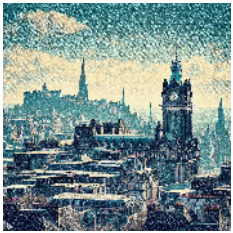
Advanced Topics in Machine Learning (deep generative modelling)

Lecture 8: Diffusion models II

Nikolay Malkin

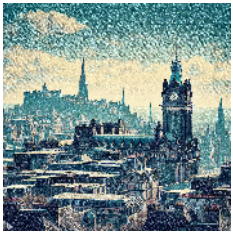
17 March 2026

Dynamics-based generative models in practice



‘Edinburgh from Calton Hill, pointillist style’

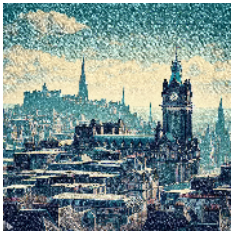
Dynamics-based generative models in practice



'Edinburgh from Calton Hill, pointillist style'

[Janner et al., ICML'22]

Dynamics-based generative models in practice

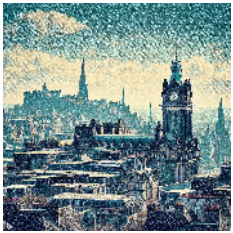


'Edinburgh from Calton Hill, pointillist style'

[Graikos et al., NeurIPS'22]

[Janner et al., ICML'22]

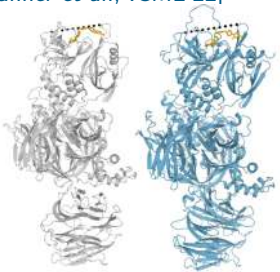
Dynamics-based generative models in practice



'Edinburgh from Calton Hill, pointillist style'

[Graikos et al., NeurIPS'22]

[Janner et al., ICML'22]



AlphaFold 3

Dynamics-based generative models in practice

[Zhang and Gienger, 2024]

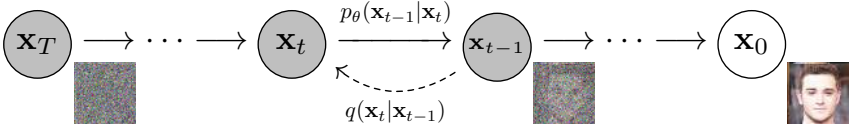
Outline of weeks 8-10

What is a diffusion model, how to understand it from different perspectives, and where to use it in practice:

Outline of weeks 8-10

What is a diffusion model, how to understand it from different perspectives, and where to use it in practice:

- **Lecture 8:** Diffusion models are hierarchical latent variable models / deep VAEs



[Ho et al., 'Denoising diffusion probabilistic models']

Outline of weeks 8-10

What is a diffusion model, how to understand it from different perspectives, and where to use it in practice:

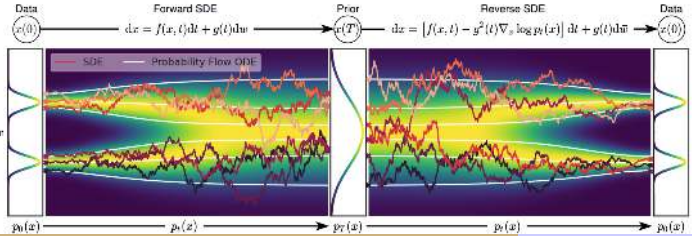
- **Lecture 8:** Diffusion models are hierarchical latent variable models / deep VAEs
- **Lecture 9:** Diffusion models are score-based models

[Song et al., 'Generative modeling by estimating...' blog post]

Outline of weeks 8-10

What is a diffusion model, how to understand it from different perspectives, and where to use it in practice:

- **Lecture 8:** Diffusion models are hierarchical latent variable models / deep VAEs
- **Lecture 9:** Diffusion models are score-based models
- **Lecture 10:** Diffusion models are continuous-time processes



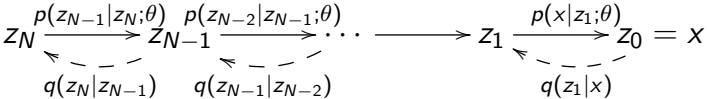
[Song et al., 'Score-based generative modeling. . . ']

Outline of Lecture 9

- Review of Lecture 8
- A three-way connection
  - Reversal of diffusion is denoising (second magic property)
  - Denoising is score matching
- Conditioning and guidance
- Unsimplifications

- Review of Lecture 8
- A three-way connection
  - Reversal of diffusion is denoising (second magic property)
  - Denoising is score matching
- Conditioning and guidance
- Unsimplifications

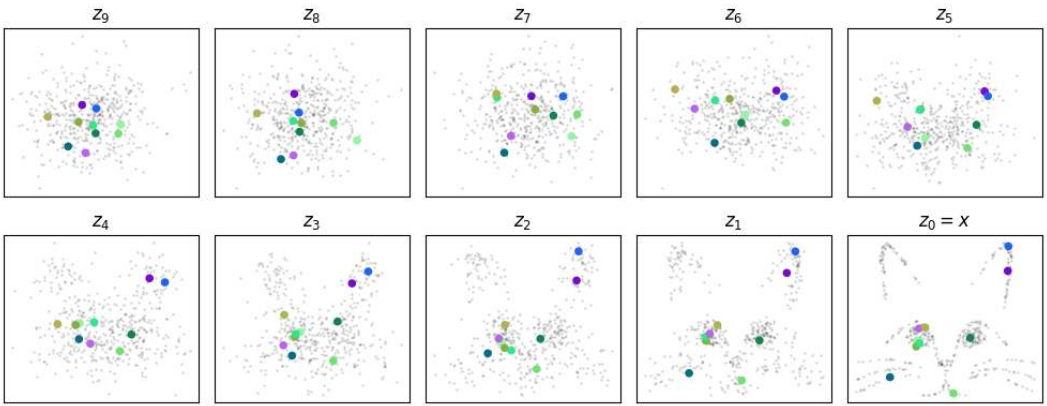
From last time: Diffusion models are hierarchical LVMs



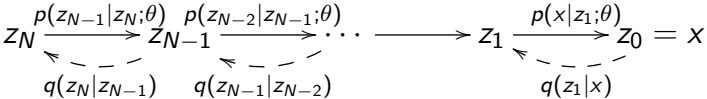
- Diffusion models (typically) fix the distributions $q(z_n \mid z_{n-1})$
  - Typically to adding isotropic Gaussian noise: $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$. . . or variants (next time)

From last time: Diffusion models are hierarchical LVMs

$$z_N \xrightarrow{p(z_{N-1}|z_N;\theta)} z_{N-1} \xrightarrow{p(z_{N-2}|z_{N-1};\theta)} \cdots \longrightarrow z_1 \xrightarrow{p(x|z_1;\theta)} z_0 = x$$



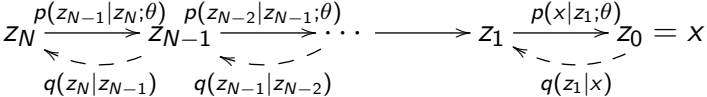
From last time: Diffusion models are hierarchical LVMs



- Diffusion models (typically) fix the distributions $q(z_n \mid z_{n-1})$
  - Typically to adding isotropic Gaussian noise: $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$. . . or variants (next time)
- Training to maximise $\log p(x^t)$ for a data sample x^t:
  - Sample the latents from q: $x = z_0 \rightsquigarrow z_1 \rightsquigarrow \cdots \rightsquigarrow z_N$
  - Gradient step on the likelihood in ancestral factorisation:

$$\log [p(z_N; \theta) p(z_{N-1} \mid z_N; \theta) \dots p(x \mid z_1; \theta)]$$

