Understanding Diffusion Objectives as the ELBO with Data Augmentation

DP Kingma & Ruiqi Gao

Google DeepMind

Oct, 2025

Presented by Scott Sun from Duke B&B

Introduction

SOTA diffusion models are optimized with objectives that look very different from the ELBO. Indeed, in the paper, the authors show that **all diffusion objectives equal the ELBO, combined with Gaussian noise perturbation**. Moreover, the objectives are weighted losses of MSE and can achieve better performance & efficiency when the weights are monotonic function of time t (i.e. #noising steps).

Goal:

1. unify all diffusion methods with one general weighting loss
2. in theory, noise scheduling' curve shape does not affect diffusion, the weight function does
3. monotonic weighting function for noise scheduling $\Rightarrow$ Diffusion $=$ ELBO w/ DA

Background: DDPM

noising/forward process: $q(x_t|x_{t-1}) \stackrel{\text{def}}{=} \mathcal{N}(x_t|\sqrt{\alpha_t}x_{t-1}, (\underbrace{1-\alpha_t}_{\beta_t})\mathbf{I}) \Leftrightarrow x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}$

$$q(x_t|x) = \mathcal{N}(x_t|\sqrt{\bar{\alpha}_t}x_{t-1}, (1-\bar{\alpha}_t)\mathbf{I}), \quad \bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i \tag{1}$$

denoising/backward process: $p_\theta(x_{t-1}|x_t) \stackrel{\text{def.}}{=} \mathcal{N}(x_{t-1}|\mu_\theta(x_t), \Sigma_t)$, where $\Sigma_t \stackrel{\text{def.}}{=} \sigma_q^2(t)\mathbf{I}$ to match the *reverse process* $q(x_{t-1}|x_t, x_0) \stackrel{\text{der.}}{=} \mathcal{N}(x_{t-1}|\mu_q(x_t, x_0), \sigma_q^2(t)\mathbf{I})$. Note μ_q and σ_q^2 are parameterized by $\{\alpha_t\}$

$$\mu_q(x_t, x_0) = \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t}x_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}x_0$$

$$\sigma_q^2(t) = \frac{(1-\alpha_t)(1-\sqrt{\bar{\alpha}_{t-1}})}{1-\bar{\alpha}_t}$$

objective:

$$\log p(x) = \log \int p(x_{0:T})dx_{1:T} \geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_{0:T})}{q(x_{1:T}|x_0)}\right] \stackrel{\text{def}}{=} \text{ELBO} \tag{2}$$

Background: DDPM (cont'd)

If we reparameterize the mean in the backward process as

$$\mu_\theta(x_t) \stackrel{\text{def}}{=} \frac{(1-\bar{\alpha}_{t-1})\sqrt{\alpha_t}}{1-\bar{\alpha}_t}\, x_t + \frac{(1-\alpha_t)\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\, \hat{x}_\theta(x_t) \tag{3}$$

The ELBO can decomposed and finally simplified as follows

$$\begin{aligned} \mathrm{ELBO}_\theta(x) &= \mathbb{E}_{q(x_1|x_0)}[\log p_\theta(x_0|x_1)] - \cancel{\mathbb{D}_{\mathrm{KL}}(q(x_T|x_0) \,\|\, p(x_T))} \\ &\quad - \sum_{t=2}^{T} \mathbb{E}_{q(x_t|x_0)}[\mathbb{D}_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0) \,\|\, p_\theta(x_{t-1}|x_t))] \qquad (4) \\ &\triangleq -\frac{1}{2}\sum_{t=1}^{T} \mathbb{E}_{q(x_t|x_0)}\left[\frac{1}{\sigma_q^2(t)} \frac{(1-\alpha_t)^2\,\bar{\alpha}_{t-1}}{(1-\bar{\alpha}_t)^2} \left\|\hat{x}_\theta(x_t; t) - x_0\right\|^2\right] \qquad (5) \end{aligned}$$

Linear schedule variance β_t from `1e-4` to `0.02` over `t=1...1e3`.
Therefore, the signal-to-noise ratio $\downarrow$

x or x_0 := original sample
x_t := noisy sample at t
p_θ := backward
q := forward or reverse

Background: VDM

In VDM, time t is mapped to interval $[0, 1]$. They generalize the framework for both discrete-time model and continuous-time model. VDM re-parameterizes the original hyperparams: $\alpha_t \overset{\text{repar.}}{\longleftarrow} \sqrt{\bar{\alpha}_t}$ and $\sigma_t^2 \overset{\text{repar.}}{\longleftarrow} 1 - \bar{\alpha}_t$ and defines SNR as follows.

