Diffusion Probabilistic Models and Fast Inference by Estimating the Optimal Reverse Covariance

Chongxuan Li

ML Group @ RUC

Gaoling School of AI, Renmin University of China

Real or fake?

Both images are generated by deep generative models!

Generative modeling

$$x_i \sim_{iid} p_D(x), \qquad i = 1,2,3 \ldots$$

where $p_D(x)$ is the underlying distribution of the data.

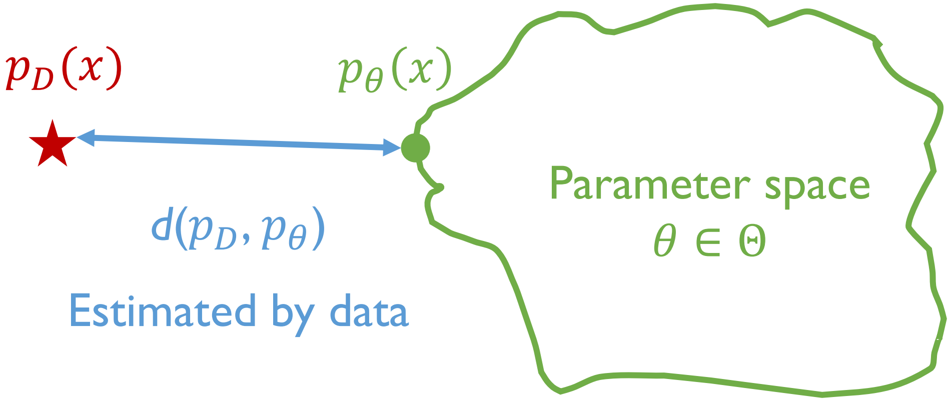
Image credit: Fisher Yu, https://www.yf.io/p/lsun

Generative modeling: representation, learning and inference

$x_i \sim_{iid} p_D(x), \qquad i = 1,2,3 \ldots$

A generative model is a joint distribution $p_\theta(x)$.

Goal of generative modeling: $p_\theta(x) \approx p_D(x)$.



Roadmap of Diffusion Probabilistic Models

2005

Score matching for EBMs

SM (Hyvärinen & Dayan, JMLR 2005)

DSM (Vincent, Neural Computation 2011)

2015

Diffusion Probabilistic Models

DPM (Sohl-Dickstein et al., ICML 2015)

2019

Score-based models

NCSN (Song & Ermon, NeurIPS 2019)

2020

Score representation of the mean in DPMs

DDPM (Ho et al., NeurIPS 2020)

DPM via SDE (Song et al., ICLR 2001)

2022

Score representation of variance in DPMs

Analytic-DPM (Bao et al., ICLR 2022)

Analytic-DPM++ (Bao et al., ICML 2022)

Diffusion Probabilistic Models

2005

Score matching for EBMs

SM (Hyvärinen & Dayan, JMLR 2005)

DSM (Vincent, Neural Computation 2011)

2015

Diffusion Probabilistic Models

DPM (Sohl-Dickstein et al., ICML 2015)

2019

Score-based models

NCSN (Song & Ermon, NeurIPS 2019)

2020

Score representation of the mean in DPMs

DDPM (Ho et al., NeurIPS 2020)

DPM via SDE (Song et al., ICLR 2001)

2022

Score representation of variance in DPMs

Analytic-DPM (Bao et al., ICLR 2022)

Analytic-DPM++ (Bao et al., ICML 2022)

Diffusion in Physics

Diffusion destroy structures along time

Image source: https://makeagif.com/gif/ink-drip-in-water-stock-footage-red-toobstock-free-stock-video-IR7ZXe

Diffusion in Physics

Diffusion destroy structures along time

What if we can reverse time?

Diffusion probabilistic models

Sohl-Dickstein et al, ICML 2015

Forward diffusion: a Markov chain with Gaussian kernel

$$q\left(\mathbf{x}^{(0)}\right) \longrightarrow q\left(\mathbf{x}^{(T)}\right) \approx \mathcal{N}\left(\mathbf{x}^{(T)}; 0, \mathbf{I}\right)$$

$$q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right) = \mathcal{N}\left(\mathbf{x}^{(t)}; \underbrace{\mathbf{x}^{(t-1)}\sqrt{1-\beta_t}}_{\text{Decay towards origin}}, \underbrace{\mathbf{I}\beta_t}_{\text{Add small noise}}\right)$$

Data distribution

Gaussian noise

Image credit: *Sohl-Dickstein*

Diffusion probabilistic models

Sohl-Dickstein et al, ICML 2015

Consider continuous diffusion

- We can get noise from any data distribution
- The forward process is reversible
- The backward process has the same functional form

Core idea: learn to map noise to data by reversing the time

Data distribution

Gaussian noise

Diffusion probabilistic models

Sohl-Dickstein et al, ICML 2015

Backward diffusion: a Markov chain with Gaussian kernel

$$p\left(\mathbf{x}^{(0)}\right) \approx q\left(\mathbf{x}^{(0)}\right) \longleftarrow p\left(\mathbf{x}^{(T)}\right) = \mathcal{N}\left(\mathbf{x}^{(T)}; 0, \mathbf{I}\right)$$

$$p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right) = \mathcal{N}\left(\mathbf{x}^{(t-1)}; \underline{f_\mu\left(\mathbf{x}^{(t)}, t\right), f_\Sigma\left(\mathbf{x}^{(t)}, t\right)}\right)$$

Learned drift and covariance functions

Data distribution

Gaussian noise

Image credit: *Sohl-Dickstein*

Training DPMs by Maximum Likelihood

Sohl-Dickstein et al, ICML 2015

$$\min_\theta KL(p_D(x_0)||p_\theta(x_0)) \Leftrightarrow \max_\theta \mathrm{E}_{p_D(x_0)}[\log p_\theta(x_0)]$$