From last time: Diffusion models are hierarchical LVMs

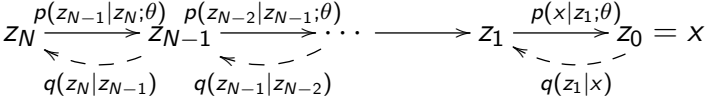


- Diffusion models (typically) fix the distributions $q(z_n \mid z_{n-1})$
  - Typically to adding isotropic Gaussian noise: $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$. . . or variants (next time)
- Training to maximise $\log p(x^t)$ for a data sample x^t:
  - Sample the latents from q: $x = z_0 \rightsquigarrow z_1 \rightsquigarrow \cdots \rightsquigarrow z_N$
  - Gradient step on the likelihood in ancestral factorisation:

$$\log [p(z_N; \theta) p(z_{N-1} \mid z_N; \theta) \dots p(x \mid z_1; \theta)]$$

- Typical assumptions:
  - $p(z_{n-1} \mid z_n; \theta)$ is Gaussian with mean computed by a NN with input n, z_n and parameters θ
  - Its variance is fixed to something related to σ_n^2 or σ_{n-1}^2

From last time: Diffusion models are hierarchical LVMs



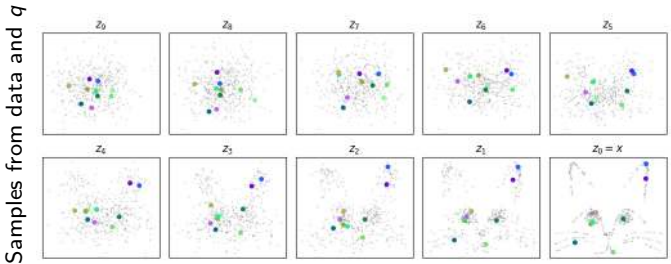
- Diffusion models (typically) fix the distributions $q(z_n \mid z_{n-1})$
  - Typically to adding isotropic Gaussian noise: $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$. . . or variants (next time)
- Training to maximise $\log p(x^t)$ for a data sample x^t:
  - Sample the latents from q: $x = z_0 \rightsquigarrow z_1 \rightsquigarrow \cdots \rightsquigarrow z_N$
  - Gradient step on the likelihood in ancestral factorisation:

$$\log [p(z_N; \theta) p(z_{N-1} \mid z_N; \theta) \dots p(x \mid z_1; \theta)]$$

- Typical assumptions:
  - $p(z_{n-1} \mid z_n; \theta)$ is Gaussian (but not always!) with mean computed by a NN with input n, z_n and parameters θ
  - Its variance is fixed to something related to σ_n^2 or σ_{n-1}^2 (but not always!)

From last time: Training a diffusion model on 2D data

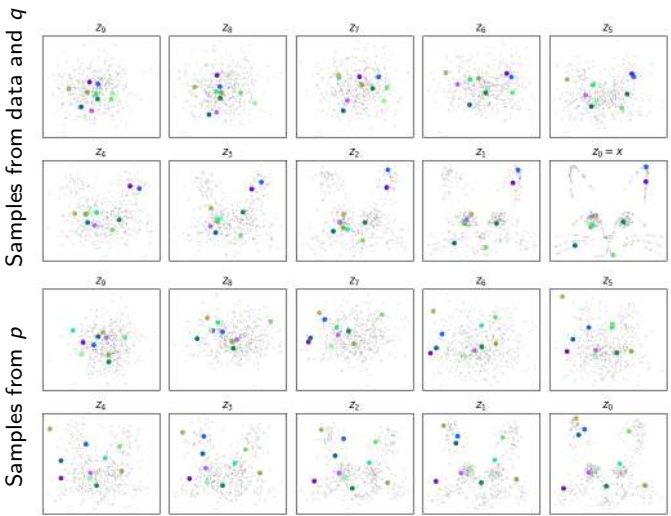
Posterior model
(noising)



From last time: Training a diffusion model on 2D data

Posterior model (noising)

Generative model (denoising / reconstruction)



From last time: First observations

- Noise distribution $p(z_N)$ not usually trained (approximately Gaussian)
  - The choice of noise schedule (σ_n) and number of steps are important (more later!)

From last time: First observations

- Noise distribution $p(z_N)$ not usually trained (approximately Gaussian)
  - The choice of noise schedule (σ_n) and number of steps are important (more later!)
- To maximise

$$\log\left[p(z_N;\theta)p(z_{N-1} \mid z_N;\theta)\dots p(x \mid z_1;\theta)\right],$$

can randomly sample **one** of the $p(z_{n-1} \mid z_n;\theta)$ for training

From last time: First observations

- Noise distribution $p(z_N)$ not usually trained (approximately Gaussian)
  - The choice of noise schedule (σ_n) and number of steps are important (more later!)
- To maximise

$$\log\left[p(z_N;\theta)p(z_{N-1} \mid z_N;\theta)\dots p(x \mid z_1;\theta)\right],$$

  can randomly sample **one** of the $p(z_{n-1} \mid z_n;\theta)$ for training
- Two magic properties to make training efficient:
  - **Simulation-free training:** We can get z_{n-1}, z_n from $z_0 = x$ without all the intermediate steps (why?)

From last time: First observations

- Noise distribution $p(z_N)$ not usually trained (approximately Gaussian)
  - The choice of noise schedule (σ_n) and number of steps are important (more later!)
- To maximise

$$\log\left[p(z_N;\theta)p(z_{N-1} \mid z_N;\theta)\dots p(x \mid z_1;\theta)\right],$$

  can randomly sample **one** of the $p(z_{n-1} \mid z_n;\theta)$ for training
- Two magic properties to make training efficient:
  - **Simulation-free training:** We can get z_{n-1}, z_n from $z_0 = x$ without all the intermediate steps (why?)
  - **Rao-Blackwellised denoising objective:** We can estimate the gradient using just z_0 and z_n (this next!)

From last time: First observations

- Noise distribution $p(z_N)$ not usually trained (approximately Gaussian)
  - The choice of noise schedule (σ_n) and number of steps are important (more later!)
- To maximise

$$\log\left[p(z_N;\theta)p(z_{N-1} \mid z_N;\theta)\dots p(x \mid z_1;\theta)\right],$$

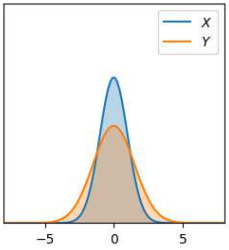
  can randomly sample **one** of the $p(z_{n-1} \mid z_n;\theta)$ for training
- Two magic properties to make training efficient:
  - **Simulation-free training:** We can get z_{n-1}, z_n from $z_0 = x$ without all the intermediate steps (why?)
  - **Rao-Blackwellised denoising objective:** We can estimate the gradient using just z_0 and z_n (this next!)
  - Generalisations may lose one or both magic properties

- Review of Lecture 8
- A three-way connection
  - Reversal of diffusion is denoising (second magic property)
  - Denoising is score matching
- Conditioning and guidance
- Unsimplifications

Review of Gaussians I

Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ are independent Gaussian random variables...

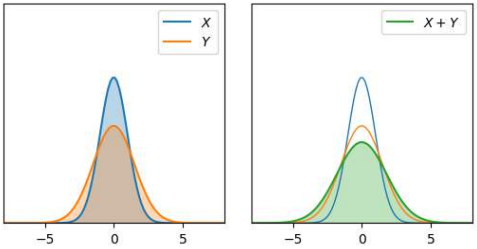
- What is the distribution of $X + Y$? Of aX ($a \in \mathbb{R}$)?



Review of Gaussians I

Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ are independent Gaussian random variables...

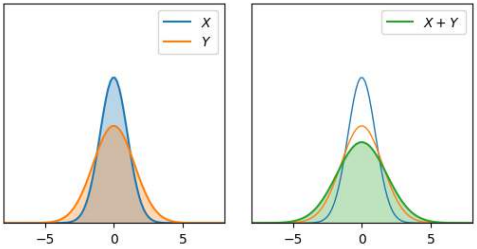
- What is the distribution of $X + Y$? Of aX ($a \in \mathbb{R}$)?
  - $X + Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_Y^2)$, $aX \sim \mathcal{N}(0, a^2\sigma_X^2)$



Review of Gaussians I

Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ are independent Gaussian random variables...

