भारतीय प्रौद्योगिकी संस्थान दिल्ली
Indian Institute of Technology Delhi

COV877
Special Module on Visual Computing

Generative AI for Visual Content Creation: Image, Video, and 3D

Score Based Diffusion Model

Instructor:

Dr. Lokender Tiwari

Research Scientist

Diffusion

Diffusion

There are several different ways to interpret diffusion models

- Score-Based
- DDPM, DDIM etc.

Score-Based Diffusion Model

- GAN
- Normalizing Flows
- VAE

Limitations

- strong restrictions on the model architecture to ensure a tractable normalizing constant for likelihood computation,
- rely on surrogate objectives to approximate maximum likelihood training (e.g. ELBO)
- require adversarial training, which is notoriously unstable and can lead to mode collapse (e.g. GAN)

New way to model probability density

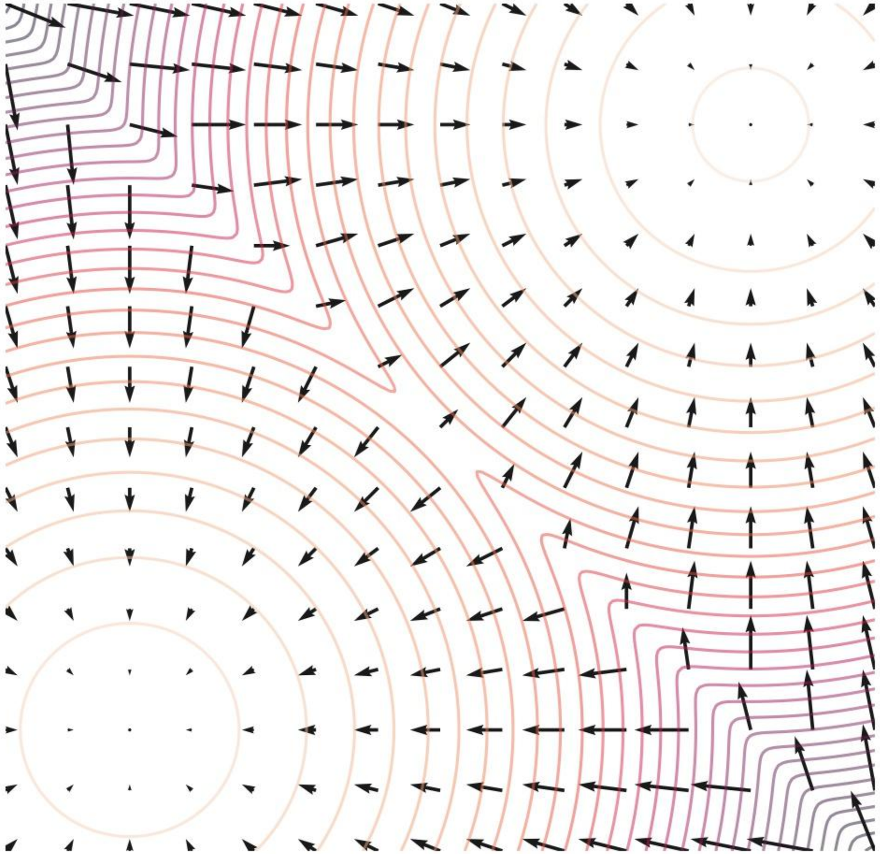
[1] https://yang-song.net/blog/2021/score/

Score-Based Diffusion Model

The key idea

- Model the gradient of the log probability density function, using quantity often known as the **score function**
- Score-based models doesn't required to have a tractable normalizing constant
  - can be directly learned by **score matching**

Score function (the vector field) and density function (contours) of a mixture of two Gaussians.



[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/

Score-Based Diffusion Model

Dataset $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ Independently drawn from data distribution $p(\mathbf{x})$

Generative Modeling : Fit a model to the data distribution so that we can generate new sample from it

How to model the probability distribution ? one way is likelihood-based modeling

real-valued function parametrized by learnable parameter

$f_\theta(\mathbf{x}) \in \mathbb{R}$

$$p_\theta(\mathbf{x}) = \frac{e^{-\boxed{f_\theta(\mathbf{x})}}}{Z_\theta}$$

also known as unnormalized probabilistic model or energy based model

$$Z_\theta > 0$$

$$\int p_\theta(\mathbf{x}) \mathrm{d}\mathbf{x} = 1$$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/

Score-Based Diffusion Model

We can train by maximizing the log-likelihood of the data

$$\max_{\theta} \sum_{i=1}^{N} \log \boxed{p_\theta(\mathbf{x}_i)}$$

- this needs to be normalized PDF
- requires normalizing constant (Intractable)

To make the likelihood training possible

- NF models restrict model architechtures
- VAE approximate the normalizing constant using variational inference in VAE or MCMC sampling

Score function removes the requirement of intractable normalizing constants

- Model the score function instead of density function

The *score function* of a distribution is given by

$$\nabla_{\mathbf{x}} \log p(\mathbf{x})$$

The model for the score-function is knowns as **score-based model**

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/

Score-Based Diffusion Model

The model for the score-function is knowns as **score-based model**

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) \qquad\qquad \mathbf{s}_\theta(\mathbf{x})$$

We learn score-based model that can be parametrized without worrying normalizing constant

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log p(\mathbf{x})$$

For example,

- independent of the normalizing constant

$$\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) = -\nabla_{\mathbf{x}} f_\theta(\mathbf{x}) - \underbrace{\nabla_{\mathbf{x}} \log Z_\theta}_{=0} = \boxed{-\nabla_{\mathbf{x}} f_\theta(\mathbf{x})}$$

$$p_\theta(\mathbf{x}) = \frac{e^{-f_\theta(\mathbf{x})}}{Z_\theta}$$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/

Score-Based Diffusion Model

We can train score-based models by minimizing the Fisher divergence

$$\mathbb{E}_{p(\mathbf{x})}[||\boxed{\nabla_{\mathbf{x}} \log p(\mathbf{x})} - \mathbf{s}_\theta(\mathbf{x})||_2^2]$$

Compare the distance between GT data score and the score-based model

- This is unknown

Solution ?
Score-matching - Family of methods that can minimize the Fisher divergence without GT data score

Score matching objective can *directly be estimated on a dataset* and optimized with stochastic gradient descent, similar to the log-likelihood objective

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] P. Vincent. , A connection between score matching and denoising autoencoders, Neural computation, Vol 23(7), pp. 1661--1674. MIT Press. 2011.
[4] Y. Song, S. Garg, J. Shi, S. Ermon., Sliced score matching: A scalable approach to density and score estimation, Uncertainty in Artificial Intelligence, pp. 574--584. 2020

Score-Based Diffusion Model

$$\mathbb{E}_{p(\mathbf{x})}[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2]$$

$$\mathbb{E}_{p_{\text{data}}}\left[\operatorname{tr}(\nabla_{\mathbf{x}}^2 \log p_\theta(\mathbf{x})) + \frac{1}{2}\|\nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})\|_2^2\right] + \text{const.}$$