KL divergence minimization   Maximum likelihood

Training DPMs by Maximum Likelihood

Sohl-Dickstein et al, ICML 2015

$$\min_{\theta} KL(p_D(x_0)||p_\theta(x_0)) \Leftrightarrow \max_{\theta} \mathrm{E}_{p_D(x_0)}[\log p_\theta(x_0)]$$

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] = \mathrm{E}_{q(x_0)}\left[\log \int p_\theta(x_{0:T}) dx_{1:T}\right] \geq \mathrm{E}_{q(x_0)} \mathrm{E}_{q(x_{1:T}|x_0)} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}$$

Jensen's inequality: equality holds when $q(x_{1:T}|x_0) = p(x_{1:T}|x_0)$

Remember that The backward process has the same functional form as the forward one.

DPM can be understood as a hierarchical VAE with tractable posterior.

Training DPMs by Maximum Likelihood

Sohl-Dickstein et al, ICML 2015

$$\mathrm{E}_{q(x_{0:T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \right]$$

ELBO

Decomposition for efficiency

Training DPMs by Maximum Likelihood

Sohl-Dickstein et al, ICML 2015

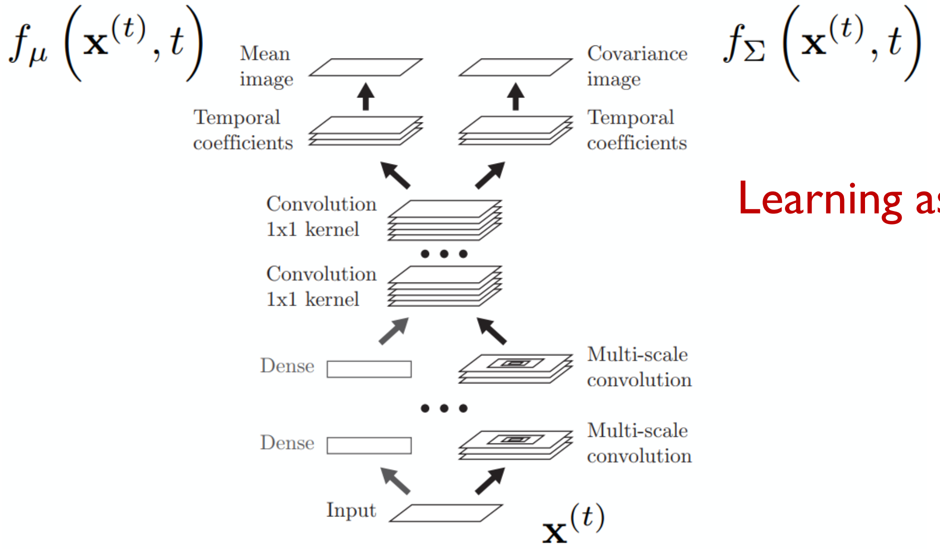
$$\mathrm{E}_{q(x_{0:T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) \right]$$

$$q(x_{t-1}|x_t, x_0) = N(\tilde{\mu}_t(x_0, x_t), \tilde{\beta}_t I)$$

$$p_\theta(x_{t-1}|x_t) = N\left(f_\mu(x_t, t), f_\Sigma(x_t, t)\right)$$

Closed-form solution

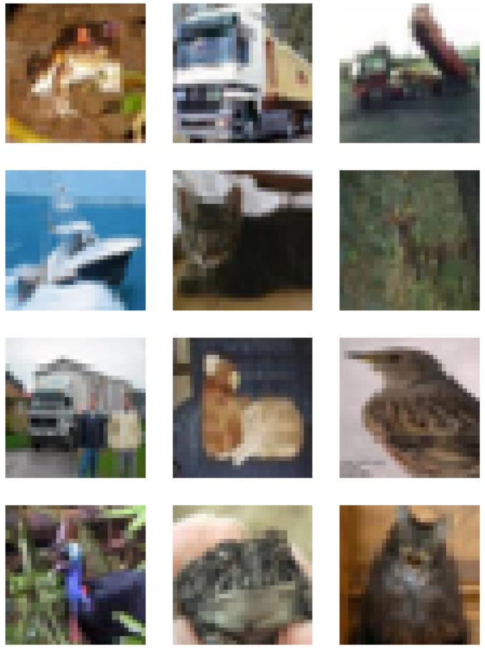
$q(x_{t-1}|x_t)$ is more natural while it is non Gaussian!