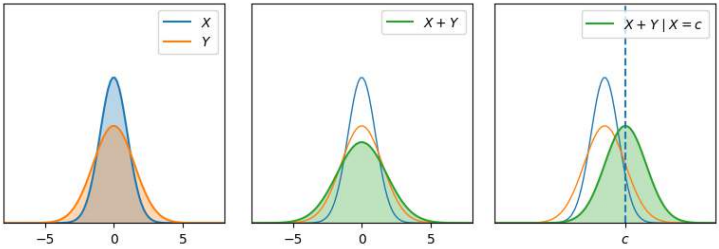
- What is the distribution of $X + Y$? Of aX ($a \in \mathbb{R}$)?
  - $X + Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_Y^2)$, $aX \sim \mathcal{N}(0, a^2\sigma_X^2)$
- What is the distribution of $X + Y$ given $X = c$?



Review of Gaussians I

Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ are independent Gaussian random variables...

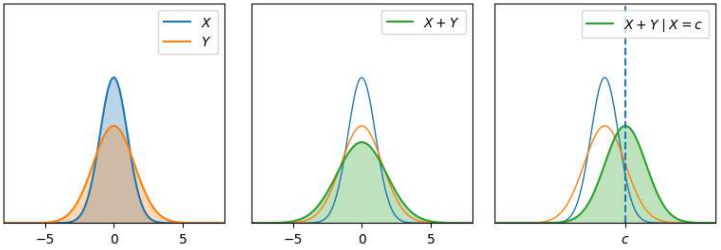
- What is the distribution of $X + Y$? Of aX ($a \in \mathbb{R}$)?
  - $X + Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_Y^2)$, $aX \sim \mathcal{N}(0, a^2\sigma_X^2)$
- What is the distribution of $X + Y$ given $X = c$?
  - $[X + Y \mid X = c] \sim \mathcal{N}(c, \sigma_Y^2)$



Review of Gaussians I

Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ are independent Gaussian random variables...

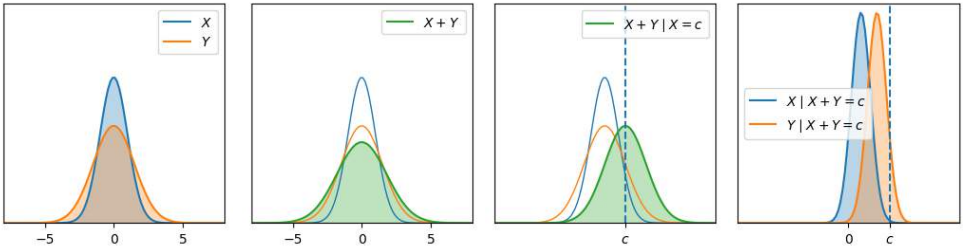
- What is the distribution of $X + Y$? Of aX ($a \in \mathbb{R}$)?
  - $X + Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_Y^2)$, $aX \sim \mathcal{N}(0, a^2\sigma_X^2)$
- What is the distribution of $X + Y$ given $X = c$?
  - $[X + Y \mid X = c] \sim \mathcal{N}(c, \sigma_Y^2)$
- What is the distribution of X given $X + Y = c$?



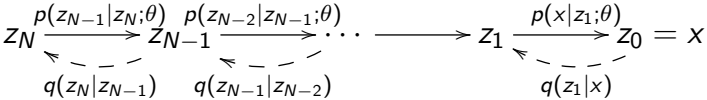
Review of Gaussians I

Suppose $X \sim \mathcal{N}(0, \sigma_X^2)$ and $Y \sim \mathcal{N}(0, \sigma_Y^2)$ are independent Gaussian random variables...

- What is the distribution of $X + Y$? Of aX ($a \in \mathbb{R}$)?
  - $X + Y \sim \mathcal{N}(0, \sigma_X^2 + \sigma_Y^2)$, $aX \sim \mathcal{N}(0, a^2\sigma_X^2)$
- What is the distribution of $X + Y$ given $X = c$?
  - $[X + Y \mid X = c] \sim \mathcal{N}(c, \sigma_Y^2)$
- What is the distribution of X given $X + Y = c$?
  - $[X \mid X + Y = c] \sim \mathcal{N}\left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_Y^2} c, \frac{1}{1/\sigma_X^2+1/\sigma_Y^2}\right)$



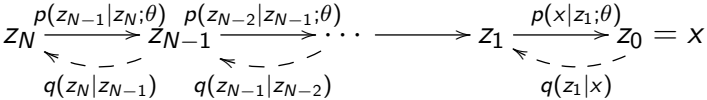
The denoising mean: Setup



Some setup:

- $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$ (or: $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$)

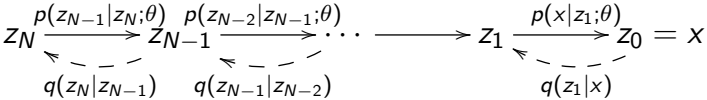
The denoising mean: Setup



Some setup:

- $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$ (or: $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$)
- $[z_n \mid z_0] \sim \mathcal{N}(z_0, (\sigma_1^2 + \cdots + \sigma_n^2) I_d)$

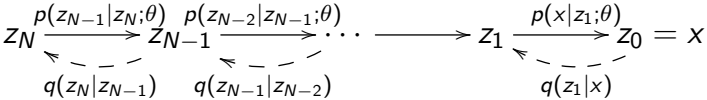
The denoising mean: Setup



Some setup:

- $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$ (or: $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$)
- $[z_n \mid z_0] \sim \mathcal{N}(z_0, (\sigma_1^2 + \cdots + \sigma_n^2) I_d) = \mathcal{N}(z_0, V_n I_d)$
- Abbreviate $V_n := \sigma_1^2 + \cdots + \sigma_n^2$

The denoising mean: Setup



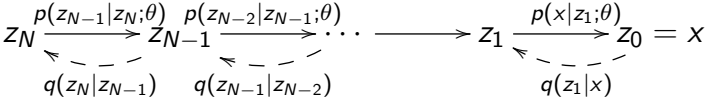
Some setup:

- $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$ (or: $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$)
- $[z_n \mid z_0] \sim \mathcal{N}(z_0, (\sigma_1^2 + \cdots + \sigma_n^2) I_d) = \mathcal{N}(z_0, V_n I_d)$
- Abbreviate $V_n := \sigma_1^2 + \cdots + \sigma_n^2$

Recall the training procedure:

- Sample $x = z_0 \sim \mathcal{D}$ and time step n
- Sample $z_0 \rightsquigarrow \cdots \rightsquigarrow z_{n-1} \rightsquigarrow z_n$ following q
- Maximize $\log p(z_{n-1} \mid z_n; \theta)$

The denoising mean: Setup



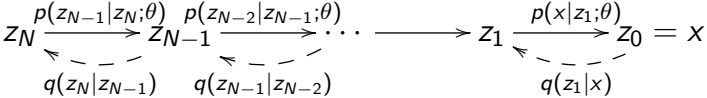
Some setup:

- $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$ (or: $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$)
- $[z_n \mid z_0] \sim \mathcal{N}(z_0, (\sigma_1^2 + \cdots + \sigma_n^2) I_d) = \mathcal{N}(z_0, V_n I_d)$
- Abbreviate $V_n := \sigma_1^2 + \cdots + \sigma_n^2$

Recall the training procedure:

- Sample $x = z_0 \sim \mathcal{D}$ and time step n
- Sample $z_0 \rightsquigarrow \cdots \rightsquigarrow z_{n-1} \rightsquigarrow z_n$ following q
- Maximize $\log p(z_{n-1} \mid z_n; \theta) = \mathcal{N}(\mu_n^\theta(z_n), \sigma_n^2)$

The denoising mean: Setup



Some setup:

- $q(z_n \mid z_{n-1}) = \mathcal{N}(z_n; z_{n-1}, \sigma_n^2 I_d)$ (or: $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$)
- $[z_n \mid z_0] \sim \mathcal{N}(z_0, (\sigma_1^2 + \cdots + \sigma_n^2) I_d) = \mathcal{N}(z_0, V_n I_d)$
- Abbreviate $V_n := \sigma_1^2 + \cdots + \sigma_n^2$

Recall the training procedure:

