# Generative Diffusion Models (7): Optimal Diffusion Variance Estimation (Part 1)

Su Jianlin

August 12, 2022

For generative diffusion models, a critical question is how to choose the variance of the generation process, as different variances significantly affect the quality of the generated results.

In "Generative Diffusion Models (2): DDPM = Autoregressive VAE", we mentioned that DDPM derived two usable results by assuming the data follows two special distributions. DDIM in "Generative Diffusion Models (4): DDIM = High-level DDPM" adjusted the generation process, treating the variance as a hyperparameter and even allowing zero-variance generation. However, the generation quality of DDIM with zero variance is generally worse than that of DDPM with non-zero variance. Furthermore, "Generative Diffusion Models (5): SDE Perspective" showed that the variances of the forward and reverse SDEs should be consistent, but this principle theoretically holds only as $\Delta t \to 0$. "Improved Denoising Diffusion Probabilistic Models" proposed treating it as a trainable parameter, though this increases training difficulty.

So, how should the variance of the generation process actually be set? Two papers from this year, "Analytic-DPM: an Analytic Estimate of the Optimal Reverse Variance in Diffusion Probabilistic Models" and "Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models", provide a relatively perfect answer to this question. Let's explore their results together.

## 1 Uncertainty

In fact, these two papers come from the same team with largely the same authors. The first paper (referred to as Analytic-DPM) derives an analytical solution for the unconditional variance based on DDIM. The second paper (referred to as Extended-Analytic-DPM) relaxes the assumptions of the first and proposes an optimization method for conditional variance. This article first introduces the results of the first paper.

In "Generative Diffusion Models (4): DDIM = High-level DDPM", we derived that for a given $p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \bar{\alpha}_t \boldsymbol{x}_0, \bar{\beta}_t^2 \boldsymbol{I})$, the general solution for the corresponding $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$ is:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}\boldsymbol{x}_t + \gamma_t \boldsymbol{x}_0, \sigma_t^2 \boldsymbol{I}\right) \tag{1}$$

where $\gamma_t = \bar{\alpha}_{t-1} - \frac{\bar{\alpha}_t \sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t}$, and $\sigma_t$ is the adjustable standard deviation parameter. In DDIM, the subsequent process is to use $\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ to estimate $\boldsymbol{x}_0$, and then assume:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \approx p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0 = \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)) \tag{2}$$

However, from a Bayesian perspective, this treatment is quite improper because predicting $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$ cannot be perfectly accurate; it carries a certain degree of uncertainty. Therefore, we

should use a probability distribution rather than a deterministic function to describe it. In fact, strictly speaking:

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \int p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) p(\boldsymbol{x}_0|\boldsymbol{x}_t) d\boldsymbol{x}_0 \tag{3}$$

The exact $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$ is usually unobtainable, but since we only need a rough approximation here, we use a normal distribution $\mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \bar{\sigma}_t^2 \boldsymbol{I})$ to approximate it (we will discuss how to approximate it later). With this approximate distribution, we can write:

$$\begin{aligned}
\boldsymbol{x}_{t-1} &= \frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t} \boldsymbol{x}_t + \gamma_t \boldsymbol{x}_0 + \sigma_t \boldsymbol{\varepsilon}_1 \\
&\approx \frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t} \boldsymbol{x}_t + \gamma_t \big(\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) + \bar{\sigma}_t \boldsymbol{\varepsilon}_2\big) + \sigma_t \boldsymbol{\varepsilon}_1 \\
&= \left(\frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t} \boldsymbol{x}_t + \gamma_t \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\right) + \underbrace{\big(\sigma_t \boldsymbol{\varepsilon}_1 + \gamma_t \bar{\sigma}_t \boldsymbol{\varepsilon}_2\big)}_{\sim \sqrt{\sigma_t^2 + \gamma_t^2 \bar{\sigma}_t^2}\boldsymbol{\varepsilon}}
\end{aligned} \tag{4}$$

where $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. It can be seen that $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ is closer to a normal distribution with mean $\frac{\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}}{\bar{\beta}_t} \boldsymbol{x}_t + \gamma_t \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ and covariance $\left(\sigma_t^2 + \gamma_t^2 \bar{\sigma}_t^2\right) \boldsymbol{I}$. While the mean is consistent with previous results, the variance now includes an additional term $\gamma_t^2 \bar{\sigma}_t^2$. Thus, even if $\sigma_t = 0$, the corresponding variance is not zero. This extra term is the correction for the optimal variance proposed in the first paper.

