Denoising diffusion probabilistic models and normalizing flows

June 18, 2025

Reminder

1. **Solve problems**: In total, $8 - 10$ questions shall be solved.
2. **one course project**: Select 1 topic to conduct numerical experiment, and write a short but complete report.
    - length: no requirement. 4-6 pages recommended.
    - problem description and goal
    - method and algorithm
    - details about numerical experiment
3. give a 5-10 min presentation.

Remark:

1. Two persons can work together on the course report. In this case, the project should be more comprehensive. Each should contribute equally to the experiment and to the writing.
2. General code of conduct for scientific writing applies (e.g. use of materials of this course, online materials, and ChatGPT)!

Part 1: Recall the previous lecture

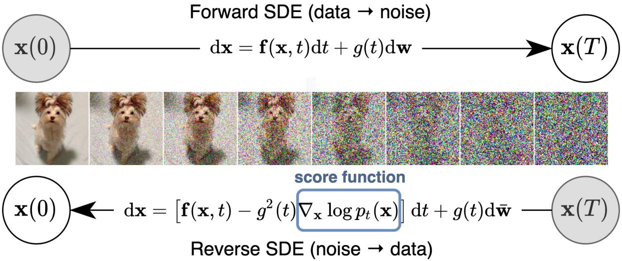
Generative modeling

Different approaches:

1. Variational AutoEncoders (VAEs)
2. Generative Adversarial Networks (GANs)
3. Normalizing Flows (NFs)
4. Diffusion generative models
5. Flow-matching generative models

Score-based diffusion models

1. choose a forward process X_t
   linear SDE $\Longrightarrow p(\cdot, T)$ is (approximately) Gaussian p_{prior}.
2. learn SDE of the backward process Y_t
3. generate new data: sample $Y_0 \sim p_{\text{prior}}$ and simulate the backward process to get Y_T



Flow-mathcing generative models

1. Target density $p_1 = p_{\text{target}}$ on $\mathbb{R}^d$.
2. Prior density p_0 on $\mathbb{R}^d$, typically a Gaussian density, is chosen.
3. Define $p(\cdot, t)$ as the probability density of

$$X_t = (1-t)X_0 + tX_1\,, \quad \text{where } X_0 \sim p_0 \text{ and } X_1 \sim p_1\,. \tag{1}$$

Idea: learn an ODE

$$\frac{dY_t}{dt} = u(Y_t, t)\,, \quad t \in [0,1] \tag{2}$$

such that, when $Y_0 \sim p_0$, then $Y_t \sim p(\cdot, t)$ for any $t \in [0,1]$.

Part 2: Denoising diffusion probabilistic models (DDPMs)

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at https://github.com/hojonathanho/diffusion

citation statistics: $>$ 23000.

Basics

1. Gaussian density with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d\times d}$:

$$\mathcal{N}(x;\mu,\Sigma) = (2\pi)^{-\frac{d}{2}}(\det\Sigma)^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^d. \tag{3}$$

Basics

1. Gaussian density with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$:

$$\mathcal{N}(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} (\det \Sigma)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^d. \tag{3}$$

2. $f : \mathbb{R} \to \mathbb{R}$ is convex. Jensen's inequality, for any $g : \mathbb{R}^d \to \mathbb{R}$,

$$f\left(\mathbb{E}_\pi(g(x))\right) = f\left(\int_{\mathbb{R}^d} g(x)\pi(x)dx \right) \leq \mathbb{E}_\pi\left(f(g(x))\right). \tag{4}$$

Basics

1. Gaussian density with mean $\mu \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$:

$$\mathcal{N}(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} (\det \Sigma)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^d. \tag{3}$$

2. $f : \mathbb{R} \to \mathbb{R}$ is convex. Jensen's inequality, for any $g : \mathbb{R}^d \to \mathbb{R}$,