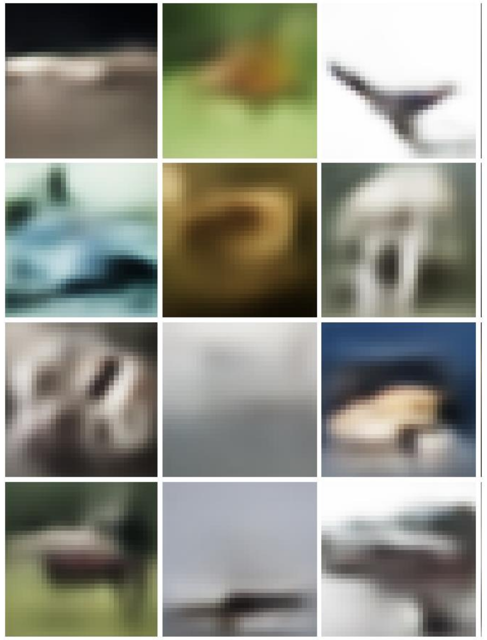
Learning as regression

Results for DPM

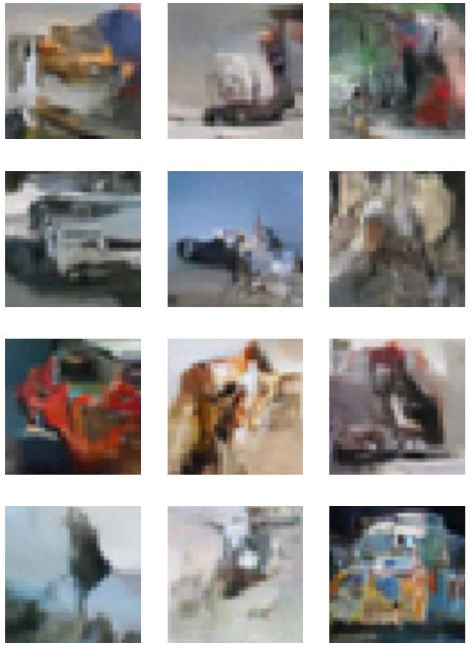
Sohl-Dickstein et al, ICML 2015



Training Data



Samples from DRAW [Gregor *et al*, 2015]



Samples from diffusion model

Diffusion Probabilistic Models

2005

Score matching for EBMs

SM (Hyvärinen & Dayan, JMLR 2005)

DSM (Vincent, Neural Computation 2011)

2015

Diffusion Probabilistic Models

DPM (Sohl-Dickstein et al., ICML 2015)

2019

Score-based models

NCSN (Song & Ermon, NeurIPS 2019)

2020

Score representation of the mean in DPMs

DDPM (Ho et al., NeurIPS 2020)

DPM via SDE (Song et al., ICLR 2001)

2022

Score representation of variance in DPMs

Analytic-DPM (Bao et al., ICLR 2022)

Analytic-DPM++ (Bao et al., ICML 2022)

Denoising diffusion probabilistic models

Jonathon et al., NeurIPS 2021

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] \geq \mathrm{E}_{q(x_{0:T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \right]$$

Regress mean with fixed variance

$$||\mu_\theta(x_t, t) - \tilde{\mu}_t(x_0, x_t)||^2$$

Denoising diffusion probabilistic models

Jonathon et al., NeurIPS 2021

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] \geq \mathrm{E}_{q(x_{0:T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) \right]$$

Regress mean with fixed variance $\|\mu_\theta(x_t, t) - \tilde{\mu}_t(x_0, x_t)\|^2$

$$\begin{aligned}
\mu_\theta(x^{(t)}, t) &\rightarrow \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} x^{(0)} + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x^{(t)} \\
&= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \frac{1}{\sqrt{\bar{\alpha}_t}} (x^{(t)} - \sqrt{1-\bar{\alpha}_t}\epsilon_t) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x^{(t)} \\
&= \frac{1}{\sqrt{\alpha_t}} \left(\frac{\beta_t}{1-\bar{\alpha}_t} x^{(t)} + \frac{\alpha_t(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} x^{(t)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right) \\
&= \frac{1}{\sqrt{\alpha_t}} \left(\frac{\beta_t + \alpha_t - \bar{\alpha}_t}{1-\bar{\alpha}_t} x^{(t)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right) = \frac{1}{\sqrt{\alpha_t}} \left(x^{(t)} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_t \right)
\end{aligned}$$

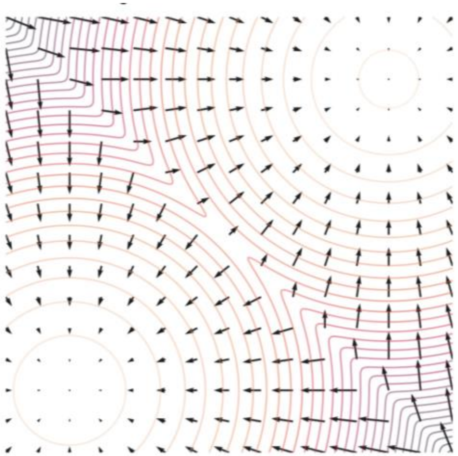
Reparameterization: $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$

Regress Gaussian noise $\|\epsilon_\theta(x_t, t) - \epsilon\|^2$

Denoising diffusion probabilistic models

Jonathon et al., NeurIPS 2021

$$E_{q(x_0)}[\log p_\theta(x_0)] \geq E_{q(x_{0:T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = E_q\left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t))\right]$$



$p_\theta(\mathbf{x})$ vs. $s_\theta(\mathbf{x})$

Regress mean with fixed variance $\|\mu_\theta(x_t, t) - \tilde{\mu}_t(x_0, x_t)\|^2$

Reparameterization: $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\,\epsilon$

Regress Gaussian noise $\|\epsilon_\theta(x_t, t) - \epsilon\|^2$

Equivalent to DSM (*Vincent, 2011*) $\|s_\theta(x_t, t) - \nabla \log q_t(x_t)\|^2$

Image credit: *Yang Song*

Denoising diffusion probabilistic models

Jonathon et al., NeurIPS 2021

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] \geq \mathrm{E}_{q(x_{0:T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)) \right]$$

