Score-based Denoising Diffusion with Non-Isotropic Gaussian Noise Models

Vikram Voleti
Mila, University of Montreal
Canada

Adam Oberman
Mila, McGill University
Canada

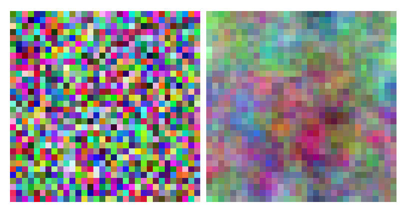
Christopher Pal
Mila, Polytechnique Montréal
CIFAR AI Chair, Service Now
Canada

Abstract

Generative models based on denoising diffusion techniques have led to an unprecedented increase in the quality and diversity of imagery that is now possible to create with neural generative models. However, most contemporary state-of-the-art methods are derived from a standard isotropic Gaussian formulation. In this work we examine the situation where non-isotropic Gaussian distributions are used. We present the key mathematical derivations for creating denoising diffusion models using an underlying non-isotropic Gaussian noise model. We also provide initial experiments with the CIFAR10 dataset to help verify empirically that this more general modelling approach can also yield high-quality samples.

1 Introduction

Score-based denoising diffusion models [16, 6, 18] have seen great success as generative models for images [4, 17], as well as other modes such as video [7, 22, 19], audio [9, 3], etc. The underlying framework relies on a noising "forward" process that adds noise to real images (or other data), and a denoising "reverse" process that iteratively removes noise. In most cases, the noise distribution used is the isotropic Gaussian i.e. noise samples are independently and identically distributed (IID) as the standard normal at each pixel.



(a) Isotropic (b) Non-isotropic

Figure 1: Gaussian noise samples.

In this work, we lay the theoretical foundations and derive the key mathematics for a non-isotropic Gaussian formulation for denoising diffusion models. It is our hope that these insights may open the door to new classes of models. One type of non-isotropic Gaussian noise arises in a family of models known as Gaussian Free Fields (GFFs) [14, 1, 2, 20] (a.k.a. Gaussian Random Fields). GFF noise can be obtained by either convolving isotropic Gaussian noise with a filter, or applying frequency masking of noise. In either case this procedure allows one to model or generate smoother and correlated types of Gaussian noise. In Figures 1 and 2, we compare examples of isotropic Gaussian noise with GFF noise obtained using a frequency space window function consisting of $w(f) = \frac{1}{f}$.

Our contributions here consist of the following: (1) deriving the key mathematics for score-based denoising diffusion models using non-isotropic multivariate Gaussian distributions, (2) examining the special case of a GFF and the corresponding non-Isotropic Gaussian noise model, and (3) showing that diffusion models trained (eg. on the CIFAR-10 dataset [10]) using a GFF noise process are also capable of yielding high-quality samples comparable to models based on isotropic Gaussian noise.

Appendix A and Appendix B contain more detailed derivations of the above equations for DDPM [6] and our NI-DDPM. See Appendix D and Appendix E for the equivalent derivations for Score Matching Langevin Dynamics (SMLD) [16, 17], and our Non-Isotropic SMLD (NI-SMLD).

NeurIPS 2022 Workshop on Score-Based Methods.

2 Isotropic Gaussian denoising diffusion models

We perform our analysis below within the Denoising Diffusion Probabilistic Models (DDPM) [6] framework, but our analysis is valid for all other types of score-based denoising diffusion models.

In DDPM, for a fixed sequence of positive scales $0 < \beta_1 < \cdots < \beta_L < 1$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, and a noise sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the cumulative "**forward**" noising process is:

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \implies \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

The "**reverse**" process involves iteratively **sampling** $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ conditioned on $\mathbf{x}_0$ i.e. $p_{t-1}(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0)$, obtained from $q_t(\mathbf{x}_t \mid \mathbf{x}_0)$ using Bayes' rule. For this, first $\boldsymbol{\epsilon}$ is estimated using a neural network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$. Then, using $\hat{\mathbf{x}}_0 = \big(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\big)/\sqrt{\bar{\alpha}_t}$ from eq. (1), $\mathbf{x}_{t-1}$ is sampled:

$$p_{t-1}(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \hat{\mathbf{x}}_0) = \mathcal{N}(\ \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0), \tilde{\beta}_t\mathbf{I}\ ) \implies \mathbf{x}_{t-1} = \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0) + \sqrt{\tilde{\beta}_t}\mathbf{z}_t \quad ;\text{ where} \tag{2}$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t \ ;\ \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \ ;\ \mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

The **objective function** to train $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ is simply an expected reconstruction loss with the true $\boldsymbol{\epsilon}$:

$$\mathcal{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}(1,\cdots,L),\mathbf{x}_0\sim p(\mathbf{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\big(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\big)\right\|_2^2\right] \tag{4}$$

From the perspective of score matching, the **score** of the DDPM forward process is:

$$\text{Score}\ \ \mathbf{s} = \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon} \tag{5}$$

Thus, the overall **score-matching objective** for a score estimation network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ is the weighted sum of the loss $\ell_{\mathbf{s}}(\boldsymbol{\theta}; t)$ for each t, the weight being the inverse of the score **variance** at t i.e. $(1-\bar{\alpha}_t)$:

$$\mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}) = \mathbb{E}_t\ (1-\bar{\alpha}_t)\ \ell_{\mathbf{s}}(\boldsymbol{\theta}; t) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\ \left\|\sqrt{1-\bar{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \boldsymbol{\epsilon}\right\|_2^2\ \right] \tag{6}$$

When the score network output is redefined as per the **score-noise relationship** in eq. (5):

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \implies \mathcal{L}_{\mathbf{s}}(\boldsymbol{\theta}) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\ \|-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t) + \boldsymbol{\epsilon}\|_2^2\ \right] = \mathcal{L}_{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) \tag{7}$$

Thus, $\mathcal{L}_{\mathbf{s}} = \mathcal{L}_{\boldsymbol{\epsilon}}$ i.e. the score-matching and noise reconstruction objectives are equivalent.

From [13], the **Expected Denoised Sample** (EDS) $\mathbf{x}_0^*(\mathbf{x}_t, t) \triangleq \mathbb{E}_{\mathbf{x}_0\sim p_t(\mathbf{x}_0|\mathbf{x}_t)}[\mathbf{x}_0]$ and the score $\mathbf{s}$, estimated optimally as $\mathbf{s}_{\boldsymbol{\theta}^*}$, are related as:

$$\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) = \mathbb{E}\left[\|\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2\right]\big(\mathbf{x}_0^*(\mathbf{x}_t, t) - \mathbf{x}_t\big) = \frac{1}{1-\bar{\alpha}_t}\big(\mathbf{x}_0^*(\mathbf{x}_t, t) - \mathbf{x}_t\big) \tag{8}$$

$$\implies \mathbf{x}_0^*(\mathbf{x}_t, t) = \mathbf{x}_t + (1-\bar{\alpha}_t)\ \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) = \mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\ \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, t) \tag{9}$$

The EDS is often used to further improve the quality of the final image at $t = 0$.

3 Non-isotropic Gaussian denoising diffusion models

We formulate the Non-Isotropic DDPM (**NI-DDPM**) using a non-isotropic Gaussian noise distribution with a positive semi-definite covariance matrix $\boldsymbol{\Sigma}$ in the place of $\mathbf{I}$. The **forward** noising process is:

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{\Sigma}) \implies \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon} \tag{10}$$

Thus, the **score** of NI-DDPM is (see Appendix B.1 for derivation):

$$\text{Score}\ \ \mathbf{s} = \nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t \mid \mathbf{x}_0) = -\boldsymbol{\Sigma}^{-1}\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{1-\bar{\alpha}_t} = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon} \tag{11}$$

The **score-matching objective** for a score estimation network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$ at each noise level t is now:

$$\ell(\boldsymbol{\theta}; t) = \mathbb{E}_{\mathbf{x}_0\sim p(\mathbf{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\ \left\|\mathbf{s}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) + \frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon}\right\|_2^2\ \right] \tag{12}$$

The **variance** of this score is:

$$\mathbb{E}\left[\|\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2\right] = \mathbb{E}\left[\left\|-\frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon}\right\|_2^2\right] = \frac{1}{1-\bar{\alpha}_t}\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[\|\boldsymbol{\epsilon}\|_2^2\right] \tag{13}$$

The **overall objective** is a weighted sum, the weight being the inverse of the score variance $(1-\bar{\alpha}_t)\boldsymbol{\Sigma}$:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}(1,\cdots,L)}\,(1-\bar{\alpha}_t)\boldsymbol{\Sigma}\,\ell(\boldsymbol{\theta};t) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)+\boldsymbol{\epsilon}\right\|_2^2\right] \tag{14}$$

Following the **score-noise relationship** in eq. (11):

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) \tag{15}$$

The **objective function** now becomes (expanding $\mathbf{s}_{\boldsymbol{\theta}}$ as per eq. (15)):

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{t\sim\mathcal{U}(1,\cdots,L),\mathbf{x}_0\sim p(\mathbf{x}_0),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\left\|-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0+\sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon},t)+\boldsymbol{\epsilon}\right\|_2^2\right] \tag{16}$$

This objective function for NI-DDPM seems like $\mathcal{L}_\epsilon$ of DDPM, but DDPM's $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ network cannot be re-used here since their forward processes are different. DDPM produces $\mathbf{x}_t$ from $\mathbf{x}_0$ using eq. (1), while NI-DDPM uses eq. (10). See Appendix B.4 for alternate formulations of the score network.

Sampling involves computing $p_{t-1}(\mathbf{x}_{t-1} \mid \mathbf{x}_t,\hat{\mathbf{x}}_0)$ (see Appendix B.6 for derivation):

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0,(1-\bar{\alpha}_t)\boldsymbol{\Sigma}) \implies \hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t,t)\big) \tag{17}$$

$$p_{t-1}(\mathbf{x}_{t-1} \mid \mathbf{x}_t,\hat{\mathbf{x}}_0) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\hat{\mathbf{x}}_0),\tilde{\beta}_t\boldsymbol{\Sigma}) \implies \mathbf{x}_{t-1} = \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\hat{\mathbf{x}}_0)+\sqrt{\tilde{\beta}_t}\sqrt{\boldsymbol{\Sigma}}\mathbf{z}_t \tag{18}$$

where $\tilde{\boldsymbol{\mu}}_t$, $\tilde{\beta}_t$ and $\mathbf{z}_t$ are the same as eq. (3).

Alternatively, [18] mentions using β_t instead of $\tilde{\beta}_t$:

$$p_{t-1}^{\beta_t}(\mathbf{x}_{t-1} \mid \mathbf{x}_t,\hat{\mathbf{x}}_0) = \mathcal{N}(\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\hat{\mathbf{x}}_0),\beta_t\boldsymbol{\Sigma}) \implies \mathbf{x}_{t-1} = \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\hat{\mathbf{x}}_0)+\sqrt{\beta_t}\sqrt{\boldsymbol{\Sigma}}\mathbf{z}_t \tag{19}$$

Alternatively, sampling using **DDIM** [15] invokes the following distribution for $\mathbf{x}_{t-1}$:

$$p_{t-1}^{\text{DDIM}}(\mathbf{x}_{t-1} \mid \mathbf{x}_t,\hat{\mathbf{x}}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0+\sqrt{1-\bar{\alpha}_{t-1}}\frac{\mathbf{x}_t-\sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_t}},\ \mathbf{0}\right) \tag{20}$$

$$\implies \mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0+\sqrt{1-\bar{\alpha}_{t-1}}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t,t) \tag{21}$$

The **Expected Denoised Sample** $\mathbf{x}_0^*(\mathbf{x}_t,t)$ and the optimal score $\mathbf{s}_{\boldsymbol{\theta}^*}$ are now related as:

$$\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t,t) = \mathbb{E}\left[\|\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2\right]\big(\mathbf{x}_0^*(\mathbf{x}_t,t)-\mathbf{x}_t\big) = \frac{1}{1-\bar{\alpha}_t}\boldsymbol{\Sigma}^{-1}\big(\mathbf{x}_0^*(\mathbf{x}_t,t)-\mathbf{x}_t\big) \tag{22}$$

$$\implies \mathbf{x}_0^*(\mathbf{x}_t,t) = \mathbf{x}_t+(1-\bar{\alpha}_t)\,\boldsymbol{\Sigma}\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t,t) = \mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\,\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t,t) \tag{23}$$

SDE formulation: Score-based diffusion models have also been analyzed as stochastic differential equations (SDEs) [18]. The SDE version of NI-DDPM, which we call Non-Isotropic Variance Preserving (NIVP) SDE, is (see Appendix B.9 for derivation):

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t+\sqrt{\beta(t)}\sqrt{\boldsymbol{\Sigma}}\,\mathrm{d}\mathbf{w} \tag{24}$$

$$\implies p_{0t}\big(\mathbf{x}(t) \mid \mathbf{x}(0)\big) = \mathcal{N}\left(\mathbf{x}(0)\,e^{-\frac{1}{2}\int_0^t\beta(s)\mathrm{d}s},\ \boldsymbol{\Sigma}(\mathbf{I}-\mathbf{I}e^{-\int_0^t\beta(s)\mathrm{d}s})\right) \tag{25}$$

Finally, Appendix A and Appendix B contain more detailed derivations of the above equations for DDPM [6] and our NI-DDPM. See Appendix D and Appendix E for the equivalent derivations for Score Matching Langevin Dynamics (SMLD) [16, 17], and our Non-Isotropic SMLD (NI-SMLD).