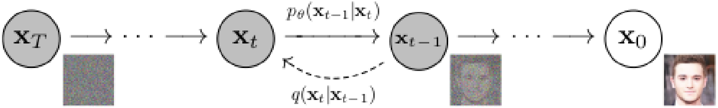
$$f\big(\mathbb{E}_\pi(g(x))\big) = f\Big(\int_{\mathbb{R}^d} g(x)\pi(x)dx\Big) \leq \mathbb{E}_\pi\big(f(g(x))\big). \tag{4}$$

3. KL divergence

$$D_{KL}(Q_1 \mid Q_2) := \mathbb{E}_{Q_1}\Big(\log \frac{Q_1}{Q_2}\Big). \tag{5}$$

- $D_{KL}(Q_1 \mid Q_2) \geq 0$
- $D_{KL}(Q_1 \mid Q_2) \neq D_{KL}(Q_2 \mid Q_1)$

Markov chains



Markov chains

1. Data distribution: $q_0(x)dx$.

Markov chains

1. Data distribution: $q_0(x)dx$.
2. Forward process: $x^{(1)}, \dots, x^{(N)} \in \mathbb{R}^d$, given $x^{(0)} \in \mathbb{R}^d$. Joint probability:

$$q(x^{(1:N)} \mid x^{(0)}) = \prod_{k=1}^{N} q(x^{(k)} \mid x^{(k-1)}),$$

where the transition density is

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d).$$

Markov chains

1. Data distribution: $q_0(x)dx$.
2. Forward process: $x^{(1)}, \dots, x^{(N)} \in \mathbb{R}^d$, given $x^{(0)} \in \mathbb{R}^d$. Joint probability:

$$q(x^{(1:N)} \mid x^{(0)}) = \prod_{k=1}^{N} q(x^{(k)} \mid x^{(k-1)}),$$

where the transition density is

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d).$$

3. Reverse process:

$$p_\theta(x^{(0:N)}) = p(x^{(N)}) \prod_{k=1}^{N} p_\theta(x^{(k-1)} \mid x^{(k)}),$$

where $p(x^{(N)})$ is a prior, θ is parameter, and

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\Big(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \Sigma_\theta(x^{(k)}, k)\Big).$$

Markov chains

1. Data distribution: $q_0(x)dx$.
2. Forward process: $x^{(1)}, \ldots, x^{(N)} \in \mathbb{R}^d$, given $x^{(0)} \in \mathbb{R}^d$. Joint probability:

$$q(x^{(1:N)} \mid x^{(0)}) = \prod_{k=1}^{N} q(x^{(k)} \mid x^{(k-1)}),$$

where the transition density is

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k} x^{(k-1)}, \beta_k \mathrm{I}_d).$$

3. Reverse process:

$$p_\theta(x^{(0:N)}) = p(x^{(N)}) \prod_{k=1}^{N} p_\theta(x^{(k-1)} \mid x^{(k)}),$$

where $p(x^{(N)})$ is a prior, θ is parameter, and

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\Big(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \Sigma_\theta(x^{(k)}, k)\Big).$$

4. $p_\theta(x^{(0)}) = \int p_\theta(x^{(0:N)})\, dx^{(1:N)}$.

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big)$$

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$
\begin{aligned}
&\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big) \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\,dx^{(1:N)}\Big)\Big]
\end{aligned}
$$

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$\begin{aligned}
&\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big)\\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\,dx^{(1:N)}\Big)\Big]\\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big)\Big]
\end{aligned}$$

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$
\begin{aligned}
&\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big) \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\,dx^{(1:N)}\Big)\Big] \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})} q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big)\Big] \\
\leq&\mathbb{E}_{q_0}\Big[\int -\log\Big(\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}\Big) q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big]
\end{aligned}
$$

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$
\begin{aligned}
&\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big) \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\,dx^{(1:N)}\Big)\Big] \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})} q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big)\Big] \\
\leq&\mathbb{E}_{q_0}\Big[\int -\log\Big(\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}\Big) q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big] \\
=&\mathbb{E}_{\mathbb{Q}}\Big(-\log\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}\Big)
\end{aligned}
$$