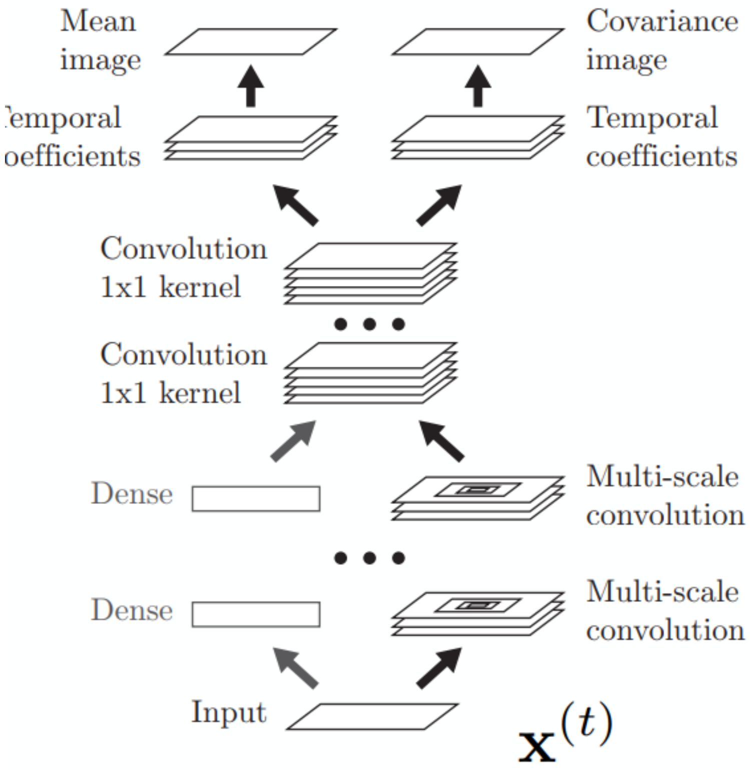
Regress mean with fixed variance $\|\mu_\theta(x_t, t) - \tilde{\mu}_t(x_0, x_t)\|^2$

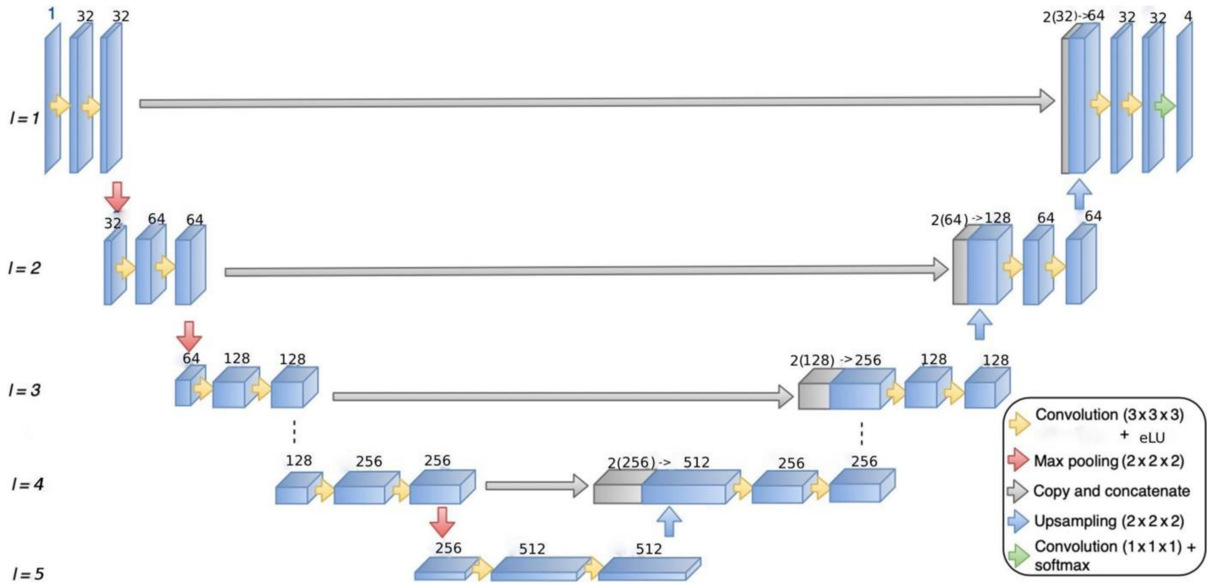
Regress Gaussian noise $\|\epsilon_\theta(x_t, t) - \epsilon\|^2$

Predicting the Gaussian noise is numerically stable and the residual is easier to learn.

Skip connections in the model

Jonathon et al., NeurIPS 2021





Similar architectures have been investigated in *Song & Ermon, NeurIPS 2019*

Results of DDPM

Jonathon et al., NeurIPS 2021

CelebA-HQ 256x256

CIFAR-10 FID = 3.17 (SOTA)

Fast inference in diffusion probabilistic models

2005

Score matching for EBMs

SM (Hyvärinen & Dayan, JMLR 2005)
DSM (Vincent, Neural Computation 2011)

2015

Diffusion Probabilistic Models

DPM (Sohl-Dickstein et al., ICML 2015)

2019

Score-based models

NCSN (Song & Ermon, NeurIPS 2019)

2020

Score representation of the mean in DPMs

DDPM (Ho et al., NeurIPS 2020)
DPM via SDE (Song et al., ICLR 2001)

2022

Score representation of variance in DPMs

Analytic-DPM (Bao et al., ICLR 2022)
Analytic-DPM++ (Bao et al., ICML 2022)

Motivation

DPM

- Learning
  - MLE (tractable posterior) / SM
- Sampling
  - 1000 number of function evaluations
- Performance
  - SOTA generation and likelihood results

GAN, VAE, FLOW

- Learning
  - Adversarial / MLE + traditional VI / det of Jacobian
- Sampling
  - Single function evaluation
- Performance
  - Competitive to SOTA

Motivation

DPM

- Learning
  - MLE (tractable posterior) / SM
- Sampling
  - 1000 number of function evaluations
- Performance
  - SOTA generation and likelihood results

GAN, VAE, FLOW

- Learning
  - Adversarial / MLE + traditional VI / det of Jacobian
- Sampling
  - Single function evaluation
- Performance
  - Competitive to SOTA

The variance matters

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] \geq \mathrm{E}_{q(x_{0\ldots T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t, x_0) || p_\theta(x_{t-1}|x_t)) \right]$$

The variance of $p_\theta(x_{t-1}|x_t)$ are set manually by considering extreme case of $q(x_0)$.