General multi-dimensional case

$$L(\boldsymbol{\theta}) \triangleq \frac{1}{2}\mathbb{E}_{p_d}[\|\mathbf{s}_m(\mathbf{x};\boldsymbol{\theta}) - \mathbf{s}_d(\mathbf{x})\|_2^2]$$

$$L(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \mathrm{C}$$

Integration by parts (Full proof is in theorem 1 of [1])

$$J(\boldsymbol{\theta}) \triangleq \mathbb{E}_{p_d}\left[\operatorname{tr}(\nabla_{\mathbf{x}}\mathbf{s}_m(\mathbf{x};\boldsymbol{\theta})) + \frac{1}{2}\|\mathbf{s}_m(\mathbf{x};\boldsymbol{\theta})\|_2^2\right]$$

Estimation of Non-Normalized Statistical Models by Score Matching

Aapo Hyvärinen AAPO.HYVARINEN@HELSINKI.FI
Helsinki Institute for Information Technology (BRU)
Department of Computer Science
FIN-00014 University of Helsinki, Finland

Editor: Peter Dayan

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] https://yang-song.net/blog/2019/ssm/
[4] Y. Song, S. Garg, J. Shi, S. Ermon., Sliced score matching: A scalable approach to density and score estimation, Uncertainty in Artificial Intelligence, pp. 574--584. 2020

Score-Based Diffusion Model - How to sample ?

Langevin Dynamics

- Provides an MCMC procedure to sample from a distribution using only its score function
- Trained score-based model $\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$
- Start from an arbitrary prior distribution and then iteratively generate

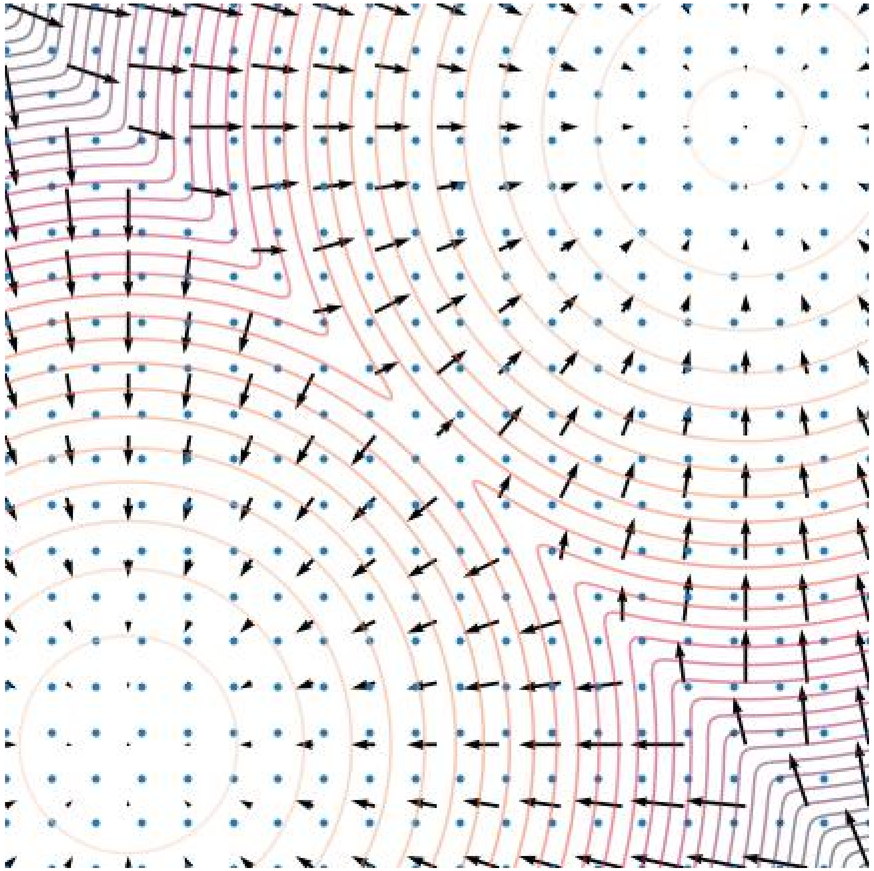
$$\mathbf{x}_0 \sim \pi(\mathbf{x})$$

$$\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \epsilon \nabla_\mathbf{x} \log p(\mathbf{x}) + \sqrt{2\epsilon}\, \mathbf{z}_i, \quad i = 0, 1, \cdots, K$$