4 Gaussian Free Field (GFF) images

A GFF image $\mathbf{g}$ can be obtained from a normal noise image $\mathbf{z}$ as follows [14] (see Appendix C for more details):

1. First, sample an $n \times n$ noise image $\mathbf{z}$ from the standard complex normal distribution with covariance matrix $\Gamma = \mathbf{I}_N$ where $N = n^2$ is the total number of pixels, and pseudo-covariance matrix $C = \mathbf{0}$: $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N, \mathbf{0})$. (In principle, real noise could be used.)
2. Apply the Discrete Fourier Transform using its $N \times N$ weights matrix $\mathbf{W}_N$: $\mathbf{W}_N \mathbf{z}$.
3. Consider a diagonal $N \times N$ matrix of the reciprocal of an index value $|k_{ij}|$ per pixel (i, j) in Fourier space : $\mathbf{K}^{-1} = [1/|k_{ij}|]_{(i,j)}$, and multiply this with the above: $\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}$.
4. Take its Inverse Discrete Fourier Transform ($\mathbf{W}_N^{-1}$) to make the raw GFF image: $\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}$. However, this results in a GFF image with a small non-unit variance.
5. Normalize the above GFF image with the standard deviation σ_N at its resolution N, so that it has unit variance (see Appendix C.1 for derivation of σ_N): $\mathbf{g}_{\text{complex}} = \frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}$
6. Extract only the real part of $\mathbf{g}_{\text{complex}}$, and normalize (see Appendix C.2 for derivation):

$$\mathbf{g} = \frac{1}{\sqrt{2N}\sigma_N}\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}\right) \tag{26}$$

See Figures 1 and 2 for examples of GFF images. Effectively, this procedure prioritizes lower frequencies over higher frequencies, thereby making the noise smoother, and hence correlated. The probability distribution of GFF images $\mathbf{g}$ can be seen as a non-isotropic multivariate Gaussian with mean $\mathbf{0}$, and a non-diagonal covariance matrix $\mathbf{\Sigma}$ (see Appendices C.1 and C.2 for derivation):

$$p(\mathbf{g}) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}); \mathbf{\Sigma} = \sqrt{\mathbf{\Sigma}}\sqrt{\mathbf{\Sigma}}^T; \sqrt{\mathbf{\Sigma}} = \frac{1}{\sqrt{2N}\sigma_N}\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\right) \implies \mathbf{g} = \sqrt{\mathbf{\Sigma}}\mathbf{z} \tag{27}$$

5 Results

We train two models on CIFAR10, one using DDPM and the other using NI-DDPM with the exact same hyperparameters (batch size, learning rate, etc.) for 300,000 iterations. We then sample 50,000 images from each, and calculate the image generation metrics of Fréchet Inception Distance (FID) [5], Precision (P), and Recall (R). Although the models were trained on 1000 steps between data and noise, we report these metrics while sampling images using 1000, and smaller steps: 100, 50, 20, 10.

Table 1: Image generation metrics FID, Precision (P), and Recall (R) for CIFAR10 using DDPM and NI-DDPM, with different generation steps.

CIFAR10	**steps**	**FID** $\downarrow$	**P** $\uparrow$	**R** $\uparrow$
DDPM	1000	6.05	0.66	0.54
	100	12.25	0.62	0.48
	50	16.61	0.60	0.43
	20	26.35	0.56	0.24
	10	44.95	0.49	0.24
NI-DDPM	1000	6.95	0.62	0.53
	100	12.68	0.60	0.49
	50	16.91	0.57	0.45
	20	30.41	0.52	0.35
	10	60.32	0.43	0.23

As can be seen from Table 1, our non-isotropic variant performs comparable to the isotropic baseline. The difference between them increases with decreasing number of steps between noise and data. This provides a reasonable proof-of-concept that non-isotropic Gaussian noise works just as well as isotropic noise when used in denoising diffusion models for image generation.

6 Conclusion

We have presented the key mathematics behind non-isotropic Gaussian DDPMs, as well as a complete example using a GFF. We then noted quantitative comparison of using GFF noise vs. regular noise on the CIFAR-10 dataset. In the appendix, we also include further derivations for non-isotropic SMLD models. GFFs are just one example of a well known class of models that are a subset of non-isotropic Gaussian distributions. In the same way that other work has examined non-Gaussian distributions such as the Gamma distribution [11], Poisson distribution [21], and Heat dissipation processes [12], we hope that our work here may lay the foundation for other new denoising diffusion formulations.

References

[1] Nathanaël Berestycki. Introduction to the gaussian free field and liouville quantum gravity. *Lecture notes*, 2015.

[2] Maury Bramson, Jian Ding, and Ofer Zeitouni. Convergence in law of the maximum of the two-dimensional discrete gaussian free field. *Communications on Pure and Applied Mathematics*, 69(1):62–123, 2016.

[3] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J. Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. In *International Conference on Learning Representations*, 2021.

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[7] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[8] Alexia Jolicoeur-Martineau, Rémi Piché-Taillefer, Rémi Tachet des Combes, and Ioannis Mitliagkas. Adversarial score matching and improved sampling for image generation. *International Conference on Learning Representations*, 2021.

[9] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto*, 2009.

[11] Eliya Nachmani, Robin San Roman, and Lior Wolf. Denoising diffusion gamma models, 2021.

[12] Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation, 2022.

[13] Saeed Saremi and Aapo Hyvarinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20:1–23, 2019.

[14] Scott Sheffield. Gaussian free fields for mathematicians. *Probability theory and related fields*, 139(3):521–541, 2007.

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *International Conference on Learning Representations*, 2021.

[16] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 2019.

[17] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems*, 2020.

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

[19] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation. In *(NeurIPS) Advances in Neural Information Processing Systems*, 2022.

[20] Wendelin Werner and Ellen Powell. Lecture notes on the gaussian free field. *arXiv preprint arXiv:2004.04720*, 2020.

[21] Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models, 2022.

[22] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022.

A Denoising Diffusion Probabilistic Models (DDPM) [6]

A.1 Forward Process

In DDPM, for a fixed sequence of positive scales $0 < \beta_1 < \cdots < \beta_L < 1$, $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$, and a noise sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the cumulative "forward" noising process is:

$$q_t(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \implies \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{28}$$

$$\implies \text{Score } \mathbf{s} = \nabla_{\mathbf{x}_t} \log q_{\bar{\alpha}_t}(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon} \tag{29}$$

The noise $\boldsymbol{\epsilon}$ is estimated using a neural network $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)$. Thus,

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)) \tag{30}$$

$[\because \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$ from eq. (28), and loss is minimized when $\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) = \boldsymbol{\epsilon}]$

A.2 Objective function for DDPM

The objective function for DDPM at noise level σ is:

$$\ell^{\text{DDPM}}(\boldsymbol{\theta}; \bar{\alpha}_t) \triangleq \frac{1}{2}\mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x})p(\mathbf{x})}\left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \frac{1}{(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0) \right\|_2^2\right] \tag{31}$$

The overall loss is the weighted sum of the losses at each step:

$$\mathcal{L}(\boldsymbol{\theta}; \{\bar{\alpha}_t\}_{t=1}^{L}) \triangleq \frac{1}{L}\sum_{t=1}^{L} \lambda(\alpha_t)\, \ell(\boldsymbol{\theta}; \bar{\alpha}_t) \tag{32}$$

The weight λ is the inverse of the variance of the score.

A.3 Variance of score for DDPM

$$\mathbb{E}\left[\|\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2\right] = \mathbb{E}\left[\left\| -\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{(1-\bar{\alpha}_t)} \right\|_2^2\right] = \mathbb{E}\left[\left\| \frac{\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}}{(1-\bar{\alpha}_t)} \right\|_2^2\right] = \frac{1}{1-\bar{\alpha}_t} \tag{33}$$

A.4 Overall objective function for DDPM

The overall objective function in [6] used $\lambda(\bar{\alpha}_t) \propto 1/\mathbb{E}\left[\|\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{x}_0)\|_2^2\right] = 1 - \bar{\alpha}_t$:

$$\begin{aligned}\mathcal{L}^{\text{DDPM}}(\boldsymbol{\theta}; \{\bar{\alpha}_t\}_{t=1}^{L}) &\triangleq \frac{1}{2L}\sum_{t=1}^{L} \mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\| \sqrt{1-\bar{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)}{\sqrt{1-\bar{\alpha}_t}} \right\|_2^2\right] \\ &= \frac{1}{2L}\sum_{t=1}^{L} \mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\| \sqrt{1-\bar{\alpha}_t}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \boldsymbol{\epsilon} \right\|_2^2\right]\end{aligned} \tag{34}$$

A.5 Smarter DDPM score estimation

A smarter score model recognizes that the score is a factor of $\boldsymbol{\epsilon}$ from eq. (29), hence only $\boldsymbol{\epsilon}$ needs to be estimated:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) \tag{35}$$

In this case, the overall objective function changes to:

$$\mathcal{L}^{\text{DDPM}}(\boldsymbol{\theta}; \{\bar{\alpha}_t\}_{t=1}^{L}) \triangleq \frac{1}{2L}\sum_{t=1}^{L} \mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\| -\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \boldsymbol{\epsilon} \right\|_2^2\right]$$

$$= \frac{1}{2L} \sum_{t=1}^{L} \mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \bar{\alpha}_t) \right\|_2^2 \right] \tag{36}$$

This is eq 14 in the DDPM paper. The DDPM paper retains conditioning of $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ on $\bar{\alpha}_t$, but SMLD omits it.

A.6 Sampling in DDPM

The reverse probability is given by:

$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1} \mid \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_i \mathbf{I}) \text{ where}$$
$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t; \quad \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{37}$$

Hence, considering $\hat{\mathbf{x}}_0$ estimated from $\mathbf{x}_t$ using eq. (30):

$$\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \sqrt{\tilde{\beta}_t}\mathbf{z}_t \tag{38}$$

[6] breaks down the reversal into 2 steps:

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t))$$
$$\mathbf{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0 + \frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \sqrt{\tilde{\beta}_t}\mathbf{z}_t \tag{39}$$

A.7 Sampling using DDIM

DDIM [15] uses this sampling:

$$q_{\bar{\alpha}_L}(\mathbf{x}_L \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_L \mid \sqrt{\bar{\alpha}_L}\mathbf{x}_0, (1-\bar{\alpha}_L)\mathbf{I}) \tag{40}$$
$$q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1} \mid \sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{1-\bar{\alpha}_t}}, \mathbf{0}\right) \tag{41}$$
$$\implies q_{\bar{\alpha}_t}(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}\right)$$

Hence:

$$\hat{\mathbf{x}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t))$$
$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) \tag{42}$$

A.8 Expected Denoised Sample

From [13], assuming isotropic Gaussian noise of variance $1-\bar{\alpha}_t$, we know that the expected denoised sample $\mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) \triangleq \mathbb{E}_{\mathbf{x}\sim q_{\bar{\alpha}_t}(\mathbf{x}|\mathbf{x}_t)}[\mathbf{x}]$ and the optimal score $\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t)$ are related as:

$$\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t) = \mathbb{E}\left[\|\nabla_{\mathbf{x}_t} \log q_{\bar{\alpha}_t}(\mathbf{x}_t \mid \mathbf{x})\|_2^2\right](\mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) - \mathbf{x}_t)$$
$$\implies \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t) = \frac{1}{1-\bar{\alpha}_t}(\mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) - \mathbf{x}_t)$$
$$\implies \mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) = \mathbf{x}_t + (1-\bar{\alpha}_t)\,\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t) = \mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) \tag{43}$$

A.9 SDE formulation : Variance Preserving (VP) SDE

The above processes can be written in terms of stochastic differential equations.