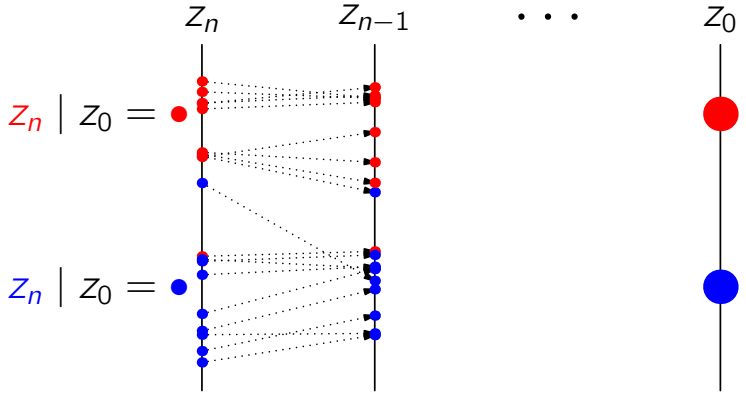
- Sample $x = z_0 \sim \mathcal{D}$ and time step n
- Sample $z_0 \rightsquigarrow \cdots \rightsquigarrow z_{n-1} \rightsquigarrow z_n$ following q
- Maximize $\log p(z_{n-1} \mid z_n; \theta) = \mathcal{N}(\mu_n^\theta(z_n), \sigma_n^2)$

The optimal value for the denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$!

The denoising mean: Simplification

Recall:

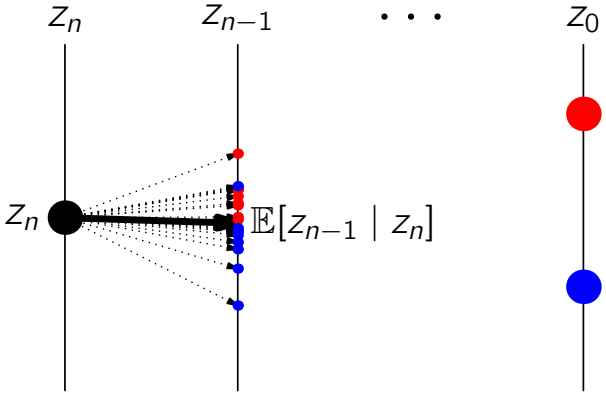
- $z_0 \sim \mathcal{D}$, $z_{n-1} = z_0 + \mathcal{N}(0, V_{n-1} I_d)$, $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$.
- The optimal denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$.



The denoising mean: Simplification

Recall:

- $z_0 \sim \mathcal{D}$, $z_{n-1} = z_0 + \mathcal{N}(0, V_{n-1} I_d)$, $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$.
- The optimal denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$.



The denoising mean: Simplification

Recall:

- $z_0 \sim \mathcal{D}$, $z_{n-1} = z_0 + \mathcal{N}(0, V_{n-1} I_d)$, $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$.
- The optimal denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$.

$$\mathbb{E}[z_{n-1} \mid z_n] = \mathbb{E}_{z_0|z_n}\mathbb{E}[z_{n-1} \mid z_n, z_0] \qquad \text{(law of total expectation)}$$

The denoising mean: Simplification

Recall:

- $z_0 \sim \mathcal{D}$, $z_{n-1} = z_0 + \mathcal{N}(0, V_{n-1} I_d)$, $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$.
- The optimal denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$.

$$\begin{aligned}
\mathbb{E}[z_{n-1} \mid z_n] &= \mathbb{E}_{z_0|z_n}\mathbb{E}[z_{n-1} \mid z_n, z_0] && \text{(law of total expectation)} \\
&= \mathbb{E}_{z_0|z_n}\left[z_0 + \frac{V_{n-1}}{V_n}(z_n - z_0)\right] && \text{(Gaussian properties!)}
\end{aligned}$$

The denoising mean: Simplification

Recall:

- $z_0 \sim \mathcal{D}$, $z_{n-1} = z_0 + \mathcal{N}(0, V_{n-1} I_d)$, $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$.
- The optimal denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$.

$$
\begin{aligned}
\mathbb{E}[z_{n-1} \mid z_n] &= \mathbb{E}_{z_0|z_n} \mathbb{E}[z_{n-1} \mid z_n, z_0] && \text{(law of total expectation)} \\
&= \mathbb{E}_{z_0|z_n} \left[z_0 + \frac{V_{n-1}}{V_n} (z_n - z_0) \right] && \text{(Gaussian properties!)} \\
&= \frac{V_{n-1}}{V_n} z_n + \frac{\sigma_n^2}{V_n} \mathbb{E}[z_0 \mid z_n] && \text{(linearity of expectation)}
\end{aligned}
$$

The denoising mean: Simplification

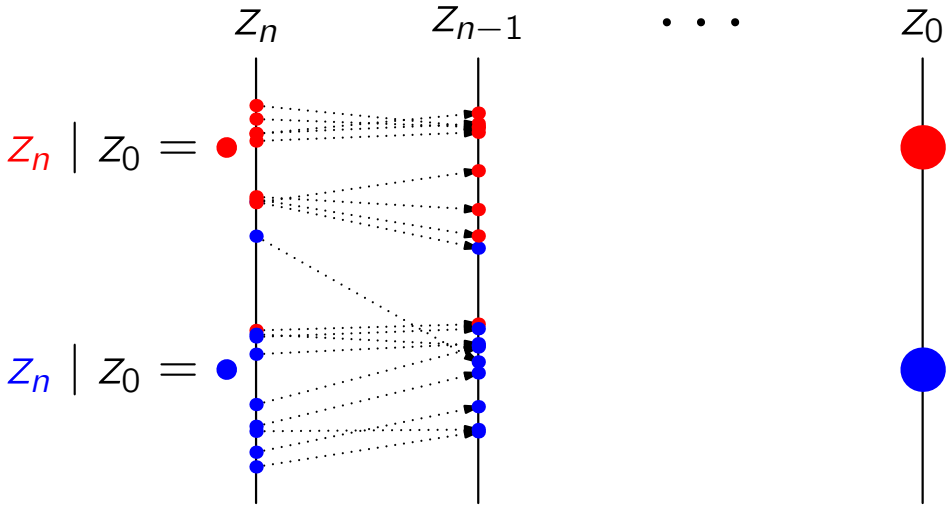
Recall:

- $z_0 \sim \mathcal{D}$, $z_{n-1} = z_0 + \mathcal{N}(0, V_{n-1} I_d)$, $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$.
- The optimal denoising mean $\mu_n^\theta(z_n)$ is $\mathbb{E}[z_{n-1} \mid z_n]$.

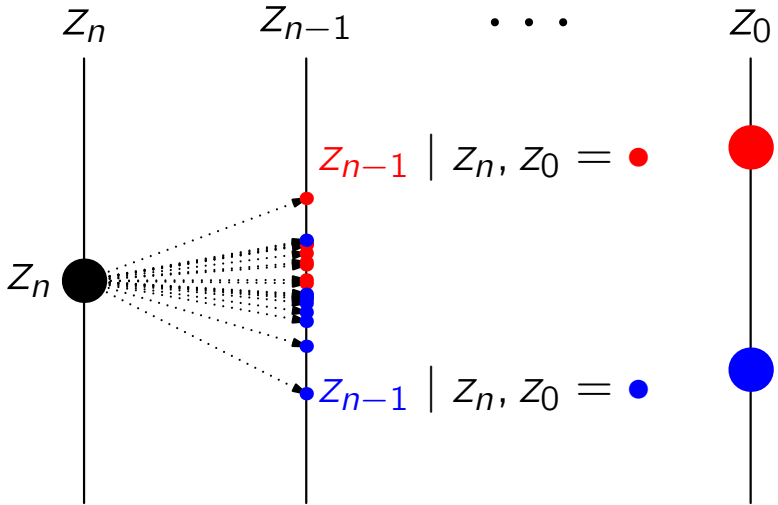
$$
\begin{aligned}
\mathbb{E}[z_{n-1} \mid z_n] &= \mathbb{E}_{z_0 | z_n} \mathbb{E}[z_{n-1} \mid z_n, z_0] && \text{(law of total expectation)} \\
&= \mathbb{E}_{z_0 | z_n} \left[z_0 + \frac{V_{n-1}}{V_n} (z_n - z_0) \right] && \text{(Gaussian properties!)} \\
&= \frac{V_{n-1}}{V_n} z_n + \frac{\sigma_n^2}{V_n} \mathbb{E}[z_0 \mid z_n] && \text{(linearity of expectation)}
\end{aligned}
$$

To estimate $\mathbb{E}[z_{n-1} \mid z_n]$, we just need to estimate $\mathbb{E}[z_0 \mid z_n]$!

