Time discretization Bounds for DDPM

References:

1. G. Reeves and H. Pfister, Information-theoretic proofs for diffusion sampling, arXiv: 2502.02305, 2025.
2. S. Chen, S. Chewi, J. Li, Y. Li, A. Salim and A. Zhang, Sampling is as easy as learning the score: theory for diffusion models with minimum data assumptions, arXiv 2209.11215, 2022.
3. H. Lee, J. Lu and Y. Tan, Convergence for score based generative modeling with polynomial complexity, arXiv: 2206.06227, 2022.
4. J. Benton, V. de Bortoli, A. Doucet and G. Deligiannidis, Nearly d-linear convergence bounds for diffusion models via stochastic localization, arXiv 2308.03686, 2023.
5. D. Guo, S. Shamai and S. Verdu, Mutual Information and minimum mean-square error in Gaussian channels, IEEE Trans. IT, 2007.

(We clsely follow [1] in this note.)

We consider the following dynamic:

$$Z_k = Z_{k-1} + \delta_k f_k(Z_{k-1}) + \sqrt{\delta_k}\,\tilde{W}_k, \quad k=1,2,\dots,n \qquad (1)$$

where $Z_0 = 0$, δ_k, $k=1,2,\dots,n$, $f_k(\cdot): \mathbb{R}^d \to \mathbb{R}^d$

$\tilde{W}_k \sim N(0, I_d)$, $k=1,2,\dots,n$ (iid).

Define the following companion process: (2)

$$Y_k = Y_{k-1} + \delta_k X_0 + \sqrt{\delta_k}\,\hat{W}_k, \quad \text{where}$$

($\hat{W}_k \sim N(0, I_d)$)

$X_0 \sim \mu_x$ (ground truth data distribution), $X_0 \perp\!\!\!\perp \{\hat{W}_k\}$ (independent)

and $Y_0 = 0$, $\hat{W}_k \sim N(0, I_d)$

Rewrite Y_k: Let $t_k = (\delta_1 + \delta_2 + \cdots + \delta_k)$ (3)

Then, $\boxed{Y_k = t_k X_0 + W_k}$, $W_k = \sum_{i=1}^{k} \sqrt{\delta_i}\,\hat{W}_i$

$$Y_1 = \delta_1 X_0 + \sqrt{\delta_1}\,\hat{W}_0 \;; \quad Y_2 = Y_1 + \delta_2 X_0 + \sqrt{\delta_2}\,\hat{W}_2$$
$$= (\delta_1+\delta_2) X_0 + \left(\sqrt{\delta_1}\,\hat{W}_0 + \sqrt{\delta_2}\,\hat{W}_0\right)$$

$$\cdots$$

i.e., $W_k \sim N(0, t_k)$ $\quad \left(\because t_k = \sum_{i=1}^{k} (\sqrt{\delta_i})^2\right)$

Further $\mathrm{Cov}(W_k, W_m) = \min\{t_k, t_m\}\, I_d$

Since $\frac{Y_k}{t_k} = X_0 + \frac{W_k}{t_k}$, ($X_0 \sim \mu_*$)

$$\left(Y_k/t_k\right) \sim \left(\mu_* * N\left(0, \tfrac{1}{t_k} I_d\right)\right) \quad k = 1, 2, \cdots, n$$

Theorem 1 [1]: Let $\Delta_k = KL\left(\mu(Y_1, Y_2, \cdots, Y_n) \,\|\, \mu(Z_1, Z_2, \cdots, Z_n)\right)$

Note: $\mu(Z)$ is the prob. dist. / measure associated with a r.v. Z (overloaded notation to denote both dist./measure)

Assume μ_* has finite second moments, and

$$E\left[\|f_k(Y_{k-1})\|^2\right] < \infty \quad \forall k. \text{ Then:}$$

$$\Delta_n \leq \frac{\delta_{max}}{2} \operatorname{tr}\left(\operatorname{Cov}(X_0)\right)$$

$$+ \sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| f_k(Y_{k-1}) - E[X_0 | Y_{k-1}] \|^2\right],$$

where $\delta_{max} = \max\{\delta_1, \delta_2, \cdots, \delta_k\}$.

Remark: Suppose we have the following:

(a) $$f_k(y_{k-1}) = E\left[X_0 \middle| Y_{k-1} = y_{k-1}\right]$$

Recall that the annealed score is affine in the conditional expectation. Thus, this selection of $f_k(\cdot)$ corresponds to a DDPM algorithm.