$$\mathbf{z}_i \sim \mathcal{N}(0, I)$$

$$\epsilon \rightarrow 0$$

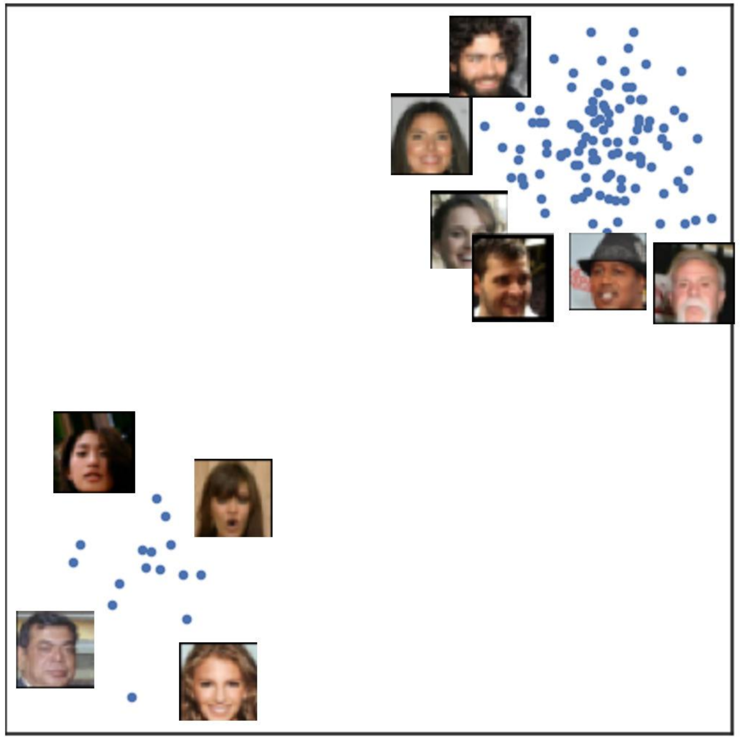
$$K \rightarrow \infty$$



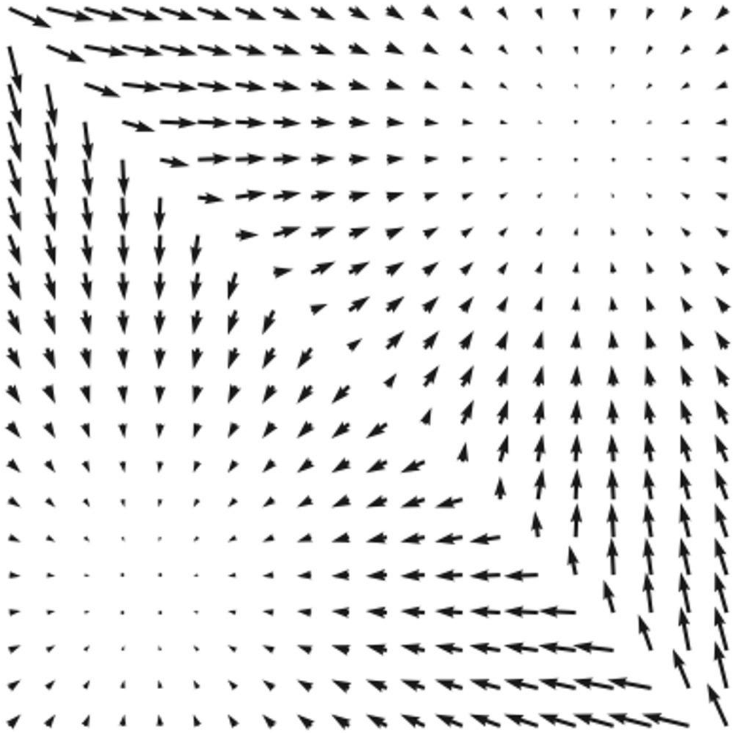
Langevin Dynamics have access to only learned score function not the actualy PDF

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] https://yang-song.net/blog/2019/ssm/
[4] Y. Song, S. Garg, J. Shi, S. Ermon., Sliced score matching: A scalable approach to density and score estimation, Uncertainty in Artificial Intelligence, pp. 574--584. 2020

Score-Based Diffusion Model - Overall Appraoch



score matching



Langevin dynamics

Data samples

$$\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\} \overset{\text{i.i.d.}}{\sim} p(\mathbf{x})$$

Scores

$$\mathbf{s}_\theta(\mathbf{x}) \approx \nabla_\mathbf{x} \log p(\mathbf{x})$$

New samples

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] https://yang-song.net/blog/2019/ssm/
[4] Y. Song, S. Garg, J. Shi, S. Ermon., Sliced score matching: A scalable approach to density and score estimation, Uncertainty in Artificial Intelligence, pp. 574--584. 2020

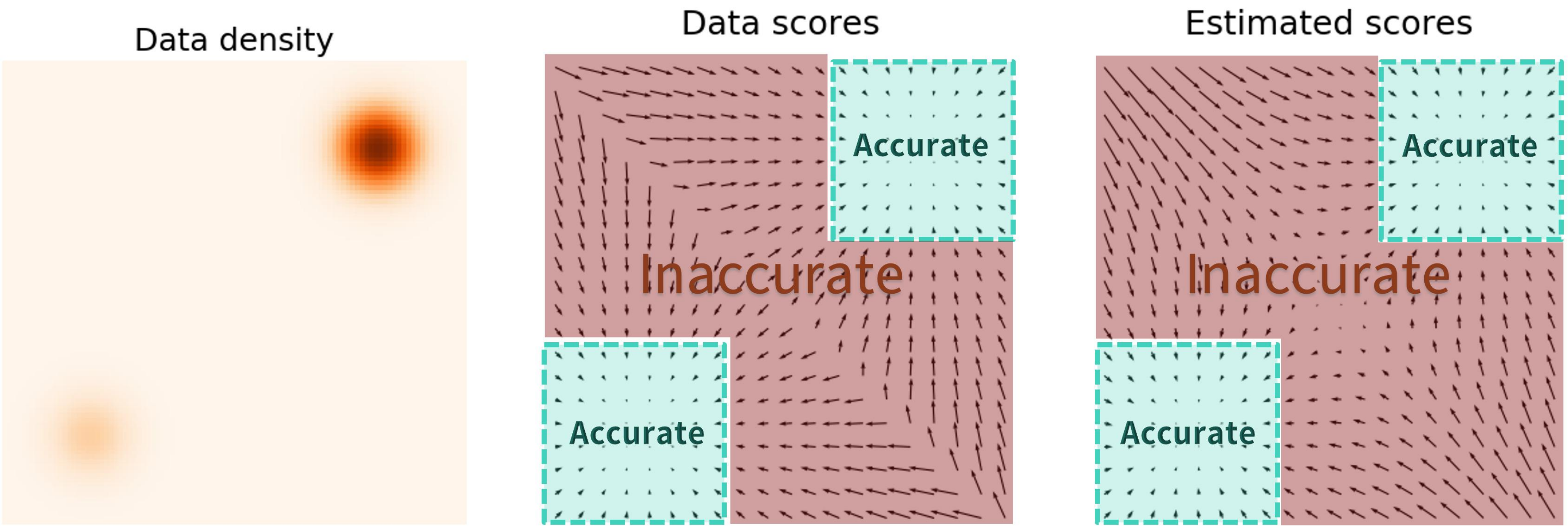
Score-Based Diffusion Model - Issues

Naive formulation ignores the low density region

$$\mathbb{E}_{p(\mathbf{x})}[\|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2] = \int \boxed{p(\mathbf{x})} \|\nabla_{\mathbf{x}} \log p(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x})\|_2^2 \mathrm{d}\mathbf{x}.$$

- Weighted

- If initial sample is in low density regions, an inaccurate score-based model will derail Langevin dynamics
  - Generated sample will not be of high quality



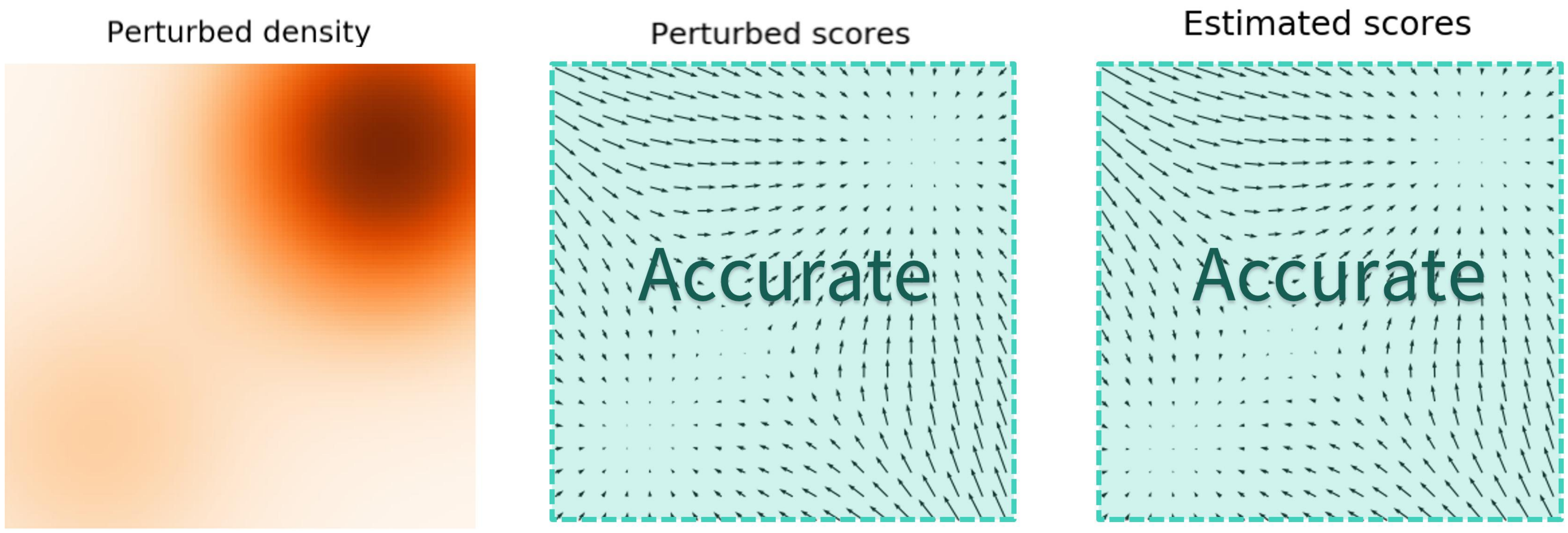
Estimated scores are accurate only in high density regions