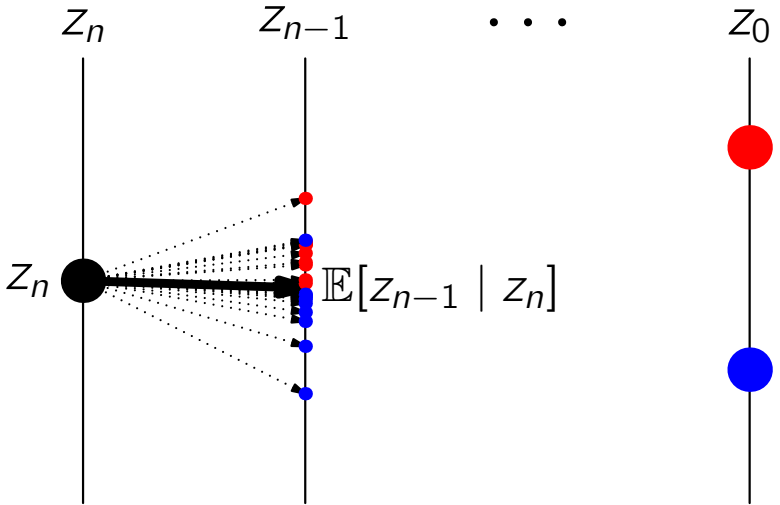
The denoising mean: Simplification



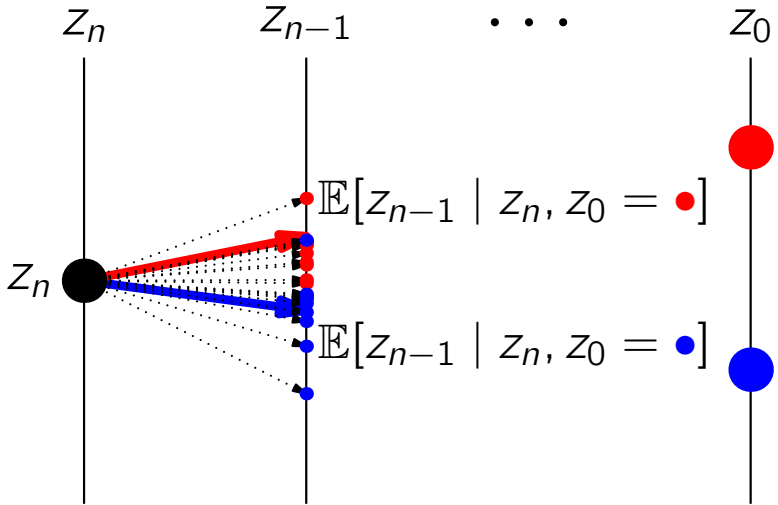
The denoising mean: Simplification



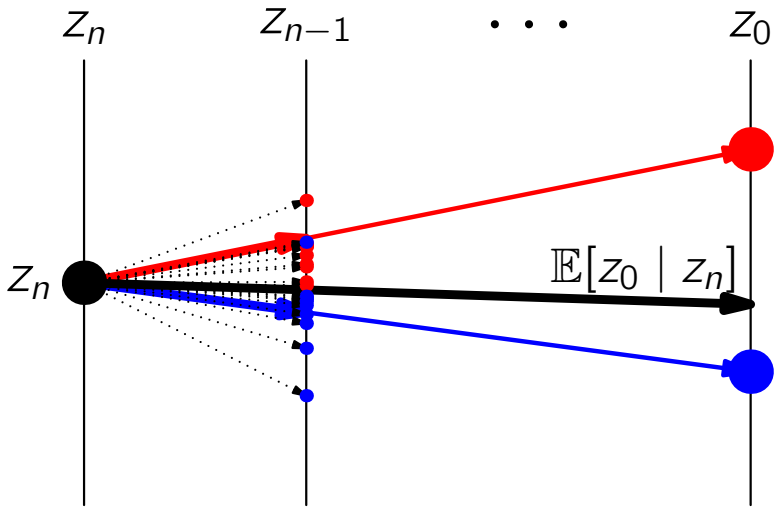
The denoising mean: Simplification



The denoising mean: Simplification



The denoising mean: Simplification



The denoising mean: Estimation

To estimate $\mathbb{E}[z_{n-1} \mid z_n]$, we just need to estimate $\mathbb{E}[z_0 \mid z_n]$!

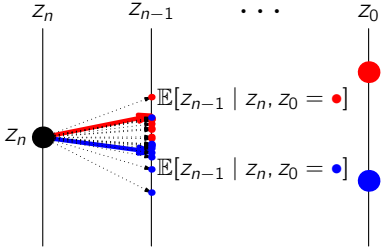
- Sample $z_0 \sim \mathcal{D}$ and time step n
- Sample $z_n = z_0 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, V_n I_d)$
- Gradient step on $\|\mu_n^\theta(z_n) - z_0\|^2$

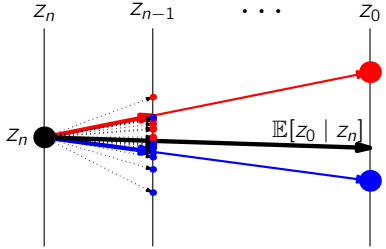
The denoising mean: Estimation

To estimate $\mathbb{E}[z_{n-1} \mid z_n]$, we just need to estimate $\mathbb{E}[z_0 \mid z_n]$!

- Sample $z_0 \sim \mathcal{D}$ and time step n
- Sample $z_n = z_0 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, V_n I_d)$
- Gradient step on $\|\mu_n^\theta(z_n) - z_0\|^2$

Stochastic regression: Each z_n can appear with many z_0, with probability proportional to $q(z_n \mid z_0)$.



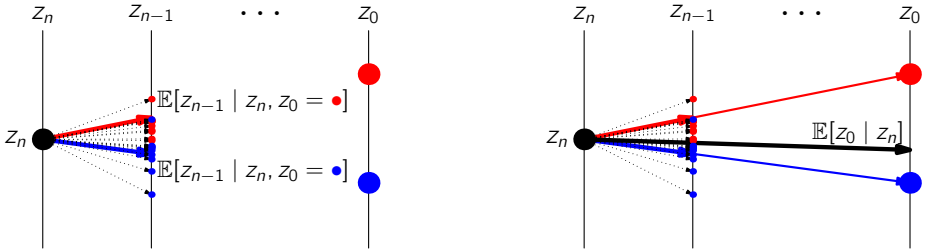


The denoising mean: Estimation

To estimate $\mathbb{E}[z_{n-1} \mid z_n]$, we just need to estimate $\mathbb{E}[z_0 \mid z_n]$!

- Sample $z_0 \sim \mathcal{D}$ and time step n
- Sample $z_n = z_0 + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, V_n I_d)$
- Gradient step on $\|\mu_n^\theta(z_n) - z_0\|^2$

Stochastic regression: Each z_n can appear with many z_0, with probability proportional to $q(z_n \mid z_0)$.



Alternatively, predict $\varepsilon = z_n - z_0$, or $\varepsilon/\sqrt{V_n}$, etc.