(b) Suppose we wish to sample from a terminal state that $\sim \mu_* * N(0, \varepsilon I_d)$ for some small $\varepsilon > 0$. Then, suppose that we choose: $\delta_k = \left(\frac{\varepsilon}{n}\right)$, $\forall\ k = 1, 2, \ldots, n$.

(a) + (b) $\Rightarrow$ $\Delta_n \leq \frac{1}{2n\varepsilon} \operatorname{tr}\left(\operatorname{Cov}(X_0)\right)$

if $X_0 \sim N(\cdot, \Sigma_d)$, then $\operatorname{tr}(\operatorname{Cov}(X_0)) \approx d$

(c) By changing f_k s.t. it returns a random value whose moments match $E[X_0^r \mid Y_{k-1} = y_{k-1}]$, then we can get $O\left(\frac{1}{n^r}\right)$; see [1] for details.

Recall: $Y_k = t_k X_0 + W_k$, $k = 1, 2, \ldots, n$.

$\hookrightarrow \sim N(0, t_k I_d)$

Then, the following surprising result holds:

Lemma 1: (a) $\{Y_k, k=1,2,\dots,n\}$ is a discrete time Markov process.

(b) $X_0 \perp\!\!\!\perp (Y_1, Y_2, \dots, Y_{k-1}) \mid Y_k$

Proof: Consider the difference sequence:

$$V_k = (Y_k - Y_{k-1}) = \delta_k X_0 + \sqrt{\delta_k}\, \hat{W}_k$$

$\hat{W}_k \hookrightarrow N(0, I_d)$

Then,

$$q_{V_1,\dots,V_n | X_0}(v_1,\dots,v_n | x_0) = \prod_{i=1}^{n} \frac{1}{(2\pi\delta_i)^{d/2}} e^{-\frac{1}{2\delta_i}\|v_i - \delta_i x_0\|^2}$$

$q_{V_1,\dots,V_n|X_0}$ ↓ density fn.

$$= \left(\prod_{i=1}^{n} \phi_i(v_i)\right) e^{\left\langle \sum_{i=1}^{k} v_i, x_0 \right\rangle - \frac{t_k}{2}\|x_0\|^2}$$

$\sum_{i=1}^{k} v_i = y_k$

$\longrightarrow$ (4)

where $\phi_i(v_i) = \frac{1}{(2\pi\delta_i)^{d/2}} e^{-\frac{1}{2\delta_i}\|v_i\|^2}$

Aside: Neyman-Fisher Factorization Thm: Let

$f(v_1, \cdots, v_n; x_0)$ be a density fn with parameter x_0. Then $T(\vec{v})$ is a **sufficient statistic** for x_0 if for any prior dist. over x_0,

$$f(x_0 \mid T(v_1, \cdots, v_n)) = f(x_0 \mid T(v_1, \cdots, v_n), v_1, \cdots, v_n)$$

Equivalently, $I(x_0; T(v_1, \cdots, v_n)) = I(x_0; (v_1, \cdots, v_n))$

A statistic $T(\vec{v})$ is sufficient if and only if $\exists$ function $h(\cdot) \geq 0$, $g(\cdot, \cdot) \geq 0$ s.t. $f(\vec{v}; x_0) = h(\vec{v})\, g(x_0, T(\vec{v}))$

$\vec{v} = (v_1, v_2, \cdots v_n)$

Observe that $Y_k = \sum_{i=1}^{k} v_i$ (by telescoping, and $Y_0 = 0$)

is a sufficient statistic for x_0 from $(v_1, \cdots, v_n)$ by Neyman-Fisher Thm. Thus, we have:

$$x_0 \perp\!\!\!\perp (v_1, v_2, \cdots, v_k) \mid Y_k$$

i.e., $x_0 \perp\!\!\!\perp (Y_1, Y_2, \cdots Y_{k-1}) \mid Y_k$ (b).

The above fact immediately leads to (a) (please see Lemma 1 in [1]). We will instead prove the M.P. property using an alternate proof, also in [1].

Let $B_k = \sqrt{\frac{t_{k+1}}{t_k}}\, Y_k - \sqrt{\frac{t_k}{t_{k+1}}}\, Y_{k+1}$ → (4)

$\left(\because Y_k = t_k x_0 + W_k\right)$

$$= \sqrt{\frac{t_{k+1}}{t_k}}\, W_k - \sqrt{\frac{t_k}{t_{k+1}}}\, W_{k+1}$$

Now, $\mathrm{Cov}(W_k, W_m) = \min\{t_k, t_m\}\, I_d$.