Forward process:

$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\mathbf{z}_{t-1}$$

$$
\begin{aligned}
\implies \mathbf{x}(t+\Delta t) &= \sqrt{1-\beta(t+\Delta t)\Delta t}\,\mathbf{x}(t) + \sqrt{\beta(t+\Delta t)\Delta t}\,\mathbf{z}(t) \\
&\approx \left(1 - \frac{1}{2}\beta(t+\Delta t)\Delta t\right)\mathbf{x}(t) + \sqrt{\beta(t+\Delta t)\Delta t}\,\mathbf{z}(t) \\
&\approx \mathbf{x}(t) - \frac{1}{2}\beta(t)\Delta t\,\mathbf{x}(t) + \sqrt{\beta(t)\Delta t}\,\mathbf{z}(t) \\
\implies \mathrm{d}\mathbf{x} &= -\frac{1}{2}\beta(t)\mathbf{x}\,\mathrm{d}t + \sqrt{\beta(t)}\,\mathrm{d}\mathbf{w} \qquad (44)
\end{aligned}
$$

Mean (from eq. 5.50 in Sarkka & Solin (2019)):

$$
\begin{aligned}
&\mathrm{d}\mathbf{x} = \mathbf{f}\,\mathrm{d}t + \mathbf{G}\,\mathrm{d}\mathbf{w} \implies \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[\mathbf{f}] \\
&\therefore \frac{\mathrm{d}\boldsymbol{\mu}_{\text{DDPM}}(t)}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[-\frac{1}{2}\beta(t)\mathbf{x}] = -\frac{1}{2}\beta(t)\mathbb{E}_{\mathbf{x}}(\mathbf{x}) = -\frac{1}{2}\beta(t)\boldsymbol{\mu}_{\text{DDPM}}(t) \\
&\implies \frac{\mathrm{d}\boldsymbol{\mu}_{\text{DDPM}}(t)}{\boldsymbol{\mu}_{\text{DDPM}}(t)} = -\frac{1}{2}\beta(t)\mathrm{d}t \implies \log\boldsymbol{\mu}_{\text{DDPM}}(t)|_0^t = -\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s \\
&\implies \log\boldsymbol{\mu}_{\text{DDPM}}(t) - \log\boldsymbol{\mu}(0) = -\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s \implies \log\frac{\boldsymbol{\mu}_{\text{DDPM}}(t)}{\boldsymbol{\mu}(0)} = -\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s \\
&\implies \boldsymbol{\mu}_{\text{DDPM}}(t) = \boldsymbol{\mu}(0)\,e^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}
\end{aligned}
$$

Covariance (from eq. 5.51 in Sarkka & Solin (2019)):

$$
\begin{aligned}
&\mathrm{d}\mathbf{x} = \mathbf{f}\,\mathrm{d}t + \mathbf{G}\,\mathrm{d}\mathbf{w} \implies \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{cov}}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[\mathbf{f}(\mathbf{x}-\boldsymbol{\mu})^T] + \mathbb{E}_{\mathbf{x}}[(\mathbf{x}-\boldsymbol{\mu})\mathbf{f}^T] + \mathbb{E}_{\mathbf{x}}[\mathbf{G}\mathbf{G}^T] \\
&\therefore \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{DDPM}}(t)}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[-\frac{1}{2}\beta(t)\mathbf{x}\mathbf{x}^T] + \mathbb{E}_{\mathbf{x}}[\mathbf{x}(-\frac{1}{2}\beta(t)\mathbf{x})^T] + \mathbb{E}_{\mathbf{x}}[\sqrt{\beta(t)}\mathbf{I}\sqrt{\beta(t)}\mathbf{I}] \\
&\qquad\qquad = -\beta(t)\boldsymbol{\Sigma}_{\text{DDPM}}(t) + \beta(t)\mathbf{I} = \beta(t)(\mathbf{I} - \boldsymbol{\Sigma}_{\text{DDPM}}(t)) \\
&\implies \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{DDPM}}(t)}{\mathbf{I} - \boldsymbol{\Sigma}_{\text{DDPM}}(t)} = \beta(t)\mathrm{d}t \implies -\log(\mathbf{I} - \boldsymbol{\Sigma}_{\text{DDPM}}(t))|_0^t = \int_0^t \beta(s)\mathrm{d}s \\
&\implies -\log(\mathbf{I} - \boldsymbol{\Sigma}_{\text{DDPM}}(t)) + \log(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{x}}(0)) = \int_0^t \beta(s)\mathrm{d}s \\
&\implies \frac{\mathbf{I} - \boldsymbol{\Sigma}_{\text{DDPM}}(t)}{\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{x}}(0)} = e^{-\int_0^t \beta(s)\mathrm{d}s} \implies \boldsymbol{\Sigma}_{\text{DDPM}}(t) = \mathbf{I} - e^{-\int_0^t \beta(s)\mathrm{d}s}(\mathbf{I} - \boldsymbol{\Sigma}_{\mathbf{x}}(0)) \\
&\implies \boldsymbol{\Sigma}_{\text{DDPM}}(t) = \mathbf{I} + e^{-\int_0^t \beta(s)\mathrm{d}s}(\boldsymbol{\Sigma}_{\mathbf{x}}(0) - \mathbf{I})
\end{aligned}
$$

For each data point $\mathbf{x}(0)$, $\boldsymbol{\mu}(0) = \mathbf{x}(0)$, $\boldsymbol{\Sigma}_{\mathbf{x}}(0) = \mathbf{0}$:

$$
\begin{aligned}
\implies \boldsymbol{\mu}_{\text{DDPM}}(t) &= \mathbf{x}(0)\,e^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}, \\
\boldsymbol{\Sigma}_{\text{DDPM}}(t) &= \mathbf{I} + e^{-\int_0^t \beta(s)\mathrm{d}s}(\mathbf{0} - \mathbf{I}) = \mathbf{I} - \mathbf{I}e^{-\int_0^t \beta(s)\mathrm{d}s} \\
\therefore p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) &= \mathcal{N}\left(\mathbf{x}(t) \mid \mathbf{x}(0)\,e^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}, \mathbf{I} - \mathbf{I}e^{-\int_0^t \beta(s)\mathrm{d}s}\right)
\end{aligned}
$$

Beta schedule, linear:

$$
\beta(t) = \beta_{\min} + t(\beta_{\max} - \beta_{\min}) \implies \int_0^t \beta(s)\mathrm{d}s = t\beta_{\min} + \frac{t^2}{2}(\beta_{\max} - \beta_{\min})
$$

B Non-isotropic DDPM (NI-DDPM)

B.1 Score for NI-DDPM

For a fixed sequence of positive scales $0 < \beta_1 < \cdots < \beta_L < 1$, $\bar{\alpha}_t = \prod_{s=1}^{i}(1-\beta_s)$,

$$q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t \mid \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{\Sigma}) \tag{45}$$

$$\implies \mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\sqrt{\boldsymbol{\Sigma}}\mathbf{z}_{t-1} \tag{46}$$

$$q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{\Sigma}) \tag{47}$$

$$\implies \mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon} \implies \boldsymbol{\epsilon} = \sqrt{\boldsymbol{\Sigma}^{-1}}\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{\sqrt{1-\bar{\alpha}_t}} \tag{48}$$

$$\implies \nabla_{\mathbf{x}_t} \log q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_0) = -\boldsymbol{\Sigma}^{-1}\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{1-\bar{\alpha}_t} = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon} \tag{49}$$

Derivation of the score value:

$$\begin{aligned}
&q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\boldsymbol{\Sigma}) \\
&= \frac{1}{(2\pi)^{D/2}((1-\bar{\alpha}_t)|\boldsymbol{\Sigma}|)^{1/2}} \exp\left(-\frac{1}{2(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)\right) \\
\implies &\log q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_0) = -\log((2\pi)^{D/2}((1-\bar{\alpha}_t)|\boldsymbol{\Sigma}|)^{1/2}) \\
&\qquad - \frac{1}{2(1-\bar{\alpha}_t)}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0) \\
\implies &\nabla_{\mathbf{x}_t} \log q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_0) = -\frac{1}{\cancel{2}(1-\bar{\alpha}_t)}\cancel{2}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0) = -\frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon}
\end{aligned}$$

B.2 Objective function for NI-DDPM

The objective function for score estimation in NI-DDPM at noise level $\bar{\alpha}_t$ is:

$$\begin{aligned}
\ell^{\text{NI-DDPM}}(\boldsymbol{\theta}; \bar{\alpha}_t) &\triangleq \frac{1}{2}\mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \boldsymbol{\Sigma}^{-1}\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{1-\bar{\alpha}_t}\right\|_2^2\right] \\
&\triangleq \frac{1}{2}\mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \frac{1}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\epsilon}\right\|_2^2\right]
\end{aligned} \tag{50}$$

B.3 Expected value of score for NI-DDPM

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_{\mathbf{x}_t} \log q_{\bar{\alpha}_t}^{\text{NI-DDPM}}(\mathbf{x}_t \mid \mathbf{x}_0)\right\|_2^2\right] = \mathbb{E}\left[\left\|-\boldsymbol{\Sigma}^{-1}\frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0}{1-\bar{\alpha}_t}\right\|_2^2\right] \\
&= \mathbb{E}\left[\left\|\boldsymbol{\Sigma}^{-1}\frac{\sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{\epsilon}}{1-\bar{\alpha}_t}\right\|_2^2\right] = \frac{1}{1-\bar{\alpha}_t}\boldsymbol{\Sigma}^{-1}\mathbb{E}\left[\|\boldsymbol{\epsilon}\|_2^2\right] = \frac{1}{1-\bar{\alpha}_t}\boldsymbol{\Sigma}^{-1}
\end{aligned} \tag{51}$$

B.4 Overall objective function for NI-DDPM

$$\mathcal{L}(\boldsymbol{\theta}; \{\bar{\alpha}_t\}_{t=1}^{L}) \triangleq \frac{1}{L}\sum_{t=1}^{L}\lambda(\alpha_t)\,\ell^{\text{NI-DDPM}}(\boldsymbol{\theta}; \bar{\alpha}_t)$$

λ inverse of the variance i.e. $\lambda(\bar{\alpha}_t) = (1-\bar{\alpha}_t)\boldsymbol{\Sigma} \implies$

$$\mathcal{L}_{\text{NI-DDPM}}(\boldsymbol{\theta}; \{\bar{\alpha}_t\}_{t=1}^{L}) \triangleq \frac{1}{2L}\sum_{t=1}^{L}\mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\|\sqrt{1-\bar{\alpha}_t}\sqrt{\boldsymbol{\Sigma}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \bar{\alpha}_t) + \sqrt{\boldsymbol{\Sigma}^{-1}}\frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\mathbf{x}_0)}{\sqrt{1-\bar{\alpha}_t}}\right\|_2^2\right]$$

$$= \frac{1}{2L}\sum_{t=1}^{L}\mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\|\sqrt{1-\bar{\alpha}_t}\sqrt{\mathbf{\Sigma}}\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\bar{\alpha}_t)+\boldsymbol{\epsilon}\right\|_2^2\right] \tag{52}$$

B.5 NI-DDPM score estimation using noise estimation

A score model that matches the actual score-noise relationship in eq. (49) is:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\bar{\alpha}_t) = -\sqrt{\mathbf{\Sigma}^{-1}}\frac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^{(1)}(\mathbf{x}_t)}{\sqrt{1-\bar{\alpha}_t}} \tag{53}$$

In this case, the overall objective function changes to:

$$\mathcal{L}_{\text{NI-DDPM}}(\boldsymbol{\theta};\{\bar{\alpha}_t\}_{t=1}^{L}) \triangleq \frac{1}{2L}\sum_{t=1}^{L}\mathbb{E}_{q_{\bar{\alpha}_t}(\mathbf{x}_t|\mathbf{x}_0)p(\mathbf{x}_0)}\left[\left\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_{\boldsymbol{\theta}}^{(1)}(\mathbf{x}_t)\right\|_2^2\right] \tag{54}$$

B.6 Sampling in NI-DDPM

We compute the parameters of the reverse process using Bishop (2006) 2.116, by additionally conditioning on $\mathbf{x}_0$:

$$\boxed{\begin{aligned} p(\mathbf{u}) &= \mathcal{N}(\mathbf{u}\mid\boldsymbol{\mu},\mathbf{\Lambda}^{-1}),\\ p(\mathbf{v}\mid\mathbf{u}) &= \mathcal{N}(\mathbf{v}\mid\mathbf{A}\mathbf{u}+\mathbf{b},\mathbf{L}^{-1})\\ \implies p(\mathbf{v}) &= \mathcal{N}(\mathbf{v}\mid\mathbf{A}\boldsymbol{\mu}+\mathbf{b},\mathbf{L}^{-1}+\mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T),\\ \implies p(\mathbf{u}\mid\mathbf{v}) &= \mathcal{N}(\mathbf{u}\mid\mathbf{C}(\mathbf{A}^T\mathbf{L}(\mathbf{v}-\mathbf{b})+\mathbf{\Lambda}\boldsymbol{\mu}),\mathbf{C}),\mathbf{C}=(\mathbf{\Lambda}+\mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}\end{aligned}}$$

Here, $\mathbf{u}=\mathbf{x}_{t-1}|\mathbf{x}_0,\mathbf{v}=\mathbf{x}_t$:

$$\begin{aligned} q_{\bar{\alpha}_{t-1}}(\mathbf{x}_{t-1}\mid\mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_{t-1}\mid\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0,(1-\bar{\alpha}_{t-1})\mathbf{\Sigma})\\ q_{\bar{\alpha}_t}(\mathbf{x}_t\mid\mathbf{x}_{t-1},\mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_t\mid\sqrt{1-\beta_t}\mathbf{x}_{t-1},\beta_t\mathbf{\Sigma})\\ &\implies \mathbf{x}=\mathbf{x}_{t-1}\mid\mathbf{x}_0,\mathbf{y}=\mathbf{x}_t\mid\mathbf{x}_0,\ \text{need}\ p(\mathbf{x}\mid\mathbf{y})=q(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{x}_0)\\ &\implies \boldsymbol{\mu}=\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0,\mathbf{\Lambda}^{-1}=(1-\bar{\alpha}_{t-1})\mathbf{\Sigma},\mathbf{A}=\sqrt{1-\beta_t},\mathbf{b}=\mathbf{0},\mathbf{L}^{-1}=\beta_t\mathbf{\Sigma}\\ &\implies \mathbf{C}=\left(\frac{1}{1-\bar{\alpha}_{t-1}}\mathbf{\Sigma}^{-1}+(1-\beta_t)\frac{1}{\beta_t}\mathbf{\Sigma}^{-1}\right)^{-1}=\left(\frac{\cancel{\beta_t}+1-\cancel{\beta_t}-\alpha_t}{(1-\bar{\alpha}_{t-1})\beta_t}\mathbf{\Sigma}^{-1}\right)^{-1}\\ &\qquad=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{\Sigma}=\tilde{\beta}_t\mathbf{\Sigma}\\ \implies \mathbf{C}(\mathbf{A}^T\mathbf{L}(\mathbf{y}-\mathbf{b})+\mathbf{\Lambda}\boldsymbol{\mu}) &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{\Sigma}\left(\sqrt{1-\beta_t}\frac{1}{\beta_t}\mathbf{\Sigma}^{-1}\mathbf{x}_t+\frac{1}{1-\bar{\alpha}_{t-1}}\mathbf{\Sigma}^{-1}\sqrt{\bar{\alpha}_{t-1}}\mathbf{x}_0\right)\\ &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0+\frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t\end{aligned}$$

