A DIFFUSION MODEL INDUCED BY MSE TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

A diffusion model for image generation transforms noise into an image via a neural denoiser. The denoiser is trained with a time-integrated, weighted mean-squared error (MSE) between the noised image and the network's prediction. The weighting is often absorbed into the noised image, yielding different parameterizations of the prediction (e.g., noise-, data-, or velocity parameterization). Thus, the denoiser is determined by the noise schedule and the chosen parameterization, whereas the generative diffusion process is specified by its noise and diffusion schedules (i.e., by both the scale and the variance-rate coefficients). In practice, the generator typically inherits only the noise schedule from the trained denoiser. In this work, guided by a principle of coherence between the MSE training objective and maximum-likelihood (ML) proximity of the induced processes, we derive a closed-form expression for the diffusion schedule given a noise schedule and a network parameterization. Widely used methods train on one (implicit) process but generate with another—often one with an optimal diffusion schedule in the ML sense, or even with zero diffusion, that is a deterministic flow. Recent empirical approaches yield diffusion schedules closer to our formula, which supports the coherence principle and suggests that it is beneficial to generate samples using the very process that is actually learned. We analyze both discrete-time and continuous-time models using elementary autoregressive arguments, yielding formulas that are simpler than those used so far. In particular, we provide a representation of the diffusion state as the sum of an explicit linear component, an unweighted pathwise integral of the denoiser, and a noise term. This representation makes it straightforward to apply classical numerical integration methods and clarifies the relation to the DPM-solver family.

1 INTRODUCTION

A typical diffusion generative model for image generation transforms noise into an image over a few-dozen to a few-hundred time steps by means of a neural network. a neural denoiser. The denoiser is trained with a time-integrated, weighted mean-squared error (MSE) between the noised image and the network's prediction. The weighting is often absorbed into the noised image, yielding different parameterizations of the prediction (e.g., noise-, data-, or velocity parameterization). Thus, the denoiser is determined by the noise schedule and the chosen parameterization, whereas the generative diffusion process is specified by its noise and diffusion schedules (i.e., by both the scale and the variance-rate coefficients). In practice, the generator typically inherits only the noise schedule from the trained denoiser. Widely used methods train on one (implicit) process but use one of two processes for generation: either one with a diffusion schedule which is optimal with respect to maximum likelihood (ML) or one with no diffusion at all, that is a deterministic flow (Ho et al., 2020; Song et al., 2021b; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Salimans & Ho, 2021; Song et al., 2021a; Kingma et al., 2021; Ho & Salimans, 2022; Rombach et al., 2022; Kingma & Gao, 2023; Esser et al., 2024).

In this work, guided by a principle of coherence between the empirically shaped MSE training objective and the main criterion for fitting distributions to data, that is ML-proximity of the induced processes, we derive a closed-form expression for the diffusion schedule given a noise schedule and a network parameterization (Proposition 3). Recent empirical approaches yield diffusion schedules closer to our formula, which supports the coherence principle and suggests that it is beneficial to generate samples using the very process that is actually learned (Ma et al., 2024; Cui et al., 2025).

We analyze both discrete-time and continuous-time models using elementary autoregressive arguments, yielding formulas that are simpler than those used so far. In particular, we provide a representation of the diffusion state as the sum of an explicit linear component, an unweighted pathwise integral of the denoiser, and a noise term (Section 3.2).

Our representation makes it straightforward to apply classical numerical integration methods. For comparison, schemes designed specifically for diffusion—such as DPM-solvers—require integrating a pathwise integral of the signal or noise estimators with an exponential weight, which is difficult. To our knowledge, an analogue of one of the most popular universal schemes, the Runge-Kutta method of order 4, has not yet been developed (Lu et al., 2023; 2022; Cui et al., 2025).

2 A DENOISER INDUCED BY MSE TRAINING

Assume that we have two positive, continuously differentiable functions of time $t \in (0,1)$, namely increasing *signal schedule* schedule α_t and decreasing *noise schedule* σ_t. Let us $\mathring{\rho}_t = \alpha_t/\sigma_t, \mathring{\lambda}_t = \log \mathring{\rho}_t$ denote *signal-to-noise ratio* oraz *log signal-to-noise ratio*, analogon. (The ring accents over the symbols indicate that these functions are special cases of more general functions, without rings, which will be defined later.) Let $t \sim U(0,1), X \sim p_x$ in $\mathbb{R}^d$ and $\varepsilon \sim \mathcal{N}(0, I_d)$ be independent. We consider the *linear noise generators*

$$\bar{Y}_t \;:=\; X + \mathring{\rho}_t^{-1}\,\varepsilon \quad \text{and} \quad \bar{Z}_t \;:=\; \alpha_t \bar{Y}_t \;=\; \alpha_t\, X + \sigma_t\, \varepsilon. \tag{1}$$

We train a *scaled denoiser* $\hat{u}_t : \mathbb{R}^d \to \mathbb{R}^d$ by fitting its parameters, denoted as a hat, according to the *mean squared error* (MSE)

$$\min_{\wedge}\; \mathbb{E}_{t,\varepsilon,X} \left\|\hat{u}_t(\alpha_t \bar{Y}_t) - u_t\right\|^2 \;=\; \min_{\wedge} \int_0^1 \mathbb{E}_{\varepsilon,X} \left\|\hat{u}_t(\alpha_t \bar{Y}_t) - u_t\right\|^2 \,\mathrm{d}t, \tag{2}$$

where $t \sim U(0,1), u_t = A_t \bar{Y}_t + S_t \varepsilon$ is a target and functions A_t, S_t are scaling schedules, with positive and continuous S_t called *parameterization*. From any $\hat{u}_t$ we recover an estimator of the *noise* from the formula for the target u_t and a *denoiser* or a *data* estimator via (1)

$$\hat{\varepsilon}_t(\bar{Y}_t) := \;\frac{\hat{u}_t(\alpha_t \bar{Y}_t) - A_t\, \bar{Y}_t}{S_t} \quad \text{and} \quad \hat{X}_t(\bar{Y}_t) := \; \bar{Y}_t - \mathring{\rho}_t^{-1} \hat{\varepsilon}_t(\bar{Y}_t). \tag{3}$$

A direct calculation shows that $\hat{u}_t - u_t = S_t(\hat{\varepsilon}_t - \varepsilon) = -\mathring{\rho}_t \sigma_t (\hat{X}_t - X)$. We can also define a target using data: $u_t = B_t \bar{Y}_t - C_t X$, then for this target learn the network, define $\hat{X}_t$, and then, using (1), define $\hat{\varepsilon}_t$ and set $S_t := C_t/\mathring{\rho}_t$.

