaged over forward trajectories,

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right) \frac{q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)} \tag{7}$$

$$= \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \frac{p\left(\mathbf{x}^{(0\cdots T)}\right)}{q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)} \tag{8}$$

$$= \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \cdot p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}. \tag{9}$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

Training

2.3. Model Probability

The probability the generative model assigns to the data is

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right). \tag{6}$$

Naively this integral is intractable – but taking a cue from annealed importance sampling and the Jarzynski equality, we instead evaluate the relative probability of the forward and reverse trajectories, averaged over forward trajectories,

$$p\left(\mathbf{x}^{(0)}\right) = \int d\mathbf{x}^{(1\cdots T)} p\left(\mathbf{x}^{(0\cdots T)}\right) \frac{q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)} \tag{7}$$

$$= \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \frac{p\left(\mathbf{x}^{(0\cdots T)}\right)}{q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)} \tag{8}$$

$$= \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \cdot p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}. \tag{9}$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

$$= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \log \left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)} \end{array} \right], \tag{11}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \log \left[p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)} \right]. \tag{12}$$

$$\log(\mathbb{E}[\mathbb{X}]) \leq (\mathbb{E}[\log \mathbb{X}])$$

Training

$$L \geq \mathbb{E}_{\mathbf{x}^{(0..T)} \sim q(\mathbf{x}^{(0..T)})} \log[p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})}]$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

$$\begin{aligned} &= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \\ &\log \left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)} \end{array} \right], \end{aligned} \tag{11}$$

which has a lower bound provided by Jensen's inequality,

$$\begin{aligned} L \geq &\int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \\ &\log \left[p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)} \right]. \end{aligned} \tag{12}$$

Training

$$L \geq \mathbb{E}_{\mathbf{x}^{(0..T)} \sim q(\mathbf{x}^{(0..T)})} \log[p(\mathbf{x}^{(T)}) \prod_{t=1}^{T} \frac{p(\mathbf{x}^{(t-1)} \mid \mathbf{x}^{(t)})}{q(\mathbf{x}^{(t)} \mid \mathbf{x}^{(t-1)})}]$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

$$= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \log \left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)} | \mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}\right)} \end{array} \right], \tag{11}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \log \left[p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}\right)} \right]. \tag{12}$$

As described in Appendix B, for our diffusion trajectories this reduces to,

$$L \geq K \tag{13}$$

$$K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}\right) \cdot D_{KL}\left(q\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right) \middle|\middle| p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)\right) + H_q\left(\mathbf{X}^{(T)} | \mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)} | \mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right). \tag{14}$$

Training

$$H(x) = \frac{1}{2} + \frac{1}{2}\log(2\pi\sigma^2)$$

$$H(\mathbf{x}^{(T)}) = \frac{1}{2} + \frac{1}{2}\log(2\pi)$$

$$H(\mathbf{x}^{(1)} \,|\, \mathbf{x}^{(0)}) = \frac{1}{2} + \frac{1}{2}\log(2\pi\beta_0)$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

$$= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \log\left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)} \end{array}\right], \tag{11}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \log\left[p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right]. \tag{12}$$

As described in Appendix B, for our diffusion trajectories this reduces to,

$$L \geq K \tag{13}$$

$$K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}\right) \cdot D_{KL}\left(q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right) \middle|\middle| p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)\right) \cancel{+ H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right)}. \tag{14}$$

Training

$$K = -\sum_{t=2}^{T} \mathbb{E}_{\mathbf{x}^{(0)} \sim \mathscr{D}_{train} \mathbf{x}^{(t)} \sim q(\mathbf{x}^{(t)} | \mathbf{x}^{(0)})} D_{KL}(q(\mathbf{x}^{(t-1)} \,|\, \mathbf{x}^{(t)}, \mathbf{x}^{(0)}) \,||\, p(\mathbf{x}^{(t-1)} \,|\, \mathbf{x}^{(t)}))$$

$$D_{KL}(p \,||\, q) = \log \frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$$

