Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models

https://github.com/baofff/Analytic-DPM ICLR 2022 outstanding paper award

Tsinghua University

Fan Bao, Chongxuan Li, Jun Zhu, Bo Zhang

Diffusion Probabilistic Models (DPMs)

Ho et al. Denoising diffusion probabilistic models (DDPM), Neurips 2020.
Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021.

- Diffusion process gradually injects noise to data
- Described by a Markov chain: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

Transition of diffusion: $q(x_n|x_{n-1}) = N(\sqrt{\alpha_n}x_{n-1}, \beta_n I)$ $\quad \alpha_n = 1 - \beta_n$

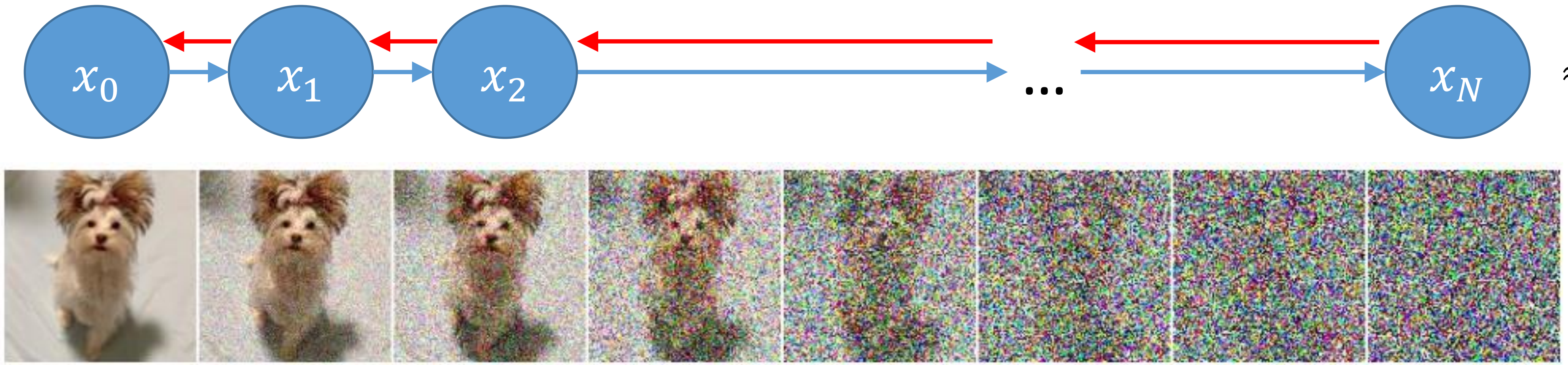
$$x_0 \to x_1 \to x_2 \to \dots \to x_N \approx N(0, I)$$

Diffusion process: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

Demo Images from *Song et al. Score-based generative modeling through stochastic differential equations, ICLR 2021.*

- Diffusion process in the reverse direction $\Leftrightarrow$ denoising process
- Reverse factorization: $q(x_0, \dots, x_N) = q(x_0|x_1) \dots q(x_{N-1}|x_N)q(x_N)$

Transition of denoising: $q(x_{n-1}|x_n) = ?$



$\approx N(0, I)$

Diffusion process: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

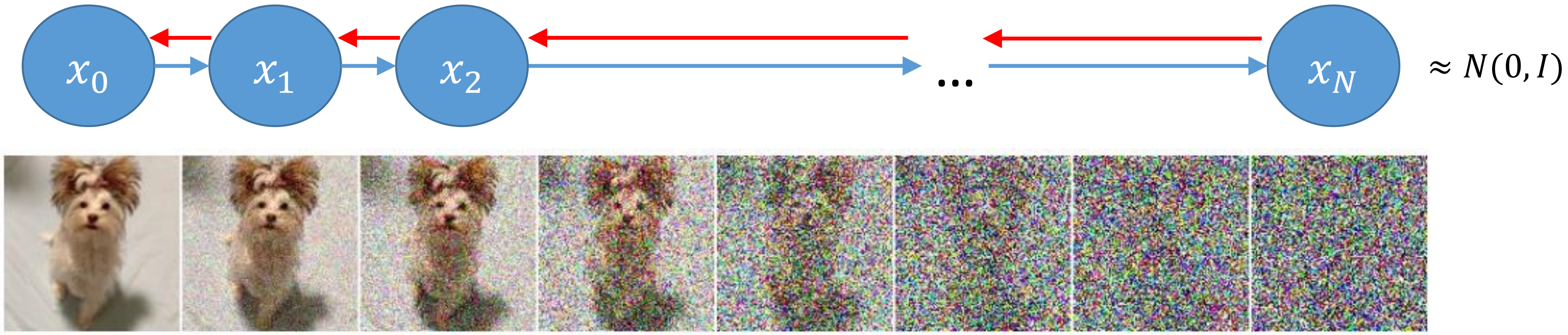
$= q(x_0|x_1) \dots q(x_{N-1}|x_N)q(x_N)$