For our purposes, the interface between the denoiser and the generator consists of, in addition to $\hat{\varepsilon}_t$ or $\hat{X}_t$, the pair $(C_t, \mathring{\rho}_t)$. These can be viewed as input and output scalings, respectively. Kingma & Gao (2023) showed that MSE-training is determined by $\mathring{\lambda}_t$ and weights equivalent to our S_t.

Popular noise schedules. Kingma & Gao (2023) demonstrated that three popular noise schedules can be derived uniformly as quantile functions of bell-shaped densities: normal, logistic and hyperbolic secant. In particular,

- *cosine*: $\alpha_t = \sin((t+.008)\pi/2.016),\ \sigma_t := \cos((t+.008)\pi/2.016),\ \lambda_t := \frac{\pi}{2} F^{-1}(t)$, where $F^{-1}(t) = \frac{2}{\pi}\log(\tan(\frac{\pi}{2}t))$ is the quatile function of the hyperbolic secant distribution (Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Kingma et al., 2021; Esser et al., 2024).
- *edm*: $\lambda_t := 1.2\Phi^{-1}(t) - 1.2$, where Φ is the standard normal cumulative distribution function (Karras et al., 2022; Esser et al., 2024).
- *linear*: $\alpha_t := t,\ \sigma_t := 1-t,\ \lambda_t := F^{-1}(t)$, where $F^{-1}(t) = \log(t/(1-t))$ is the quantile of the (standard) logistic distribution (Lipman et al., 2022; Liu et al., 2022; Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023; Ma et al., 2024; Esser et al., 2024).

Note that in our setting, unlike in the original works, a signal schedule is an increasing on $(0,1)$.

Popular parametrizations.

- *noise*: The network predicts ε, thus $S_t \equiv 1$ (Ho et al., 2020; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021).
- *data*: The network is trained to predict X, so $S_t = \mathring{\rho}_t$ (Kingma et al., 2021; Lu et al., 2022).
- *F-prediction*: $F_t := \sqrt{4+\mathring{\rho}_t^2}X - \mathring{\rho}_t^2/\sqrt{4+\mathring{\rho}_t^2}Y_t$, thus $S_t = \sqrt{\mathring{\rho}_t^{-2}+1}$ (Karras et al., 2022; Cui et al., 2025).
- *velocity*: The target is $v_t := \bar{Z}_t' = \alpha_t' X + \sigma_t'\varepsilon = \alpha_t'\bar{Y}_t - \mathring{\lambda}_t'\sigma_t\varepsilon$, then (3) yields the identity $\hat{v}_t(\alpha_t\bar{Y}_t) = \alpha_t\bar{Y}_t - \mathring{\lambda}_t'\sigma_t\hat{\varepsilon}_t(\bar{Y}_t)$, so $S_t = \mathring{\lambda}_t'\sigma_t$ (Lipman et al., 2022; Liu et al., 2022; Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023; Ma et al., 2024; Esser et al., 2024).

3 Diffusion models induced by the denoiser

A diffusion denoising model that generates from the distribution estimator p_x is a diffusion process with a drift estimated by a denoiser, starting from pure noise. In this section, we will first define the discrete-time diffusion noise. Then, by adding a denoiser, we will obtain a denoising process, and we will subsequently define an analogous process in continuous time. Finally, we will specify the generation interval and state our main problem.

3.1 Discrete diffusion models

Stationary, time-inhomogeneous autoregression. Let us fix a grid $0 < t_0 < t_1 < \cdots < t_N = t \le t_{max} < 1$, where $t_i = t_0 + i(t_{max}-t_0)/N, i = 0, 1, \ldots, N$. Let ρ_t be a positive, increasing, continuously differentiable function of time $t \in (0,1)$. This function, which we will refer to as the *diffusion schedule*, defines the cumulative relative variance of the diffusion processes. We also define $r_t = \rho_t\mathring{\rho}_t$ and $\lambda_t = \log\rho_t$.

Let $\{\xi_{t_i}\}_{i=0}^N$ and $\varepsilon \equiv \varepsilon_0$ be i.i.d. $\mathcal{N}(0, I_d)$ and for $t = t_i, s = t_{i-1}$ set

$$\varepsilon_t \;:=\; \frac{\rho_s}{\rho_t}\,\varepsilon_s \;+\; \sqrt{1-\frac{\rho_s^2}{\rho_t^2}}\,\xi_s. \tag{4}$$

The rescaled $\varepsilon_t, \varepsilon_s$ form an autoregressive process with additive noise

$$\rho_t\varepsilon_t \;=\; \rho_s\,\varepsilon_s \;+\; \sqrt{\rho_t^2-\rho_s^2}\,\xi_s. \tag{5}$$

From (5), it is clear that $\varepsilon_t \sim \mathcal{N}(0, I_d)$, and that ε_s, ξ_s are independent. The correlation and conditional variance are also easily computable

$$\mathrm{cor}(\varepsilon_{t,j}, \varepsilon_{s,k}) = \mathbf{1}(j=k)\rho_s/\rho_t,\ \mathbb{V}(\varepsilon_{t,j} \mid \varepsilon_{s,k}) = 1 - \mathbf{1}(j=k)\rho_s^2/\rho_t^2,\ \ j,k = 1,2,\ldots,d. \tag{6}$$

This indicates that ρ_t^2 represents the relative cumulative variance of the process $\{\varepsilon_{t_i}\}_{i=0}^N$. The scaled versions of ε_t and ε_s constitute an autoregressive process with additive noise.