$$\text{SNR}(t) = \alpha_t^2/\sigma_t^2 \tag{6}$$

discrete-time: using the new SNR parameterization and let $i \sim \mathcal{U}\{1, ..., T\}, s(i) = \frac{i-1}{T}$ and $t(i) = \frac{i}{T}$

$$-\text{ELBO} \triangleq \mathcal{L}_T = \frac{T}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}),\, i \sim \mathcal{U}\{1,T\}} \left[\big(\mathrm{SNR}(s) - \mathrm{SNR}(t)\big) \|x - \hat{x}_\theta(z_t; t)\|_2^2 \right] \tag{7}$$

continuous-time: when $T \to \infty$

$$\mathcal{L}_\infty = -\frac{1}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}),\, t \sim \mathcal{U}(0,1)} \left[\mathrm{SNR}'(t) \|x - \hat{x}_\theta(z_t; t)\|_2^2 \right] = \frac{1}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}),\, t \sim \mathcal{U}(0,1)} \left[\gamma'_\eta(t) \|\epsilon - \hat{\epsilon}_\theta(z_t; t)\|_2^2 \right] \tag{8}$$

VDM use a NN $\gamma_\eta(t)$ to model SNR and ensure the SNR is monotonic.
$\gamma_\eta(t) = l_1(t) + l_3(\,\text{sigmoid}(l_2(l_1(t)))\,)$
$\text{SNR}(t) = \exp(-\gamma_\eta(t))$

x or x_0 := original sample
z_t := noisy sample at t
p_θ := backward
q := forward or reverse

Method: ELBO objective

Slightly modify the notation in VDM, rewrite the forward process as follows.

$$z_t = \alpha_\lambda x + \sigma_\lambda \epsilon, \qquad \alpha_\lambda^2 + \sigma_\lambda^2 = 1 \quad \text{(variance preserving)} \tag{9}$$

where λ is the log-SNR at t s.t. $\lambda = \log \frac{\alpha_\lambda^2}{\sigma_\lambda^2}$
Then, the new expression for ELBO is

$$-\mathrm{ELBO} \triangleq \mathcal{L}(x) = \frac{1}{2} \mathbb{E}_{t \sim \mathcal{U}(0,1),\, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I})} \left[-\frac{d\lambda}{dt} \cdot \|\hat{\epsilon}_\theta(z_t; \lambda_t) - \epsilon\|_2^2 \right] \tag{10}$$

$\boxed{\mathrm{SNR}(t) = \exp(-\gamma_\eta(t))}$

Method: diffusion objective as a weighted loss

Generalizing the loss by adding a weighting term

$$\mathcal{L}_w(x) = \frac{1}{2}\,\mathbb{E}_{t\sim\mathcal{U}(0,1),\,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[w(\lambda_t) \cdot \left(-\frac{d\lambda}{dt}\right) \cdot \|\hat{\epsilon}_\theta(z_t;\lambda_t) - \epsilon\|_2^2 \right] \tag{11}$$

Loss function	**Implied weighting** $w(\lambda)$	**Monotonic?**
ELBO [Kingma et al., 2021, Song et al., 2021a]	1	✓
IDDPM ($\boldsymbol{\epsilon}$-prediction with 'cosine' schedule) [Nichol and Dhariwal, 2021]	$\mathrm{sech}(\lambda/2)$	
EDM [Karras et al., 2022] (Appendix D.1)	$\mathcal{N}(\lambda; 2.4, 2.4^2) \cdot (e^{-\lambda} + 0.5^2)$	
$\mathbf{v}$-prediction with 'cosine' schedule [Salimans and Ho, 2022] (Appendix D.2)	$e^{-\lambda/2}$	✓
Flow Matching with OT path (FM-OT) [Lipman et al., 2022] (Appendix D.3)	$e^{-\lambda/2}$	✓
InDI [Delbracio and Milanfar, 2023] (Appendix D.4)	$e^{-\lambda}\mathrm{sech}^2(\lambda/4)$	✓
P2 weighting with 'cosine' schedule [Choi et al., 2022] (Appendix D.5)	$\mathrm{sech}(\lambda/2)/(1+e^{\lambda})^{\gamma}, \gamma = 0.5$ or 1	
Min-SNR-γ [Hang et al., 2023] (Appendix D.6)	$\mathrm{sech}(\lambda/2) \cdot \min(1, \gamma e^{-\lambda})$	

Figure: of note, it's about the monotonicity w.r.t. λ

Method: invariance to noise scheduling

With change of variable from t to λ in the integration (note: $\lambda_{\text{max}} = \lambda_{t=0}$ and $\lambda_{\text{min}} = \lambda_{t=1}$)

$$\mathcal{L}_w(\mathbf{x}) = \frac{1}{2}\int_{\lambda_{\text{min}}}^{\lambda_{\text{max}}} w(\lambda)\, \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda;\lambda) - \boldsymbol{\epsilon}\|_2^2\right] d\lambda \tag{12}$$

Thus, **in theory**, the noise scheduling only matters in terms of the boundary values but not the shape of the curve. Only the weighting function $w(x)$ matters!