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] https://yang-song.net/blog/2019/ssm/
[4] Y. Song, S. Garg, J. Shi, S. Ermon., Sliced score matching: A scalable approach to density and score estimation, Uncertainty in Artificial Intelligence, pp. 574--584. 2020

Score-Based Diffusion Model - Noise conditioning

Solution ?

- **Perturb** the data points with noise, and use them to train the score-based models



What is the right amount of noise ?

High Noise : can cover more low density regions, over corrupts data (distroy original distribution)

Low Noise : may not cover all low density regions, roughly retains original distribution

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] https://yang-song.net/blog/2019/ssm/
[4] Y. Song, S. Garg, J. Shi, S. Ermon., Sliced score matching: A scalable approach to density and score estimation, Uncertainty in Artificial Intelligence, pp. 574--584. 2020

Score-Based Diffusion Model - Noise conditioning

Solution ?

- Use multiple scale of noise perturbations

$\sigma_1 < \sigma_2 < \cdots < \sigma_L$ Multi scale noise

$p_{\sigma_i}(\mathbf{x})$ Perturbed data distribution

$\mathbf{x} \sim p(\mathbf{x})$ $\quad \mathbf{x} + \sigma_i \mathbf{z} \quad \mathbf{z} \sim \mathcal{N}(0, I)$ sampling from perturbed distribution

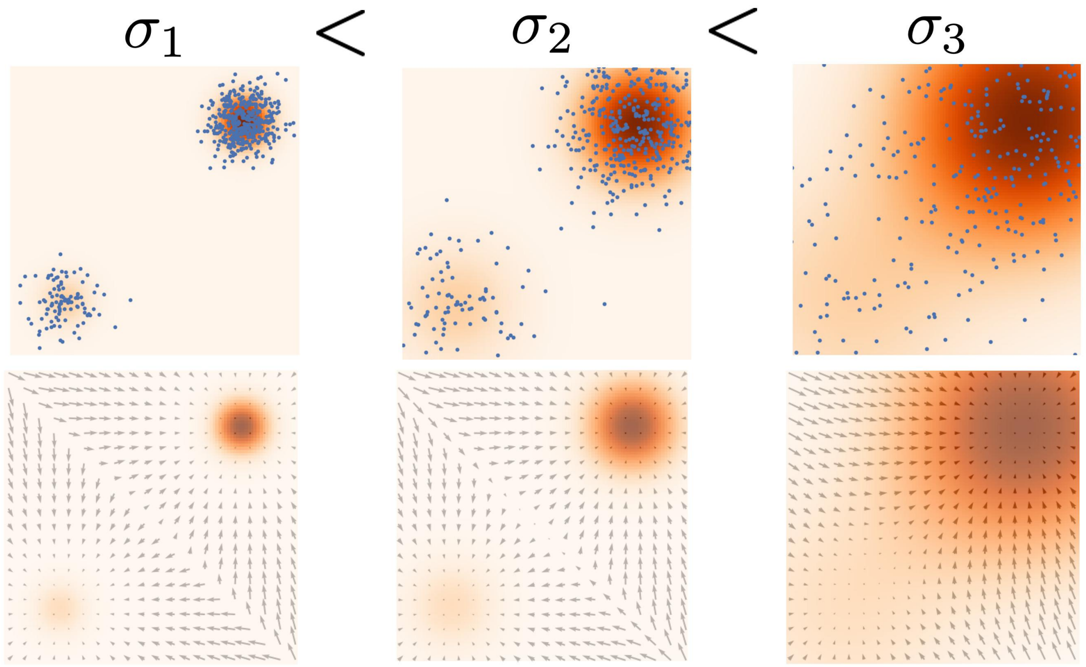
$\mathbf{s}_\theta(\mathbf{x}, i) \approx \nabla_\mathbf{x} \log p_{\sigma_i}(\mathbf{x})$ Noise Conditional Score-Based Model [3]

$$\sum_{i=1}^{L} \lambda(i) \mathbb{E}_{p_{\sigma_i}(\mathbf{x})}[\|\nabla_\mathbf{x} \log p_{\sigma_i}(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x}, i)\|_2^2]$$

Weighted sum of Fisher divergences for all noise scales

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems

Score-Based Diffusion Model - Noise conditioning



jointly estimate the score functions for all

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems

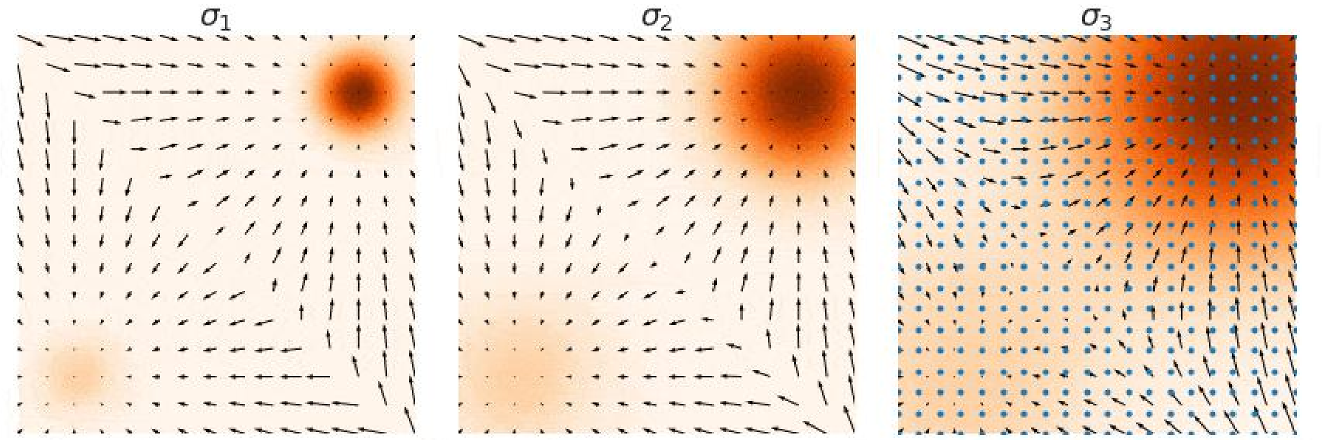
Score-Based Diffusion Model - Sampling from Noise Conditional Score Model

Annealed Langevin Dynamics

noise scale decreases (anneals) gradually over time

Note : In this algo σ_L is the smallest noise and σ_1 is the largest (opposite of what discussed in previous slides)

We first sample from noisiest score model using Langevin dynamics



Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T.$

Initialize $\tilde{\mathbf{x}}_0$

for $i \leftarrow 1$ to L **do**

$\quad \alpha_i \leftarrow \epsilon \cdot \sigma_i^2 / \sigma_L^2 \qquad \triangleright\ \alpha_i$ is the step size.