Denoising diffusion models. Assume that $\{\varepsilon_{t_i}\}_{i=0}^N$ are independent of X. Let us define

$$Y_t \;:=\; X + \mathring{\rho}_t^{-1}\,\varepsilon_t \;\text{ and }\; Z_t \;:=\; \alpha_t\,X + \sigma_t\,\varepsilon_t. \tag{7}$$

It is clear that for any diffusion schedule ρ_t, the random variables Y_t, Z_t are distributed identically to the linear noise generators $\bar{Y}_t, \bar{Z}_t$ defined in (1) at times $t = t_i$. By substituting the expressions for $\varepsilon_t, \varepsilon_s$ in terms of Y_t, Y_s lub Z_t, Z_s (from (7)) into (5), we get

$$r_tY_t \;=\; r_sY_s \;+\; (r_t - r_s)\,X \;+\; \sqrt{\rho_t^2-\rho_s^2}\,\xi_s. \tag{8}$$

$$\frac{\rho_t}{\sigma_t}\,Z_t \;=\; \frac{\rho_s}{\sigma_s}\,Z_s \;+\; (r_t - r_s)\,X \;+\; \sqrt{\rho_t^2-\rho_s^2}\,\xi_s. \tag{9}$$

Thus, the function α_t, which was previously shown to be an internal learning function, is now independently found to be unnecessary for generation. It is sufficient to generate the process Y_t and, if necessary, scale it at the end of generation to obtain $Z_{t_N} = \alpha_{t_N} Y_{t_N}$. However, the value of α_{t_N} can be chosen in many ways, so in this section, we only consider the process Y_t. This is not a true generator because it uses X, so we will call it an *oracle*.

By substituting the prediction induced by the denoiser, $\hat{X}_s \equiv \hat{X}_s(\hat{Y}_s)$, for X in the oracle process (8), we obtain the *generator*

$$\hat{Y}_{t_0} := \mathring{\rho}_{t_0}^{-1} \varepsilon_{t_0} \text{ and } r_t \hat{Y}_t \;:=\; r_s\, \hat{Y}_s \;+\; (r_t - r_s)\, \hat{X}_s \;+\; \sqrt{\rho_t^2 - \rho_s^2}\, \xi_s. \tag{10}$$

Note that in our setting, unlike in many works on generative diffusion models, time in the oracle and the generator runs forward.

3.2 Continuous diffusion models.

By induction, from (10) for any grid points $t_i < t_j$ we obtain

$$r_{t_j} \hat{Y}_{t_j} \;=\; r_{t_i} \hat{Y}_{t_i} \;+\; (r_{t_{i+1}} - r_{t_i}) \hat{X}_{t_i} + \ldots + (r_{t_j} - r_{t_{j-1}}) \hat{X}_{t_{j-1}} + \sqrt{\rho_{t_j}^2 - \rho_{t_i}^2}\; \xi^*_{t_i}, \tag{11}$$

where $\hat{Y}_{t_i}, \xi^*_{t_i}$ are independent and

$$\xi^*_{t_i} := \frac{\sqrt{\rho_{t_{i+1}}^2 - \rho_{t_i}^2}\; \xi_{t_i} + \ldots + \sqrt{\rho_{t_j}^2 - \rho_{t_{j-1}}^2}\; \xi_{t_{j-1}}}{\sqrt{\rho_{t_j}^2 - \rho_{t_i}^2}} \sim \mathcal{N}(0, I_d).$$

Assuming $t \mapsto \hat{X}_t(\hat{Y}_t)$ is continuous and N approaches infinity in (11), we arrive at a new representation of the diffusion state as the sum of an explicit linear component, an unweighted pathwise integral of the denoiser, and a noise term

$$r_t \hat{Y}_t \;=\; r_s \hat{Y}_s \;+\; \int_{r_s}^{r_t} \hat{X}_r(\hat{Y}_r)\, \mathrm{d}r \;+\; \sqrt{\rho_t^2 - \rho_s^2}\; \xi^*_s, \tag{12}$$

where $\hat{Y}_s, \xi^*_s$ are independent and $\xi^*_s \sim \mathcal{N}(0, I_d)$.

Universal diffusion generator. From (12) we obtain the generator

$$\hat{Y}_t \;=\; \frac{r_s}{r_t}\, \hat{Y}_s \;+\; \frac{1}{r_t} \text{APPROX}\left[\int_{r_s}^{r_t} \hat{X}_r(\hat{Y}_r)\, \mathrm{d}r \right] \;+\; \mathring{\rho}_t^{-1} \sqrt{1 - \frac{\rho_s^2}{\rho_t^2}}\; \xi^*_s, \tag{13}$$

where APPROX denotes any numerical ODE integration method.

Universal schemes as the Euler-Maruyama suffer from integrating the rapidly-changing linear term, whereas schemes designed specifically for diffusion, such as DPM-solvers, treat the linear term analytically, but they require integrating $\hat{\varepsilon}_t$ or $\hat{X}_t$ with an exponential weight, which is difficult (Lu et al., 2023; 2022; Cui et al., 2025). In particular, DPM solvers of order 1-3 are analogous to the Runge-Kutta methods, but—to our knowledge—an analogue of one of the most popular universal schemes, the Runge-Kutta method of order 4, has not yet been developed. For comparison, our method is both universal and specific to diffusion.