- Approximate diffusion process in the reverse direction

Model transition: $p(x_{n-1}|x_n) = N(\mu_n(x_n), \sigma_n^2 I)$

$\downarrow$ approximate

Transition of denoising: $q(x_{n-1}|x_n) = ?$



Diffusion process: $q(x_0, \dots, x_N) = q(x_0)q(x_1|x_0) \dots q(x_N|x_{N-1})$

$= q(x_0|x_1) \dots q(x_{N-1}|x_N)q(x_N)$

The model: $p(x_0, \dots, x_N) = p(x_0|x_1) \dots p(x_{N-1}|x_N)p(x_N)$

- We hope $q(x_0, \dots, x_N) \approx p(x_0, \dots, x_N)$ $p(x_{n-1}|x_n) = N(\mu_n(x_n), \sigma_n^2 I)$
- Achieved by minimizing their KL divergence (i.e., maximizing the ELBO)

KL divergence ELBO Noise prediction (L_{simple})

$$\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{0:N})||p(x_{0:N})) \Leftrightarrow \max_{\mu_n(\cdot),\sigma_n^2} \mathrm{E}_q \log \frac{p(x_{0:N})}{q(x_{1:N}|x_0)} \Rightarrow \min_{\epsilon_n(\cdot)} \mathrm{E}_n \mathrm{E}_{x_0,\epsilon} \|\epsilon_n(x_n) - \epsilon\|^2$$

$$\Updownarrow$$

Score matching

$$\min_{s_n(\cdot)} \mathrm{E}_n \bar{\beta}_n \mathrm{E}_{q_n(x_n)} \|s_n(x_n) - \nabla \log q_n(x_n)\|^2$$

Parameterization of $\mu_n(\cdot)$ in DDPM:

$$\mu_n(x_n) = \underbrace{\frac{1}{\sqrt{\alpha_n}}\left(x_n + \beta_n s_n(x_n)\right)}_{\text{Score function form}} = \underbrace{\frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}} \epsilon_n(x_n)\right)}_{\text{Noise prediction form}}$$

DDPM only optimizes the model mean.
Use handcrafted model variance, e.g., $\sigma_n^2 = \beta_n$

How OpenAI deals with the variance?

OpenAI. Improved Denoising Diffusion Probabilistic Models, ICML 2021
OpenAI. Diffusion Models Beat GANs on Image Synthesis, NeurIPS 2021
OpenAI. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models, ICML 2022

$$\Sigma_\theta(x_t, t) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t)$$

Train the variance to maximize ELBO $\mathrm{E}_q \log \frac{p(x_{0:N})}{q(x_{1:N}|x_0)}$

Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models

- Can we directly find the optimal solution for $\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{0:N})||p(x_{0:N}))$?
- Yes!!!

Theorem 1. *(Score representation of the optimal solution to KL minimization)*

The optimal solution to $\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{0:N})||p(x_{0:N}))$ *is*

$$\mu_n^*(x_n) = \frac{1}{\sqrt{\alpha_n}}\underbrace{(x_n + \beta_n \nabla \log q_n(x_n))}_{\text{Score function form}} = \frac{1}{\sqrt{\alpha_n}}\underbrace{\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}}\mathrm{E}[\epsilon|x_n]\right)}_{\text{Noise prediction form}},$$

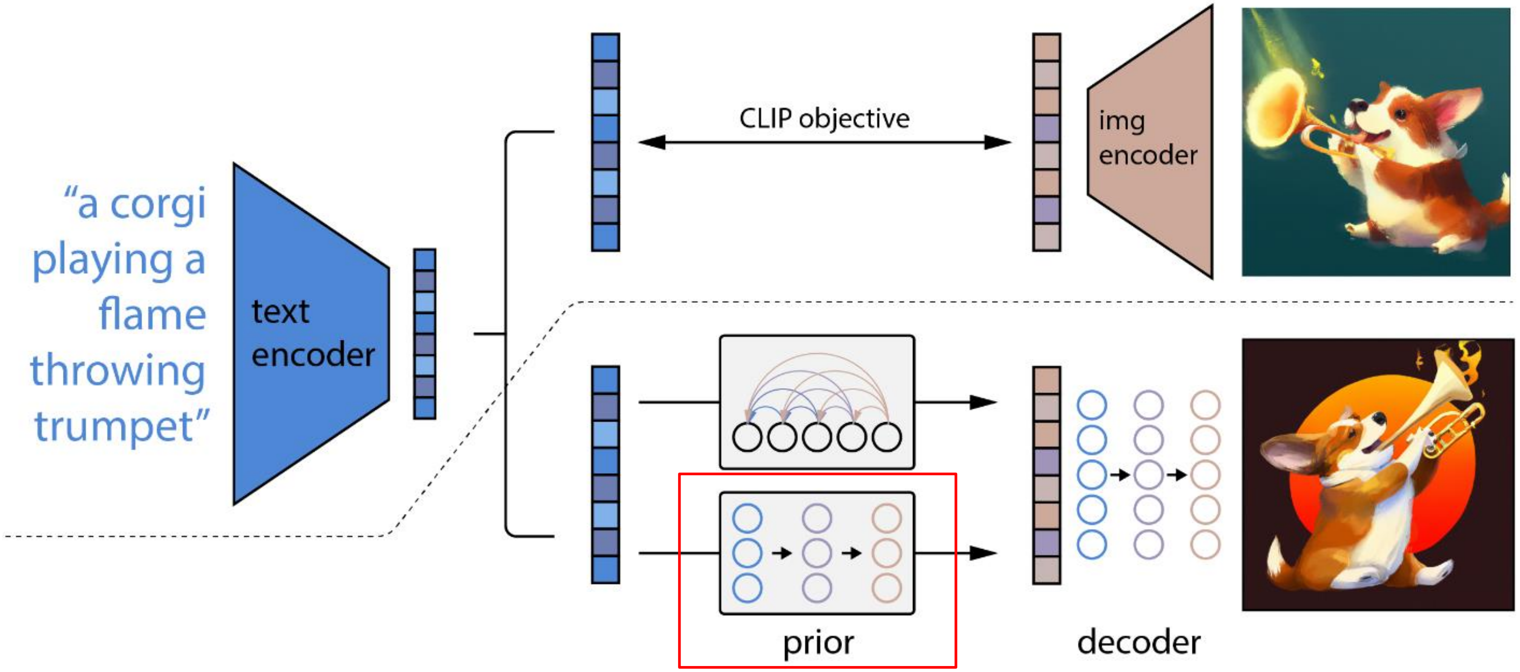
$$\sigma_n^{*2} = \frac{\beta_n}{\alpha_n}(1 - \beta_n \underbrace{\mathrm{E}_{q_n(x_n)}\frac{\|\nabla \log q_n(x_n)\|^2}{d}}_{\text{Score function form}}) = \frac{\beta_n}{\alpha_n}(1 - \frac{\beta_n}{\bar{\beta}_n}\underbrace{\mathrm{E}_{q_n(x_n)}\frac{\|\mathrm{E}[\epsilon|x_n]\|^2}{d}}_{\text{Noise prediction form}}).$$