$\quad$ **for** $t \leftarrow 1$ to T **do**

$\quad\quad$ Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$

$$\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2} \mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\, \mathbf{z}_t$$

$\quad$ **end for**

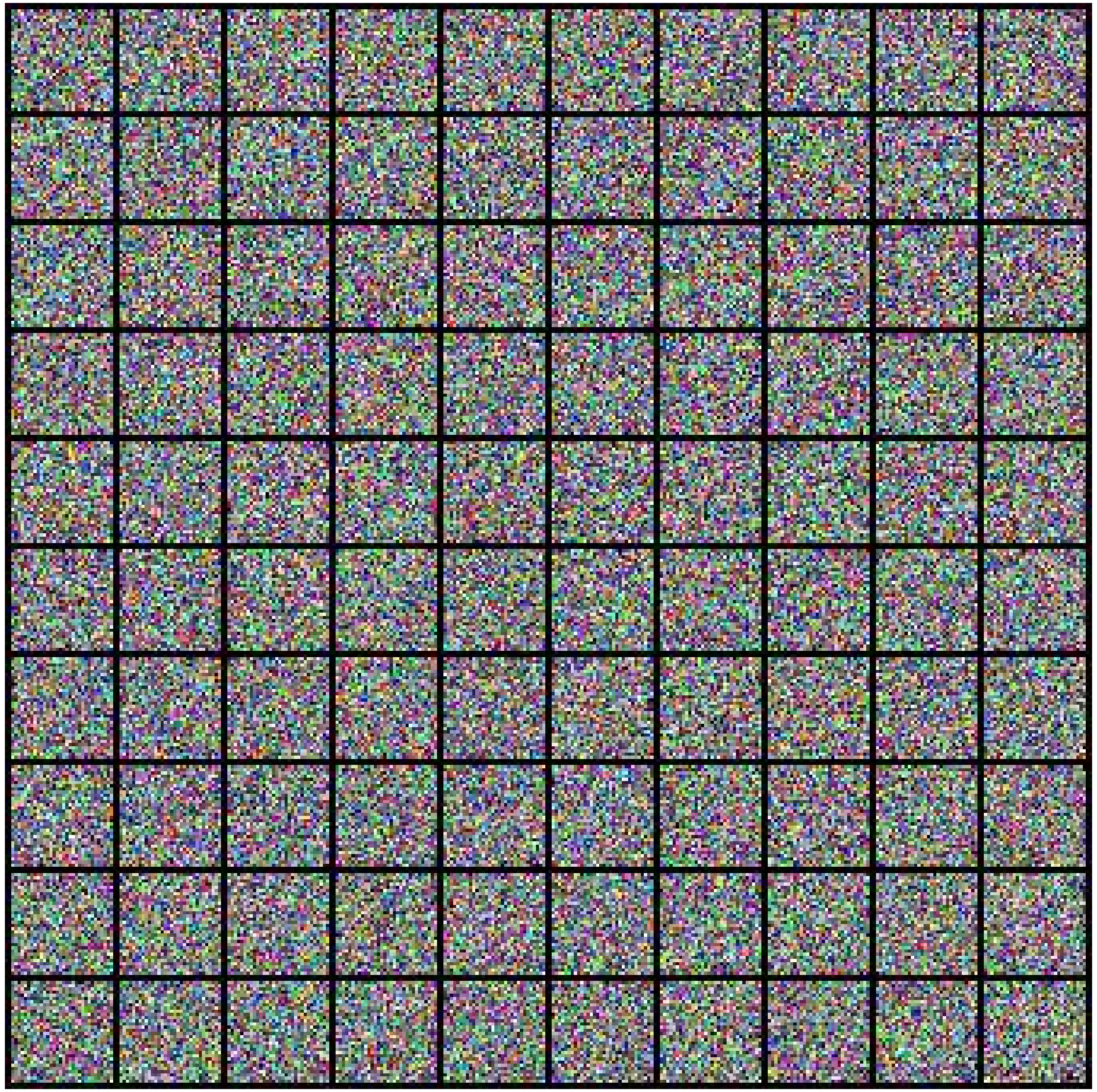
$\quad \tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$