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$
\begin{aligned}
&\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big)\\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\,dx^{(1:N)}\Big)\Big]\\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\,|\,x^{(0)})} q(x^{(1:N)}\,|\,x^{(0)})\,dx^{(1:N)}\Big)\Big]\\
\leq&\mathbb{E}_{q_0}\Big[\int -\log\Big(\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\,|\,x^{(0)})}\Big) q(x^{(1:N)}\,|\,x^{(0)})\,dx^{(1:N)}\Big]\\
=&\mathbb{E}_{\mathbb{Q}}\Big(-\log\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\,|\,x^{(0)})}\Big)\\
=&\mathbb{E}_{\mathbb{Q}}\Big(-\log p(x^{(N)}) - \sum_{k=1}^{N}\log\frac{p_\theta(x^{(k-1)}\,|\,x^{(k)})}{q(x^{(k)}\,|\,x^{(k-1)})}\Big) =: L\,,
\end{aligned}
\tag{6}
$$

where $\mathbb{E}_{\mathbb{Q}}$ is expectation w.r.t. the density $q(x^{(0:N)})$ of the forward process.

Variational bound on negative log-likelihood

Using Jensen's inequality for the convex function $f(x) = -\log(x)$,

$$
\begin{aligned}
&\mathbb{E}_{q_0}\big(-\log p_\theta(x^{(0)})\big) \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int p_\theta(x^{(0:N)})\,dx^{(1:N)}\Big)\Big] \\
=&\mathbb{E}_{q_0}\Big[-\log\Big(\int \frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big)\Big] \\
\leq&\mathbb{E}_{q_0}\Big[\int -\log\Big(\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}\Big)q(x^{(1:N)}\mid x^{(0)})\,dx^{(1:N)}\Big] \\
=&\mathbb{E}_{\mathbb{Q}}\Big(-\log\frac{p_\theta(x^{(0:N)})}{q(x^{(1:N)}\mid x^{(0)})}\Big) \\
=&\mathbb{E}_{\mathbb{Q}}\Big(-\log p(x^{(N)}) - \sum_{k=1}^{N}\log\frac{p_\theta(x^{(k-1)}\mid x^{(k)})}{q(x^{(k)}\mid x^{(k-1)})}\Big) =: L\,,
\end{aligned}
\tag{6}
$$

where $\mathbb{E}_{\mathbb{Q}}$ is expectation w.r.t. the density $q(x^{(0:N)})$ of the forward process.

Idea: optimize θ by minimizing the upper bound L.

Analyzing L, (1)

Goal: to derive a simplified objective based on L.

Analyzing L, (1)

Goal: to derive a simplified objective based on L.

Transition density of the forward process:

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d)\,.$$

Analyzing L, (1)

Goal: to derive a simplified objective based on L.

Transition density of the forward process:

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d)\,.$$

Lemma

For $k > 0$, $q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d)$, where

$$\alpha_k = 1 - \beta_k\,, \quad \text{and} \quad \bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i\,. \tag{7}$$

Analyzing L, (1)

Goal: to derive a simplified objective based on L.

Transition density of the forward process:

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d) .$$

Lemma

For $k > 0$, $q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d)$, where

$$\alpha_k = 1 - \beta_k \,, \quad \text{and} \quad \bar{\alpha}_k = \prod_{i=1}^{k} \alpha_i \,. \tag{7}$$

Proof.

Prove by induction. □

Analyzing L, (2)

1. Variational bound: $L = \mathbb{E}_{\mathbb{Q}}\Big(-\log p(x^{(N)}) - \sum_{k=1}^{N} \log \frac{p_\theta(x^{(k-1)} \mid x^{(k)})}{q(x^{(k)} \mid x^{(k-1)})} \Big)$.
2. Bayes' theorem $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$

$$\implies \quad q(x^{(k)} \mid x^{(k-1)}) = q(x^{(k-1)} \mid x^{(k)}, x^{(0)}) \frac{q(x^{(k)}|x^{(0)})}{q(x^{(k-1)}|x^{(0)})} \,.$$

Analyzing L, (2)