Thus, the parameters of the distribution of the reverse process are:

$$\begin{aligned} &\therefore q(\mathbf{x}_{t-1}\mid\mathbf{x}_t,\mathbf{x}_0)=\mathcal{N}(\mathbf{x}_{t-1}\mid\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0),\tilde{\beta}_t\mathbf{\Sigma})\text{ , where}\\ &\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0)=\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0+\frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t;\quad \tilde{\beta}_t=\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \qquad (56)\\ &\implies \mathbf{x}_{t-1}=\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\hat{\mathbf{x}}_0)+\sqrt{\tilde{\beta}_t}\sqrt{\mathbf{\Sigma}}\boldsymbol{\epsilon}_t\text{ , where}\\ &\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\hat{\mathbf{x}}_0^{(1)})=\frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t-\sqrt{1-\bar{\alpha}_t}\sqrt{\mathbf{\Sigma}}\boldsymbol{\epsilon}^*\right)\right)+\frac{\sqrt{1-\beta_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t\\ &\qquad=\frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\frac{\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_t-\frac{\sqrt{\bar{\alpha}_{t-1}}}{\sqrt{\bar{\alpha}_t}}\frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\sqrt{\mathbf{\Sigma}}\boldsymbol{\epsilon}^*+\frac{\sqrt{1-\beta_t}}{1-\bar{\alpha}_t}\left(1-\bar{\alpha}_{t-1}\right)\mathbf{x}_t\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{1-\beta_t}} \frac{\beta_t}{1-\bar{\alpha}_t} \mathbf{x}_t + \frac{\sqrt{1-\beta_t}}{1-\bar{\alpha}_t} \left(1 - \frac{\bar{\alpha}_t}{1-\beta_t}\right) \mathbf{x}_t - \frac{1}{\sqrt{1-\beta_t}} \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}^* \\
&= \frac{1}{\sqrt{1-\beta_t}} \left(\frac{\beta_t}{1-\bar{\alpha}_t} \mathbf{x}_t + \frac{1-\beta_t}{1-\bar{\alpha}_t} \left(1 - \frac{\bar{\alpha}_t}{1-\beta_t}\right) \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}^* \right) \\
&= \frac{1}{\sqrt{1-\beta_t}} \left(\frac{\cancel{\beta_t} + \cancel{1-\beta_t - \bar{\alpha}_t}}{\cancel{1-\bar{\alpha}_t}} \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}^* \right) \\
\implies \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \hat{\mathbf{x}}_0) &= \frac{1}{\sqrt{1-\beta_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}^{(1)}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) \right), \qquad (57)
\end{aligned}
$$

$$
\begin{aligned}
\implies \mathbf{x}_{t-1} &= \frac{1}{\sqrt{1-\beta_t}} \left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}} \sqrt{\mathbf{\Sigma}}\, \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) \right) + \sqrt{\tilde{\beta}_t} \sqrt{\mathbf{\Sigma}}\, \boldsymbol{\epsilon}_t \qquad (58) \\
&= \frac{1}{\sqrt{1-\beta_t}} \left(\mathbf{x}_t + \beta_t \mathbf{\Sigma}\, \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t) \right) + \sqrt{\tilde{\beta}_t} \sqrt{\mathbf{\Sigma}}\, \boldsymbol{\epsilon}_t, \qquad (59)
\end{aligned}
$$

B.7 Sampling using DDIM

$$
\begin{aligned}
q_{\bar{\alpha}_L}(\mathbf{x}_L \mid \mathbf{x}_0) &= \mathcal{N}(\mathbf{x}_L \mid \sqrt{\bar{\alpha}_L} \mathbf{x}_0, (1-\bar{\alpha}_L)\mathbf{\Sigma}) \qquad (60) \\
q(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) &= \mathcal{N}\left(\mathbf{x}_{t-1} \mid \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0}{\sqrt{1-\bar{\alpha}_t}}, \mathbf{0} \right) \qquad (61) \\
\implies q_{\bar{\alpha}_t}(\mathbf{x}_t \mid \mathbf{x}_0) &= \mathcal{N}\left(\mathbf{x}_t \mid \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{\Sigma} \right)
\end{aligned}
$$

$$
\begin{aligned}
\implies \hat{\mathbf{x}}_0 &= \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t)) \\
\mathbf{x}_{t-1} &= \sqrt{\bar{\alpha}_{t-1}} \hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-1}} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) \qquad (62)
\end{aligned}
$$

B.8 Expected Denoised Sample

From [13], assuming isotropic Gaussian noise of covariance $(1-\bar{\alpha}_t)\mathbf{\Sigma}$, we know that the expected denoised sample $\mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) \triangleq \mathbb{E}_{\mathbf{x} \sim q_{\bar{\alpha}_t}(\mathbf{x}|\mathbf{x}_t)}[\mathbf{x}]$ and the optimal score $\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t)$ are related as:

$$
\begin{aligned}
\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t) &= \frac{1}{1-\bar{\alpha}_t} \mathbf{\Sigma}^{-1} (\mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) - \mathbf{x}_t) \\
\implies \mathbf{x}^*(\mathbf{x}_t, \bar{\alpha}_t) &= \mathbf{x}_t + (1-\bar{\alpha}_t)\mathbf{\Sigma}\, \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_t, \bar{\alpha}_t) = \mathbf{x}_t - \sqrt{1-\bar{\alpha}_t} \sqrt{\mathbf{\Sigma}}\, \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_t) \qquad (63)
\end{aligned}
$$

B.9 SDE formulation : Non-Isotropic Variance Preserving (NIVP) SDE

Forward process:

$$
\begin{aligned}
\mathbf{x}_t &= \sqrt{1-\beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}_{t-1} \\
\implies \mathbf{x}(t+\Delta t) &= \sqrt{1-\beta(t+\Delta t)\Delta t}\, \mathbf{x}(t) + \sqrt{\beta(t+\Delta t)\Delta t}\, \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}(t) \\
&\approx \left(1 - \frac{1}{2}\beta(t+\Delta t)\Delta t\right) \mathbf{x}(t) + \sqrt{\beta(t+\Delta t)\Delta t}\, \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}(t) \\
&\approx \mathbf{x}(t) - \frac{1}{2}\beta(t)\Delta t\, \mathbf{x}(t) + \sqrt{\beta(t)\Delta t}\, \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}(t) \\
\implies \mathrm{d}\mathbf{x} &= -\frac{1}{2}\beta(t)\mathbf{x}\, \mathrm{d}t + \sqrt{\beta(t)} \sqrt{\mathbf{\Sigma}}\, \mathrm{d}\mathbf{w} \qquad (64)
\end{aligned}
$$

Mean (from eq. 5.50 in Sarkka & Solin (2019)):

$$
\begin{aligned}
&\mathrm{d}\mathbf{x} = \mathbf{f}\, \mathrm{d}t + \mathbf{G}\, \mathrm{d}\mathbf{w} \implies \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[\mathbf{f}] \\
&\therefore \frac{\mathrm{d}\boldsymbol{\mu}_{\text{NI-DDPM}}(t)}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[-\frac{1}{2}\beta(t)\mathbf{x}] = -\frac{1}{2}\beta(t)\mathbb{E}_{\mathbf{x}}(\mathbf{x}) = -\frac{1}{2}\beta(t)\boldsymbol{\mu}_{\text{NI-DDPM}}(t)
\end{aligned}
$$

$$\implies \frac{\mathrm{d}\boldsymbol{\mu}_{\text{NI-DDPM}}(t)}{\boldsymbol{\mu}_{\text{NI-DDPM}}(t)} = -\frac{1}{2}\beta(t)\mathrm{d}t \implies \log \boldsymbol{\mu}_{\text{NI-DDPM}}(t)|_0^t = -\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s$$
$$\implies \log \boldsymbol{\mu}_{\text{NI-DDPM}}(t) - \log \boldsymbol{\mu}(0) = -\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s \implies \log \frac{\boldsymbol{\mu}_{\text{NI-DDPM}}(t)}{\boldsymbol{\mu}(0)} = -\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s$$
$$\implies \boldsymbol{\mu}_{\text{NI-DDPM}}(t) = \boldsymbol{\mu}(0)\, e^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}$$

Covariance (from eq. 5.51 in Sarkka & Solin (2019)):

$$\mathrm{d}\mathbf{x} = \mathbf{f}\,\mathrm{d}t + \mathbf{G}\,\mathrm{d}\mathbf{w} \implies \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{cov}}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[\mathbf{f}(\mathbf{x}-\boldsymbol{\mu})^T] + \mathbb{E}_{\mathbf{x}}[(\mathbf{x}-\boldsymbol{\mu})\mathbf{f}^T] + \mathbb{E}_{\mathbf{x}}[\mathbf{G}\mathbf{G}^T]$$
$$\therefore \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{NI-DDPM}}(t)}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[-\frac{1}{2}\beta(t)\mathbf{x}\mathbf{x}^T] + \mathbb{E}_{\mathbf{x}}[\mathbf{x}(-\frac{1}{2}\beta(t)\mathbf{x})^T] + \mathbb{E}_{\mathbf{x}}[\sqrt{\beta(t)}\sqrt{\boldsymbol{\Sigma}}\sqrt{\beta(t)}\sqrt{\boldsymbol{\Sigma}}]$$
$$= -\beta(t)\boldsymbol{\Sigma}_{\text{NI-DDPM}}(t) + \beta(t)\boldsymbol{\Sigma} = \beta(t)(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t))$$
$$\implies \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{NI-DDPM}}(t)}{\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t)} = \beta(t)\mathrm{d}t \implies -\log(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t))|_0^t = \int_0^t \beta(s)\mathrm{d}s$$
$$\implies -\log(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t)) + \log(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\mathbf{x}}(0)) = \int_0^t \beta(s)\mathrm{d}s$$
$$\implies \frac{\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t)}{\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\mathbf{x}}(0)} = e^{-\int_0^t \beta(s)\mathrm{d}s} \implies \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t) = \boldsymbol{\Sigma} - e^{-\int_0^t \beta(s)\mathrm{d}s}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\mathbf{x}}(0))$$
$$\implies \boldsymbol{\Sigma}_{\text{NI-DDPM}}(t) = \boldsymbol{\Sigma} + e^{-\int_0^t \beta(s)\mathrm{d}s}(\boldsymbol{\Sigma}_{\mathbf{x}}(0) - \boldsymbol{\Sigma})$$

For each data point $\mathbf{x}(0)$, $\boldsymbol{\mu}(0) = \mathbf{x}(0)$, $\boldsymbol{\Sigma}_{\mathbf{x}}(0) = \mathbf{0}$:

$$\implies \boldsymbol{\mu}_{\text{NI-DDPM}}(t) = \mathbf{x}(0)\, e^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s},$$
$$\boldsymbol{\Sigma}_{\text{NI-DDPM}}(t) = \boldsymbol{\Sigma} + e^{-\int_0^t \beta(s)\mathrm{d}s}(\mathbf{0} - \boldsymbol{\Sigma}) = \boldsymbol{\Sigma}(\mathbf{I} - \mathbf{I}e^{-\int_0^t \beta(s)\mathrm{d}s})$$
$$\therefore p_{0t}(\mathbf{x}(t) \mid \mathbf{x}(0)) = \mathcal{N}\left(\mathbf{x}(t) \mid \mathbf{x}(0)\, e^{-\frac{1}{2}\int_0^t \beta(s)\mathrm{d}s}, \boldsymbol{\Sigma}(\mathbf{I} - \mathbf{I}e^{-\int_0^t \beta(s)\mathrm{d}s})\right)$$

C GFF

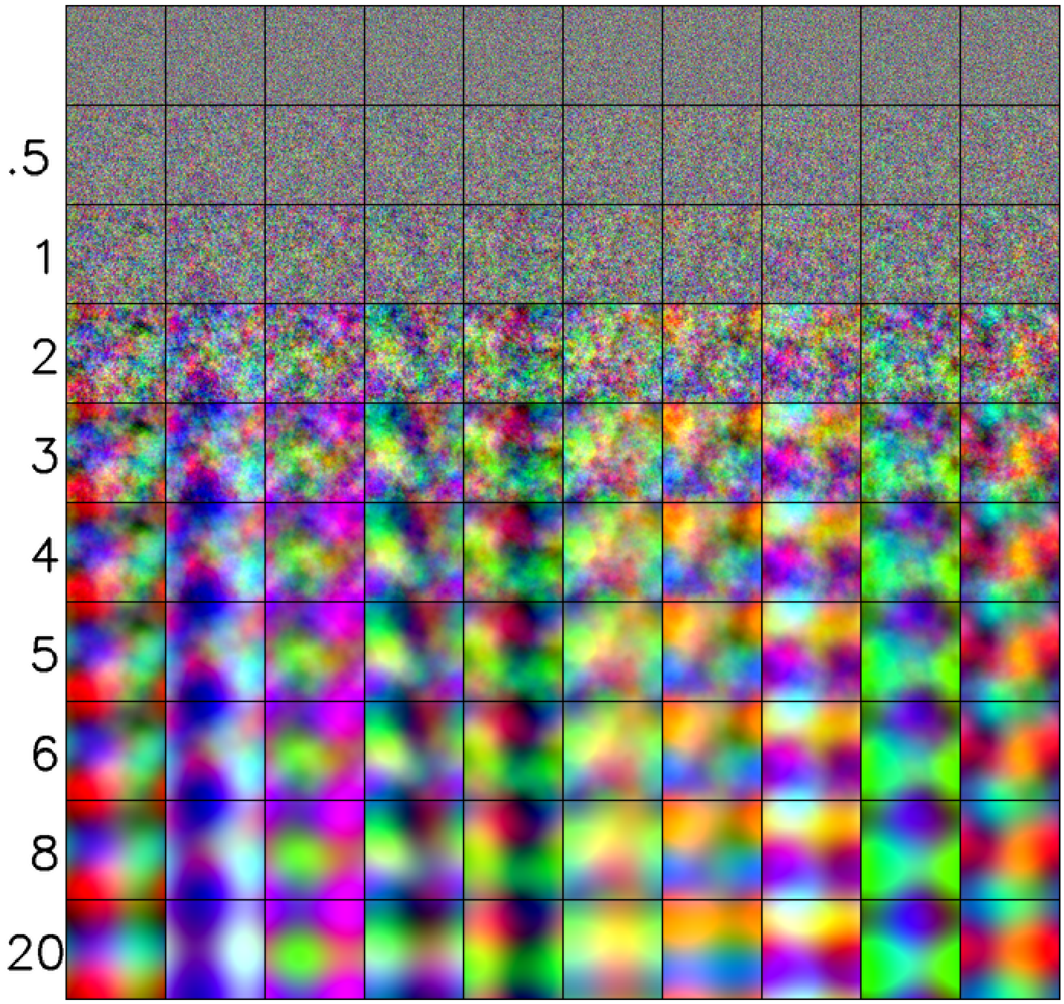


Figure 2: (left to right) 10 GFF images (each varying downwards) as a function of the power γ of the index (mentioned on the left).