- Review of Lecture 8
- A three-way connection
  - Reversal of diffusion is denoising (second magic property)
  - Denoising is score matching
- Conditioning and guidance
- Unsimplifications

Review of Gaussians II

A Gaussian in $\mathbb{R}^d$ with mean μ and spherical variance $\sigma^2 I_d$ has density

$$\mathcal{N}(z; \mu, \sigma^2 I_d) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|z-\mu\|^2\right)$$

Review of Gaussians II

A Gaussian in $\mathbb{R}^d$ with mean μ and spherical variance $\sigma^2 I_d$ has density

$$\mathcal{N}(z; \mu, \sigma^2 I_d) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|z-\mu\|^2\right)$$

The (Stein) score of a density p is $\nabla \log p$; for a Gaussian:

$$\nabla_z \log \mathcal{N}(z; \mu, \sigma^2 I_d) = \nabla_z \left(-\frac{1}{2\sigma^2}\|z-\mu\|^2 + (\text{terms without } z)\right) = -\frac{z-\mu}{\sigma^2}$$

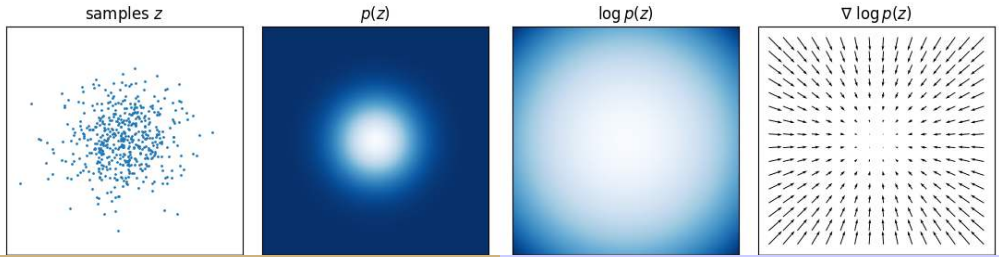
Review of Gaussians II

A Gaussian in $\mathbb{R}^d$ with mean μ and spherical variance $\sigma^2 I_d$ has density

$$\mathcal{N}(z; \mu, \sigma^2 I_d) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{1}{2\sigma^2}\|z-\mu\|^2\right)$$

The (Stein) score of a density p is $\nabla \log p$; for a Gaussian:

$$\nabla_z \log \mathcal{N}(z; \mu, \sigma^2 I_d) = \nabla_z \left(-\frac{1}{2\sigma^2}\|z-\mu\|^2 + (\text{terms without } z)\right) = -\frac{z-\mu}{\sigma^2}$$



A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\nabla_z \log p_n(z_n)$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\nabla_z \log p_n(z_n) = \frac{\nabla p_n(z_n)}{p_n(z_n)}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\nabla_z \log p_n(z_n) = \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\begin{aligned} \nabla_z \log p_n(z_n) &= \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)} \\ &= \sum_{z_0 \in \mathcal{D}} \underbrace{\frac{\frac{1}{|\mathcal{D}|} \mathcal{N}(z_n \mid z_0; V_n I_d)}{p_n(z_n)}}_{} \underbrace{\nabla \log \mathcal{N}(z_n; z_0; V_n I_d)}_{} \end{aligned}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\begin{aligned}
\nabla_z \log p_n(z_n) &= \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)} \\
&= \sum_{z_0 \in \mathcal{D}} \underbrace{\frac{\frac{1}{|\mathcal{D}|} \mathcal{N}(z_n \mid z_0; V_n I_d)}{p_n(z_n)}}_{\text{posterior over } z_0 \text{ given } z_n} \underbrace{\nabla \log \mathcal{N}(z_n; z_0; V_n I_d)}
\end{aligned}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\nabla_z \log p_n(z_n) = \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)}$$

$$= \sum_{z_0 \in \mathcal{D}} \underbrace{\frac{\frac{1}{|\mathcal{D}|} \mathcal{N}(z_n \mid z_0; V_n I_d)}{p_n(z_n)}}_{\text{posterior over } z_0 \text{ given } z_n} \underbrace{\nabla \log \mathcal{N}(z_n; z_0; V_n I_d)}_{\text{score of one Gaussian}}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\begin{aligned}
\nabla_z \log p_n(z_n) &= \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)} \\
&= \sum_{z_0 \in \mathcal{D}} \underbrace{\frac{\frac{1}{|\mathcal{D}|} \mathcal{N}(z_n \mid z_0; V_n I_d)}{p_n(z_n)}}_{\text{posterior over } z_0 \text{ given } z_n} \underbrace{\nabla \log \mathcal{N}(z_n; z_0; V_n I_d)}_{\text{score of one Gaussian}} \\
&= \mathbb{E}_{z_0 | z_n} \left[\nabla \log \mathcal{N}(z_n; z_0; V_n I_d) \right]
\end{aligned}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\begin{aligned}
\nabla_z \log p_n(z_n) &= \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)} \\
&= \sum_{z_0 \in \mathcal{D}} \underbrace{\frac{\frac{1}{|\mathcal{D}|} \mathcal{N}(z_n \mid z_0; V_n I_d)}{p_n(z_n)}}_{\text{posterior over } z_0 \text{ given } z_n} \underbrace{\nabla \log \mathcal{N}(z_n; z_0; V_n I_d)}_{\text{score of one Gaussian}} \\
&= \mathbb{E}_{z_0 | z_n} \left[\nabla \log \mathcal{N}(z_n; z_0; V_n I_d) \right] \\
&= \mathbb{E}_{z_0 | z_n} \left[\frac{z_0 - z_n}{V_n} \right]
\end{aligned}$$

A magic property of scores of mixtures

If $z_0 \sim \mathcal{D}$ and $z_n = z_0 + \mathcal{N}(0, V_n I_d)$, then z_n has a mixture density:

$$p_n(z_n) = \frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \mathcal{N}(z_n; z_0, V_n I_d);$$

Now some algebra:

$$\begin{aligned}
\nabla_z \log p_n(z_n) &= \frac{\nabla p_n(z_n)}{p_n(z_n)} = \frac{\frac{1}{|\mathcal{D}|} \sum_{z_0 \in \mathcal{D}} \nabla \mathcal{N}(z_n; z_0; V_n I_d)}{p_n(z_n)} \\
&= \sum_{z_0 \in \mathcal{D}} \underbrace{\frac{\frac{1}{|\mathcal{D}|} \mathcal{N}(z_n \mid z_0; V_n I_d)}{p_n(z_n)}}_{\text{posterior over } z_0 \text{ given } z_n} \underbrace{\nabla \log \mathcal{N}(z_n; z_0; V_n I_d)}_{\text{score of one Gaussian}} \\
&= \mathbb{E}_{z_0 | z_n} \left[\nabla \log \mathcal{N}(z_n; z_0; V_n I_d) \right] \\
&= \mathbb{E}_{z_0 | z_n} \left[\frac{z_0 - z_n}{V_n} \right] = \frac{1}{V_n} \left(\mathbb{E}[z_0 \mid z_n] - z_n \right) \qquad \text{(familiar?)}
\end{aligned}$$

Denoising is score matching

We have just shown:

$$\mathbb{E}[z_{n-1} \mid z_n] = \frac{1}{V_n}\left(V_{n-1} z_n + \sigma_n^2 \mathbb{E}[z_0 \mid z_n]\right)$$

$$\nabla_z \log p(z_n) = \frac{1}{V_n}\left(\mathbb{E}[z_0 \mid z_n] - z_n\right)$$

Denoising is score matching

We have just shown:

$$\mathbb{E}[z_{n-1} \mid z_n] = \frac{1}{V_n}\left(V_{n-1} z_n + \sigma_n^2 \mathbb{E}[z_0 \mid z_n]\right)$$

$$\nabla_z \log p(z_n) = \frac{1}{V_n}\left(\mathbb{E}[z_0 \mid z_n] - z_n\right)$$

- Learning to denoise one step $\longleftrightarrow$ learning to denoise all the way $\longleftrightarrow$ learning the score of the noised data

Denoising is score matching

We have just shown:

$$\mathbb{E}[z_{n-1} \mid z_n] = \frac{1}{V_n}\left(V_{n-1} z_n + \sigma_n^2 \mathbb{E}[z_0 \mid z_n]\right)$$

$$\nabla_z \log p(z_n) = \frac{1}{V_n}\left(\mathbb{E}[z_0 \mid z_n] - z_n\right)$$

- Learning to denoise one step $\longleftrightarrow$ learning to denoise all the way $\longleftrightarrow$ learning the score of the noised data
- Diffusion models are denoising autoencoders