1. Variational bound: $L = \mathbb{E}_{\mathbb{Q}}\Big(-\log p(x^{(N)}) - \sum_{k=1}^{N} \log \frac{p_\theta(x^{(k-1)} \mid x^{(k)})}{q(x^{(k)} \mid x^{(k-1)})}\Big)$.
2. Bayes' theorem $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$

$$\Longrightarrow \quad q(x^{(k)} \mid x^{(k-1)}) = q(x^{(k-1)} \mid x^{(k)}, x^{(0)}) \frac{q(x^{(k)}|x^{(0)})}{q(x^{(k-1)}|x^{(0)})}.$$

3. Decomposition of L:

Proposition

We have $L = L_N + \sum_{k=2}^{N} L_{k-1} + L_0$, where

$$\begin{aligned} L_0 &= -\mathbb{E}_{\mathbb{Q}}\Big(\log p_\theta(x^{(0)}|x^{(1)})\Big), \\ L_{k-1} &= \mathbb{E}_{\mathbb{Q}}\Big[D_{KL}\Big(q\big(x^{(k-1)} \mid x^{(k)}, x^{(0)}\big) \,\Big|\, p_\theta\big(x^{(k-1)} \mid x^{(k)}\big)\Big)\Big], \quad k = 2, \ldots, N, \\ L_N &= \mathbb{E}_{\mathbb{Q}}\Big[D_{KL}\Big(q\big(x^{(N)}|x^{(0)}\big) \,\Big|\, p(x^{(N)})\Big)\Big]. \end{aligned}$$

Analyzing L, (3)

Transition density

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d)\,.$$

Analyzing L, (3)

Transition density

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d)\,.$$

$$q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d)\,,$$

Analyzing L, (3)

Transition density

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d)\,.$$

$$q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d)\,,$$

Bayes' theorem:

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

Analyzing L, (3)

Transition density

$$q(x^{(k)} \mid x^{(k-1)}) = \mathcal{N}(x^{(k)}; \sqrt{1-\beta_k}x^{(k-1)}, \beta_k \mathrm{I}_d)\,.$$

$$q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d)\,,$$

Bayes' theorem:

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

Putting together,

Lemma

We have $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$, *where*

$$\begin{aligned} \widetilde{\mu}_k(x^{(k)}, x^{(0)}) &= \frac{\sqrt{\bar{\alpha}_{k-1}}\beta_k}{1-\bar{\alpha}_k}x^{(0)} + \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k}x^{(k)}\,, \\ \widetilde{\beta}_k &= \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k\,. \end{aligned} \tag{8}$$

Analyzing L, (4)

1. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$.

Analyzing L, (4)

1. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$.
2. Choices of $p_\theta(x^{(k-1)} \mid x^{(k)})$:

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\left(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\right),$$

where $\sigma_k^2 = \beta_k$, or $\sigma_k^2 = \widetilde{\beta}_k = \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k$.

Analyzing L, (4)

1. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$.
2. Choices of $p_\theta(x^{(k-1)} \mid x^{(k)})$:

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\left(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\right),$$

where $\sigma_k^2 = \beta_k$, or $\sigma_k^2 = \widetilde{\beta}_k = \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k$.

Lemma

For $2 \leq k \leq N$, we have

$$\begin{aligned} L_{k-1} =& \mathbb{E}_{\mathbb{Q}}\left[D_{KL}\left(q\left(x^{(k-1)} \mid x^{(k)}, x^{(0)}\right) \middle| p_\theta\left(x^{(k-1)} \mid x^{(k)}\right)\right)\right] \\ =& \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\left(\left|\mu_\theta(x^{(k)}, k) - \widetilde{\mu}_k(x^{(k)}, x^{(0)})\right|^2\right) + C. \end{aligned}$$

Analyzing L, (4)

1. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$.
2. Choices of $p_\theta(x^{(k-1)} \mid x^{(k)})$:

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\left(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\right),$$

where $\sigma_k^2 = \beta_k$, or $\sigma_k^2 = \widetilde{\beta}_k = \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k$.