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

$$= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \log \left[\begin{array}{c} \int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)} | \mathbf{x}^{(0)}\right) \cdot \\ p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}\right)} \end{array} \right], \tag{11}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \log \left[p\left(\mathbf{x}^{(T)}\right) \prod_{t=1}^{T} \frac{p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)}\right)} \right]. \tag{12}$$

As described in Appendix B, for our diffusion trajectories this reduces to,

$$L \geq K \tag{13}$$

$$K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)}, \mathbf{x}^{(t)}\right) \cdot D_{KL}\left(q\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right) \middle|\middle| p\left(\mathbf{x}^{(t-1)} | \mathbf{x}^{(t)}\right)\right) \cancel{+ H_q\left(\mathbf{X}^{(T)} | \mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)} | \mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right)}. \tag{14}$$

Training

$$K = -\sum_{t=2}^{T} \mathbb{E}_{\mathbf{x}^{(0)}\sim\mathscr{D}_{train}\mathbf{x}^{(t)}\sim q(\mathbf{x}^{(t)}|\mathbf{x}^{(0)})} D_{KL}(q(\mathbf{x}^{(t-1)}\,|\,\mathbf{x}^{(t)},\mathbf{x}^{(0)})\,||\,p(\mathbf{x}^{(t-1)}\,|\,\mathbf{x}^{(t)}))$$

$$D_{KL}(p\,||\,q) = \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\sigma_q^2} - \frac{1}{2}$$

We don't know the mean μ_q and standard deviation σ_q for $(q(\mathbf{x}^{(t-1)}\,|\,\mathbf{x}^{(t)},\mathbf{x}^{(0)})$!

2.4. Training

Training amounts to maximizing the model log likelihood,

$$L = \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \log p\left(\mathbf{x}^{(0)}\right) \tag{10}$$

$$= \int d\mathbf{x}^{(0)} q\left(\mathbf{x}^{(0)}\right) \cdot \log\left[\begin{array}{c}\int d\mathbf{x}^{(1\cdots T)} q\left(\mathbf{x}^{(1\cdots T)}|\mathbf{x}^{(0)}\right)\cdot \\ p\left(\mathbf{x}^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\end{array}\right], \tag{11}$$

which has a lower bound provided by Jensen's inequality,

$$L \geq \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \cdot \log\left[p\left(\mathbf{x}^{(T)}\right)\prod_{t=1}^{T}\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right]. \tag{12}$$

As described in Appendix B, for our diffusion trajectories this reduces to,

$$L \geq K \tag{13}$$

$$K = -\sum_{t=2}^{T}\int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)},\mathbf{x}^{(t)}\right)\cdot D_{KL}\left(q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)\middle|\middle|p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)\right) \cancel{+ H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right).} \tag{14}$$

Training

$$p(x\,|\,y) = \frac{p(y\,|\,x)p(x)}{p(y)}$$

We don’t know the mean μ_q and standard deviation σ_q for $q(\mathbf{x}^{(t-1)}\,|\,\mathbf{x}^{(t)}, \mathbf{x}^{(0)})$!

$$\frac{1}{\sigma\sqrt{(2\pi)}}\exp^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

$$q(\mathbf{x}^{(t-1)}\,|\,\mathbf{x}^{(t)}, \mathbf{x}^{(0)}) = \frac{q(\mathbf{x}^{(t)}\,|\,\mathbf{x}^{(t-1)}, \mathbf{x}^{(0)})q(\mathbf{x}^{(t-1)}\,|\,\mathbf{x}^{(0)})}{q(\mathbf{x}^{(t)}\,|\,\mathbf{x}^{(0)})}$$

$$= \frac{1}{\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}\sqrt{(2\pi)}}\exp^{[-\frac{1}{2}(\frac{\mathbf{x}^{(t-1)} - (\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}^{(0)} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}^{(t)})}{\sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t}})^2]}$$