3 key steps in proof:
- Moment matching
- Law of total variance
- Score representation of moments of $q(x_0|x_n)$

See a more general version of Theorem 1 for more general $q(x_{0:N})$ in the full paper
See extension to score-based SDE (Song et al.) in the full paper

How OpenAI deals with the variance after Analytic-DPM?

OpenAI. Hierarchical Text-Conditional Image Generation with CLIP Latents (DALLE2)



The diffusion prior uses Analytic-DPM to calculates the optimal variance, instead of learning the variance

	AR prior	Diffusion prior	64	$64 \rightarrow 256$	$256 \rightarrow 1024$
Diffusion steps	-	1000	1000	1000	1000
Noise schedule	-	cosine	cosine	cosine	linear
Sampling steps	-	64	250	27	15
Sampling variance method	-	analytic [2]	learned [34]	DDIM [47]	DDIM [47]

Recall…

Parameterization of $\mu_n(\cdot)$ in DDPM:

$$\mu_n(x_n) = \frac{1}{\sqrt{\alpha_n}}\left(x_n + \beta_n s_n(x_n)\right) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}}\epsilon_n(x_n)\right)$$

Score matching

Noise prediction

Optimal

$$\mu_n^*(x_n) = \frac{1}{\sqrt{\alpha_n}}\left(x_n + \beta_n \nabla \log q_n(x_n)\right) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}}\mathrm{E}[\epsilon|x_n]\right)$$

The parameterization in DDPM is consistent with the optimal solution

Proof

$$\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{0:N})||p(x_{0:N}))$$

$$\Updownarrow$$

$$\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{n-1}|x_n)||p(x_{n-1}|x_n)), \qquad n = 1, \ldots, N$$

The problem becomes:
Use a Gaussian distribution to approximate a target distribution, which is exactly the moment matching.

Transition of denoising: $q(x_{n-1}|x_n)$

- Mean: $\mathrm{E}[x_{n-1}|x_n]$
- Covariance: $\mathrm{Cov}[x_{n-1}|x_n]$

Model transition: $p(x_{n-1}|x_n) = N(\mu_n(x_n), \sigma_n^2 I)$

- Mean: $\mu(x_n)$
- Variance: σ_n^2

Proof

$$\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{0:N})||p(x_{0:N}))$$

$$\Updownarrow$$

$$\min_{\mu_n(\cdot),\sigma_n^2} KL(q(x_{n-1}|x_n)||p(x_{n-1}|x_n)), \qquad n = 1, \dots, N$$

The problem becomes:
Use a Gaussian distribution to approximate a target distribution, which is exactly the moment matching.

Transition of denoising: $q(x_{n-1}|x_n)$

- Mean: $\mathrm{E}[x_{n-1}|x_n]$
- Covariance: $\mathrm{Cov}[x_{n-1}|x_n]$

Model transition: $p(x_{n-1}|x_n) = N(\mu_n(x_n), \sigma_n^2 I)$

- Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$
- Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \mathrm{E}[\mathrm{E}[x_{n-1}|x_n, x_0]|x_n]$

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \mathrm{E}[\mathrm{Cov}(x_{n-1}|x_n, x_0)|x_n] + \mathrm{Cov}(\mathrm{E}[x_{n-1}|x_n, x_0]|x_n)$

$x_{n-1}|x_n, x_0 \sim N(\tilde{\mu}_n(x_n, x_0), \tilde{\beta}_n)$ A result in DDPM paper

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{7}$$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \mathrm{E}[\overset{\tilde{\mu}_n(x_n, x_0)}{\overset{\shortparallel}{\mathrm{E}[x_{n-1}|x_n, x_0]}}|x_n]$

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \mathrm{E}[\underset{\tilde{\beta}_n}{\underset{\shortparallel}{\mathrm{Cov}(x_{n-1}|x_n, x_0)}}|x_n] + \mathrm{Cov}(\overset{\tilde{\mu}_n(x_n, x_0)}{\overset{\shortparallel}{\mathrm{E}[x_{n-1}|x_n, x_0]}}|x_n)$