Lemma

For $2 \leq k \leq N$, we have

$$\begin{aligned} L_{k-1} =& \mathbb{E}_{\mathbb{Q}}\left[D_{KL}\left(q\left(x^{(k-1)} \mid x^{(k)}, x^{(0)}\right) \middle| p_\theta\left(x^{(k-1)} \mid x^{(k)}\right)\right)\right] \\ =& \frac{1}{2\sigma_k^2}\mathbb{E}_{\mathbb{Q}}\left(\left|\mu_\theta(x^{(k)}, k) - \widetilde{\mu}_k(x^{(k)}, x^{(0)})\right|^2\right) + C\,. \end{aligned}$$

Conclusion: $\mu_\theta(x^{(k)}, k)$ should predict $\widetilde{\mu}_k$.

Analyzing L, (5)

1. Reparametrization:

$$\begin{aligned} & q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k}x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d) \\ \implies \quad & x^{(k)}(x^{(0)}, \epsilon) = \sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon\,, \text{ where } \epsilon \sim \mathcal{N}(0, \mathrm{I}_d)\,. \end{aligned}$$

2. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k\mathrm{I}_d)$, where $\widetilde{\mu}_k(x^{(k)}, x^{(0)}) = \frac{\sqrt{\bar{\alpha}_{k-1}}\beta_k}{1-\bar{\alpha}_k}x^{(0)} + \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k}x^{(k)}$.

Analyzing L, (5)

1. Reparametrization:

$$
\begin{aligned}
& q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar{\alpha}_k} x^{(0)}, (1-\bar{\alpha}_k)\mathrm{I}_d) \\
\Longrightarrow \quad & x^{(k)}(x^{(0)}, \epsilon) = \sqrt{\bar{\alpha}_k} x^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon\,, \ \text{where } \epsilon \sim \mathcal{N}(0, \mathrm{I}_d)\,.
\end{aligned}
$$

2. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$, where $\widetilde{\mu}_k(x^{(k)}, x^{(0)}) = \frac{\sqrt{\bar{\alpha}_{k-1}}\beta_k}{1-\bar{\alpha}_k} x^{(0)} + \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k} x^{(k)}$.

$$
\begin{aligned}
& L_{k-1} - C \\
=& \mathbb{E}_{\mathbb{Q}} \left(\frac{1}{2\sigma_k^2} \left| \mu_\theta(x^{(k)}, k) - \widetilde{\mu}_k(x^{(k)}, x^{(0)}) \right|^2 \right) \\
=& \mathbb{E}_{x^{(0)}, \epsilon} \left[\frac{1}{2\sigma_k^2} \left| \mu_\theta\left(x^{(k)}(x^{(0)}, \epsilon), k\right) - \widetilde{\mu}_k\left(x^{(k)}(x^{(0)}, \epsilon), \frac{1}{\sqrt{\bar{\alpha}_k}} \left(x^{(k)}(x^{(0)}, \epsilon) - \sqrt{1-\bar{\alpha}_k}\epsilon \right) \right) \right|^2 \right] \\
=& \mathbb{E}_{x^{(0)}, \epsilon} \left[\frac{1}{2\sigma_k^2} \left| \mu_\theta\left(x^{(k)}(x^{(0)}, \epsilon), k\right) - \frac{1}{\sqrt{\alpha_k}} \left(x^{(k)}(x^{(0)}, \epsilon) - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}} \epsilon \right) \right|^2 \right].
\end{aligned}
$$

Analyzing L, (5)

1. Reparametrization:

$$\begin{aligned} & q(x^{(k)}|x^{(0)}) = \mathcal{N}(x^{(k)}; \sqrt{\bar\alpha_k}x^{(0)}, (1-\bar\alpha_k)\mathrm{I}_d) \\ \implies \quad & x^{(k)}(x^{(0)},\epsilon) = \sqrt{\bar\alpha_k}x^{(0)} + \sqrt{1-\bar\alpha_k}\epsilon\,, \text{ where } \epsilon \sim \mathcal{N}(0,\mathrm{I}_d)\,. \end{aligned}$$