A GFF image can be obtained from a noise image in the following way:

1. First, sample an $n{\times}n$ noise image $\mathbf{z}$ from the standard complex normal distribution with covariance matrix $\Gamma = \mathbf{I}_N$ where $N = n^2$ is the total number of pixels, and pseudo-covariance matrix $C = \mathbf{0}$: $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N, \mathbf{0})$.
   The standard **complex** normal distribution is one where the real part $\mathbf{x}$ and imaginary part $\mathbf{y}$ are each distributed as the standard normal distribution with variance $\frac{1}{2}\mathbf{I}_N$. Let $\mathbf{\Sigma_{ab}}$ be the covariance matrix between $\mathbf{a}$ and $\mathbf{b}$. We know that $\mathbf{\Sigma_{xx}} = \mathbf{\Sigma_{yy}} = \frac{1}{2}\mathbf{I}_N$, and $\mathbf{\Sigma_{xy}} = \mathbf{\Sigma_{yx}} = \mathbf{0}_N$ Then:
$$\Gamma = \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^H] = \mathbf{\Sigma_{xx}} + \mathbf{\Sigma_{yy}} + i(\mathbf{\Sigma_{yx}} - \mathbf{\Sigma_{xy}}) = \mathbf{I}_N \tag{65}$$
$$C = \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^T] = \mathbf{\Sigma_{xx}} - \mathbf{\Sigma_{yy}} + i(\mathbf{\Sigma_{yx}} + \mathbf{\Sigma_{xy}}) = \mathbf{0}_N \tag{66}$$
2. Apply the Discrete Fourier Transform using the $N{\times}N$ weights matrix $\mathbf{W}_N$: $\mathbf{W}_N\mathbf{z}$.
3. Consider a diagonal $N{\times}N$ matrix of the reciprocal of an index value k_{ij} per pixel (i, j) in Fourier space : $\mathbf{K}^{-1} = [1/|k_{ij}|]_{(i,j)}$, and multiply this with the above: $\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}$.
4. Take its Inverse Discrete Fourier Transform ($\mathbf{W}_N^{-1}$) to make the raw GFF image: $\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}$. However, this results in a GFF image with a small variance.

5. Normalize the image with the standard deviation σ_N corresponding to its resolution N:

$$\mathbf{g} = \frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z} \quad \Longleftrightarrow \quad \mathbf{z} = \sigma_N \mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N\mathbf{g} \tag{67}$$

6. Extract only the real part of $\mathbf{g}$, and normalize accordingly (refer Appendix C.2):

$$\mathbf{g}_{\text{real}} = \frac{1}{\sqrt{2N}\sigma_N}\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}\right) \iff \mathbf{z} = \sqrt{2N}\sigma_N\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N\mathbf{g}_{\text{real}}\right) \tag{68}$$

C.1 Probability distribution of GFF

Let the probability distribution of GFF images be $\mathcal{G}$. This can be seen as a non-isotropic multivariate Gaussian with a non-diagonal covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{g}_{\text{real}} \sim \mathcal{G} = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{0}_N, \mathbf{\Sigma}) \tag{69}$$

We know from the properties of Discrete Fourier Transform (following the normalization convention of the Pytorch / Numpy implementation) that:

$$\mathbf{W}_N = \mathbf{W}_N^T; \mathbf{W}_N^{-1} = {\mathbf{W}_N^{-1}}^T; \mathbf{W}_N^{-1} = \frac{1}{N}\mathbf{W}_N^* = \frac{1}{N}\mathbf{W}_N^H \tag{70}$$

$\boldsymbol{\mu}$ is given by:

$$\begin{aligned}\boldsymbol{\mu} &= \mathbb{E}_{\mathbf{g}}[\mathbf{g}] = \mathbb{E}_{\mathbf{z}}\left[\frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}\right] = \frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbb{E}_{\mathbf{z}}\left[\mathbf{z}\right] \\ \implies \boldsymbol{\mu} &= \mathbf{0}_N \qquad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}] = \mathbf{0}_N]\end{aligned} \tag{71}$$

$\mathbf{\Sigma}$ is given by:

$$\begin{aligned}\mathbf{\Sigma} &= \mathbb{E}_{\mathbf{g}}[\ \mathbf{g}\ \mathbf{g}^T\ ] \\ &= \mathbb{E}_{\mathbf{z}}\left[\frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}\ \left(\frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}\right)^T\right] \\ &= \mathbb{E}_{\mathbf{z}}\left[\frac{1}{\sigma_N^2}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{z}\ \mathbf{z}^T\mathbf{W}_N^T\mathbf{K}^{-1}\mathbf{W}_N^{-T}\right] \qquad [\because \mathbf{K}^{-1} \text{ is diagonal }] \\ &= \frac{1}{\sigma_N^2}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^T]\mathbf{W}_N\mathbf{K}^{-1}\mathbf{W}_N^{-1} \\ &= \frac{1}{\sigma_N^2}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{W}_N\mathbf{K}^{-1}\mathbf{W}_N^{-1} \qquad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^T] = \mathbf{I}_N] \\ &= \left(\frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\right)\left(\frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\right)^T \qquad [\because (\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N)^T = \mathbf{W}_N\mathbf{K}^{-1}\mathbf{W}_N^{-1}]\end{aligned}$$

σ_N^2 is such that the variance of each variable is 1, i.e. each diagonal element of $\mathbf{\Sigma}$ is 1. Thus,

$$\sigma_N^2 = \text{Var}[\mathbf{\Sigma}_N] \quad \text{where } \mathbf{\Sigma}_N = \mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\mathbf{W}_N\mathbf{K}^{-1}\mathbf{W}_N^{-1} \tag{72}$$

Hence,

$$\sqrt{\mathbf{\Sigma}} = \frac{1}{\sigma_N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N, \tag{73}$$

$$\mathbf{\Sigma} = \sqrt{\mathbf{\Sigma}}\sqrt{\mathbf{\Sigma}}^T, \tag{74}$$

$$\sqrt{\mathbf{\Sigma}^{-1}} = \sigma_N\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N \tag{75}$$

$$\mathbf{\Sigma}^{-1} = \sqrt{\mathbf{\Sigma}^{-1}}^T\sqrt{\mathbf{\Sigma}^{-1}}, \tag{76}$$

C.2 Derivation of GFF $\boldsymbol{\Sigma}$ for real $\mathbf{g}$ and complex $\mathbf{z}$

z is complex, g is real : $\boldsymbol{\mu}$ is given by:

$$
\begin{aligned}
\boldsymbol{\mu} &= \mathbb{E}_{\mathbf{g}_{\text{real}}}[\mathbf{g}_{\text{real}}] = \mathbb{E}_{\mathbf{g}}[\ \frac{1}{2\sigma_N}(\mathbf{g}+\mathbf{g}^*)\ ] = \frac{1}{2\sigma_N}\left(\mathbb{E}_{\mathbf{g}}[\mathbf{g}] + \mathbb{E}_{\mathbf{g}}[\mathbf{g}^*]\right) \\
&= \frac{1}{2\sigma_N}\left(\mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z}] + \mathbb{E}_{\mathbf{z}}[(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^*]\right) \\
&= \frac{1}{2\sigma_N}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbb{E}_{\mathbf{z}}[\mathbf{z}] + \mathbf{W}_N^{-1*}\mathbf{K}^{-1}\mathbb{E}_{\mathbf{z}}[\mathbf{z}^*]\right) \\
\implies \boldsymbol{\mu} &= \mathbf{0}_N \qquad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}] = \mathbb{E}_{\mathbf{z}}[\mathbf{z}^*] = \mathbf{0}_N]
\end{aligned} \tag{77}
$$

z is complex, g is real : $\boldsymbol{\Sigma}$ is given by:

$$
\begin{aligned}
\boldsymbol{\Sigma} &= \mathbb{E}_{\mathbf{g}_{\text{real}}}[\ \mathbf{g}_{\text{real}}\ \mathbf{g}_{\text{real}}^T\ ] \\
&= \mathbb{E}_{\mathbf{g}}[\ \frac{1}{2\sigma_N}(\mathbf{g}+\mathbf{g}^*)\ \frac{1}{2\sigma_N}(\mathbf{g}+\mathbf{g}^*)^T\ ] \\
&= \frac{1}{4\sigma_N^2}\mathbb{E}_{\mathbf{g}}[\ (\mathbf{g}+\mathbf{g}^*)\ (\mathbf{g}^T+\mathbf{g}^H)\ ] \\
&= \frac{1}{4\sigma_N^2}\mathbb{E}_{\mathbf{g}}[\ \mathbf{g}\mathbf{g}^T + \mathbf{g}\mathbf{g}^H + \mathbf{g}^*\mathbf{g}^T + \mathbf{g}^*\mathbf{g}^H\ ] \\
&= \frac{1}{4\sigma_N^2}\left(\mathbb{E}_{\mathbf{g}}[\mathbf{g}\mathbf{g}^T] + \mathbb{E}_{\mathbf{g}}[\mathbf{g}\mathbf{g}^H] + \mathbb{E}_{\mathbf{g}}[\mathbf{g}^*\mathbf{g}^T] + \mathbb{E}_{\mathbf{g}}[\mathbf{g}^*\mathbf{g}^H]\right) \\
\mathbb{E}_{\mathbf{g}}[\mathbf{g}\mathbf{g}^T] &= \mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z}\ (\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^T] \\
&= \mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z}\ \mathbf{z}^T\mathbf{K}^{-1}\mathbf{W}_N^{-T}] \quad [\because \mathbf{K}^{-1} \text{ is diagonal }] \\
&= \mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^T]\mathbf{K}^{-1}\mathbf{W}_N^{-T} \\
&= \mathbf{0}_N \qquad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^T] = \mathbf{0}_N \text{ (eq. (66))}] \\
\mathbb{E}_{\mathbf{g}}[\mathbf{g}\mathbf{g}^H] &= \mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z}\ (\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^H] \\
&= \mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z}\ \mathbf{z}^H\mathbf{K}^{-1}\mathbf{W}_N^{-H}] \quad [\because \boldsymbol{K}^{-1} \text{ is real diagonal }] \\
&= \mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^H]\mathbf{K}^{-1}\frac{1}{N}\mathbf{W}_N \quad [\because \mathbf{W}_N^{-1} = \frac{1}{N}\mathbf{W}_N^H \text{ (eq. (70))}] \\
&= \frac{1}{N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N \qquad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^H] = \mathbf{I}_N \text{ (eq. (65))}] \\
\mathbb{E}_{\mathbf{g}}[\mathbf{g}^*\mathbf{g}^T] &= \mathbb{E}_{\mathbf{z}}[(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^*\ (\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^T] \\
&= \mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1*}\mathbf{K}^{-1}\mathbf{z}^*\ \mathbf{z}^T\mathbf{K}^{-1}\mathbf{W}_N^{-1}] \\
&= \frac{1}{N}\mathbf{W}_N\mathbf{K}^{-1}\mathbb{E}_{\mathbf{z}}[\mathbf{z}^*\mathbf{z}^T]\mathbf{K}^{-1}\mathbf{W}_N^{-1} \quad [\because \mathbf{W}_N^{-1} = \frac{1}{N}\mathbf{W}_N^* \text{ (eq. (70))}] \\
&= \frac{1}{N}\mathbf{W}_N\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N^{-1} \quad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}^*\mathbf{z}^T] = \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^H]^* = \mathbf{I}_N \text{ (eq. (65))}] \\
\mathbb{E}_{\mathbf{g}}[\mathbf{g}^*\mathbf{g}^H] &= \mathbb{E}_{\mathbf{z}}[(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^*\ (\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{z})^H] \\
&= \mathbb{E}_{\mathbf{z}}[\mathbf{W}_N^{-1*}\mathbf{K}^{-1}\mathbf{z}^*\ \mathbf{z}^H\mathbf{K}^{-1}\mathbf{W}_N^{-H}] \\
&= \mathbf{W}_N^{-1*}\mathbf{K}^{-1}\mathbb{E}_{\mathbf{z}}[\mathbf{z}^*\mathbf{z}^H]\mathbf{K}^{-1}\mathbf{W}_N^{-H} \\
&= \mathbf{0}_N \qquad [\because \mathbb{E}_{\mathbf{z}}[\mathbf{z}^*\mathbf{z}^H] = \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^T]^* = \mathbf{0}_N \text{ (eq. (65))}] \\
\implies \boldsymbol{\Sigma} &= \frac{1}{4\sigma_N^2}\left(\mathbf{0} + \frac{1}{N}\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N + \frac{1}{N}\mathbf{W}_N\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N^{-1} + \mathbf{0}\right) \\
&= \frac{1}{4N\sigma_N^2}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N + \mathbf{W}_N\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N^{-1}\right) \\
&= \frac{1}{4N\sigma_N^2}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N + (N\mathbf{W}_N^{-1*})\mathbf{K}^{-1}\mathbf{K}^{-1}(\frac{1}{N}\mathbf{W}_N^*)\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2N\sigma_N^2}\left(\frac{1}{2}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N + (\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N)^*\right)\right) \\
\implies \boldsymbol{\Sigma} &= \frac{1}{2N\sigma_N^2}\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{K}^{-1}\mathbf{W}_N\right), \quad (78) \\
\sqrt{\boldsymbol{\Sigma}} &= \frac{1}{\sqrt{2N}\sigma_N}\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\right), \quad (79) \\
\boldsymbol{\Sigma}^{-1} &= 2N\sigma_N^2\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}\mathbf{K}\mathbf{W}_N\right), \quad (80) \\
\sqrt{\boldsymbol{\Sigma}^{-1}} &= \sqrt{2N}\sigma_N\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N\right) \quad (81)
\end{aligned}
$$