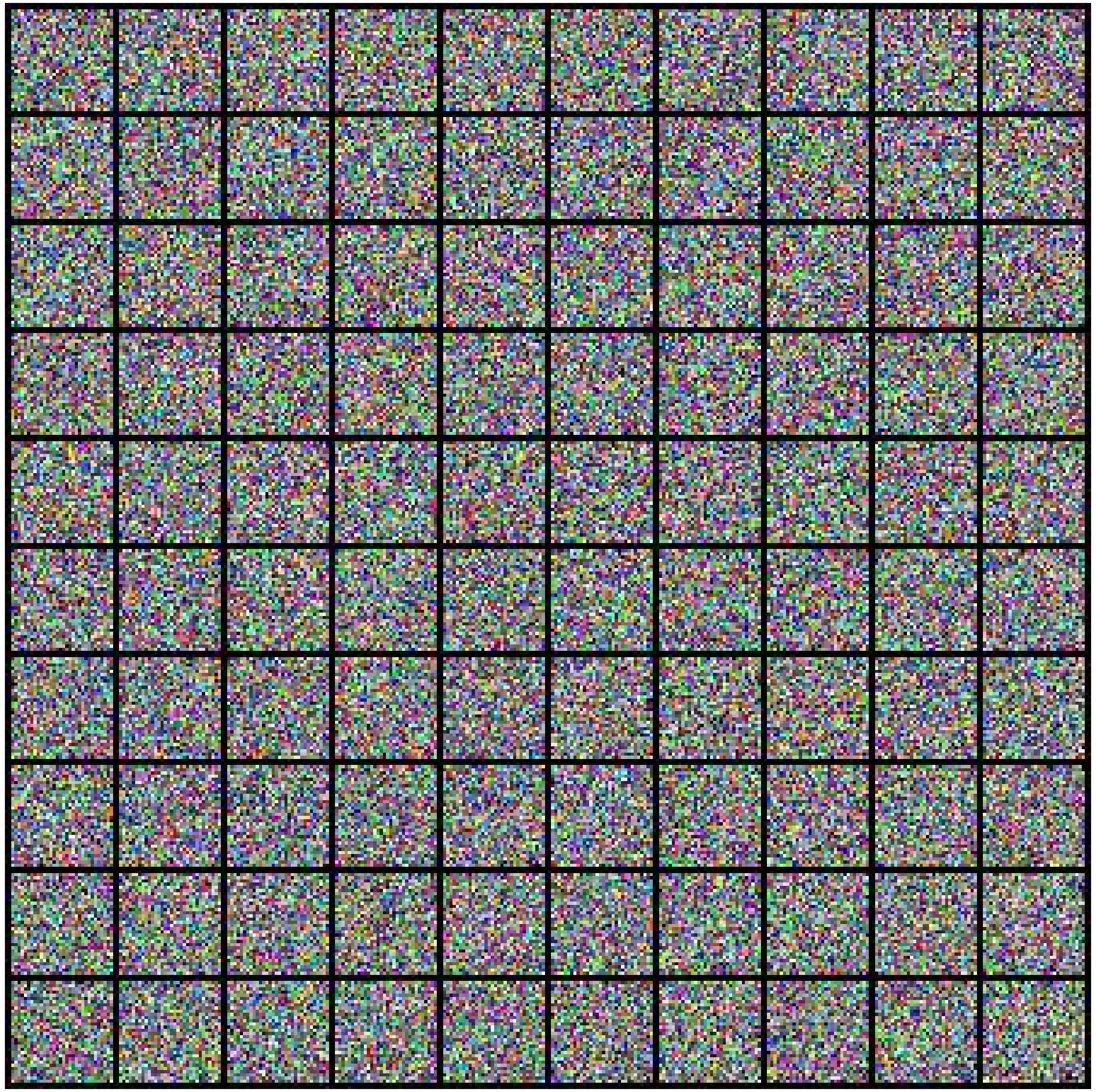
end for

return $\tilde{\mathbf{x}}_T$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems

Score-Based Diffusion Model - Results from NCSN [3]





[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems

Score-Based Diffusion Model - with SDE

- Noise conditional perturbation of data distribution is crucial

- If we increase the number of noise scales to large number (infinity)
  - The perturbation process become a continuous time stochastic process

- Stochastic processes (Diffusion process is one of them)
  - are solutions to stochastic differential equations (SDEs)

$g(t) \in \mathbb{R}$

- real-valued function
- infinitesimal white noise

$$\mathrm{d}\mathbf{x} = \boxed{\mathbf{f}(\mathbf{x}, t)}\mathrm{d}t + \boxed{g(t)}\boxed{\mathrm{d}\mathbf{w}}$$

- vector-valued function
  $\mathbf{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Score-Based Diffusion Model - with SDE

$g(t) \in \mathbb{R}$

- real-valued function
- infinitesimal white noise

$$\mathrm{d}\mathbf{x} = \boxed{\mathbf{f}(\mathbf{x}, t)}\mathrm{d}t + \boxed{g(t)}\boxed{\mathrm{d}\mathbf{w}}$$

- vector-valued function
  $\mathbf{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$

The solution to the SDE is the continous collection of random variable $\{\mathbf{x}(t)\}_{t\in[0,T]}$

Think of these as stochastic trajectories w.r.t the growing time index ***t*** from **0** to ***T***

$p_t(\mathbf{x})$ probability density function of $\mathbf{x}(t)$

$t \in [0, T]$ analogous to $i = 1, 2, \cdots, L$ of noise scales

$p_t(\mathbf{x}) = p_{\sigma_i}(\mathbf{x})$

$p_0(\mathbf{x}) = p(\mathbf{x})$ $p_T(\mathbf{x}) = p_{\sigma_L}(\mathbf{x})$ In the case of finite noise scale

$p_T(\mathbf{x}) = \pi(\mathbf{x})$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Score-Based Diffusion Model - SDE with Gaussian Noise

$g(t) \in \mathbb{R}$

- real-valued function
- infinitesimal white noise

$$\mathrm{d}\mathbf{x} = \boxed{\mathbf{f}(\mathbf{x}, t)}\mathrm{d}t + \boxed{g(t)}\boxed{\mathrm{d}\mathbf{w}}$$

- vector-valued function

$\mathbf{f}(\cdot, t) : \mathbb{R}^d \to \mathbb{R}^d$

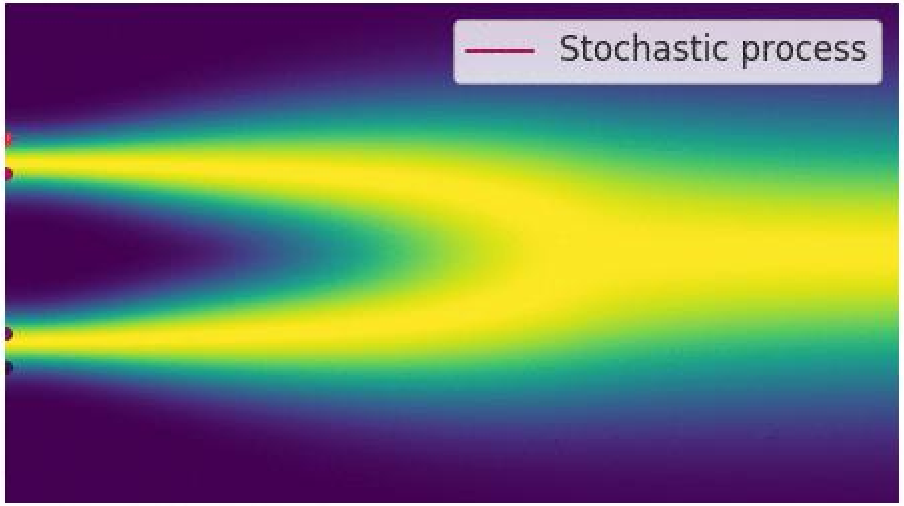
SDE to perturb data with a Gaussian noise of mean zero and exponentially growing variance analogous to $\mathcal{N}(0, \sigma_1^2 I), \mathcal{N}(0, \sigma_2^2 I), \cdots, \mathcal{N}(0, \sigma_L^2 I)$

$\sigma_1 < \sigma_2 < \cdots < \sigma_L$ geometric progression

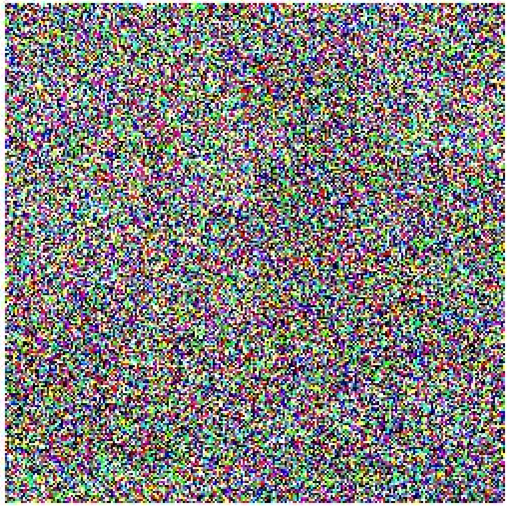
$$\mathrm{d}\mathbf{x} = e^t \mathrm{d}\mathbf{w}$$