$x_{n-1}|x_n, x_0 \sim N(\tilde{\mu}_n(x_n, x_0), \tilde{\beta}_n)$ A result in DDPM paper

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{7}$$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \mathrm{E}[\tilde{\mu}_n(x_n, x_0)|x_n] = \tilde{\mu}_n(x_n, \boxed{\mathrm{E}[x_0|x_n]})$ ← x_0 prediction form

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \mathrm{E}\left[\tilde{\beta}_n I \middle| x_n\right] + \mathrm{Cov}(\tilde{\mu}_n(x_n, x_0)|x_n) = \tilde{\beta}_n I + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\boxed{\mathrm{Cov}(x_0|x_n)}$

$x_{n-1}|x_n, x_0 \sim N(\tilde{\mu}_n(x_n, x_0), \tilde{\beta}_n)$ A result in DDPM paper

$$\text{where} \quad \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \quad \text{and} \quad \tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{7}$$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \mathrm{E}[\tilde{\mu}_n(x_n, x_0)|x_n] = \tilde{\mu}_n(x_n, \boxed{\mathrm{E}[x_0|x_n]})$ ← x_0 prediction form

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \mathrm{E}\left[\tilde{\beta}_n I \middle| x_n\right] + \mathrm{Cov}(\tilde{\mu}_n(x_n, x_0)|x_n) = \tilde{\beta}_n I + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\boxed{\mathrm{Cov}(x_0|x_n)}$

x_0 prediction form to noise prediction form:

$$x_n = \sqrt{\bar{\alpha}_n}x_0 + \sqrt{\bar{\beta}_n}\epsilon \rightarrow \boxed{x_n - \sqrt{\bar{\alpha}_n}x_0 = \sqrt{\bar{\beta}_n}\epsilon}$$

Take expectation

Take covariance

$$\boxed{x_n - \sqrt{\bar{\alpha}_n}\mathrm{E}[x_0|x_n] = \sqrt{\bar{\beta}_n}\mathrm{E}[\epsilon|x_n]}$$

$$\boxed{\bar{\alpha}_n\mathrm{Cov}(x_0|x_n) = \bar{\beta}_n\mathrm{Cov}(\epsilon|x_n)}$$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

x_0 prediction form

noise prediction form

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \tilde{\mu}_n(x_n, \mathrm{E}[x_0|x_n]) = \tilde{\mu}_n\left(x_n, \frac{1}{\sqrt{\bar{\alpha}_n}}(x_n - \sqrt{\bar{\beta}_n}\mathrm{E}[\epsilon|x_n])\right)$

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \tilde{\beta}_n I + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\mathrm{Cov}(x_0|x_n) = \tilde{\beta}_n I + \frac{\beta_n^2}{\bar{\beta}_n \alpha_n}\mathrm{Cov}(\epsilon|x_n)$

x_0 prediction form to noise prediction form:

$x_n = \sqrt{\bar{\alpha}_n}x_0 + \sqrt{\bar{\beta}_n}\epsilon \rightarrow x_n - \sqrt{\bar{\alpha}_n}x_0 = \sqrt{\bar{\beta}_n}\epsilon$

Take expectation

Take covariance

$x_n - \sqrt{\bar{\alpha}_n}\mathrm{E}[x_0|x_n] = \sqrt{\bar{\beta}_n}\mathrm{E}[\epsilon|x_n]$

$\bar{\alpha}_n\mathrm{Cov}(x_0|x_n) = \bar{\beta}_n\mathrm{Cov}(\epsilon|x_n)$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

x_0 prediction form

noise prediction form

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \tilde{\mu}_n(x_n, \mathrm{E}[x_0|x_n]) = \tilde{\mu}_n\left(x_n, \frac{1}{\sqrt{\bar{\alpha}_n}}(x_n - \sqrt{\bar{\beta}_n}\mathrm{E}[\epsilon|x_n])\right)$

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \tilde{\beta}_n I + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\mathrm{Cov}(x_0|x_n) = \tilde{\beta}_n I + \frac{\beta_n^2}{\bar{\beta}_n\alpha_n}\mathrm{Cov}(\epsilon|x_n)$

Calculate the optimal mean:

$$\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n] = \tilde{\mu}_n\left(x_n, \frac{1}{\sqrt{\bar{\alpha}_n}}(x_n - \sqrt{\bar{\beta}_n}\mathrm{E}[\epsilon|x_n])\right) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}}\mathrm{E}[\epsilon|x_n]\right)$$

Proof

Optimal Mean: $\mu^*(x_n) = \mathrm{E}[x_{n-1}|x_n]$

Optimal Variance: $\sigma_n^{*2} = \mathrm{E}[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}]$

x_0 prediction form

noise prediction form

Law of total expectation (全期望公式): $\mathrm{E}[x_{n-1}|x_n] = \tilde{\mu}_n(x_n, \mathrm{E}[x_0|x_n]) = \tilde{\mu}_n\left(x_n, \frac{1}{\sqrt{\bar{\alpha}_n}}(x_n - \sqrt{\bar{\beta}_n}\mathrm{E}[\epsilon|x_n])\right)$

Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \tilde{\beta}_n I + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\mathrm{Cov}(x_0|x_n) = \tilde{\beta}_n I + \frac{\beta_n^2}{\bar{\beta}_n\alpha_n}\mathrm{Cov}(\epsilon|x_n)$

Calculate the optimal variance:

$\mathrm{Cov}(\epsilon|x_n) = \mathrm{E}[\epsilon\epsilon^\top|x_n] - \mathrm{E}[\epsilon|x_n]\mathrm{E}[\epsilon^\top|x_n]$ // the expansion of covariance

$\mathrm{E}[\mathrm{Cov}(\epsilon|x_n)] = \mathrm{E}[\epsilon\epsilon^\top] - \mathrm{E}\left[\mathrm{E}[\epsilon|x_n]\mathrm{E}[\epsilon^\top|x_n]\right] = I - \mathrm{E}\left[\mathrm{E}[\epsilon|x_n]\mathrm{E}[\epsilon^\top|x_n]\right]$ // taking expectation

$$\sigma_n^{*2} = \mathrm{E}\left[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}\right] = \tilde{\beta}_n + \frac{\beta_n^2}{\bar{\beta}_n\alpha_n}\mathrm{E}\left[\frac{\mathrm{tr}(\mathrm{Cov}[\epsilon|x_n])}{d}\right] = \tilde{\beta}_n + \frac{\beta_n^2}{\bar{\beta}_n\alpha_n}\left(1 - \mathrm{E}\left[\frac{\|\mathrm{E}[\epsilon|x_n]\|^2}{d}\right]\right) = \frac{\beta_n}{\alpha_n}\left(1 - \frac{\beta_n}{\bar{\beta}_n}\mathrm{E}\frac{\|\mathrm{E}[\epsilon|x_n]\|^2}{d}\right)$$

- Now we have finished the proof
- There is also some byproduct…

Byproduct (bounds of σ_n^{*2})

$\leq \tilde{\beta}_n + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\left(\frac{b-a}{2}\right)^2$ (assume x_0 is bounded in $[a, b]$)

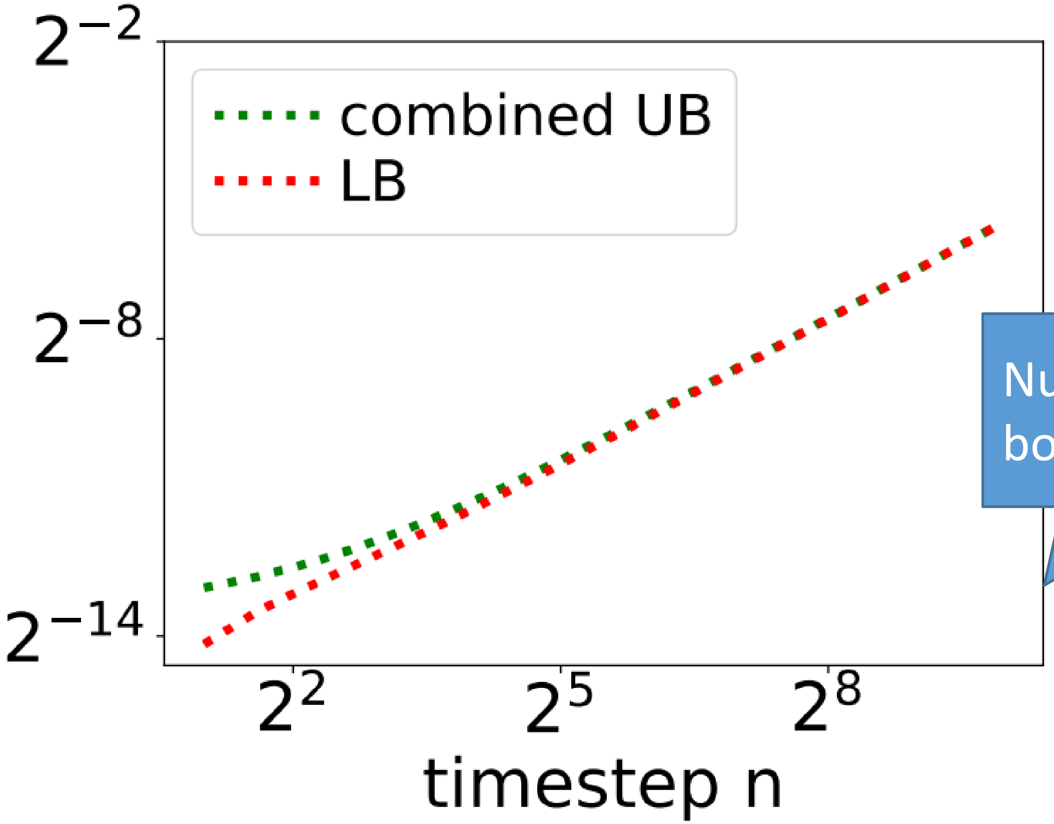
Law of total variance (全方差公式): $\mathrm{Cov}(x_{n-1}|x_n) = \tilde{\beta}_n I + \frac{\bar{\alpha}_{n-1}\beta_n^2}{\bar{\beta}_n^2}\mathrm{Cov}(x_0|x_n) = \tilde{\beta}_n I + \frac{\beta_n^2}{\bar{\beta}_n \alpha_n}\mathrm{Cov}(\epsilon|x_n)$