C.3 Log probability of transformation

$$
\begin{aligned}
\mathbf{g} &= \sqrt{\boldsymbol{\Sigma}}\mathbf{z} \\
\implies \log p(\mathbf{g}) &= \log p(\mathbf{z}) - \log\left|\det\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{z}}\right| \\
&= \log p(\mathbf{z}) - \log\left|\det\sqrt{\boldsymbol{\Sigma}}\right| \\
&= \log p(\mathbf{z}) - \log\left|\det\frac{1}{\sqrt{2N}\sigma_N}\text{Real}\left(\mathbf{W}_N^{-1}\mathbf{K}^{-1}\mathbf{W}_N\right)\right| \\
&= \log p(\mathbf{z}) - \frac{1}{\sqrt{2N}\sigma_N}\log\left|\det\mathbf{K}^{-1}\right|
\end{aligned}
$$

This is useful for building (normalizing) flows using non-isotropic Gaussian noise.

C.4 Varying $\mathbf{K}$

The index matrix $\mathbf{K}$ involves computation of an index value k_{ij} per pixel (i, j). However, this index value could be raised to any power γ i.e. $|k_{ij}|^\gamma$. The effect of varying γ can be seen in Figure 2 : greater the γ, the more correlated are neighbouring pixels.

D Score Matching Langevin Dynamics (SMLD) [16, 17]

D.1 Score for SMLD

For isotropic Gaussian noise as in SMLD,

$$q_{\sigma_t}^{\text{SMLD}}(\mathbf{x}_i \mid \mathbf{x}) = \mathcal{N}(\mathbf{x}_i \mid \mathbf{x}, \sigma_i^2 \mathbf{I}) \implies \mathbf{x}_i = \mathbf{x} + \sigma_i \boldsymbol{\epsilon} \tag{82}$$

$$\implies \nabla_{\mathbf{x}_i} \log q_{\sigma_i}^{\text{SMLD}}(\mathbf{x}_i \mid \mathbf{x}) = -\frac{1}{\sigma_i^2}(\mathbf{x}_i - \mathbf{x}) = -\frac{1}{\sigma_i}\boldsymbol{\epsilon} \tag{83}$$

D.2 Objective function for SMLD

The objective function for SMLD at noise level σ is:

$$\ell^{\text{SMLD}}(\boldsymbol{\theta}; \sigma_i) \triangleq \frac{1}{2}\mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x})p(\mathbf{x})}\left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \frac{1}{\sigma_i^2}(\mathbf{x}_i - \mathbf{x}) \right\|_2^2\right] \tag{84}$$

D.3 Variance of actual score for SMLD

$$\mathbb{E}\left[\left\|\nabla_{\mathbf{x}_i} \log q_{\sigma_i}^{\text{SMLD}}(\mathbf{x}_i \mid \mathbf{x})\right\|_2^2\right] = \mathbb{E}\left[\left\| -\frac{(\mathbf{x}_i - \mathbf{x})}{\sigma_i^2} \right\|_2^2\right] = \mathbb{E}\left[\left\| \frac{\sigma_i \boldsymbol{\epsilon}}{\sigma_i^2} \right\|_2^2\right] = \frac{1}{\sigma_i^2}\mathbb{E}\left[\|\boldsymbol{\epsilon}\|_2^2\right] = \frac{1}{\sigma_i^2} \tag{85}$$

D.4 Overall objective function for SMLD

[16, 17] chose a geometric series of σ_i's, i.e. $\sigma_{i-1}/\sigma_i = \gamma$. The overall objective function was a weighted combination of the objectives at different noise levels, the weight being $\lambda(\sigma_i) = \sigma_i^2$:

$$\begin{aligned}
\mathcal{L}^{\text{SMLD}}(\boldsymbol{\theta}; \{\sigma_i\}_{i=1}^L) &\triangleq \frac{1}{2L}\sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x})p(\mathbf{x})}\left[\left\| \sigma_i \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \frac{(\tilde{\mathbf{x}} - \mathbf{x})}{\sigma_i} \right\|_2^2\right] \\
&= \frac{1}{2L}\sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x})p(\mathbf{x})}\left[\|\sigma_i \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \boldsymbol{\epsilon}\|_2^2\right]
\end{aligned} \tag{86}$$

D.5 Unconditional SMLD score estimation

Song et. al. discovered that empirically the estimated score was proportional to $\frac{1}{\sigma}$. So an unconditional score model is:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) = -\frac{1}{\sigma_i}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_i) \tag{87}$$

In this case, the overall objective function changes to:

$$\begin{aligned}
\mathcal{L}^{\text{SMLD}}(\boldsymbol{\theta}; \{\sigma_i\}_{i=1}^L) &\triangleq \frac{1}{2L}\sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x})p(\mathbf{x})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_i)\|_2^2\right] \\
&= \frac{1}{2L}\sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x})p(\mathbf{x})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x} + \sigma_i \boldsymbol{\epsilon})\|_2^2\right]
\end{aligned} \tag{88}$$

D.6 Sampling in SMLD

$i = 0$ corresponds to data, and $i = L$ corresponds to noise. Hence, $i = L, \cdots, 0$ is the time order for sampling.

Using ALS from [16, 17]:

$\mathbf{x}_L^M \sim \mathcal{N}(\mathbf{x} \mid \mathbf{0}, \sigma_{\max}\mathbf{I})$

$$\left.\begin{aligned} & \mathbf{x}_i^M = \mathbf{x}_{i+1}^0 \\ & \alpha_i = \epsilon\sigma_i^2/\sigma_{\min}^2 \\ & \mathbf{x}_i^{m-1} \leftarrow \mathbf{x}_i^m + \alpha_i \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i^m, \sigma_i) + \sqrt{2\alpha_i}\boldsymbol{\epsilon}_i^{m-1}, m = M, \cdots, 0 \\ \implies & \mathbf{x}_i^{m-1} \leftarrow \mathbf{x}_i^m - \tfrac{\alpha_i}{\sigma_i}\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_i^m) + \sqrt{2\alpha_i}\boldsymbol{\epsilon}_i^{m-1}, m = M, \cdots, 0 \end{aligned}\right\} i = L, \cdots, 1 \tag{89}$$

Using Consistent Annealed Sampling [8]:

$$\begin{aligned} & \alpha_i = \epsilon\sigma_i^2/\sigma_{\min}^2 = \eta\sigma_i^2;\ \beta = \sqrt{1 - \gamma^2(1 - \epsilon/\sigma_{\min}^2)^2};\ \gamma = \sigma_i/\sigma_{i-1}; \sigma_i > \sigma_{i-1} \\ & \mathbf{x}_{i-1} \leftarrow \mathbf{x}_i + \alpha_i \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \sigma_i) + \beta\sigma_{i-1}\boldsymbol{\epsilon}_{i-1}, i = L, \cdots, 1 \\ \implies & \mathbf{x}_{i-1} \leftarrow \mathbf{x}_i - \eta\sigma_i \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_i) + \beta\sigma_{i-1}\boldsymbol{\epsilon}_{i-1}, i = L, \cdots, 1 \end{aligned} \tag{90}$$

D.7 Expected Denoised Sample

From [13], assuming isotropic Gaussian noise, we know that the expected denoised sample $\mathbf{x}^*(\mathbf{x}_i, \sigma_i) \triangleq \mathbb{E}_{\mathbf{x} \sim q_{\sigma_i}(\mathbf{x}|\mathbf{x}_i)}[\mathbf{x}]$ and the optimal score $\mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \sigma_i)$ are related as:

$$\begin{aligned} & \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \sigma_i) = \frac{1}{\sigma_i^2}(\mathbf{x}^*(\mathbf{x}_i, \sigma_i) - \mathbf{x}_i) \\ \implies & \mathbf{x}^*(\mathbf{x}_i, \sigma_i) = \mathbf{x}_i + \sigma_i^2 \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \sigma_i) = \mathbf{x}_i - \sigma_i \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\mathbf{x}_i) \end{aligned} \tag{91}$$

D.8 SDE formulation : Variance Exploding (VE) SDE

Forward process:

$$\begin{aligned} \mathbf{x}_i &= \mathbf{x}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\boldsymbol{\epsilon}_{i-1} \\ \implies \mathbf{x}(t + \Delta t) &= \mathbf{x}(t) + \sqrt{(\sigma^2(t + \Delta t) - \sigma^2(t))\Delta t}\, \boldsymbol{\epsilon}(t) \\ &\approx \mathbf{x}(t) + \sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}\Delta t}\, \mathbf{w}(t) \\ \implies \mathrm{d}\mathbf{x} &= \sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}\, \mathrm{d}\mathbf{w} \end{aligned} \tag{92}$$

$$\begin{aligned} & \mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + \mathbf{L}(\mathbf{x}, t)\mathrm{d}\mathbf{w} \implies \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[\mathbf{f}(\mathbf{x}, t)], \\ & \frac{\mathrm{d}\mathbf{Cov}[\mathbf{x}]}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}[\mathbf{f}(\mathbf{x}, t)(\mathbf{x} - \boldsymbol{\mu})^T] + \mathbb{E}_{\mathbf{x}}[(\mathbf{x} - \boldsymbol{\mu})\mathbf{f}(\mathbf{x}, t)^T] + \mathbb{E}_{\mathbf{x}}[\mathbf{L}(\mathbf{x}, t)\mathbf{Q}\mathbf{L}^T(\mathbf{x}, t)] \end{aligned}$$

where $\mathbf{w}$ is Brownian motion, $\mathbf{Q}$ is the PSD of $\mathbf{w}$. For GFF noise, $\mathbf{Q} = \boldsymbol{\Sigma}$.