2. $q(x^{(k-1)}|x^{(k)}, x^{(0)}) = \mathcal{N}(x^{(k-1)}; \widetilde{\mu}_k(x^{(k)}, x^{(0)}), \widetilde{\beta}_k \mathrm{I}_d)$, where $\widetilde{\mu}_k(x^{(k)}, x^{(0)}) = \frac{\sqrt{\bar\alpha_{k-1}}\beta_k}{1-\bar\alpha_k}x^{(0)} + \frac{\sqrt{\alpha_k}(1-\bar\alpha_{k-1})}{1-\bar\alpha_k}x^{(k)}$.

$$\begin{aligned} & L_{k-1} - C \\ =& \mathbb{E}_{\mathbb{Q}}\left(\frac{1}{2\sigma_k^2}\left|\mu_\theta(x^{(k)},k) - \widetilde{\mu}_k(x^{(k)},x^{(0)})\right|^2\right) \\ =& \mathbb{E}_{x^{(0)},\epsilon}\left[\frac{1}{2\sigma_k^2}\left|\mu_\theta\left(x^{(k)}(x^{(0)},\epsilon),k\right) - \widetilde{\mu}_k\left(x^{(k)}(x^{(0)},\epsilon), \frac{1}{\sqrt{\bar\alpha_k}}\left(x^{(k)}(x^{(0)},\epsilon) - \sqrt{1-\bar\alpha_k}\epsilon\right)\right)\right|^2\right] \\ =& \mathbb{E}_{x^{(0)},\epsilon}\left[\frac{1}{2\sigma_k^2}\left|\mu_\theta\left(x^{(k)}(x^{(0)},\epsilon),k\right) - \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)}(x^{(0)},\epsilon) - \frac{\beta_k}{\sqrt{1-\bar\alpha_k}}\epsilon\right)\right|^2\right]. \end{aligned}$$

Hence, the parametrization: $\mu_\theta(x^{(k)},k) = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar\alpha_k}}\epsilon_\theta(x^{(k)},k)\right)$.

Analyzing L, (6)

Transition density:

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\left(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\right),$$

with $\mu_\theta(x^{(k)}, k) = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}} \epsilon_\theta(x^{(k)}, k)\right)$.

Analyzing L, (6)

Transition density:

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\left(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\right),$$

with $\mu_\theta(x^{(k)}, k) = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x^{(k)}, k)\right)$.

Sampling $x^{(k-1)} \sim p_\theta(x^{(k-1)}|x^{(k)})$:

$$x^{(k-1)} = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x^{(k)}, k)\right) + \sigma_k z\,, \text{ where } z \sim \mathcal{N}(0, \mathrm{I}_d)\,. \tag{9}$$

Analyzing L, (6)

Transition density:

$$p_\theta(x^{(k-1)} \mid x^{(k)}) = \mathcal{N}\left(x^{(k-1)}; \mu_\theta(x^{(k)}, k), \sigma_k^2 \mathrm{I}_d\right),$$

with $\mu_\theta(x^{(k)}, k) = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x^{(k)}, k)\right)$.

Sampling $x^{(k-1)} \sim p_\theta(x^{(k-1)}|x^{(k)})$:

$$x^{(k-1)} = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x^{(k)}, k)\right) + \sigma_k z\,, \text{ where } z \sim \mathcal{N}(0, \mathrm{I}_d)\,. \tag{9}$$

$$\begin{aligned} L_{k-1} - C =& \mathbb{E}_{x^{(0)},\epsilon}\left[\frac{1}{2\sigma_k^2}\left|\mu_\theta\left(x^{(k)}(x^{(0)}, \epsilon), k\right) - \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)}(x^{(0)}, \epsilon) - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon\right)\right|^2\right] \\ =& \mathbb{E}_{x^{(0)},\epsilon}\left[\frac{\beta_k^2}{2\sigma_k^2\alpha_k(1-\bar{\alpha}_k)}\left|\epsilon_\theta(\sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon, k) - \epsilon\right|^2\right]. \end{aligned}$$

Loss and algorithm

Simplified loss:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{k,x^{(0)},\epsilon}\left[\left|\epsilon_\theta(\sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon, k) - \epsilon\right|^2\right]. \tag{10}$$