From the generator to SDE and back. The equation (10) for the simplest generator is equivalent to

$$\hat{Y}_t - \hat{Y}_s = \Big(\frac{r_s - r_t}{\Delta t\; r_t} \hat{Y}_s + \frac{r_t - r_s}{\Delta t\; r_t} \hat{X}_s \Big) \Delta t + \mathring{\rho}_t^{-1} \sqrt{\frac{\rho_t^2 - \rho_s^2}{\Delta t\; \rho_t^2}} \sqrt{\Delta t}\; \xi^*_s. \tag{14}$$

A direct manipulation with Taylor expansions yields, for $s = t - \Delta t$,

$$\frac{r_t - r_s}{\Delta t\, r_t} = (\log r_t)'(1 + \delta_1) \text{ and } \frac{\rho_t - \rho_s}{\Delta t\, \rho_t} \frac{\rho_t + \rho_s}{\rho_t} = 2(\log \rho_t)'(1 + \delta_2), \tag{15}$$

where $|\delta_1|, |\delta_2| = \mathcal{O}(1/N)$. Hence (14) takes the form

$$\hat{Y}_t - \hat{Y}_s = \ (\log r_t)' \left(\hat{X}_t - \hat{Y}_t\right) \Delta t + \mathring{\rho}_t^{-1} \sqrt{2(\log \rho_t)'}\ \sqrt{\Delta t}\ \xi_s^* + \mathcal{O}_P(\Delta t). \tag{16}$$

Let W_t be a standard d-dimensional Wiener process and assume that $\hat{X}_t$ is sufficiently regular. Then the difference equation (16) converges to the Itô SDE

$$\mathrm{d}\hat{Y}_t = \ (\log r_t)' \left(\hat{X}_t - \hat{Y}_t\right) \mathrm{d}t \ + \ \mathring{\rho}_t^{-1} \sqrt{2(\log \rho_t)'}\ \mathrm{d}W_t, \tag{17}$$

$$= \ (\lambda_t' + \mathring{\lambda}_t') \left(\hat{X}_t - \hat{Y}_t\right) \mathrm{d}t \ + \ \mathring{\rho}_t^{-1} \sqrt{2\lambda_t'}\ \mathrm{d}W_t, \tag{18}$$

$$= \ \mathring{\rho}_t^{-1} (\lambda_t' + \mathring{\lambda}_t')\, \hat{\varepsilon}_t\, \mathrm{d}t \ + \ \mathring{\rho}_t^{-1} \sqrt{2\lambda_t'}\ \mathrm{d}W_t. \tag{19}$$

By substituting $\tilde{Y}_t = (r_t/r_s)\hat{Y}_t$ for $\hat{Y}_t$ in (17) we get

$$r_s \mathrm{d}\tilde{Y}_t = \ r_t' \hat{X}_t\, \mathrm{d}t \ + \ \rho_t \sqrt{2(\log \rho_t)'}\ \mathrm{d}W_t \tag{20}$$

By integrating (20) and returning to $\hat{Y}_t$, we obtain (12), which, with the simplest discretization, takes the form of (10).

Note that the substitution leading to formula (20) uses the method of variation of constants—this approach is used to derive DPM solvers. Thanks to the construction of the diffusion process by (12), SDEs are not needed.

3.3 Generation interval and the main problem

Comparing the formulas for popular noise schedules with the formula for the generator (10), we see that $\mathring{\rho}_0 = 0$ oraz $\mathring{\rho}_0 = \infty$. This means we do not start at the moment when the generator and oracle have the same distribution, nor do we reach the point where the oracle has the distribution p_x. To precisely define the generation task, we need to specify the start t_0 and end t_{max} of the generation. From formula (10), it is also clear that the function λ_t is not needed for generation, only its quotients. Equivalently, it is sufficient to calculate λ_t from the integral formula based on the derivative λ_t', hereafter referred to as the *diffusion rate*, by arbitrarily setting $\lambda_{t_{max}}$. At this point, we can define the main problem of our work.

Main Problem. Determine λ_t' based on the training of the denoiser $(\mathring{\rho}_t, S_t)$ (the input and output scales of the prediction) and the generation interval $0 < t_0 < t_{max} < 1$.

4 A difusion model induced by the penalized maximum likelihood

We need a measure of proximity between the oracle process and the generator process to choose the diffusion schedule. The processes are defined by distributions, and MSE does not determine the proximity between them, so we will use the most popular measure for this purpose, that is the maximum likelihood or, equivalently, the Kullback-Leibler divergence.

4.1 Divergence decompositions

As in the previous sections, we start with the processes $Z_t = \alpha_t Y_t$ and $\hat{Z}_t = \alpha_t \hat{Y}_t$ to see that it is enough to consider only Y_t and $\hat{Y}_t$. Let $t_0 < t_1 < \cdots < t_N = t$ be a discretization of the time interval $[t_0, t]$, set $s = t_{N-1}$. For $x \sim p_x$ we denote latent variables along the path by z_{t_i} and write $z_{t_i:t} = (z_{t_i}, z_{t_{i+1}}, \ldots, z_t)$. From the definition of the oracle process (9) it follows that

$$p_t(z_t|z_s, x) = \mathcal{N}\Big(z_t|\mu_s(z_s, x), \sigma_t \sqrt{1 - \rho_s^2/\rho_t^2}\, I_d\Big),$$

$$\text{where} \quad \mu_s(z_s, x) := \frac{\sigma_t \rho_s}{\sigma_s \rho_t} z_s + \frac{\sigma_s}{\rho_t}(\rho_t \mathring{\rho}_t - \rho_s \mathring{\rho}_s) X,$$

$$= \frac{\alpha_t}{\alpha_s} z_s - \sigma_t \Big(\frac{\mathring{\rho}_t}{\mathring{\rho}_s} - \frac{\rho_s}{\rho_t}\Big)\varepsilon_s.$$

Analogously

$$\hat{p}_t(z_t|z_s) = \mathcal{N}\Big(z_t|\hat{\mu}_s(z_s), \sigma_t\sqrt{1-\rho_s^2/\rho_t^2}I_d\Big),$$
$$\text{where}\quad \hat{\mu}_s(z_s) := \frac{\sigma_t\rho_s}{\sigma_s\rho_t}z_s + \frac{\sigma_s}{\rho_t}(\rho_t\mathring{\rho}_t - \rho_s\mathring{\rho}_s)\hat{X}_s,$$
$$= \frac{\alpha_t}{\alpha_s}z_s - \sigma_t\Big(\frac{\mathring{\rho}_t}{\mathring{\rho}_s} - \frac{\rho_s}{\rho_t}\Big)\hat{\varepsilon}_s.$$