Mean and Covariance:

$$\begin{aligned} & \frac{\mathrm{d}\boldsymbol{\mu}_{\text{SMLD}}(t)}{\mathrm{d}t} = \mathbf{0} \implies \boldsymbol{\mu}_{\text{SMLD}}(t) = \boldsymbol{\mu}(0) \\ & \frac{\mathrm{d}\boldsymbol{\Sigma}_{\text{SMLD}}(t)}{\mathrm{d}t} = \mathbb{E}_{\mathbf{x}}\left[\sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}\sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}\right] = \frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t} \end{aligned}$$

E Non-isotropic SMLD (NI-SMLD)

E.1 Score for NI-SMLD

$$q_{\sigma_i}^{\text{SMLD}}(\mathbf{x}_i \mid \mathbf{x}) = \mathcal{N}(\mathbf{x}_i \mid \mathbf{x}, \sigma_i^2 \mathbf{\Sigma}) \implies \mathbf{x}_i = \mathbf{x} + \sigma_i \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon} \implies \boldsymbol{\epsilon} = \sqrt{\mathbf{\Sigma}^{-1}} \frac{\mathbf{x}_i - \mathbf{x}}{\sigma_i} \tag{93}$$

$$\implies \nabla_{\mathbf{x}_i} \log q_{\sigma_i}^{\text{SMLD}}(\mathbf{x}_i \mid \mathbf{x}) = -\mathbf{\Sigma}^{-1} \frac{\mathbf{x}_i - \mathbf{x}}{\sigma_i^2} = -\sqrt{\mathbf{\Sigma}^{-1}} \frac{\boldsymbol{\epsilon}}{\sigma_i} \tag{94}$$

E.2 Objective function for NI-SMLD

The objective function for SMLD at noise level σ is:

$$\ell^{\text{NI-SMLD}}(\boldsymbol{\theta}; \sigma_i) \triangleq \frac{1}{2} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x}) p(\mathbf{x})} \left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \mathbf{\Sigma}^{-1} \frac{\mathbf{x}_i - \mathbf{x}}{\sigma_i^2} \right\|_2^2 \right] \tag{95}$$

$$= \frac{1}{2} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x}) p(\mathbf{x})} \left[\left\| \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \frac{1}{\sigma_i} \sqrt{\mathbf{\Sigma}^{-1}} \boldsymbol{\epsilon} \right\|_2^2 \right] \tag{96}$$

E.3 Expected value of score for NI-SMLD

$$\mathbb{E}\left[\left\| \nabla_{\mathbf{x}_i} \log q_{\sigma_i}^{\text{NI-SMLD}}(\mathbf{x}_i \mid \mathbf{x}) \right\|_2^2 \right] = \mathbb{E}\left[\left\| -\mathbf{\Sigma}^{-1} \frac{\mathbf{x}_i - \mathbf{x}}{\sigma_i^2} \right\|_2^2 \right]$$

$$= \mathbb{E}\left[\left\| \mathbf{\Sigma}^{-1} \frac{\sigma_i \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}}{\sigma_i^2} \right\|_2^2 \right] = \frac{1}{\sigma_i^2} \mathbf{\Sigma}^{-1} \mathbb{E}\left[\|\boldsymbol{\epsilon}\|_2^2 \right] = \frac{1}{\sigma_i^2} \mathbf{\Sigma}^{-1} \dim(\boldsymbol{\epsilon}) \tag{97}$$

E.4 Overall objective function for NI-SMLD

$\implies \lambda(\sigma_i) = \sigma_i^2 \mathbf{\Sigma}$

$$\mathcal{L}^{\text{NI-SMLD}}(\boldsymbol{\theta}; \{\sigma_i\}_{i=1}^L) \triangleq \frac{1}{2L} \sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x}) p(\mathbf{x})} \left[\left\| \sigma_i \sqrt{\mathbf{\Sigma}} \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \sqrt{\mathbf{\Sigma}^{-1}} \frac{(\tilde{\mathbf{x}} - \mathbf{x})}{\sigma_i} \right\|_2^2 \right]$$

$$= \frac{1}{2L} \sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x}) p(\mathbf{x})} \left[\left\| \sigma_i \sqrt{\mathbf{\Sigma}} \mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) + \boldsymbol{\epsilon} \right\|_2^2 \right] \tag{98}$$

E.5 Unconditional NI-SMLD score estimation

An unconditional score model is:

$$\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_i, \sigma_i) = -\sqrt{\mathbf{\Sigma}^{-1}} \frac{1}{\sigma_i} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_i) \tag{99}$$

In this case, the overall objective function changes to:

$$\mathcal{L}^{\text{GFF-SMLD}}(\boldsymbol{\theta}; \{\sigma_i\}_{i=1}^L) \triangleq \frac{1}{2L} \sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x}) p(\mathbf{x})} \left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_i)\|_2^2 \right]$$

$$= \frac{1}{2L} \sum_{i=1}^{L} \mathbb{E}_{q_{\sigma_i}(\mathbf{x}_i|\mathbf{x}) p(\mathbf{x})} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x} + \sigma_i \sqrt{\mathbf{\Sigma}} \boldsymbol{\epsilon}) \right\|_2^2 \right] \tag{100}$$

E.6 Sampling in NI-SMLD

$i = 0$ corresponds to data, and $i = L$ corresponds to noise. Hence, $i = L, \cdots, 0$ is the time order for sampling.

Forward : $\mathbf{x}_i = \mathbf{x}_{i-1} + \sqrt{\sigma_i^2 - \sigma_{i-1}^2}\sqrt{\mathbf{\Sigma}}\boldsymbol{\epsilon}_{i-1}$

Reverse:

From Song et. al., ALS:

$$\begin{aligned} &\mathbf{x}_L^0 \sim \mathcal{N}(\mathbf{x} \mid \mathbf{0}, \sigma_{\max}\sqrt{\mathbf{\Sigma}}) \\ &\left.\begin{aligned} &\mathbf{x}_i^0 = \mathbf{x}_{i+1}^M \\ &\mathbf{x}_i^{m+1} \leftarrow \mathbf{x}_i^m + \alpha_i \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i^m, \sigma_i) + \sqrt{2\alpha_i}\sqrt{\mathbf{\Sigma}}\boldsymbol{\epsilon}_i^{m+1}, m = 1, \cdots, M \end{aligned}\right\} i = L, \cdots, 1 \\ &\alpha_i = \epsilon\sigma_i^2/\sigma_L^2 \end{aligned} \tag{101}$$

From Alexia et. al., CAS:

$$\begin{aligned} &\mathbf{x}_{i-1} \leftarrow \mathbf{x}_i + \alpha_i \mathbf{s}_{\boldsymbol{\theta}^*}(\mathbf{x}_i, \sigma_i) + \beta\sigma_{i-1}\sqrt{\mathbf{\Sigma}}\boldsymbol{\epsilon}_{i-1}, i = L, \cdots, 1 \\ &\alpha_i = \epsilon\sigma_t^2/\sigma_{\min}^2;\ \beta = \sqrt{1 - \gamma^2(1 - \epsilon/\sigma_{\min}^2)^2};\ \gamma = \sigma_t/\sigma_{t-1}; \sigma_t > \sigma_{t-1} \end{aligned} \tag{102}$$

E.7 Expected Denoised Sample

From [13], assuming isotropic Gaussian noise of covariance $\sigma^2\mathbf{\Sigma}$, we know that the expected denoised sample $\mathbf{x}^*(\tilde{\mathbf{x}}, \sigma) \triangleq \mathbb{E}_{\mathbf{x}\sim q_\sigma(\mathbf{x}|\tilde{\mathbf{x}})}[\mathbf{x}]$ and the optimal score $\mathbf{s}_{\boldsymbol{\theta}^*}(\tilde{\mathbf{x}}, \sigma)$ are related as:

$$\begin{aligned} \mathbf{s}_{\boldsymbol{\theta}^*}(\tilde{\mathbf{x}}, \sigma) &= \frac{1}{\sigma^2}\mathbf{\Sigma}^{-1}(\mathbf{x}^*(\tilde{\mathbf{x}}, \sigma) - \tilde{\mathbf{x}}) \\ \implies \mathbf{x}^*(\tilde{\mathbf{x}}, \sigma) &= \tilde{\mathbf{x}} + \sigma^2\mathbf{\Sigma}\, \mathbf{s}_{\boldsymbol{\theta}^*}(\tilde{\mathbf{x}}, \sigma) = \tilde{\mathbf{x}} - \sigma\sqrt{\mathbf{\Sigma}}\, \boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\tilde{\mathbf{x}}) \end{aligned} \tag{103}$$

E.8 Initial noise scale for NI-SMLD

Let $\hat{p}_{\sigma_1}(\mathbf{x}) \triangleq \frac{1}{N}\sum_{i=1}^N p^{(i)}(\mathbf{x})$, where $p^{(i)}(\mathbf{x}) \triangleq \mathcal{N}(\mathbf{x} \mid \mathbf{x}^{(i)}, \sigma_1^2\mathbf{\Sigma})$. With $r^{(i)}(\mathbf{x}) \triangleq \frac{p^{(i)}(\mathbf{x})}{\sum_{k=1}^N p^{(k)}(\mathbf{x})}$, the score function is $\nabla_{\mathbf{x}} \log \hat{p}_{\sigma_1}(\mathbf{x}) = \sum_{i=1}^N r^{(i)}(\mathbf{x})\nabla_{\mathbf{x}} \log p^{(i)}(\mathbf{x})$.

We know that:

$$\mathcal{N}(\mathbf{x} \mid \mathbf{x}^{(i)}, \sigma_1^2\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2\sigma_1^2}(\mathbf{x} - \mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)})\right)$$

$$\begin{aligned} &\mathbb{E}_{p^{(i)}(\mathbf{x})}[r^{(j)}(\mathbf{x})] = \int \frac{p^{(i)}(\mathbf{x})p^{(j)}(\mathbf{x})}{\sum_{k=1}^N p^{(k)}(\mathbf{x})}\mathrm{d}\mathbf{x} \leq \int \frac{p^{(i)}(\mathbf{x})p^{(j)}(\mathbf{x})}{p^{(i)}(\mathbf{x}) + p^{(j)}(\mathbf{x})}\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\int \frac{2}{\frac{1}{p^{(i)}(\mathbf{x})} + \frac{1}{p^{(j)}(\mathbf{x})}}\mathrm{d}\mathbf{x} \leq \frac{1}{2}\int \sqrt{p^{(i)}(\mathbf{x})p^{(j)}(\mathbf{x})}\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\left(-\frac{1}{4\sigma_1^2}\left((\mathbf{x} - \mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)}) + (\mathbf{x} - \mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(j)})\right)\right)\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\left(-\frac{1}{4\sigma_1^2}\left((\mathbf{x} - \mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)}) + (\mathbf{x} - \mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)} + \mathbf{x}^{(i)} - \mathbf{x}^{(j)})\right)\right)\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\left(-\frac{1}{4\sigma_1^2}\left((\mathbf{x} - \mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)}) + (\mathbf{x} - \mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)})\right.\right. \\ &\qquad\qquad \left.\left. + (\mathbf{x} - \mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})\right)\right)\mathrm{d}\mathbf{x} \\ &= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\left(-\frac{1}{4\sigma_1^2}\left((\mathbf{x} - \mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)}) + (\mathbf{x} - \mathbf{x}^{(i)} + \mathbf{x}^{(i)} - \mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}^{(i)})\right.\right. \\ &\qquad\qquad \left.\left. + (\mathbf{x} - \mathbf{x}^{(i)} + \mathbf{x}^{(i)} - \mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mathbf{x}^{(j)})\right)\right)\mathrm{d}\mathbf{x} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\Bigg(-\frac{1}{4\sigma_1^2}\Big((\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}^{(i)}) + (\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}^{(i)}) \\
&\qquad + (\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}^{(i)}) + (\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}) \\
&\qquad + (\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\Big)\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\Bigg(-\frac{1}{2\sigma_1^2}\Big((\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}^{(i)}) \\
&\qquad + (\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}) \\
&\qquad + \frac{1}{2}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\Big)\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\Bigg(-\frac{1}{2\sigma_1^2}\Big((\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}^{(i)}) \\
&\qquad + 2(\mathbf{x}-\mathbf{x}^{(i)})^T\mathbf{\Sigma}^{-1}\frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2} \\
&\qquad + \frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2}^T\mathbf{\Sigma}^{-1}\frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2} - \frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2}^T\mathbf{\Sigma}^{-1}\frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2} \\
&\qquad + \frac{1}{2}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\Big)\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\int \exp\Bigg(-\frac{1}{2\sigma_1^2}\Big((\mathbf{x}-\mathbf{x}^{(i)}+\frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{x}^{(i)}+\frac{(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})}{2}) \\
&\qquad + \frac{1}{4}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\Big)\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\exp\Big(-\frac{1}{8\sigma_1^2}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\Big) \\
&\qquad \int \frac{1}{(2\pi)^{D/2}\sigma_1^D|\mathbf{\Sigma}|^{1/2}}\exp\Big(-\frac{1}{2\sigma_1^2}\Big((\mathbf{x}-\frac{(\mathbf{x}^{(i)}+\mathbf{x}^{(j)})}{2})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\frac{(\mathbf{x}^{(i)}+\mathbf{x}^{(j)})}{2})\Big)\mathrm{d}\mathbf{x} \\
&= \frac{1}{2}\exp\Big(-\frac{1}{8\sigma_1^2}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\Big) \\
&\Longrightarrow \frac{1}{\sigma_1^2}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}) \approx 1 \\
&\Longrightarrow (\sqrt{\mathbf{\Sigma}^{-1}}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)}))^T(\sqrt{\mathbf{\Sigma}^{-1}}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})) \approx \sigma_1^2 \\
&\Longrightarrow \left\|\sqrt{\mathbf{\Sigma}^{-1}}(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\right\|_2 \approx \sigma_1 \\
&\Longrightarrow \left\|\sigma_N\ \mathrm{Real}\big(\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N(\mathbf{x}^{(i)}-\mathbf{x}^{(j)})\big)\right\|_2 \approx \sigma_1 \\
&\Longrightarrow \left\|\sigma_N\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N\mathbf{x}^{(i)} - \sigma_N\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N\mathbf{x}^{(j)}\right\|_2 \approx \sigma_1
\end{aligned}
$$

For CIFAR10, this $\sigma_1 \approx 20$ for NI-SMLD (whereas for SMLD $\sigma_1 \approx 50$).

E.9 Other noise scales

$$
\begin{aligned}
p_{\sigma_i}(r) &= \mathcal{N}(r \mid m_i, s_i^2), \text{where } m_i \triangleq \sqrt{D}\sigma_i; s_i^2 \triangleq \sigma_i^2/2 \\
\mathcal{I}_i &\triangleq (m_i - 3s_i, m_i + 3s_i) \\
p_{\sigma_i}(r \in \mathcal{I}_{i-1}) &= \Phi\left(\frac{(m_{i-1}+3s_{i-1})-m_i}{s_i}\right) - \Phi\left(\frac{(m_{i-1}-3s_{i-1})-m_i}{s_i}\right) \\
&= \Phi\left(\frac{\sqrt{2}}{\sigma_i}(\sqrt{D}\sigma_{i-1}+\frac{3\sigma_{i-1}}{\sqrt{2}}-\sqrt{D}\sigma_i)\right) - \Phi\left(\frac{\sqrt{2}}{\sigma_i}(\sqrt{D}\sigma_{i-1}-\frac{3\sigma_{i-1}}{\sqrt{2}}-\sqrt{D}\sigma_i)\right)
\end{aligned}
$$

$$= \Phi\left(\frac{1}{\sigma_i}(\sqrt{2D}(\sigma_{i-1}-\sigma_i)+3\sigma_{i-1})\right) - \Phi\left(\frac{1}{\sigma_i}(\sqrt{2D}(\sigma_{i-1}-\sigma_i)-3\sigma_{i-1})\right)$$
$$= \Phi\left(\sqrt{2D}(\gamma-1)+3\gamma\right) - \Phi\left(\sqrt{2D}(\gamma-1)-3\gamma\right) \approx 0.5$$

Hence, the value of γ remains (almost) the same:

$\gamma = 1.0375867506951884$ ($\sigma_1 = 20, \sigma_L = 0.01, L = 207$).