- Standard Gaussian $q(x_0)$ corresponds to β_t
- Single point $q(x_0)$ corresponds to the $\tilde{\beta}_t$

Can we find the optimal covariance w.r.t. the ELBO with a minimal assumption on data?

Analytic-DPM

Bao et al, ICLR 2022

- An equivalent decomposition of the objective

$$\mathrm{E}_{q(x_0)}[\log p_\theta(x_0)] \geq \mathrm{E}_{q(x_{0\ldots T})} \log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)} = \mathrm{E}_q \left[L_T + L_0 - \sum KL(q(x_{t-1}|x_t)||p_\theta(x_{t-1}|x_t)) \right]$$

- Why change the decomposition?
  - $q(x_{t-1}|x_t, x_0)$ uses ground truth data, which is not available for $p_\theta(x_{t-1}|x_t)$ (generation)
- Challenge: $q(x_{t-1}|x_t)$ is NOT Gaussian in general and does not have an analytic solution
  - $q(x_{t-1}|x_t) = \int q(x_{t-1}|x_t, x_0)q(x_0)dx_0$

Score representation of the optimal mean and covariance

Bao et al, ICLR 2022

Main Theorem. The optimal mean and covariance w.r.t. the ELBO can be written as:

$$\mu_t^*(x_t) = \frac{1}{\sqrt{1-\beta_t}}(x_t + \beta_t \nabla \log q_t(x_t)), \qquad \sigma_t^{*2} = \frac{\beta_t}{1-\beta_t}\left(1 - \beta_t \mathrm{E}_{q_t(x_t)} \frac{\|\nabla \log q_t(x_t)\|^2}{d}\right).$$

Key steps in the proof:

- Moment matching: $\min_{p \text{ is exponential family}} KL(q||p)$ is equivalent to matching the moments of q to p
- Low of total variance: conditional expectation (moments) of $q(x_{t-1}|x_t)$ can be represented conditional expectation of $q(x_0|x_t)$
- The conditional expectation of $q(x_0|x_t)$ can be represented by score of $q_t(x_t)$ if $q(x_t|x_0)$ is Gaussian.

Score representation of the optimal mean and covariance

Bao et al, ICLR 2022

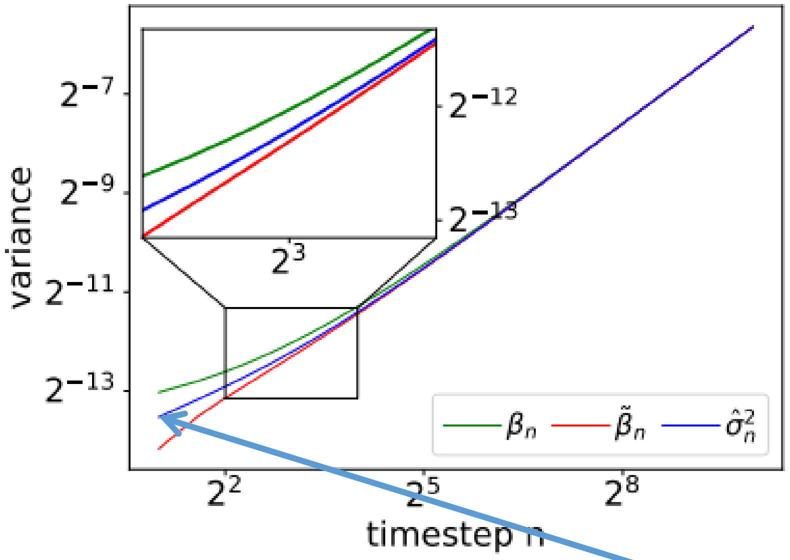
Main Theorem. The optimal mean and covariance w.r.t. the ELBO can be written as:

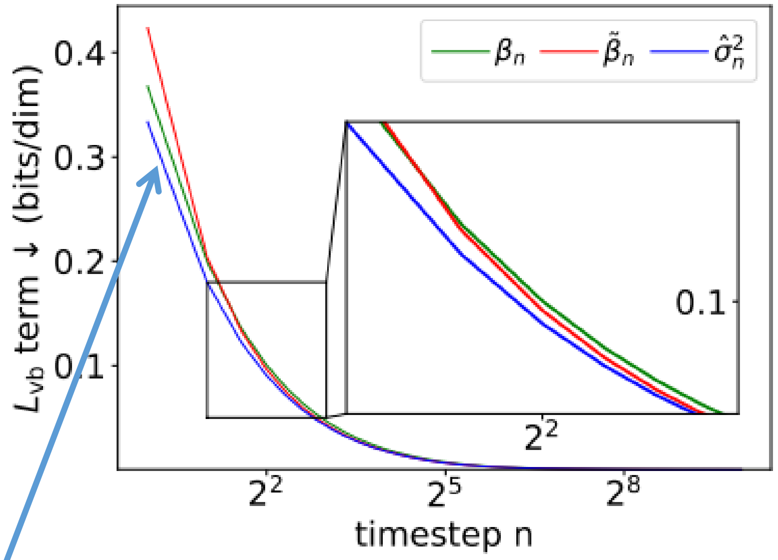
$$\mu_t^*(x_t) = \frac{1}{\sqrt{1-\beta_t}}(x_t + \beta_t \nabla \log q_t(x_t)), \qquad \sigma_t^{*2} = \frac{\beta_t}{1-\beta_t}(1 - \beta_t \mathrm{E}_{q_t(x_t)} \frac{\|\nabla \log q_t(x_t)\|^2}{d}).$$