The Kullback-Leibler divergence between these two normal distributions is

$$\mathbb{D}[p_t(.|z_s,x) \mid \hat{p}_t(.|z_s)] := \mathbb{E}_{Z_t}\log\big[p_t(Z_t|z_s,x)/\hat{p}_t(Z_t|z_s)\big] \tag{21}$$
$$= \frac{\|\mu_s(z_s,x) - \hat{\mu}_s(z_s)\|^2}{2\sigma_t^2(1-\rho_s^2/\rho_t^2)} = w_s^N\,\|\hat{\varepsilon}_s(z_s) - \varepsilon_s\|^2, \tag{22}$$
$$\text{where}\quad w_s^N := \frac{(\mathring{\rho}_t/\mathring{\rho}_s - \rho_s/\rho_t)^2}{2(1-\rho_s^2/\rho_t^2)}. \tag{23}$$

We define the following conditional and joint distributions

$$p^N_{t_1:t}(z_{t_1:t}|z_{t_0},x) := p_t(z_t|z_s,x)\ldots p_{t_1}(z_{t_1}|z_{t_0},x),$$
$$\hat{p}^N_{t_1:t}(z_{t_1:t}|z_{t_0}) := \hat{p}_t(z_t|z_s)\ldots\hat{p}_{t_1}(z_{t_1}|z_{t_0}),$$
$$p_{t_0,x}(z_{t_0},x) := p_{t_0}(z_{t_0}|x)\,p_x(x),$$
$$\overline{p}_{t_0,t}(z_{t_0},x|z_t) := p_{t_0}(z_{t_0})\,\overline{p}_t(x|z_t),$$
$$p^N_{t_0:t,x}(z_{t_0:t},x) := p^N_{t_1:t}(z_{t_1:t}|z_{t_0},x)p_{t_0,x}(z_{t_0},x),$$
$$\hat{p}^N_{t_0:t,x}(z_{t_0:t},x) := \hat{p}^N_{t_1:t}(z_{t_1:t}|z_{t_0})\overline{p}_{t_0,t}(z_{t_0},x|z_t).$$

The distribution $\overline{p}_t$ represents the "reconstruction error," which determines how well the image was recovered from the final z_t representation. The overline symbol indicates its parameters, and its effect on the divergence between the joint distributions is often called the *bias*, denoted below as $\bar{\mathcal{B}}$. Two KL decompositions that we shall use are

$$\mathbb{D}\big[p^N_{t_0:t,x}\,\big\|\,\hat{p}^N_{t_0:t,x}\big] = \mathbb{D}\big[p_x\,\big\|\,\hat{p}^N_x\big]\ +\ \mathbb{E}_X\,\mathbb{D}\big[p^N_{t_0:t}(.|X)\,\big\|\,\hat{p}^N_{t_0:t}(.|X)\big], \tag{24}$$

$$\mathbb{D}\big[p^N_{t_0:t,x}\,\big\|\,\hat{p}^N_{t_0:t,x}\big] = \hat{\mathcal{L}}^N(t_0,t,\mathring{\lambda},\lambda)\ +\ \bar{\mathcal{B}}(t_0,t), \tag{25}$$

where

$$\hat{\mathcal{L}}^N(t_0,t,\mathring{\lambda},\lambda) := \mathbb{E}_{Z_{t_0},X}\,\mathbb{D}\big[p^N_{t_1:t}(.|Z_{t_0},X)\,\big\|\,\hat{p}^N_{t_1:t}(.|Z_{t_0})\big] \tag{26}$$
$$\bar{\mathcal{B}}(t_0,t) := \mathbb{D}\big[p_{t_0,x}\,\big\|\,\overline{p}_{t_0,t}\big]. \tag{27}$$

Both of these decompositions together imply that the *diffusion loss*, denoted as $\hat{\mathcal{L}}^N$, is the objective function for (implicitly) a penalized negative log-likelihood of the estimator for the distribution p_x induced by the denoiser with parameters $\wedge$.

Proposition 1.

$$\hat{\mathcal{L}}^N(t_0,t,\mathring{\lambda},\lambda) = \sum_{i=0}^{N} w^N_{t_i}\mathbb{E}_{\varepsilon,x}\,\|\hat{\varepsilon}_{t_i} - \varepsilon\|^2. \tag{28}$$

Proposition 2. *Assuming $t \mapsto \hat{\varepsilon}_t(\hat{Y}_t)$ is continuous, we have*

$$\hat{\mathcal{L}}^N(t_0,t,\mathring{\lambda},\lambda) = \hat{\mathcal{L}}(t_0,t,\mathring{\lambda},\lambda) + \mathcal{O}(1/N), \tag{29}$$

where

$$\hat{\mathcal{L}}(t_0,t,\mathring{\lambda},\lambda) \equiv \hat{\mathcal{L}}(t_0,t,\mathring{\lambda}',\lambda') := \int_{t_0}^{t}\frac{\big(\lambda'_\tau + \mathring{\lambda}'_\tau\big)^2}{4\,\lambda'_\tau}\,\mathbb{E}_{\varepsilon,X}\,\big\|\hat{\varepsilon}_\tau(Y_\tau) - \varepsilon\big\|^2\,d\tau. \tag{30}$$

Propositions 1-2 are proven in Appendix A.

4.2 PENALIZED MAXIMUM LIKELIHOOD

Observe that the weights under the integral in (30) are of the form

$$\frac{\left(\lambda'_t + \mathring{\lambda}'_t\right)^2}{4\,\lambda'_t} = \mathring{\lambda}'_t + \frac{1}{4}\chi^2(\lambda'_t, \mathring{\lambda}'_t), \text{ where } \chi^2(\lambda'_t, \mathring{\lambda}'_t) := \frac{\left(\lambda'_t - \mathring{\lambda}'_t\right)^2}{\lambda'_t} \tag{31}$$

is the well known χ^2-distance. So weights determining the diffusion process in $\hat{\mathcal{L}}$ are, up to a constant, scaled distance between λ'_t and $\mathring{\lambda}'_t$.