## 2 Mean Optimization

Now we discuss how to use $\mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \bar{\sigma}_t^2 \boldsymbol{I})$ to approximate the true $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$, which essentially means finding the mean and covariance of $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$.

For the mean $\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$, it depends on $\boldsymbol{x}_t$, so a model is needed to fit it, and training the model requires a loss function. Using:

$$\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}] = \operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{x}} \left[\|\boldsymbol{x} - \boldsymbol{\mu}\|^2\right] \tag{5}$$

we get:

$$\begin{aligned}
\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) &= \mathbb{E}_{\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}[\boldsymbol{x}_0] \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)} \left[\|\boldsymbol{x}_0 - \boldsymbol{\mu}\|^2\right] \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_t \sim p(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)} \left[\|\boldsymbol{x}_0 - \boldsymbol{\mu}(\boldsymbol{x}_t)\|^2\right] \\
&= \operatorname*{argmin}_{\boldsymbol{\mu}(\boldsymbol{x}_t)} \mathbb{E}_{\boldsymbol{x}_0 \sim \tilde{p}(\boldsymbol{x}_0)} \mathbb{E}_{\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\|\boldsymbol{x}_0 - \boldsymbol{\mu}(\boldsymbol{x}_t)\|^2\right]
\end{aligned} \tag{6}$$

This is the loss function used to train $\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$. If we introduce the parameterization as before:

$$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) = \frac{1}{\bar{\alpha}_t} \Big(\boldsymbol{x}_t - \bar{\beta}_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\Big) \tag{7}$$

we obtain the loss function form used in DDPM training: $\Big\|\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, t)\Big\|^2$. The result for mean optimization is consistent with previous work, with no changes.

## 3 Variance Estimation 1

Similarly, by definition, the covariance matrix should be:

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{x}_t) =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))^\top\right] \\ =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[((\boldsymbol{x}_0 - \boldsymbol{\mu}) - (\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}))((\boldsymbol{x}_0 - \boldsymbol{\mu}) - (\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}))^\top\right] \\ =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)(\boldsymbol{x}_0 - \boldsymbol{\mu}_0)^\top\right] - (\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0)(\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0)^\top\end{aligned} \tag{8}$$

where $\boldsymbol{\mu}_0$ can be any constant vector, corresponding to the translation invariance of covariance.

The above estimates the full covariance matrix, but it is not exactly what we want, because we currently want to use $\mathcal{N}(\boldsymbol{x}_0; \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \bar{\sigma}_t^2\boldsymbol{I})$ to approximate $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$. The designed covariance matrix $\bar{\sigma}_t^2\boldsymbol{I}$ has two characteristics:

1. Independent of $\boldsymbol{x}_t$: To eliminate the dependence on $\boldsymbol{x}_t$, we average over all $\boldsymbol{x}_t$, i.e., $\boldsymbol{\Sigma}_t = \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}[\boldsymbol{\Sigma}(\boldsymbol{x}_t)]$;
2. Multiple of the identity matrix: This means we only need to consider the diagonal part and average the diagonal elements, i.e., $\bar{\sigma}_t^2 = \mathrm{Tr}(\boldsymbol{\Sigma}_t)/d$, where $d = \dim(\boldsymbol{x})$.

Thus we have:

$$\begin{aligned}\bar{\sigma}_t^2 =& \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\frac{\|\boldsymbol{x}_0 - \boldsymbol{\mu}_0\|^2}{d}\right] - \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\frac{\|\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\|^2}{d}\right] \\ =& \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\|\boldsymbol{x}_0 - \boldsymbol{\mu}_0\|^2\right] - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\|\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\|^2\right] \\ =& \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\left[\|\boldsymbol{x}_0 - \boldsymbol{\mu}_0\|^2\right] - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\|\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\|^2\right]\end{aligned} \tag{9}$$

This is an analytical form for $\bar{\sigma}_t^2$ provided by the author. Once $\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ is trained, the above expression can be approximately calculated by sampling a batch of $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$.