- The optimal mean representation coincides with existing work
- The optimal covariance representation depends on the score as well
- DSM proves that matching a noisy score is equivalent to matching the noise
  - The score estimation by the noise perdition network in DDPM $\nabla \log q_t(x_t) \approx -\frac{1}{\sqrt{\beta_t}} \epsilon_\theta(x_t, t)$
- We can estimate the optimal covariance without additional training

Differences between Analytic-DPM and DDPM

Bao et al, ICLR 2022





Comparing the variances

Analytic-DPM

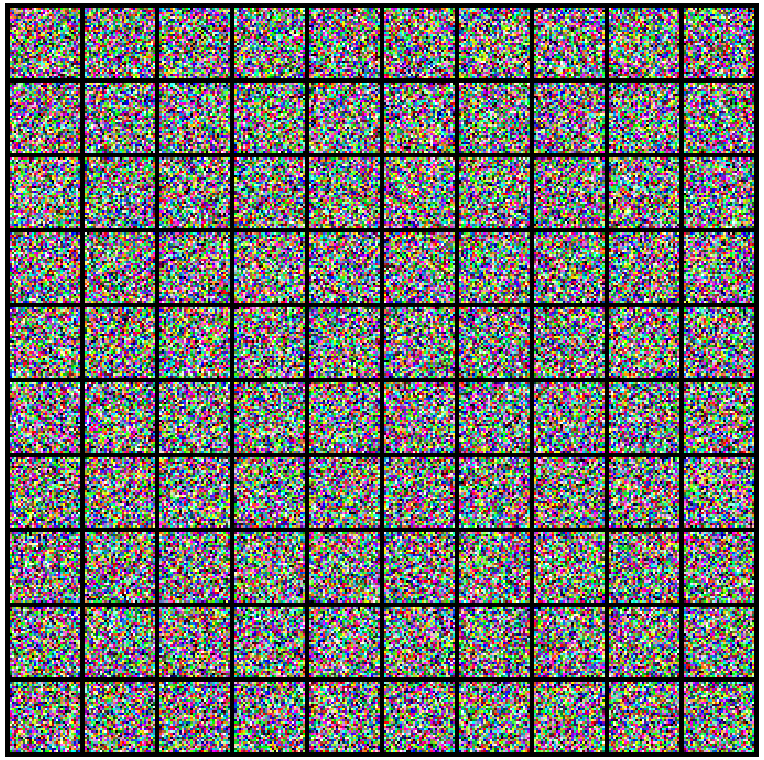
Comparing the ELBO

Very different near data

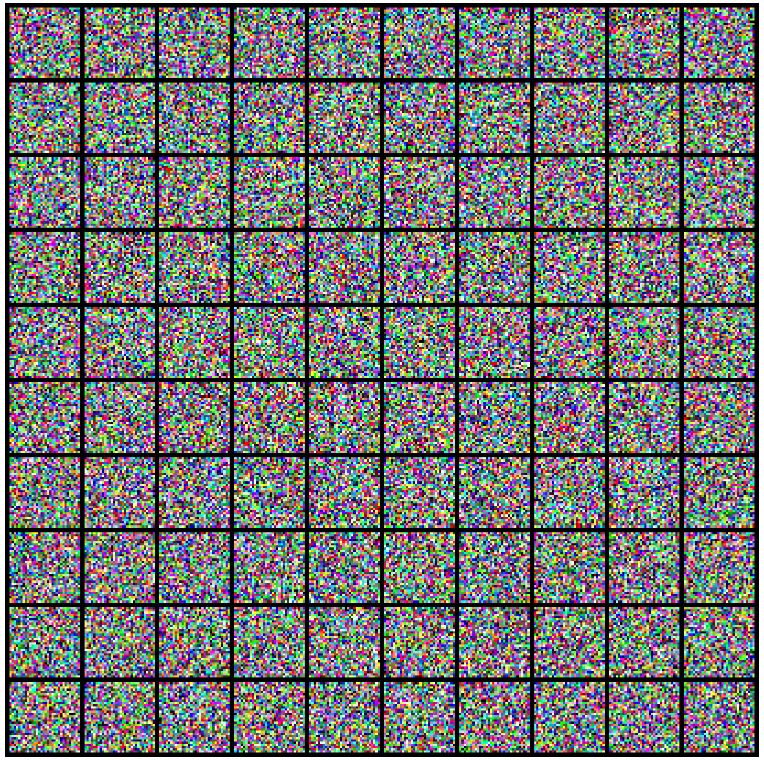
A tighter ELBO

Quality-efficiency trade-off: $20\times$ to $80\times$ speed up with the same sample quality