Denoising is score matching

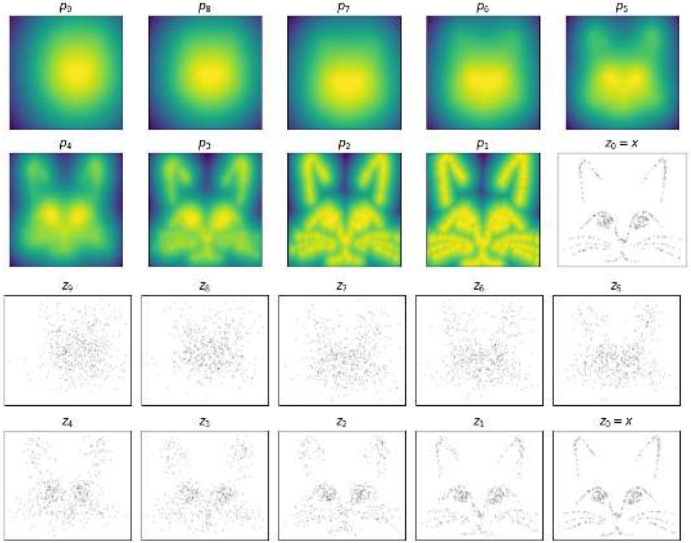
We have just shown:

$$\mathbb{E}[z_{n-1} \mid z_n] = \frac{1}{V_n}\left(V_{n-1} z_n + \sigma_n^2 \mathbb{E}[z_0 \mid z_n]\right)$$

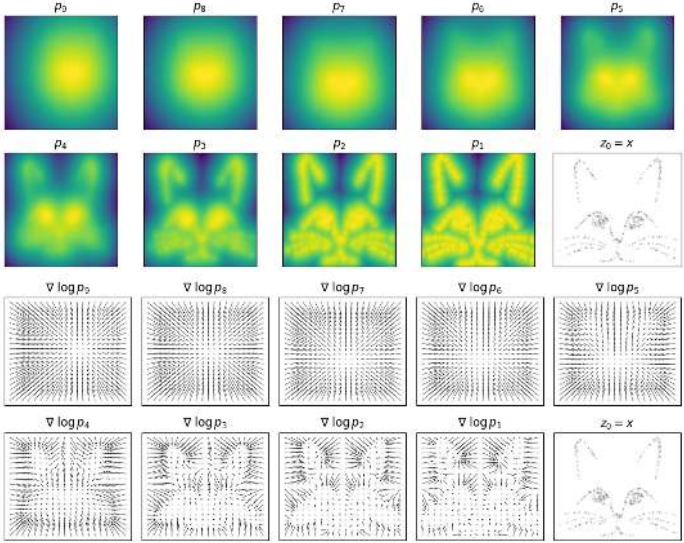
$$\nabla_z \log p(z_n) = \frac{1}{V_n}\left(\mathbb{E}[z_0 \mid z_n] - z_n\right)$$

- Learning to denoise one step $\longleftrightarrow$ learning to denoise all the way $\longleftrightarrow$ learning the score of the noised data
- Diffusion models are denoising autoencoders
- This is important in the continuous-time case (next time)...

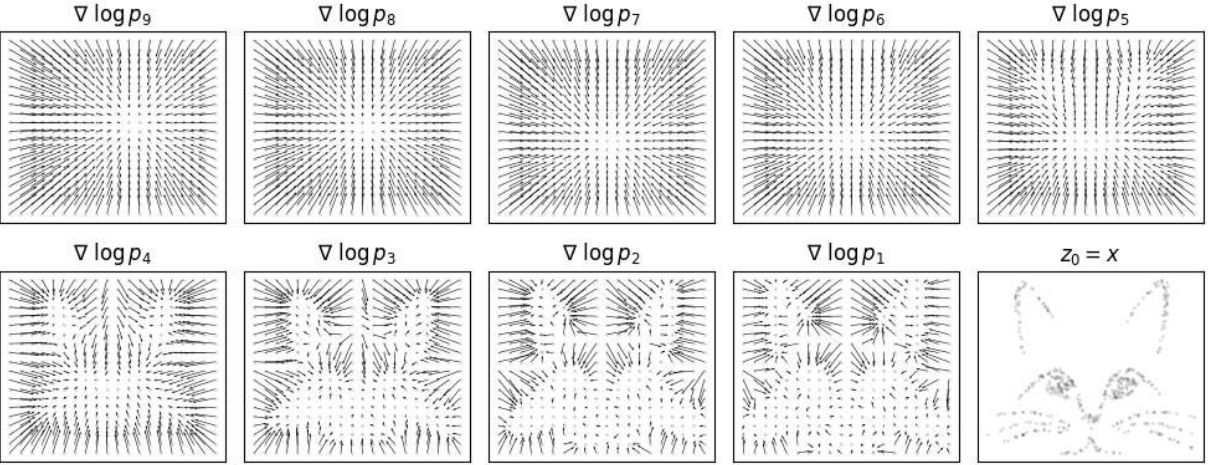
More than one way to score a (2D) cat



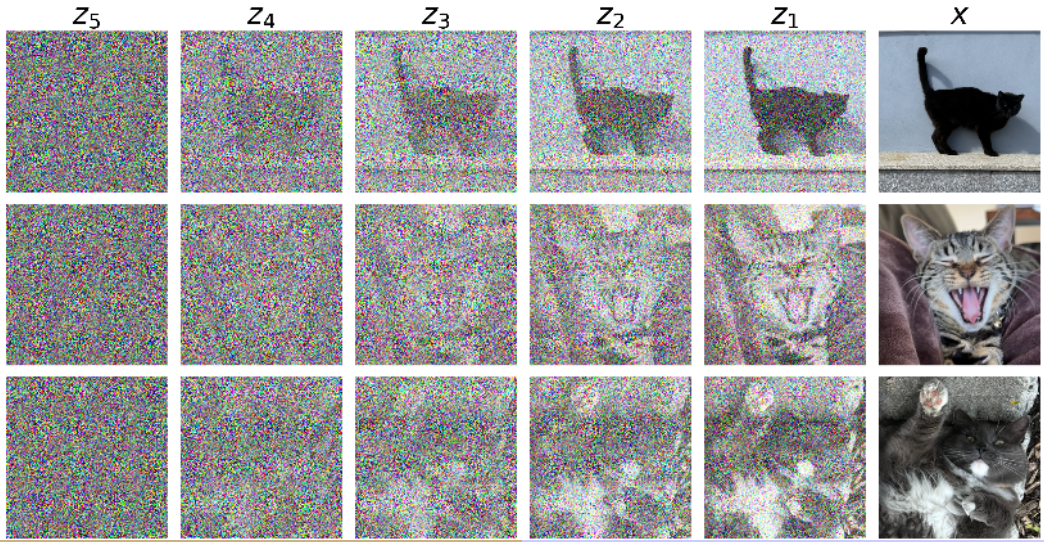
More than one way to score a (2D) cat



More than one way to score a (2D) cat



More than one way to score a real cat



- Review of Lecture 8
- A three-way connection
  - Reversal of diffusion is denoising (second magic property)
  - Denoising is score matching
- Conditioning and guidance
- Unsimplifications

Conditional generation with diffusion models

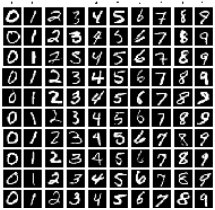
A denoiser can be conditioned on some context y: $p(z_{n-1} \mid z_n, y; \theta)$

DALL-E 3
$y =$ "Edinburgh from Calton Hill, pointillist style"

Conditional generation with diffusion models

A denoiser can be conditioned on some context y: $p(z_{n-1} \mid z_n, y; \theta)$

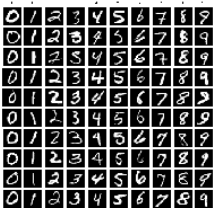
DALL-E 3
$y =$ "Edinburgh from Calton Hill, pointillist style"



Conditional generation with diffusion models

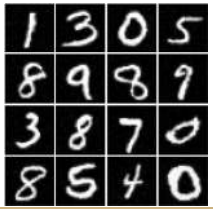
A denoiser can be conditioned on some context y: $p(z_{n-1} \mid z_n, y; \theta)$