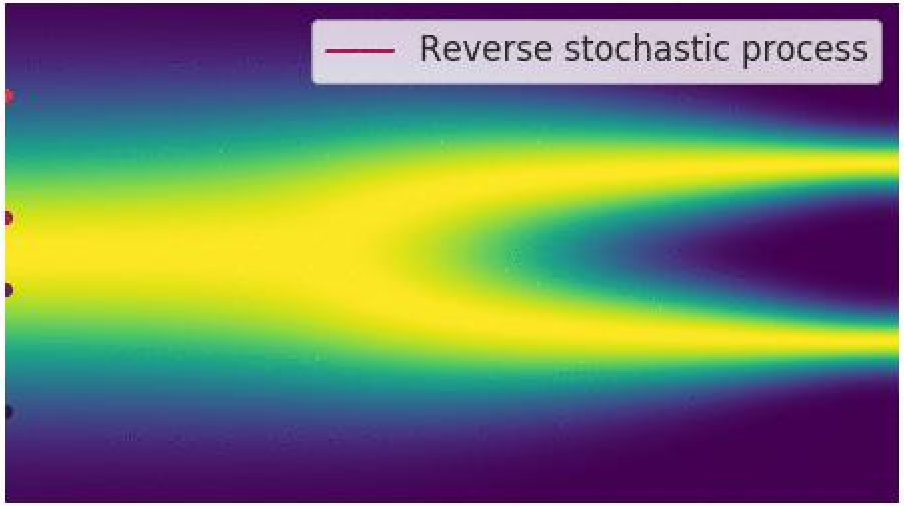
[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Score-Based Diffusion Model - Reverse SDE for Sample Generation



Perturbing data to noise with a continuous-time stochastic process





Generate data from noise by reversing the perturbation procedure.

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021

Score-Based Diffusion Model - Reverse SDE for Sample Generation

- Similar to the annealed Langevin dynamics to reverse the perturbation process

- Analogously reverse the perturbation process by reversing the SDE
- Every SDE has a corresponding reverse SDE

- negative infinitesimal time step
- since SDE needs to eb solved backward in time ***T*** to **0**

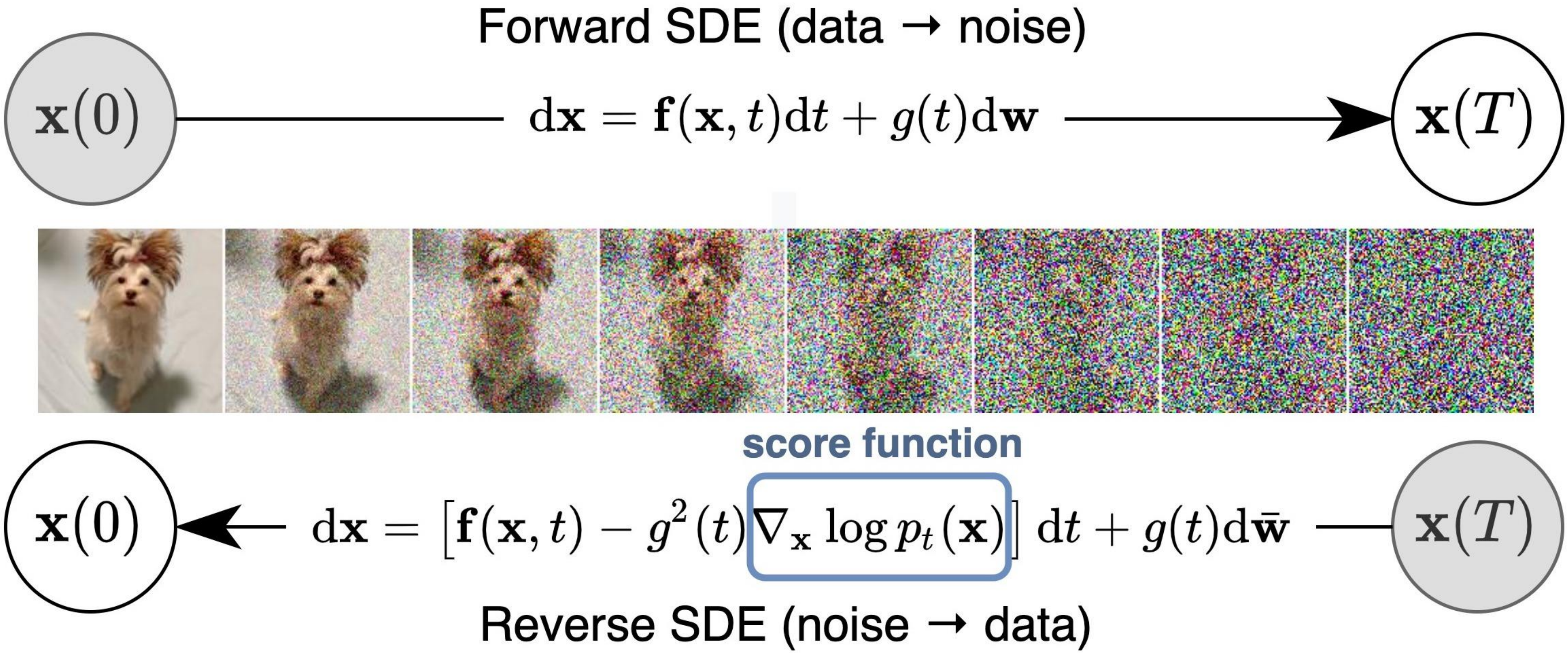
$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}]\boxed{\mathrm{d}t} + g(t)\mathrm{d}\mathbf{w}$$

- This is exaclty the score function

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}$$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Score-Based Diffusion Model - Reverse SDE for Sample Generation



[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Score-Based Diffusion Model - Estimate Reverse SDE using Score Matching

Time-Dependent Score-Based Modeling $\mathbf{s}_\theta(\mathbf{x}, t)$

$$\mathbf{s}_\theta(\mathbf{x}, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$$

analogous to

$$\mathbf{s}_\theta(\mathbf{x}, i) \approx \nabla_{\mathbf{x}} \log p_{\sigma_i}(\mathbf{x})$$

$$\mathbb{E}_{t \in \mathcal{U}(0,T)} \mathbb{E}_{p_t(\mathbf{x})} [\lambda(t) \| \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) - \mathbf{s}_\theta(\mathbf{x}, t) \|_2^2]$$

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t) \mathbf{s}_\theta(\mathbf{x}, t)] \mathrm{d}t + g(t) \mathrm{d}\mathbf{w}$$

Once trained plug it into the reverse SDE

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Score-Based Diffusion Model - Generating New Samples

$$d\mathbf{x} = \sigma^t d\mathbf{w}, \quad t \in [0, 1]$$

$$d\mathbf{x} = -\sigma^{2t} \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) dt + \sigma^t d\bar{\mathbf{w}}$$

$$d\mathbf{x} = -\sigma^{2t} s_\theta(\mathbf{x}, t) dt + \sigma^t d\bar{\mathbf{w}}$$

Use numerical methods to solve reverse-time SDE (one such method is Euler-Maruyama)
It simply discretize the SDE

replacing dt with Δt and $d\mathbf{w}$ with $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, g^2(t)\Delta t\mathbf{I})$

$p_1 \approx \mathbf{N}\left(\mathbf{x}; \mathbf{0}, \frac{1}{2}(\sigma^2 - 1)\mathbf{I}\right)$ draw first sample from the prior distribution, then solve reverse-tiem SDE

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t + \sigma^{2t} s_\theta(\mathbf{x}_t, t)\Delta t + \sigma^t \sqrt{\Delta t}\mathbf{z}_t$$

$$\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Euler-Maruyama is similar to Langevin dynamics : both update by following score functions perturbed with Gaussian noise.