Bao et al, ICLR 2022



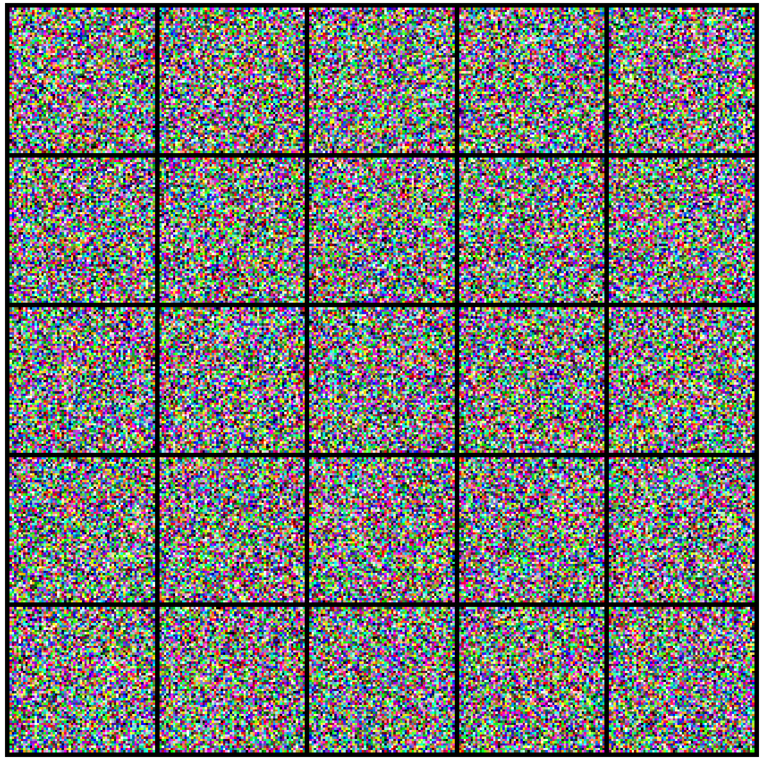
Original DDPM in 1000 steps



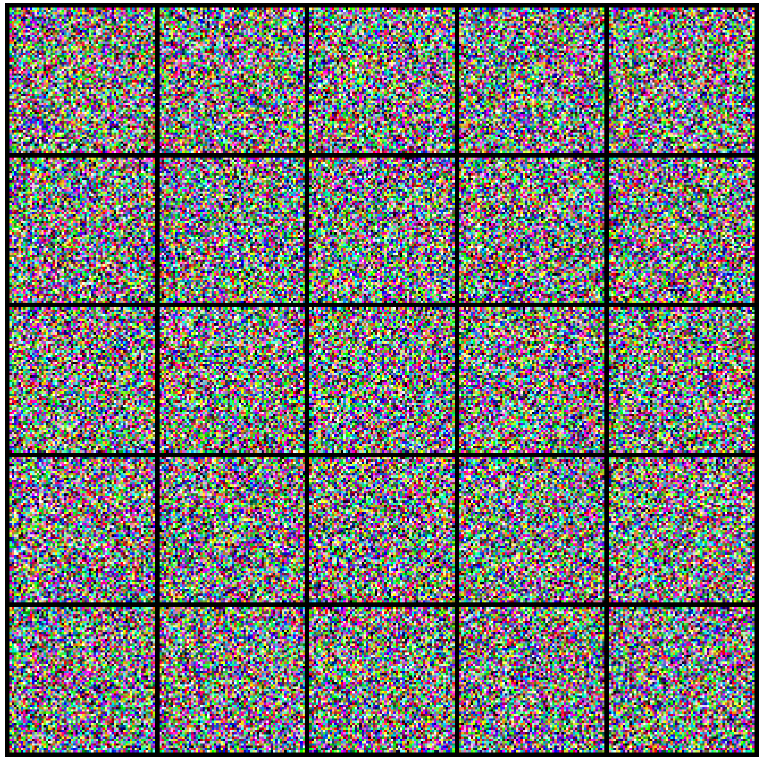
Analytic-DDPM in 50 steps

Quality-efficiency trade-off: $20\times$ to $80\times$ speed up with the same sample quality

Bao et al, ICLR 2022



Original DDPM in 1000 steps



Analytic-DDPM in 50 steps

Analytic-DPM++

Bao et al, ICML 2022

- From scalar to full covariance
- Learn a time dependent covariance in an analytical form by predicting the square of noise as well
- Correcting the bias of the score model: the optimal covariance given imperfect mean

$$|\sigma_n^{*2} - \hat{\sigma}_n^2| = \underbrace{\left(\sqrt{\frac{\overline{\beta}_n}{\alpha_n}} - \sqrt{\overline{\beta}_{n-1} - \lambda_n^2}\right)^2 \overline{\beta}_n}_{\text{Coefficient}} \underbrace{\left|\boxed{\Gamma_n} - \boxed{\mathbb{E}_{q_n(\boldsymbol{x}_n)} \frac{\|\nabla_{\boldsymbol{x}_n} \log q_n(\boldsymbol{x}_n)\|^2}{d}}\right|}_{\text{Approximation error}}$$

Estimate with a model True value with the score

Analytic-DPM++

Bao et al, ICML 2022

- Much better results in 10 steps

	CIFAR10 (VP SDE)					
# TIMESTEPS K	10	25	50	100	200	1000
EULER-MARUYAMA	292.20	170.17	90.79	47.46	21.92	**2.55**
ANCESTRAL SAMPLING	235.28	129.29	68.52	31.99	12.81	2.72
PROBABILITY FLOW	107.74	21.34	7.78	4.33	3.27	2.82
A-DPM	35.10	11.57	6.54	4.71	3.61	2.98
NPR-DPM	33.70	10.44	5.83	3.97	3.05	3.04
SN-DPM	**25.30**	**7.34**	**4.46**	**3.27**	**2.83**	2.71

What's next?