However, **in reality**, when we use MC estimator, the scheduling matters.

We can further rewrite the w-loss

$$\mathcal{L}_w(\mathbf{x}) = \frac{1}{2}\,\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\, \lambda\sim p(\lambda)}\left[\frac{w(\lambda)}{p(\lambda)}\,\|\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda;\lambda) - \boldsymbol{\epsilon}\|_2^2\right] \tag{13}$$

which clarifies the role of $p(\lambda)$ as an importance sampling distribution

Given $t \sim U(0,1)$, pdf $p(\lambda) = -\frac{dt}{d\lambda}$

Method: weighted loss = ELBO w/ DA

Theorem 1.: If the weighting $w(\lambda_t)$ is monotonic, then the weighted diffusion objective is equivalent to the ELBO with data augmentation (additive noise).

proof part 1

With monotonic $w(\lambda_t)$ we mean that w is a monotonically increasing function of t, and therefore a monotonically decreasing function of λ.

We'll use shorthand notation $\mathcal{L}(t;\mathbf{x})$ for the KL divergence between the joint distributions of the forward process $q(\mathbf{z}_{t,\ldots 1}|\mathbf{x})$ and the reverse model $p(\mathbf{z}_{t,\ldots 1})$, for the subset of timesteps from t to 1:

$$\boxed{\mathcal{L}(t;\mathbf{x}) := D_{KL}(q(\mathbf{z}_{t,\ldots,1}|\mathbf{x})||p(\mathbf{z}_{t,\ldots,1}))} \tag{7}$$

In Appendix A.1, we prove that[2]:

$$\frac{d}{dt}\mathcal{L}(t;\mathbf{x}) = \frac{1}{2}\frac{d\lambda}{dt}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I})}\left[||\boldsymbol{\epsilon} - \hat{\boldsymbol{\epsilon}}_{\boldsymbol{\theta}}(\mathbf{z}_\lambda;\lambda)||_2^2\right] \tag{8}$$

As shown in Appendix A.1, this allows us to rewrite the weighted loss of Equation 4 as simply:

$$\mathcal{L}_w(\mathbf{x}) = -\int_0^1 \frac{d}{dt}\mathcal{L}(t;\mathbf{x})\, w(\lambda_t)\, dt \tag{9}$$

In Appendix A.2, we prove that using integration by parts, the weighted loss can then be rewritten as:

$$\mathcal{L}_w(\mathbf{x}) = \int_0^1 \frac{d}{dt}w(\lambda_t)\,\mathcal{L}(t;\mathbf{x})\, dt + w(\lambda_{\max})\,\mathcal{L}(0;\mathbf{x}) + \text{constant} \tag{10}$$

proof part 2

Now, assume that $w(\lambda_t)$ is a monotonically increasing function of $t \in [0, 1]$. Also, without loss of generality, assume that $w(\lambda_t)$ is normalized such that $w(\lambda_1) = 1$. We can then further simplify to an expected KL divergence:

$$\boxed{\mathcal{L}_w(\mathbf{x}) = \mathbb{E}_{p_w(t)}\left[\mathcal{L}(t; \mathbf{x})\right] + \text{constant}} \tag{11}$$

where $p_w(t)$ is a probability distribution determined by the weighting function, namely $p_w(t) := (d/dt\ w(\lambda_t))$, with support on $t \in [0, 1]$. The probability distribution $p_w(t)$ has Dirac delta peak of typically very small mass $w(\lambda_{\text{max}})$ at $t = 0$.

Note that:

$$\begin{aligned} \mathcal{L}(t; \mathbf{x}) &= D_{KL}(q(\mathbf{z}_{t,\dots,1}|\mathbf{x})||p(\mathbf{z}_{t,\dots,1})) & (12) \\ &\geq D_{KL}(q(\mathbf{z}_t|\mathbf{x})||p(\mathbf{z}_t)) = -\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x})}[\log p(\mathbf{z}_t)] + \text{constant}. & (13) \end{aligned}$$

More specifically, $\mathcal{L}(t; \mathbf{x})$ equals the expected negative ELBO of noise-perturbed data, plus a constant; see Section C for a detailed derivation.

This concludes our proof of Theorem 1. ∎

Recommendation

Is it worth reading? Yes.

- unifies all diffusion objective & comprehensive discuss the conversion in Appx
- derivations are thorough and detailed

Is it worth implementing? Yes.

- it is worth while to read through the paper; the paper provides a holistic review of diffusion as a variational method and unifies all objectives in a clear & clean way!