(wheras ealier $\gamma = 1.0375591319992028$ ($\sigma_1 = 50, \sigma_L = 0.01, L = 232$)

E.10 Configuring annealed Langevin dynamics

Let $\gamma = \frac{\sigma_{i-1}}{\sigma_i}$. For $\alpha = \epsilon \cdot \frac{\sigma_i^2}{\sigma_L^2}$, we have $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \text{Var}[\mathbf{x}_T])$, where

$$\frac{\text{Var}[\mathbf{x}_T]}{\sigma_i^2} = \gamma^2 \mathbf{P}^T \mathbf{\Sigma} \mathbf{P}^T + \frac{2\epsilon}{\sigma_L^2} \sum_{t=0}^{T-1} (\mathbf{P}^t \mathbf{\Sigma} \mathbf{P}^t) \tag{104}$$

Proof:

First, the conditions we know are

$$\mathbf{x}_0 \sim p_{\sigma_{i-1}}(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \mathbf{0}, \sigma_{i-1}^2 \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}\sigma_{i-1}^D |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2\sigma_{i-1}^2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right),$$
$$\nabla_{\mathbf{x}} \log p_{\sigma_i}(\mathbf{x}_t) = \nabla_{\mathbf{x}}\left(-\log(\text{const.}) - \frac{1}{2\sigma_i^2}\mathbf{x}_t^T\mathbf{\Sigma}^{-1}\mathbf{x}_t\right) = -\frac{1}{\sigma_i^2}\mathbf{\Sigma}^{-1}\mathbf{x}_t,$$
$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \alpha\nabla_{\mathbf{x}} \log p_{\sigma_i}(\mathbf{x}_t) + \sqrt{2\alpha}\mathbf{g}_t = \mathbf{x}_t - \alpha\frac{1}{\sigma_i^2}\mathbf{\Sigma}^{-1}\mathbf{x}_t + \sqrt{2\alpha}\mathbf{g}_t,$$

where $\mathbf{g}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\alpha = \epsilon\frac{\sigma_i^2}{\sigma_L^2}$. Therefore, the variance of $\mathbf{x}_t$ satisfies

$$\text{Var}[\mathbf{x}_t] = \begin{cases} \sigma_{i-1}^2\mathbf{\Sigma} & \text{if } t=0 \\ \text{Var}[\big(\mathbf{I} - \frac{\alpha}{\sigma_i^2}\mathbf{\Sigma}^{-1}\big)\mathbf{x}_{t-1}] + 2\alpha\mathbf{\Sigma} & \text{otherwise.} \end{cases}$$

$$\begin{aligned}
\text{Var}[\mathbf{A}\mathbf{x}] &= \mathbf{A}\text{Var}[\mathbf{x}]\mathbf{A}^T \\
\implies \text{Var}[\mathbf{x}_t] &= \big(\mathbf{I} - \frac{\alpha}{\sigma_i^2}\mathbf{\Sigma}^{-1}\big)\text{Var}[\mathbf{x}_{t-1}]\big(\mathbf{I} - \frac{\alpha}{\sigma_i^2}\mathbf{\Sigma}^{-1}\big) + 2\alpha\mathbf{\Sigma}
\end{aligned}$$

$$\text{Let } \mathbf{P} = \mathbf{I} - \frac{\alpha}{\sigma_i^2}\mathbf{\Sigma}^{-1} = \mathbf{I} - \frac{\epsilon}{\sigma_L^2}\mathbf{\Sigma}^{-1}$$

$$\begin{aligned}
\implies \text{Var}[\mathbf{x}_t] &= \mathbf{P}\text{Var}[\mathbf{x}_{t-1}]\mathbf{P} + 2\alpha\mathbf{\Sigma} \\
&= \mathbf{P}(\mathbf{P}\text{Var}[\mathbf{x}_{t-2}]\mathbf{P} + 2\alpha\mathbf{\Sigma})\mathbf{P} + 2\alpha\mathbf{\Sigma} \\
&= \mathbf{P}\mathbf{P}\text{Var}[\mathbf{x}_{t-2}]\mathbf{P}\mathbf{P} + 2\alpha(\mathbf{P}\mathbf{\Sigma}\mathbf{P} + \mathbf{\Sigma}) \\
&= \mathbf{P}^{(2)}\text{Var}[\mathbf{x}_{t-2}]\mathbf{P}^{(2)} + 2\alpha(\mathbf{P}\mathbf{\Sigma}\mathbf{P} + \mathbf{\Sigma}) \\
\implies \text{Var}[\mathbf{x}_T] &= \mathbf{P}^{(T)}\text{Var}[\mathbf{x}_0]\mathbf{P}^{(T)} + 2\alpha\sum_{t=0}^{T-1}(\mathbf{P}^{(t)}\mathbf{\Sigma}\mathbf{P}^{(t)}) \\
&= \sigma_{i-1}^2\mathbf{P}^{(T)}\mathbf{\Sigma}\mathbf{P}^{(T)} + 2\epsilon\frac{\sigma_i^2}{\sigma_L^2}\sum_{t=0}^{T-1}(\mathbf{P}^{(t)}\mathbf{\Sigma}\mathbf{P}^{(t)}) \\
\implies \frac{\text{Var}[\mathbf{x}_T]}{\sigma_i^2} &= \gamma^2\mathbf{P}^{(T)}\mathbf{\Sigma}\mathbf{P}^{(T)} + \frac{2\epsilon}{\sigma_L^2}\sum_{t=0}^{T-1}(\mathbf{P}^{(t)}\mathbf{\Sigma}\mathbf{P}^{(t)})
\end{aligned}$$

Hence, we choose ϵ s.t. $\frac{Var[\mathbf{x}_T]}{\sigma_i^2} \approx 1$:

$\epsilon = 3.1\text{e}{-7}$ for $T = 5$, $\epsilon = 2.0\text{e}{-6}$ for $T = 1$

(whereas earlier $\epsilon = 6.2\text{e}{-6}$ for $T = 5$)

E.11 Calculus of Variations

Alexia et. al. discovered in Appendix E that the unconditional score model's estimate of the score in the case of a single data point $\mathbf{x}_0$ is:

$$\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) = \frac{1}{L}\sum_{i=1}^{L}\left(\frac{\mathbb{E}_{\mathbf{x}\sim q_{\sigma_i}(\mathbf{x}|\tilde{\mathbf{x}})}[\mathbf{x}] - \tilde{\mathbf{x}}}{\sigma_i}\right) = \frac{\mathbf{x}_0 - \tilde{\mathbf{x}}}{\sigma_H} \tag{105}$$

$$\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \sigma_i) = \frac{1}{\sigma_i}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) = \frac{\mathbf{x}_0 - \tilde{\mathbf{x}}}{\sigma_i \sigma_H} \neq \frac{\mathbf{x}_0 - \tilde{\mathbf{x}}}{\sigma_i^2} \tag{106}$$

where $\frac{1}{\sigma_H} = \frac{1}{L}\sum_{i=1}^{L}\frac{1}{\sigma_i}$, i.e. σ_H is the harmonic mean of the σ_is used to train.

In our case,

$$\begin{aligned}
\frac{\partial \mathcal{L}_2}{\partial \mathbf{s}} &= \int\int q_\sigma(\tilde{\mathbf{x}}, \mathbf{x}, \sigma)\left(\mathbf{s}(\tilde{\mathbf{x}}) + \sqrt{\boldsymbol{\Sigma}^{-1}}\frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma}\right)\mathrm{d}\mathbf{x}\mathrm{d}\sigma = 0 \\
\iff \mathbf{s}(\tilde{\mathbf{x}})q(\tilde{\mathbf{x}}) &= \sqrt{\boldsymbol{\Sigma}^{-1}}\int\int q_\sigma(\tilde{\mathbf{x}}, \mathbf{x})p(\sigma)\left(\frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma}\right)\mathrm{d}\mathbf{x}\mathrm{d}\sigma \\
\iff \mathbf{s}(\tilde{\mathbf{x}})q(\tilde{\mathbf{x}}) &= \sqrt{\boldsymbol{\Sigma}^{-1}}\mathbb{E}_{\sigma\sim p(\sigma)}\left[\int q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})q(\tilde{\mathbf{x}})\left(\frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma}\right)\mathrm{d}\mathbf{x}\right] \\
\iff \mathbf{s}(\tilde{\mathbf{x}}) &= \sqrt{\boldsymbol{\Sigma}^{-1}}\mathbb{E}_{\sigma\sim p(\sigma)}\left[\int q_\sigma(\tilde{\mathbf{x}} \mid \mathbf{x})\left(\frac{\tilde{\mathbf{x}} - \mathbf{x}}{\sigma}\right)\mathrm{d}\mathbf{x}\right] \\
\iff \mathbf{s}(\tilde{\mathbf{x}}) &= \sqrt{\boldsymbol{\Sigma}^{-1}}\mathbb{E}_{\sigma\sim p(\sigma)}\left[\frac{\mathbb{E}_{\mathbf{x}\sim q_\sigma(\mathbf{x}|\tilde{\mathbf{x}})}[\mathbf{x}] - \mathbf{x}}{\sigma}\right]
\end{aligned}$$

In the case of a single data point $\mathbf{x}_0$:

$$\mathbf{s}(\tilde{\mathbf{x}}) = \sqrt{\boldsymbol{\Sigma}^{-1}}\frac{\mathbf{x}_0 - \mathbf{x}}{\sigma_H}$$

$$\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \sigma_i) = \frac{1}{\sigma_i}\sqrt{\boldsymbol{\Sigma}^{-1}}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) = \boldsymbol{\Sigma}^{-1}\frac{\mathbf{x}_0 - \tilde{\mathbf{x}}}{\sigma_i\sigma_H} \neq \boldsymbol{\Sigma}^{-1}\frac{\mathbf{x}_0 - \tilde{\mathbf{x}}}{\sigma_i^2} \tag{107}$$

Hence, we correct for it while sampling:

$$\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}, \sigma_i) = \frac{\sigma_H}{\sigma_i^2}\sqrt{\boldsymbol{\Sigma}^{-1}}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}) = \frac{\sigma_H\sqrt{\mathrm{Var}[\boldsymbol{\Sigma}_N]}}{\sigma_i^2}\mathrm{Real}(\mathbf{W}_N^{-1}\mathbf{K}\mathbf{W}_N\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}})) \tag{108}$$

E.12 beta for CAS

$$\begin{aligned}
&\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta\sigma_t^2 s^*(\mathbf{x}_t, \sigma_t) + \sigma_{t+1}\beta\mathbf{g} \qquad (\eta = \epsilon/\sigma_L^2) \\
\implies &\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta\boldsymbol{\Sigma}^{-1}(EDS - \mathbf{x}_t) + \sigma_{t+1}\beta\mathbf{g} \\
\implies &\mathbf{x}_{t+1} \leftarrow (\mathbf{I} - \eta\boldsymbol{\Sigma}^{-1})\mathbf{x}_t + \eta\boldsymbol{\Sigma}^{-1}EDS + \sigma_{t+1}\beta\mathbf{g}
\end{aligned}$$

Noise component is:

$$(\mathbf{I} - \eta\boldsymbol{\Sigma}^{-1})\sigma_t\mathbf{g}_t + \sigma_{t+1}\beta\mathbf{g}$$

Variance of noise component is:

$$\sigma_t^2(\mathbf{I} - \eta\boldsymbol{\Sigma}^{-1})\mathrm{Var}[\mathbf{g}_t](\mathbf{I} - \eta\boldsymbol{\Sigma}^{-1})^T + \sigma_{t+1}^2\beta^2\mathrm{Var}[\mathbf{g}]$$

Each diagonal element of Variance of noise component is:

$$\begin{aligned}
&\sigma_t^2(1-\eta)^2 + \sigma_{t+1}^2\beta^2 = \sigma_{t+1}^2\left(\gamma^2(1-\eta)^2 + \beta^2\right) = \sigma_{t+1}^2 \\
\implies &\beta = \sqrt{1 - \gamma^2(1-\eta)^2} \qquad \text{where } \gamma = \sigma_t/\sigma_{t+1}
\end{aligned}$$

In this case, the signal-to-noise ratio is:

$$\mathbb{E}\left[\left\|\frac{\eta\sigma_t^2\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \sigma_t)}{\sigma_{t+1}\beta\mathbf{g}}\right\|_2\right] = \frac{\eta\gamma\sigma_t}{\beta}\frac{\mathbb{E}\left[\|\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \sigma_t)\|_2\right]}{\mathbb{E}\left[\|\mathbf{g}\|_2\right]} \approx \frac{\eta\gamma}{\beta}$$