$$\sigma_n^{*2} = \mathrm{E}\left[\frac{\mathrm{tr}(\mathrm{Cov}[x_{n-1}|x_n])}{d}\right] = \tilde{\beta}_n + \frac{\beta_n^2}{\bar{\beta}_n\alpha_n}\mathrm{E}\left[\frac{\mathrm{tr}(\mathrm{Cov}[\epsilon|x_n])}{d}\right] = \tilde{\beta}_n + \frac{\beta_n^2}{\bar{\beta}_n\alpha_n}\left(1 - \mathrm{E}\left[\frac{\|\mathrm{E}[\epsilon|x_n]\|^2}{d}\right]\right) = \frac{\beta_n}{\alpha_n}\left(1 - \frac{\beta_n}{\bar{\beta}_n}\mathrm{E}\frac{\|\mathrm{E}[\epsilon|x_n]\|^2}{d}\right)$$

$\geq \tilde{\beta}_n$

$\leq \frac{\beta_n}{\alpha_n}$

- Tightness of bounds of the optimal variance σ_n^{*2}
- Empirically, we clip our estimate using these bounds



Numerically, these bounds are tight

- **Analytic** estimate of the optimal mean

optimal mean to KL: $\mu^*(x_n) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\bar{\beta}_n}} \boxed{\mathrm{E}[\epsilon|x_n]}\right)$

$$\begin{array}{c} \wr\wr \\ \boxed{\epsilon_n(x_n)} \end{array}$$

DDPM is the analytic estimate of the optimal mean

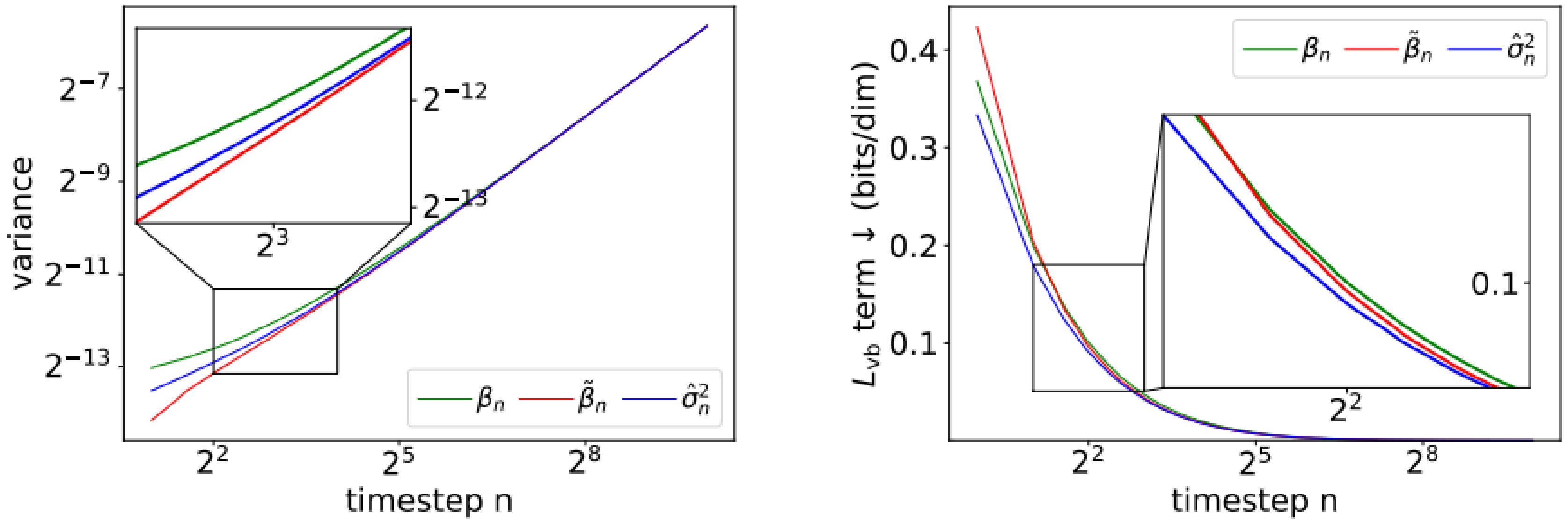
- **Analytic** estimate of the optimal variance

optimal variance to KL: $\sigma_n^{*2} = \frac{\beta_n}{\alpha_n}\left(1 - \frac{\beta_n}{\bar{\beta}_n} \boxed{\mathrm{E}\frac{\|\mathrm{E}[\epsilon|x_n]\|^2}{d}}\right)$

$\wr\wr$

$\epsilon_n(x_n) \approx \mathrm{E}[\epsilon|x_n]$+ Monte Carlo: $\Lambda_n = \boxed{\frac{1}{M}\sum_{m=1}^{M}\frac{\left\|\epsilon_n(x_{n,m})\right\|^2}{d}}, \ x_{n,m} \sim q_n(x_n)$

Analytic estimate of the optimal variance: $\hat{\sigma}_n^2 = \frac{\beta_n}{\alpha_n}\left(1 - \frac{\beta_n}{\bar{\beta}_n}\Lambda_n\right)$