Since the function $\hat{\mathcal{L}}$ is (implicitly) a penalized ML objective, we do not change its meaning or difficulty of its calculation, if we add a simple penalty to the weights and define (explicitly) the *penalized maximum likelihood* objective

$$\hat{\mathcal{L}}_c(t_0, t, \mathring{\lambda}, \lambda) \equiv \hat{\mathcal{L}}_c(t_0, t, \mathring{\lambda}', \lambda') := \int_{t_0}^{t} \left[\mathring{\lambda}'_\tau + \frac{\chi^2(\lambda'_t, \mathring{\lambda}'_t) + c_t\lambda'_t}{4}\right] \mathbb{E}_{\varepsilon,X} \left\|\hat{\varepsilon}_\tau(Y_\tau) - \varepsilon\right\|^2 d\tau, \tag{32}$$

for some positive, continuous function c_t. Indeed, it is easy to check that the optimal diffusion rate for such penalized weights is

$$\lambda'_{c,t} := \mathring{\lambda}'_t/\sqrt{1 + c_t}. \tag{33}$$

Penalized maximum likelihood covers many important approaches: if $c_t = 0$, then we obtain the maximum (joint) likelihood solution; if $c_t \to \infty$, then $\lambda'_{c,t} \to 0$, and a diffusion process converges to a deterministic flow. Our MSE-induced diffusion is a tradeoff between these extremes.

5 A DIFFUSION MODEL INDUCED BY MSE TRAINING

We want the losses to agree not only globally on the interval $[t_0, t_{max}]$, but also on each of its subintervals. Let us imagine a scenario where the group optimizing the reconstruction error improves its method and decreases t_{max}, or when it becomes possible to start the generation process for a larger t_0. It could also be that we should generate diffusions in stages using different samplers, and our sampler might only care about optimality for a certain subinterval. Below we will formulate an appropriate condition, but first let us define MSE for each initial interval (t_0, t)

$$\hat{\mathcal{M}}(t_0, t, \mathring{\lambda}, S) := \int_{t_0}^{t} \mathbb{E}_{\varepsilon,X} S_t^2 \left\|\hat{\varepsilon}_t(\bar{Y}_t) - \varepsilon_t\right\|^2 \mathrm{d}t = \int_{t_0}^{t} \mathbb{E}_{\varepsilon,X} \left\|\hat{u}_t(\alpha_t \bar{Y}_t) - u_t\right\|^2 \mathrm{d}t. \tag{34}$$

We will say that the diffusion process defined by $(t_0, t, \mathring{\lambda}', \lambda'_c)$ is *coherent with MSE* if and only if the following condition is satisfied

Coherence Principle. There exist a constant $M \equiv M(t_0, t_{max}, \mathring{\lambda}, S)$ such that $\forall t \in [t_0, t_{max}]$ we have

$$\hat{\mathcal{L}}(t_0, t, \mathring{\lambda}', \lambda'_c) = M\hat{\mathcal{M}}(t_0, t, \mathring{\lambda}, S). \tag{35}$$

The loss $\hat{\mathcal{L}}$ (without subscript c) is invariant to data scaling, because it is the expected divergence, whereas MSE depends on data scaling. Therefore, to compare the two functions, we need an appropriate normalization, that is some constant M.

Proposition 3. *Let us define*

$$M := \max_{t \in [t_0, t_{max}]} \mathring{\lambda}_t / S_t^2 \quad \textit{and} \quad S_{t,M} := \sqrt{M} S_t \quad \forall t \in [t_0, t_{max}]. \tag{36}$$

Then the coherence principle holds with M iff the diffusion rate is

$$\lambda'_{t,c} = \left(S_{t,M} - \sqrt{S_{t,M}^2 - \mathring{\lambda}'_t}\right)^2. \tag{37}$$

The diffusion process with a parameter $\check{\lambda}_t{}' \equiv \lambda'_{t,c}$ is called the *MSE-induced diffusion*.

Proof of Proposition 3. Let us fix t and simplify notation $\beta := \lambda'_t, \beta_c := \lambda'_{c,t}, \mathring{\beta} := \mathring{\lambda}'_t, s := S_{t,M}$. The coherence condition implies that the integrals $\hat{\mathcal{L}}$ and $M\hat{\mathcal{M}}$ agree on the initial intervals, which is equivalent to the equality of the integrands. Therefore

$$\frac{(\beta + \mathring{\beta})^2}{4\,\beta} = M S_t^2 = s^2. \tag{38}$$

The definition of the constant M implies that $s^2 \geq \mathring{\beta}$, thus equation (38) has 2 roots

$$\beta_- = \left(s^2 - \sqrt{s^2 - \mathring{\beta}}\right)^2 \quad \text{and} \quad \beta_+ = \left(s^2 + \sqrt{s^2 - \mathring{\beta}}\right)^2. \tag{39}$$

Observe that $\beta_-\beta_+ = \mathring{\beta}^2$, so $\beta_- \leq \mathring{\beta} \leq \beta_+$. By comparison with (33) $\beta_- = \beta_{c_-}, \beta_+ = \beta_{c_+}$, we obtain formulas for the penalty constants $c_- = (\mathring{\beta}/\beta_-)^2 - 1,\ \ c_+ = (\mathring{\beta}/\beta_+)^2 - 1$ and $c_+ \leq 0 \leq c_-$. Hence only β_- optimizes the penalized maximum likelihood objective. Sufficiency is obvious. □

Example. Consider the logistic noise schedule with the velocity parametrization: $\alpha_t := t, \sigma_t := 1 - t, \mathring{\lambda}'_t = 1/[t(1-t)], S_t = 1/t$. Thus $M = t_{max}/(1 - t_{max})$ and

$$\check{\lambda}_t{}' = \left(\sqrt{M} - \sqrt{M - t/(1-t)}\right)^2 \Big/ t^2.$$