Loss and algorithm

Simplified loss:

$$L_{\text{simple}}(\theta) = \mathbb{E}_{k,x^{(0)},\epsilon}\left[\left|\epsilon_\theta(\sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon, k) - \epsilon\right|^2\right]. \tag{10}$$

Algorithm Training

1: **repeat**
2: $x^{(0)} \sim q_0(x^{(0)})$
3: $k \sim \mathrm{Uniform}(\{1, 2, \ldots, N\})$
4: $\epsilon \sim \mathcal{N}(0, \mathrm{I}_d)$.
5: take gradient descent step on
6: $\nabla_\theta\left|\epsilon_\theta(\sqrt{\bar{\alpha}_k}x^{(0)} + \sqrt{1-\bar{\alpha}_k}\epsilon, k) - \epsilon\right|^2$
7: **until** converged

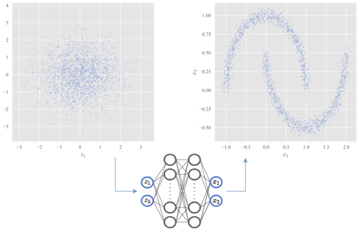
Algorithm Sampling

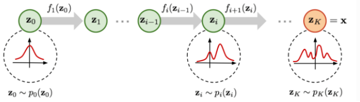
$x^{(N)} \sim \mathcal{N}(0, \mathrm{I}_d)$
for $k = N, \dots, 1$ **do**
  $z \sim \mathcal{N}(0, \mathrm{I}_d)$ if $k > 1$, else $z = 0$
  $x^{(k-1)} = \frac{1}{\sqrt{\alpha_k}}\left(x^{(k)} - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(x^{(k)}, k)\right) + \sigma_k z$
end for
return $x^{(0)}$

Part 3: Normalizing flows

Idea

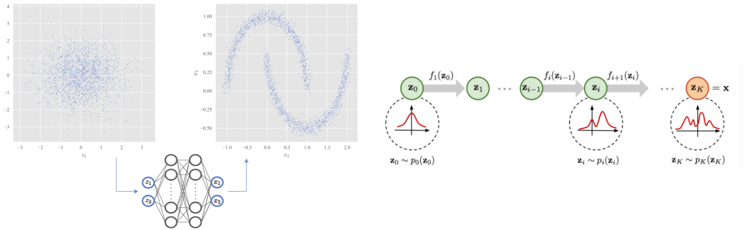
Express data $x \in \mathbb{R}^d$ as a transformation of (Gaussian) $z \sim p_z(z)$.





Idea

Express data $x \in \mathbb{R}^d$ as a transformation of (Gaussian) $z \sim p_z(z)$.



1. "normalizing": f^{-1} changes data distribution to normal distribution.
2. "flow": f is often composition of $f_1, f_2, \cdots, f_K$.

Change of variables

Let $z \sim p_z(z)$ and $x = f(z)$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is invertible and differentiable. Then,

$$p_x(x) = p_z(f^{-1}(x)) \left| \det J_{f^{-1}}(x) \right| . \tag{11}$$

Change of variables

Let $z \sim p_z(z)$ and $x = f(z)$, where $f : \mathbb{R}^d \to \mathbb{R}^d$ is invertible and differentiable. Then,

$$p_x(x) = p_z(f^{-1}(x)) \left| \det J_{f^{-1}}(x) \right| . \tag{11}$$

1. target $p_x^*(x)$
2. transformation $x = f(z; \theta)$.

Goal: to find θ such that $p_x(x; \theta)$ is close to $p_x^*(x)$.