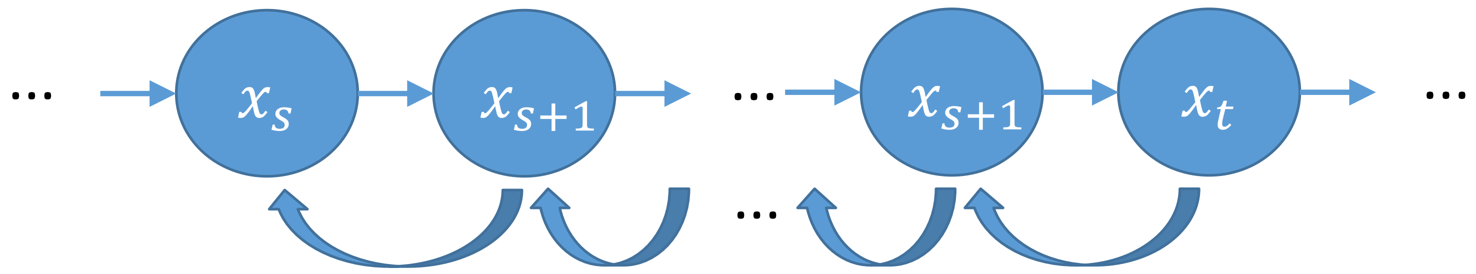


(a) Variance (b) Terms in L_{vb}

Figure 1: Comparing our analytic estimate $\hat{\sigma}_n^2$ and prior works with handcrafted variances β_n and $\tilde{\beta}_n$. (a) compares the values of the variance for different timesteps. (b) compares the term in L_{vb} corresponding to each timestep. The value of L_{vb} is the area under the corresponding curve.

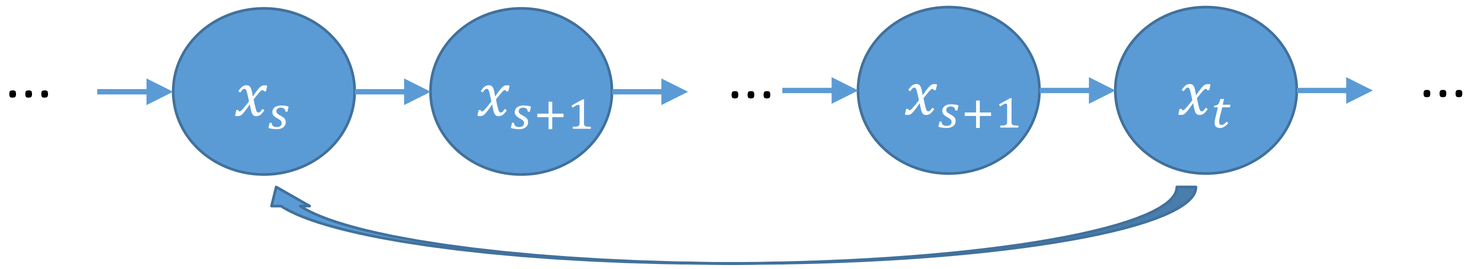
Fast inference

Original: one step



Fast inference

Fast: multiple step



$$\mu^*_{s|t}(x_t) = \frac{1}{\sqrt{\alpha_{t|s}}}\left(x_t + \beta_{t|s}\nabla \log q_t(x_t)\right) = \frac{1}{\sqrt{\alpha_{t|s}}}\left(x_t - \frac{\beta_{t|s}}{\sqrt{\overline{\beta}_{t|s}}}\mathrm{E}[\epsilon|x_t]\right),$$

$$\sigma^{*2}_{s|t} = \frac{\beta_{t|s}}{\alpha_{t|s}}\left(1 - \beta_{t|s}\mathrm{E}_{q_t(x_t)}\frac{\|\nabla \log q_t(x_t)\|^2}{d}\right) = \frac{\beta_{t|s}}{\alpha_{t|s}}\left(1 - \frac{\beta_{t|s}}{\overline{\beta}_{t|s}}\mathrm{E}_{q_t(x_t)}\frac{\|\mathrm{E}[\epsilon|x_t]\|^2}{d}\right).$$

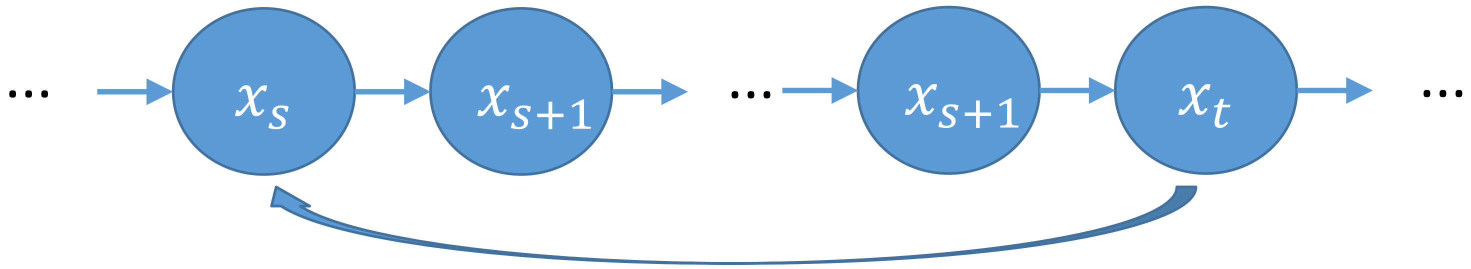
Fast inference

Formalize as KL minimization w.r.t. time steps
The objective is also **analytic**!
Can be solved by **dynamic programming**!