In this case, we obtain a compact form for $\check{\lambda}_t$

$$\check{\lambda}_t \;=\; -\log\Big(\frac{t}{1-t}\Big) - \log\left(\frac{1+g_t}{1-g_t}\right) + \frac{2\,t_{max}}{(1-t_{max})}\,\frac{1-t}{t}\,\Big(g_t - 1\Big) + const,$$

where

$$g_t := \sqrt{\frac{t_{max} - t}{t_{max}(1-t)}}.$$

Discrete time. It seems that in the discrete model there is no natural parameter that would also be associated with λ_t. In our research, the parameter $\eta_t := \sqrt{1 - \rho_s^2/\rho_t^2}$ proved to be convenient. Rewriting equation (4) we see that η_t^2 is the proportion of new noise ξ_s to the total noise ε_t

$$\varepsilon_t \;=\; \sqrt{1-\eta_t^2}\;\varepsilon_s \;+\; \eta_t\;\xi_s.$$

From (23)

$$w_t^N \equiv w_t^N(\eta_t) = \frac{1}{\eta_t^2}\Big(\sqrt{\gamma_t} \;-\; \sqrt{1-\eta_t^2}\Big)^2, \;\text{ where }\; \gamma_t := \mathring{\rho}_t^2/\mathring{\rho}_s^2.$$

It can be easily checked that

$$\mathring{\eta}_t := \arg\min_{\eta_t} w_t^2(\eta_t) = \sqrt{1 - \gamma_t^{-1}},$$

and

$$\check{\eta}_t = \frac{\gamma_t - 1}{\sqrt{\gamma_t S_{t,M}^2} + \sqrt{S_{t,M}^2 + 1 - \gamma_t}}.$$

MSE-induced diffusion and recent empirical diffusion schedules. In Figure (1) we illustrate comparison of recent diffusion schedules in three popular scenarios (noise, parametrization). We normalize λ_t because then $((t_{max} - t_0)/N)\lambda'_t \approx \eta_t^2$, while η_t^2 has an easy interpretation. The results in different scenarios are very similar, so it might be worth considering other bell-shaped densities like Student's t or Tukey-lambda. Interestingly, MSE-induced diffusion differs significantly from a deterministic flow only at the very end of the generation, when the noise is the lowest.

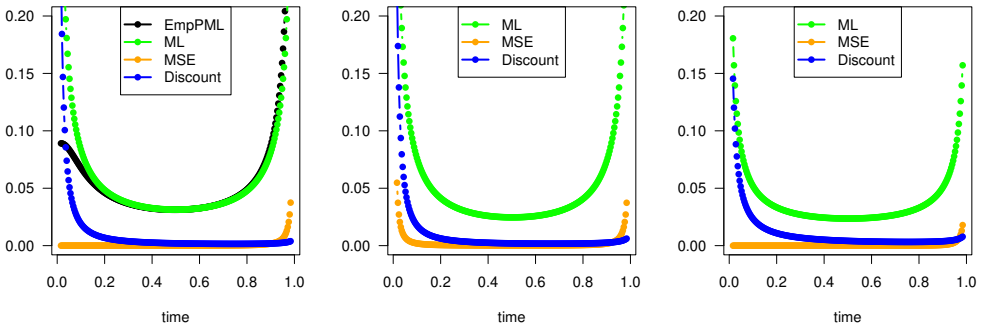

Figure 1: This study compares several diffusion schedules $((t_{max} - t_0)/N)\lambda'_t$ in different (noise, parametrization) scenarios: MSE-induced diffusion (MSE); maximum likelihood (ML); maximum likelihood multiplied by a discount function, ER-SDE-4 (Cui et al., 2025) (Discount); and empirically fitted penalized ML (empPML) (Ma et al., 2024). Noise schedule and parametrization from left: (a) linear and velocity, (b) edm and F-prediction, (c) cosine and noise. In all cases N=250, $t_0 = 0.016, t_{max} = 1 - t_0$

6 Discussion

Two-state domain. From a formal standpoint, a diffusion process can have a degenerate scale, $\mathring{\lambda}_t \equiv C$, but it must have a diffusion schedule λ_t. It is therefore interesting that many seminal works on generative diffusion models do not include λ_t or an equivalent parameter defining the process's cumulative relative variance (Ho et al., 2020; Song et al., 2021b; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Salimans & Ho, 2021; Song et al., 2021a; Kingma et al., 2021; Ho & Salimans, 2022; Rombach et al., 2022; Kingma & Gao, 2023; Esser et al., 2024). One might think that what these papers do provide—time-reversible processes with $\lambda_t = \mathring{\lambda}_t$—would be entirely sufficient. However, if so, why are deterministic flows so popular, resulting in a two-state domain? It seems that the main reason for the lack of λ_t is the difficulty in setting it, as seen in papers like Karras et al. (2022). In this work, we show that there is a natural choice between 0 and $\mathring{\lambda}_t$ that is consistent with empirically motivated training using a weighted MSE and with the standard criterion for fitting distributions to data, namely maximum likelihood.

Scaling. Implementing a diffusion model using α_t and σ_t has become common practice, despite the mostly simulation-based arguments of Karras et al. (2022), that α_t is unnecessary. In our work, we specify these arguments: α_t is merely an input scaling in the denoiser, which is not needed for generation or in the context of maximum likelihood analysis. Our research indicates that the natural scale for the process values is $\lambda_t + \mathring{\lambda}_t$, while the natural scale for the process arguments is $\rho_t + \mathring{\rho}_t$.

Time interval. From a theoretical standpoint, we see no difference between score-based models that generate processes on $(0, \infty)$ and stochastic interpolants that work on $[0, 1]$. It is important that $\mathring{\rho}_t$ and ρ_t take on positive values within the closed interval of actual generation. This is necessary to make the analysis realistic, which is clearly visible in the proofs of global convergence for numerical ODE solvers. As long as the limits $0 < t_0 < t_{max}$ for the main noise schedules do not depend on the number of steps, there are no problems. However, when we begin to consider more realistic scenarios, such as $t_0 \equiv t_0(N) \to 0$ and $t_{max} \equiv t_{max}(N) \to 1$, we see that the convergence is violated by the conditions $\sigma_t \to 0$ or $\sigma_t \to 1$.

Open problems. We transform the generative process to one with additive noise. This allows us to replace the Gaussian noise with noise originating from a α-stable distribution. Upper bounds on KL divergence exist for these processes, so there may also be formulas analogous to our MSE-induced diffusion. An interesting problem seems to be training a noise schedule and a parameterization in alternation with a diffusion schedule.