2005 — **Score matching for EBMs**

SM (Hyvärinen & Dayan, JMLR 2005)

DSM (Vincent, Neural Computation 2011)

2015 — **Diffusion Probabilistic Models**

DPM (Sohl-Dickstein et al., ICML 2015)

2019 — **Score-based models**

NCSN (Song & Ermon, NeurIPS 2019)

2020 — **Score representation of the mean in DPMs**

DDPM (Ho et al., NeurIPS 2020)

DPM via SDE (Song et al., ICLR 2001)

2022 — **Score representation of variance in DPMs**

Analytic-DPM (Bao et al., ICLR 2022)

Analytic-DPM++ (Bao et al., ICML 20

Faster inference

Lu et al, arxiv 2022

- Simple form of the probability ODE by reparameterization and exploiting semi-linearity
- Customized higher-order ODE solver for high quality generation in 10 steps

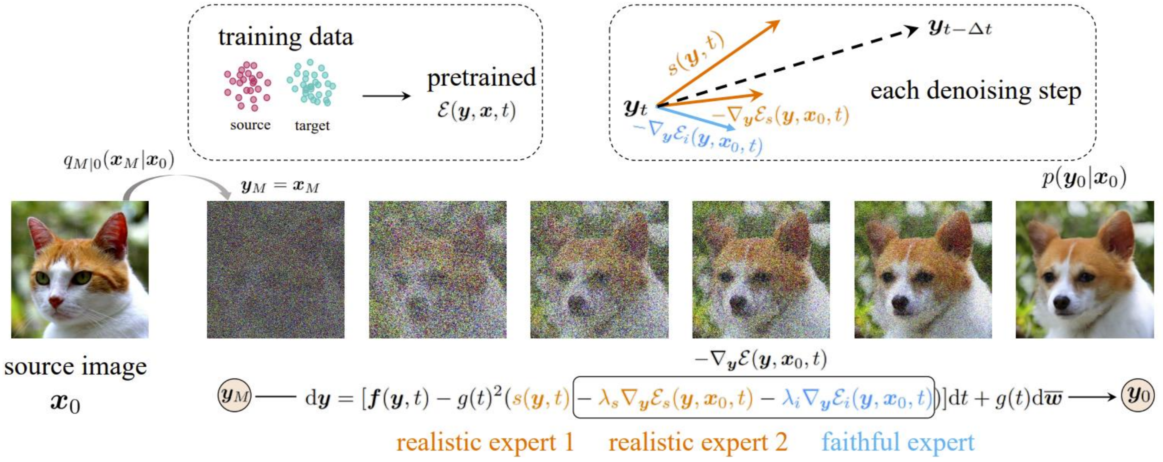
NFE = 10 NFE = 15 NFE = 20 NFE = 100 NFE = 10

(a) DDIM [19] (b) DPM-Solver (ours)

Controllable generation

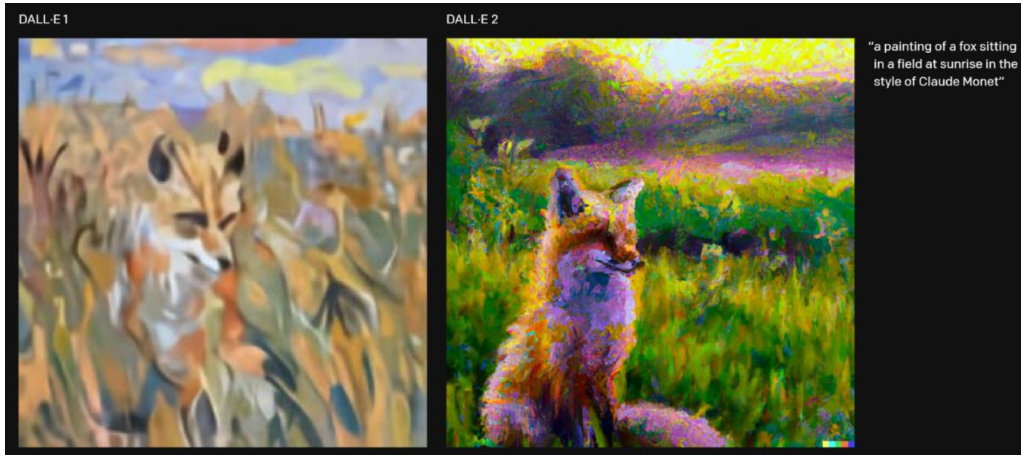
Min Zhao et al., 2022

- Using energy functions trained separately
- In the formulation of Product of Experts
- Evaluated in unpaired I2I translation



Large scale diffusion models

DALL·E 2



Method	Zero-shot FID
DALL·E	28
GLIDE	12.24
DALL·E 2-AR	10.63
DALL·E 2-Diffusion	**10.39**

With Analytic-DPM

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

a teddy bear on a skateboard in times square

Imagination of AI: zero shot text to image generation

Summary of the talk

- Diffusion models gradually map a Gaussian to data by a Markov chain
- Score representations of both the mean and variance of diffusion models are effective
- Faster, controllable and larger diffusion models are on the way

Thank you!

Email: chongxuanli@ruc.edu.cn

Homepage: https://zhenxuan00.github.io/

References

- Hyvärinen A, Dayan P. Estimation of non-normalized statistical models by score matching[J]. Journal of Machine Learning Research, 2005, 6(4).
- Vincent P. A connection between score matching and denoising autoencoders[J]. Neural computation, 2011, 23(7): 1661-1674.
- Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International Conference on Machine Learning. PMLR, 2015: 2256-2265.
- Song Y, Ermon S. Generative modeling by estimating gradients of the data distribution[J]. Advances in Neural Information Processing Systems, 2019.
- Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020.
- Song Y, Sohl-Dickstein J, Kingma D P, et al. Score-based generative modeling through stochastic differential equations[J]. ICLR 2021.
- **Bao F, Li C, Zhu J, et al. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models[J]. ICLR 2022.**
- **Bao F, Li C, Sun J, et al. Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models[J]. ICML 2022.**
- **Lu C, et al. DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. Arxiv preprint 2022.**
- **Zhao M, et al. EGSDE: Unpaired Image-to-Image Translation via Energy-Guided Stochastic Differential Equations. Arxiv preprint 2022.**