Loss function

$$p_x(x;\theta) = p_z(f^{-1}(x;\theta))|\det J_{f^{-1}}(x;\theta)|.$$

Loss function

$p_x(x;\theta) = p_z(f^{-1}(x;\theta))|\det J_{f^{-1}}(x;\theta)|.$

(forward) KL divergence:

$$
\begin{aligned}
&D_{KL}\Big(p_x^*(x) \mid p_x(x;\theta)\Big) \\
=&\mathbb{E}_{x\sim p_x^*(x)}\Big(\log\frac{p_x^*(x)}{p_x(x;\theta)}\Big) \\
=&-\mathbb{E}_{x\sim p_x^*(x)}\Big[\log p_z(f^{-1}(x;\theta)) + \log|\det J_{f^{-1}}(x;\theta)|\Big] + C\,.
\end{aligned}
$$

Loss function

$p_x(x;\theta) = p_z(f^{-1}(x;\theta))|\det J_{f^{-1}}(x;\theta)|.$

(forward) KL divergence:

$$
\begin{aligned}
& D_{KL}\Big(p_x^*(x) \mid p_x(x;\theta)\Big) \\
=& \mathbb{E}_{x\sim p_x^*(x)}\Big(\log \frac{p_x^*(x)}{p_x(x;\theta)}\Big) \\
=& -\mathbb{E}_{x\sim p_x^*(x)}\Big[\log p_z(f^{-1}(x;\theta)) + \log|\det J_{f^{-1}}(x;\theta)|\Big] + C\,.
\end{aligned}
$$

Training objective in practice:

$$
\mathrm{Loss}(\theta) = -\frac{1}{N}\sum_{n=1}^{N}\Big[\log p_z(f^{-1}(x_n;\theta)) + \log|\det J_{f^{-1}}(x_n;\theta)|\Big]\,.
$$

Normalizing flow models

Requirements on f:

1. input and output dimensions are the same;
2. f is invertible;
3. computing f^{-1} and $\det J_{f^{-1}}$ is efficient.

Normalizing flow models

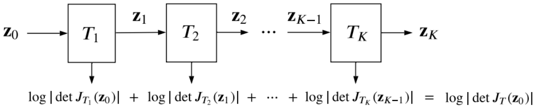
Requirements on f:

1. input and output dimensions are the same;
2. f is invertible;
3. computing f^{-1} and $\det J_{f^{-1}}$ is efficient.

If $f_1, f_2, \cdots, f_K$ satisfy the above requirements, so does the composition $f = f_K \circ \cdots \circ f_1$, and

$$\log|\det J_{f^{-1}}(x)| = \log\left|\prod_{k=1}^{K} \det J_{f_k^{-1}}(z_k)\right| = \sum_{k=1}^{K} \log|\det J_{f_k^{-1}}(z_k)|, \tag{12}$$

where $z_K = x$ and $z_{k-1} = f_k^{-1}(z_k), \quad k = K, \ldots, 1.$



Nonlinear Independent Components Estimation (NICE) model

$x = (x_1, x_2) = f(z)$ is defined as

$$
\begin{aligned}
x_1 &= z_1 \,, \\
x_2 &= z_2 + \mu_\theta(z_1) \,,
\end{aligned}
$$

where $z = (z_1, z_2)$, $z_1 \in \mathbb{R}^{d'}$ and $z_2 \in \mathbb{R}^{d-d'}$, and $\mu_\theta : \mathbb{R}^{d'} \to \mathbb{R}^{d-d'}$ is a neural network. The inverse of f is

$$
\begin{aligned}
z_1 &= x_1 \,, \\
z_2 &= x_2 - \mu_\theta(x_1) \,.
\end{aligned}
$$

The Jacobian determinant of f is one.

Real Non-Volume Preserving (RealNVP) model

$x = (x_1, x_2) = f(z)$ is defined as

$$\begin{aligned} x_1 &= z_1 \,, \\ x_2 &= \exp(\sigma_\theta(z_1)) \odot z_2 + \mu_\theta(z_1) \,, \end{aligned}$$

where $\odot$ denotes elementwise product, and $\sigma_\theta, \mu_\theta : \mathbb{R}^{d'} \to \mathbb{R}^{d-d'}$. The inverse of f is

$$\begin{aligned} z_1 &= x_1 \,, \\ z_2 &= \exp(-\sigma_\theta(x_1)) \odot (x_2 - \mu_\theta(x_1)) \,, \end{aligned}$$

The Jacobian determinant is $\det J_{f^{-1}}(x) = \exp(-\sum_{j=1}^{d-d'} (\sigma_\theta(x_1))_j)$.