REFERENCES

Marco S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Marco S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Qinpeng Cui, Xinyi Zhang, Qiqi Bao, and Qingmin Liao. Elucidating the solution space of extended reverse-time sde for diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 243–252. IEEE, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in neural information processing systems*, volume 34, pp. 8780–8794, 2021.

Patrick Esser, Suneet Kulal, Andreas Blattmann, Rameen Entezari, Jürgen Müller, Himanshu Saini, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, pp. 23–40, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pp. 6840–6851, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Diederik Kingma and Rui Gao. Understanding diffusion objectives as the elbo with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in neural information processing systems*, volume 34, pp. 21696–21707, 2021.

Yaron Lipman, Ricky TQ Chen, Hadar Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chen Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2023.

Nic Ma, Matt Goldstein, Marco S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Sifan Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40, 2024.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

Robin Rombach, Andreas Blattmann, Dan Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021.

Yang Song, Chris Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in neural information processing systems*, volume 34, pp. 1415–1428, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

A Appendix

Proof of Proposition 1. By the chain rule for KL along the grid $t_0 < t_1 < \cdots < t_N = t$ we obtain

$$\begin{aligned}
&\mathbb{D}\big[p^N_{t_1:t}(.|z_{t_0},x)\,\big\|\,\hat{p}^N_{t_1:t}(.|z_{t_0})\big] \\
&\quad= \mathbb{D}\big[p_{t_1}(.|z_{t_0},x)\,\big\|\,\hat{p}_{t_1}(.|z_{t_0})\big] + \mathbb{E}_{Z_{t_1},X}\,\mathbb{D}\big[p_{t_2}(.|Z_{t_1},X)\,\big\|\,\hat{p}_{t_2}(.|Z_{t_1})\big] + \cdots \\
&\qquad+\ \mathbb{E}_{Z_{t_{N-1}},X}\,\mathbb{D}\big[p_t(.|Z_{t_{N-1}},X)\,\big\|\,\hat{p}_t(.|Z_{t_{N-1}})\big]. \qquad (40)
\end{aligned}$$

In our setting, from (23) this can be rewritten in terms of denoising errors with weights $w^N_{t_j}$:

$$(40)\ = w^N_{t_1}\,\|\hat{\varepsilon}_{t_0}-\varepsilon_{t_0}\|^2 + w^N_{t_2}\,\mathbb{E}_{\varepsilon_{t_1}}\big\|\hat{\varepsilon}_{t_1}-\varepsilon_{t_1}\big\|^2 + \cdots + w^N_{t_N}\,\mathbb{E}_{\varepsilon_{t_{N-1}}}\big\|\hat{\varepsilon}_{t_{N-1}}-\varepsilon_{t_{N-1}}\big\|^2. \qquad (41)$$

So

$$\mathbb{E}_{Z_{t_0},X}\mathbb{D}\big[p^N_{t_1:t}(.|Z_{t_0},X)\,\big\|\,\hat{p}^N_{t_1:t}(.|Z_{t_0})\big] = \sum_{i=0}^{N}\mathbb{E}_{\varepsilon_{t_i},X}\,w^N_{t_i}\,\|\hat{\varepsilon}_{t_i}-\varepsilon_{t_i}\|^2 = \sum_{i=0}^{N}\mathbb{E}_{\varepsilon,X}\,w^N_{t_i}\,\|\hat{\varepsilon}_{t_i}-\varepsilon\|^2.$$

□

Proof of Proposition 2. A direct manipulation with Taylor expansions yields, for $s = t - \Delta\tau$,

$$\begin{aligned}
\bar{w}^N_{t-\Delta\tau} := \frac{w^N_{t-\Delta\tau}}{\Delta\tau} &= \left(\frac{\mathring{\rho}_t-\mathring{\rho}_s}{\Delta\tau\,\mathring{\rho}_s}+\frac{\rho_t-\rho_s}{\Delta\tau\,\rho_t}\right)^2 \Bigg/ \left(2\,\frac{\rho_t-\rho_s}{\Delta\tau\,\rho_t}\,\frac{\rho_t+\rho_s}{\rho_t}\right) \\
&= \left(\frac{\mathring{\lambda}'_t}{\lambda_t}(1+\delta_1)+\frac{\lambda'_t}{\lambda_t}(1+\delta_2)\right)^2 \Bigg/ \left(4\,\frac{\lambda'_t}{\lambda_t}(1+\delta_3)\right) \\
&= \bar{w}_t + \mathcal{O}(1/N),\ \text{where}\ \bar{w}_t := \frac{\left(\lambda'_t+\mathring{\lambda}'_t\right)^2}{4\,\lambda'_t}\ \text{and}\ |\delta_1|,|\delta_2|,|\delta_3| = \mathcal{O}(1/N). \qquad (42)
\end{aligned}$$

For $\tau \in [t_0, t]$ define $t^N(\tau) := \min\{t_i : \tau \geq t_i\}$. We have

$$\max_{t_0\leq\tau\leq t}\left(\bar{w}^N_{t^N(\tau)}\left\|\hat{\varepsilon}_{t^N(\tau)}-\varepsilon\right\|^2 - \bar{w}_\tau\,\|\hat{\varepsilon_\tau}-\varepsilon\|^2\right) = \mathcal{O}(1/N), \qquad (43)$$

and consequently

$$\begin{aligned}
\hat{\mathcal{L}}^N(t_0,t,\mathring{\lambda},\lambda) &= \mathbb{E}_{\varepsilon,X}\left(\sum_{i=0}^{N}\frac{w^N_{t_i}}{\Delta\tau}\left\|\hat{\varepsilon}_{t_i}-\varepsilon\right\|^2\Delta\tau\right) \\
&= \mathbb{E}_{\varepsilon,X}\left(\int_{t_0}^{t}\bar{w}^N_{t^N(\tau)}\left\|\hat{\varepsilon}_{t^N(\tau)}-\varepsilon\right\|^2 d\tau\right) = \int_{t_0}^{t}\bar{w}_\tau\,\mathbb{E}_{\varepsilon,X}\big\|\hat{\varepsilon}_\tau-\varepsilon\big\|^2 d\tau + \mathcal{O}(1/N). \quad \square
\end{aligned}$$