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Score-Based Diffusion Model - Sample Results

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Score-Based Diffusion Model - Sample Results

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Probability Flow ODE

Sampling based on Langevin Dynamics and SDE solvers does not provide way to compte exact log-likelihood of score based generative models

Any SDE can be converted to an ODE (Ordinary differential equation)

- The corresponding ODE of an SDE is named **probability flow ODE**
- This probability flow ODE formulation has several unique advantages.
- Probability flow ODE becomes a special case of a neural ODE, when score function is approximated by a neural model
  - It is an example of continuous normalizing flows, since the probability flow ODE converts a data distribution to a prior noise distribution
- it shares the same marginal distributions as the SDE and is fully invertible.

Neural Ordinary Differential Equations

Ricky T. Q. Chen*, Yulia Rubanova*, Jesse Bettencourt*, David Duvenaud
University of Toronto, Vector Institute
{rtqichen, rubanova, jessebett, duvenaud}@cs.toronto.edu

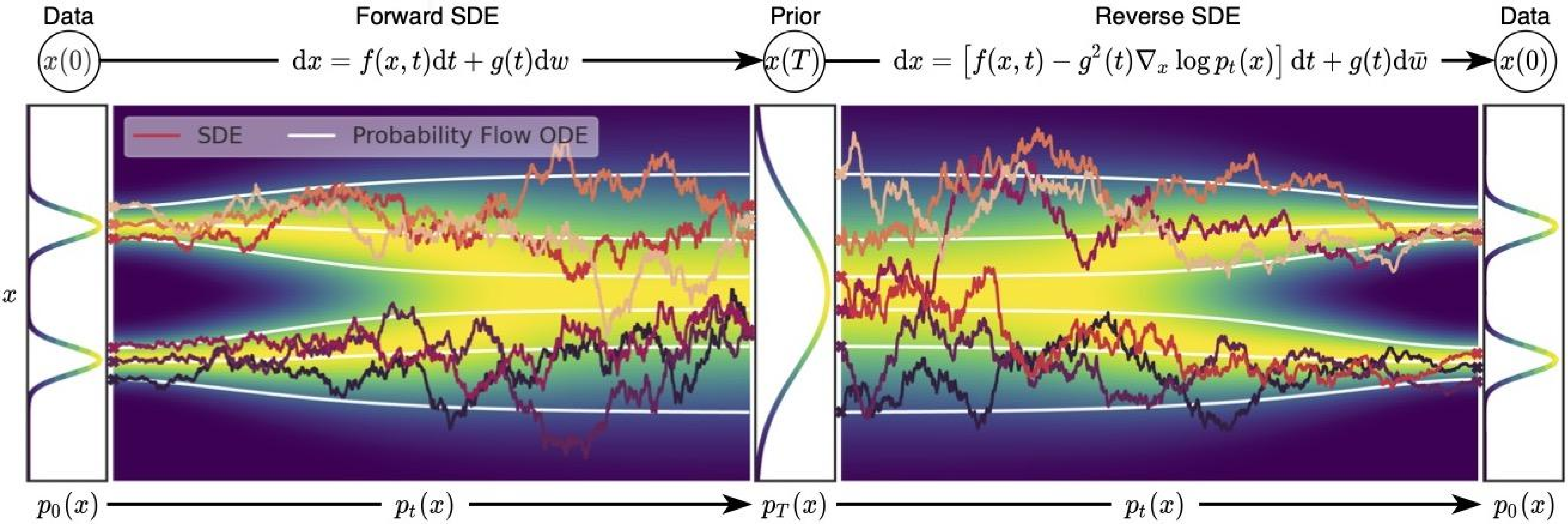
FFJORD: FREE-FORM CONTINUOUS DYNAMICS FOR SCALABLE REVERSIBLE GENERATIVE MODELS

Will Grathwohl*†‡, Ricky T. Q. Chen*†, Jesse Bettencourt†, Ilya Sutskever‡, David Duvenaud†

ABSTRACT

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications

Probability Flow ODE

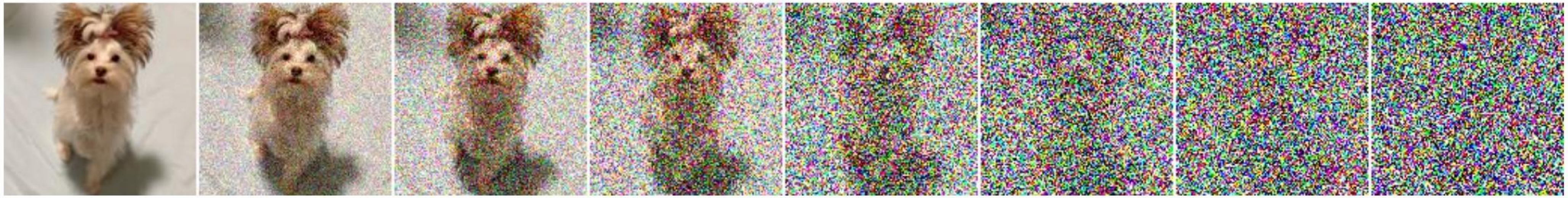


$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g^2(t) \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] \mathrm{d}t.$$

Score-Based Diffusion Model - Reverse SDE for Sample Generation

Forward SDE (data → noise)

$$\mathbf{x}(0) \longrightarrow \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w} \longrightarrow \mathbf{x}(T)$$



score function

$$\mathbf{x}(0) \longleftarrow \mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\right]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}} \longleftarrow \mathbf{x}(T)$$

Reverse SDE (noise → data)

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2}g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right]\mathrm{d}t.$$

[1] A. Hyvarinen. *Estimation of non-normalized statistical models by score matching* Journal of Machine Learning Research, Vol 6(Apr), pp. 695--709. 2005.
[2] https://yang-song.net/blog/2021/score/
[3] Y. Song, S. Ermon., Generative Modeling by Estimating Gradients of the Data Distribution, Advances in Neural Information Processing Systems
[4] Y. Song, J. Sohl-Dickstein, D.P. Kingma, A. Kumar, S. Ermon, B. Poole., Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021
[5] B.D. Anderson., Reverse-time diffusion equation models, Stochastic Processes and their Applications