DALL-E 3
y="Edinburgh from Calton Hill, pointillist style"



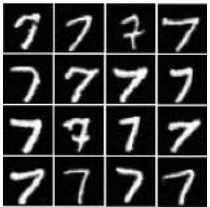
But can we condition an existing diffusion model, given a classifier?

diffusion model



$+$ classifier $p(7 \mid x)$ $\rightarrow$

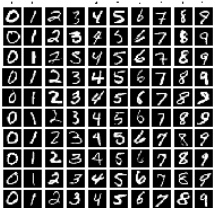
conditional samples



Conditional generation with diffusion models

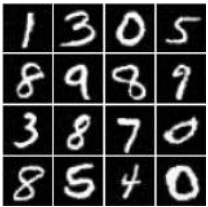
A denoiser can be conditioned on some context y: $p(z_{n-1} \mid z_n, y; \theta)$

DALL-E 3
$y =$ "Edinburgh from Calton Hill, pointillist style"



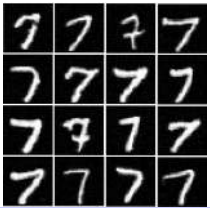
But can we condition an existing diffusion model, given a classifier? Yes, but...

diffusion model



$+$ classifier $p(7 \mid x)$ $\rightarrow$

conditional samples



Classifier guidance

Bayes' rule in logarithmic form:

$$\log p(z \mid y) = \log p(z) + \log p(y \mid z) - \log p(y)$$

Classifier guidance

Bayes' rule in logarithmic form:

$$\log p(z \mid y) = \log p(z) + \log p(y \mid z) - \log p(y)$$

Bayes' rule for the score:

$$\nabla_z \log p(z \mid y) = \nabla_z \log p(z) + \nabla_z \log p(y \mid z)$$

Classifier guidance

Bayes' rule in logarithmic form:

$$\log p(z \mid y) = \log p(z) + \log p(y \mid z) - \log p(y)$$

Bayes' rule for the score:

$$\nabla_z \log p(z \mid y) = \nabla_z \log p(z) + \nabla_z \log p(y \mid z)$$

In particular, if z_n is noisy data and y is a label, we can use a classifier (likelihood of y given z_n) to guide the denoising:

$$\underbrace{\nabla_z \log p_n(z_n \mid y)}_{\text{conditional score}} = \underbrace{\nabla_z \log p_n(z_n)}_{\text{unconditional score}} + \underbrace{\nabla_z \log p_n(y \mid z_n)}_{\text{classifier gradient}}$$

Classifier guidance

Bayes' rule in logarithmic form:

$$\log p(z \mid y) = \log p(z) + \log p(y \mid z) - \log p(y)$$

Bayes' rule for the score:

$$\nabla_z \log p(z \mid y) = \nabla_z \log p(z) + \nabla_z \log p(y \mid z)$$

In particular, if z_n is noisy data and y is a label, we can use a classifier (likelihood of y given z_n) to guide the denoising:

$$\underbrace{\nabla_z \log p_n(z_n \mid y)}_{\text{conditional score}} = \underbrace{\nabla_z \log p_n(z_n)}_{\text{unconditional score}} + \underbrace{\nabla_z \log p_n(y \mid z_n)}_{\text{classifier gradient}}$$

- **NB!** Requires us to have classifiers $p_n(y \mid z_n)$ trained on **noised data** z_n, not just clean data $x = z_0$

Classifier guidance

Bayes' rule in logarithmic form:

$$\log p(z \mid y) = \log p(z) + \log p(y \mid z) - \log p(y)$$

Bayes' rule for the score:

$$\nabla_z \log p(z \mid y) = \nabla_z \log p(z) + \nabla_z \log p(y \mid z)$$

In particular, if z_n is noisy data and y is a label, we can use a classifier (likelihood of y given z_n) to guide the denoising:

$$\underbrace{\nabla_z \log p_n(z_n \mid y)}_{\text{conditional score}} = \underbrace{\nabla_z \log p_n(z_n)}_{\text{unconditional score}} + \underbrace{\nabla_z \log p_n(y \mid z_n)}_{\text{classifier gradient}}$$

- **NB!** Requires us to have classifiers $p_n(y \mid z_n)$ trained on **noised data** z_n, not just clean data $x = z_0$
- But given only $p(y \mid z_0)$ trained on clean data, conditioning is **intractable**

- Review of Lecture 8
- A three-way connection
  - Reversal of diffusion is denoising (second magic property)
  - Denoising is score matching
- Conditioning and guidance
- Unsimplifications

More general noising processes

Apologies are in order...

- We assumed $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$ (variance-exploding process)
- In practice, one often takes $z_n = c_n z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$, where c_n are constants (discrete-time Ornstein-Uhlenbeck process – next time!)

More general noising processes

Apologies are in order...

- We assumed $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$ (variance-exploding process)
- In practice, one often takes $z_n = c_n z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$, where c_n are constants (discrete-time Ornstein-Uhlenbeck process – next time!)

But this is reducible to the variance-exploding process by a change of variables:

$$z'_n := \frac{z_n}{c_n}$$

More general noising processes

Apologies are in order...

- We assumed $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$ (variance-exploding process)
- In practice, one often takes $z_n = c_n z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$, where c_n are constants (discrete-time Ornstein-Uhlenbeck process – next time!)

But this is reducible to the variance-exploding process by a change of variables:

$$z'_n := \frac{z_n}{c_n} \quad z'_n = z_{n-1} + \mathcal{N}(0, (\sigma_n^2 / c_n^2) I_d)$$

More general noising processes

Apologies are in order...

- We assumed $z_n = z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$ (variance-exploding process)
- In practice, one often takes $z_n = c_n z_{n-1} + \mathcal{N}(0, \sigma_n^2 I_d)$, where c_n are constants (discrete-time Ornstein-Uhlenbeck process – next time!)

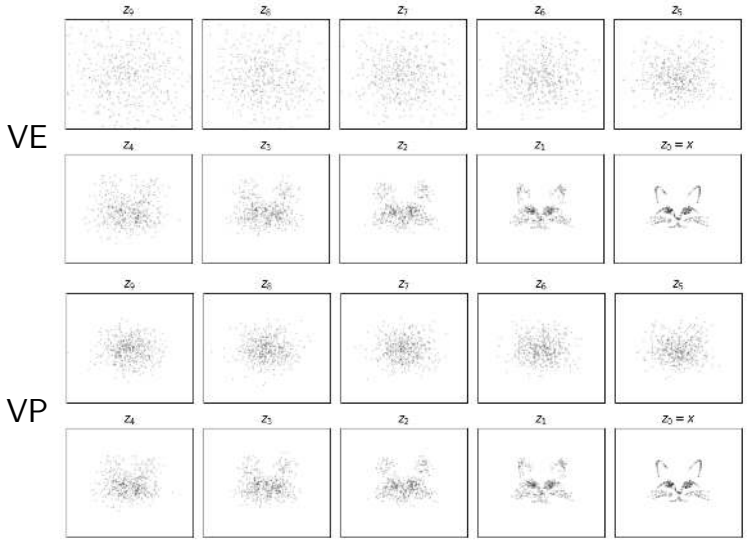
But this is reducible to the variance-exploding process by a change of variables:

$$z'_n := \frac{z_n}{c_n} \quad z'_n = z_{n-1} + \mathcal{N}(0, (\sigma_n^2/c_n^2) I_d)$$

A common choice is variance-preserving: $c_n = \sqrt{1 - \sigma_n^2}$.

- If z_{n-1}'s components have unit variance, so does z_n

More general noising processes



Conclusion and looking forward

Next time:

- To continuous time via score matching
  - Training and sampling with varying N

Conclusion and looking forward

Next time:

- To continuous time via score matching
  - Training and sampling with varying N
- The probability flow ODE
  - Likelihood estimation revisited

Conclusion and looking forward

Next time:

- To continuous time via score matching
  - Training and sampling with varying N
- The probability flow ODE
  - Likelihood estimation revisited
- If time: survey of flow matching and (Schrödinger) bridge models