In particular, if we take $\boldsymbol{\mu}_0 = \mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}[\boldsymbol{x}_0]$, then it can be written as:

$$\bar{\sigma}_t^2 = \mathbb{V}ar[\boldsymbol{x}_0] - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\|\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) - \boldsymbol{\mu}_0\|^2\right] \tag{10}$$

Here $\mathbb{V}ar[\boldsymbol{x}_0]$ is the pixel-level variance of all training data $\boldsymbol{x}_0$. If each pixel value of $\boldsymbol{x}_0$ is within the interval $[a, b]$, then its variance obviously will not exceed $\left(\frac{b-a}{2}\right)^2$, leading to the inequality:

$$\bar{\sigma}_t^2 \leq \mathbb{V}ar[\boldsymbol{x}_0] \leq \left(\frac{b-a}{2}\right)^2 \tag{11}$$

## 4 Variance Estimation 2

The previous solution is one that the author considers relatively intuitive. The original Analytic-DPM paper provides a slightly different solution, which the author finds less intuitive. By substituting equation (7), we can obtain:

$$\begin{aligned}\boldsymbol{\Sigma}(\boldsymbol{x}_t) =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))(\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t))^\top\right] \\ =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left(\left(\boldsymbol{x}_0 - \frac{\boldsymbol{x}_t}{\bar{\alpha}_t}\right) + \frac{\bar{\beta}_t}{\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right)\left(\left(\boldsymbol{x}_0 - \frac{\boldsymbol{x}_t}{\bar{\alpha}_t}\right) + \frac{\bar{\beta}_t}{\bar{\alpha}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right)^\top\right] \\ =& \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left(\boldsymbol{x}_0 - \frac{\boldsymbol{x}_t}{\bar{\alpha}_t}\right)\left(\boldsymbol{x}_0 - \frac{\boldsymbol{x}_t}{\bar{\alpha}_t}\right)^\top\right] - \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)^\top \\ =& \frac{1}{\bar{\alpha}_t^2}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0)(\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0)^\top\right] - \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)^\top\end{aligned} \tag{12}$$

Now, if we average both sides over $\boldsymbol{x}_t \sim p(\boldsymbol{x}_t)$, we have:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)^\top\right]\\=&\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)^\top\right]\end{aligned} \tag{13}$$

Recall that $p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \bar{\alpha}_t\boldsymbol{x}_0, \bar{\beta}_t^2\boldsymbol{I})$, so $\bar{\alpha}_t\boldsymbol{x}_0$ is actually the mean of $p(\boldsymbol{x}_t|\boldsymbol{x}_0)$. Thus, $\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)^\top\right]$ is simply the covariance matrix of $p(\boldsymbol{x}_t|\boldsymbol{x}_0)$, which is $\bar{\beta}_t^2\boldsymbol{I}$. Therefore:

$$\mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)\left(\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0\right)^\top\right] = \mathbb{E}_{\boldsymbol{x}_0\sim \tilde{p}(\boldsymbol{x}_0)}\left[\bar{\beta}_t^2\boldsymbol{I}\right] = \bar{\beta}_t^2\boldsymbol{I} \tag{14}$$

Then:

$$\boldsymbol{\Sigma}_t = \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}[\boldsymbol{\Sigma}(\boldsymbol{x}_t)] = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(\boldsymbol{I} - \mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)^\top\right]\right) \tag{15}$$

Taking the trace of both sides and dividing by $d$, we get:

$$\bar{\sigma}_t^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left(1 - \frac{1}{d}\mathbb{E}_{\boldsymbol{x}_t\sim p(\boldsymbol{x}_t)}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t,t)\|^2\right]\right) \le \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2} \tag{16}$$

This provides another estimate and upper bound, which is the original result of Analytic-DPM.

# 5 Experimental Results

The experimental results in the original paper show that the variance correction made by Analytic-DPM mainly provides a significant improvement when the number of diffusion steps is small. Therefore, it is quite meaningful for accelerating diffusion models:

| Model \ # timesteps $K$ | 10 | 25 | 50 | 100 | 200 | 400 | 1000 |
|---|---|---|---|---|---|---|---|
| CIFAR10 (LS) | | | | | | | |
| DDPM, $\sigma_n^2 = \tilde{\beta}_n$ | 44.45 | 21.83 | 15.21 | 10.94 | 8.23 | 6.43 | 5.11 |
| DDPM, $\sigma_n^2 = \beta_n$ | 233.41 | 125.05 | 66.28 | 31.36 | 12.96 | 4.86 | †**3.04** |
| Analytic-DDPM | **34.26** | **11.60** | **7.25** | **5.40** | **4.01** | **3.62** | 4.03 |
| DDIM | 21.31 | 10.70 | 7.74 | 6.08 | 5.07 | 4.61 | 4.13 |
| Analytic-DDIM | **14.00** | **5.81** | **4.04** | **3.55** | **3.39** | **3.50** | **3.74** |
| CIFAR10 (CS) | | | | | | | |
| DDPM, $\sigma_n^2 = \tilde{\beta}_n$ | 34.76 | 16.18 | 11.11 | 8.38 | 6.66 | 5.65 | 4.92 |
| DDPM, $\sigma_n^2 = \beta_n$ | 205.31 | 84.71 | 37.35 | 14.81 | 5.74 | **3.40** | **3.34** |
| Analytic-DDPM | **22.94** | **8.50** | **5.50** | **4.45** | **4.04** | 3.96 | 4.31 |
| DDIM | 34.34 | 16.68 | 10.48 | 7.94 | 6.69 | 5.78 | 4.89 |
| Analytic-DDIM | **26.43** | **9.96** | **6.02** | **4.88** | **4.92** | **5.00** | **4.66** |
| CelebA 64x64 | | | | | | | |
| DDPM, $\sigma_n^2 = \tilde{\beta}_n$ | 36.69 | 24.46 | 18.96 | 14.31 | 10.48 | 8.09 | 5.95 |
| DDPM, $\sigma_n^2 = \beta_n$ | 294.79 | 115.69 | 53.39 | 25.65 | 9.72 | **3.95** | **3.16** |
| Analytic-DDPM | **28.99** | **16.01** | **11.23** | **8.08** | **6.51** | 5.87 | 5.21 |
| DDIM | 20.54 | 13.45 | 9.33 | 6.60 | 4.96 | 4.15 | 3.40 |
| Analytic-DDIM | **15.62** | **9.22** | **6.13** | **4.29** | **3.46** | **3.38** | **3.13** |

Figure 1: Analytic-DPM shows significant improvement mainly when the number of diffusion steps is small.

The author also tried the Analytic-DPM correction on previously implemented code. The reference code is:

**Github: https://github.com/bojone/Keras-DDPM/blob/main/adpm.py**

Figure 2: DDPM generation results with 10 diffusion steps

Figure 3: Analytic-DDPM generation results with 10 diffusion steps

When the number of diffusion steps is 10, the comparison between DDPM and Analytic-DDPM is shown below:

As can be seen, when the number of diffusion steps is small, the generation results of DDPM are quite smooth, giving a sense of "heavy skin smoothing." In contrast, the results of Analytic-DDPM appear more realistic, although they also bring additional noise. In terms of evaluation metrics, Analytic-DDPM performs better.

## 6 Nitpicking

So far, we have completed the introduction to Analytic-DPM. The derivation process involves some techniques but is not overly complex; at least the logic is clear. If readers find it difficult to understand, they might want to look at the derivation in the original paper's appendix, which spans 7 pages and 13 lemmas. Seeing that might make the derivation in this article seem much more friendly!

Admittedly, I admire the authors' insight in obtaining this analytical solution for variance. However, from a hindsight perspective, Analytic-DPM took some "detours" in its derivation and results, making it feel "too convoluted" and "too clever," thus lacking some heuristic value. One of the most obvious characteristics is that the original paper's results are all expressed using $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$. This brings three problems: first, it makes the derivation process particularly unintuitive, making it hard to understand "how they thought of it"; second, it requires readers to have additional knowledge of score matching results, increasing the difficulty of understanding; finally, in practice, $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ must be converted back to $\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ or $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, adding an extra step.

The starting point of the derivation in this article is that we are estimating the parameters of a normal distribution. For a normal distribution, moment estimation is the same as maximum likelihood estimation, so we can directly estimate the corresponding mean and variance. In terms of results, there is no need to force the form to align with $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ or score matching,

because the baseline model for Analytic-DPM is DDIM, and DDIM itself does not start from score matching. Adding a connection to score matching is beneficial neither to theory nor to experiments. Directly aligning with $\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$ or $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is more intuitive in form and easier to convert into experimental forms.

## 7 Summary

This article shared the optimal variance estimation results for diffusion models from the Analytic-DPM paper. it provides a directly usable analytical formula for optimal variance estimation, allowing us to improve generation quality without retraining. The author simplified the derivation of the original paper using his own logic and conducted simple experimental verification.