$$\min_{\tau_1,\cdots,\tau_K} D_{\mathrm{KL}}(q(\boldsymbol{x}_0,\boldsymbol{x}_{\tau_1},\cdots,\boldsymbol{x}_{\tau_K})||p^*(\boldsymbol{x}_0,\boldsymbol{x}_{\tau_1},\cdots,\boldsymbol{x}_{\tau_K})) = \frac{d}{2}\sum_{k=2}^{K} J(\tau_{k-1},\tau_k)+c,$$

where $J(\tau_{k-1},\tau_k) = \log(\sigma^{*2}_{\tau_{k-1}|\tau_k}/\lambda^2_{\tau_{k-1}|\tau_k})$

How to choose the time steps?



$$\mu^*_{s|t}(x_t) = \frac{1}{\sqrt{\alpha_{t|s}}}(x_t+\beta_{t|s}\nabla\log q_t(x_t)) = \frac{1}{\sqrt{\alpha_{t|s}}}\left(x_t - \frac{\beta_{t|s}}{\sqrt{\bar{\beta}_{t|s}}}\mathrm{E}[\epsilon|x_t]\right),$$

$$\sigma^{*2}_{s|t} = \frac{\beta_{t|s}}{\alpha_{t|s}}(1-\beta_{t|s}\mathrm{E}_{q_t(x_t)}\frac{\|\nabla\log q_t(x_t)\|^2}{d}) = \frac{\beta_{t|s}}{\alpha_{t|s}}(1-\frac{\beta_{t|s}}{\bar{\beta}_{t|s}}\mathrm{E}_{q_t(x_t)}\frac{\|\mathrm{E}[\epsilon|x_t]\|^2}{d}).$$

Empirical performance after combining all techniques:

- Density estimation: 1000 steps -> 25-50 steps + better performance
- Sample quality: 1000 steps -> 50-100 steps + comparable performance

More works…

Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models
Bao et al., ICML 2022

What is the optimal diagonal covariance $\Sigma_n(x_n) = \mathrm{diag}(\sigma_n^2(x_n))$?

Theorem 1. Suppose $\Sigma_n(x_n) = \mathrm{diag}(\sigma_n^2(x_n))$. The optimal solution is

$$\mu_n^*(x_n) = \frac{1}{\sqrt{\alpha_n}}\left(x_n - \frac{\beta_n}{\sqrt{\overline{\beta}_n}}\mathbf{E}_{q(x_0|x_n)}[\epsilon_n]\right),$$

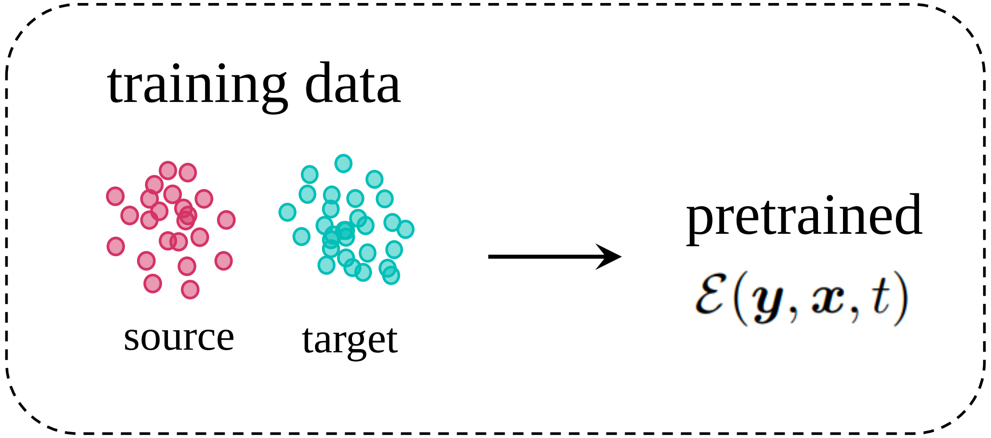
$$\sigma_n^*(x_n)^2 = \frac{\overline{\beta}_{n-1}}{\overline{\beta}_n}\beta_n + \frac{\beta_n^2}{\overline{\beta}_n\alpha_n}(\underbrace{\mathbf{E}_{q(x_0|x_n)}[\epsilon_n^2]}_{\approx h_n(x_n)} - \underbrace{\mathbf{E}_{q(x_0|x_n)}[\epsilon_n]^2}_{\approx \hat{\epsilon}_n(x_n)^2}).$$

predict SN: $\min_{h_n}\mathbf{E}_{q(x_0,x_n)}||h_n(x_n) - \boxed{\epsilon_n^2}||^2$

squared noise (SN)

Energy-Guided Stochastic Differential Equations (EGSDE)

Zhao et al, NeurIPS 2022



source image

$\boldsymbol{x}_0$

$$\boldsymbol{y}_M \longrightarrow \mathrm{d}\boldsymbol{y} = [\boldsymbol{f}(\boldsymbol{y}, t) - g(t)^2(\boldsymbol{s}(\boldsymbol{y}, t) - \nabla_{\boldsymbol{y}}\mathcal{E}(\boldsymbol{y}, \boldsymbol{x}_0, t))]\mathrm{d}t + g(t)\mathrm{d}\overline{\boldsymbol{w}} \longrightarrow \boldsymbol{y}_0$$

Following the SDE and decreasing the energy at the same time

Thanks!