$\therefore\ m > k:\ \mathrm{Cov}(B_k, W_m) = \sqrt{\frac{t_{k+1}}{t_k}} \cdot t_k \cdot I_d - \sqrt{\frac{t_k}{t_{k+1}}}\, t_{k+1} \cdot I_d$

$= 0$

$\therefore\ B_k \perp\!\!\!\perp (W_{k+1}, W_{k+2}, \ldots, W_n)$.

$\therefore$ Rewriting (4), we have:

$$Y_k = \frac{t_k}{t_{k+1}} Y_{k+1} + \sqrt{\frac{t_k}{t_{k+1}}} B_k ,$$

with $B_k \sim N(0, s_{k+1} I_d)$.

$\Rightarrow$ Reversed process $\{Y_n, Y_{n-1}, \ldots, Y_1\}$ is Markov (indep. Gaussian increments).

$\Updownarrow$ (Useful fact!)

Forward process $\{Y_1, Y_2, \ldots, Y_n\}$ is Markov. (a)

(Note: The reversed process could be time inhomogeneous even when forward is time homogeneous or vice-versa. However, as long as we show one of these processes is Markovian, the other is also Markovian.).

Lemma 2 Under same assumptions in Thm 1, we have:

$$\Delta_n = \sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| X_0 - E[X_0|Y_{k-1}] \|^2\right]$$

(mean-square estimation error)

$$- I(X_0; Y_n) + \sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| f_k(Y_{k-1}) - E[X_0|Y_{k-1}] \|^2\right]$$

$I(X_0; Y_n)$: mutual information

f_k: can be made to '0' by appropriate choice of f_k.

where $I(X_0; Y_n) = KL\left(\mu(x, y_n) \,\|\, \mu(x)\mu(y_n)\right)$.

Proof:- $\{Y_k\}$ a M.P. with trans. dist. $p_k(\cdot|\cdot)$

$\{Z_k\}$ a M.P. with trans dist. $q_k(\cdot|\cdot)$.

By definition, and Markovian property, we have:

$$\Delta_n = KL\left(\mu(Y_1, \cdots, Y_n) \,\|\, \mu(Z_1, \cdots Z_n)\right)$$

$$= \sum_{k=1}^{n} E\left[\ln \frac{p_k(Y_k|Y_{k-1})}{q_k(Y_k|Y_{k-1})}\right] \longrightarrow (5)$$

We now derive expressions for the numerator and denominator in the expectation above.

Recall from ①: $z_k = z_{k-1} + \delta_k f_k(z_{k-1}) + \sqrt{\delta_k}\,\tilde{w}_k$

$$\therefore \quad q_k(y_k \mid y_{k-1}) = \phi_k\left(y_k - y_{k-1} - \delta_k f_k(y_{k-1})\right)$$

↳ $N(0, \delta_k I_d)$.

$$\therefore \quad -\ln q_k(y_k \mid y_{k-1}) = \frac{d}{2}\ln(2\pi\delta_k) + \frac{\|y_k - y_{k-1} - \delta_k f_k(y_{k-1})\|^2}{2\delta_k}$$

i.e.,
$$-\ln q_k(y_k \mid y_{k-1}) - \underbrace{\frac{d}{2}\ln(2\pi\delta_k) - \frac{d}{2}}_{= -\frac{d}{2}\ln(2\pi e \delta_k)}$$

$$= \frac{1}{2\delta_k}\|y_k - y_{k-1} - \delta_k f_k(y_{k-1})\|^2 - \frac{d}{2}$$

Also recall: $Y_k = Y_{k-1} + \delta_k X_0 + \sqrt{\delta_k}\,\hat{w}_k$ (from ②).

↳ $N(0, I_d)$.

$\therefore$ We have:

$$-E\left[\ln q_k(Y_k \mid Y_{k-1})\right] - \frac{d}{2}\ln(2\pi e \delta_k)$$

$$= \frac{1}{2\delta_k} E\left[\|\underbrace{Y_k - Y_{k-1}}_{\text{use ②}} - \delta_k f_k(Y_{k-1})\|^2\right] - \frac{d}{2}$$

$$= \frac{1}{2\delta_k} E\left[\left\| \delta_k x_0 + \sqrt{\delta_k}\, \hat{W}_k - \delta_k f_k(y_{k-1}) \right\|^2 \right] - \frac{d}{2}$$

$$= \frac{1}{2\delta_k} \Bigg(E\left[\left\| \delta_k x_0 - \delta_k f_k(y_{k-1}) \right\|^2 \right] + \underbrace{E\left[\left\| \sqrt{\delta_k}\, \hat{W}_k \right\|^2 \right]}_{= \delta_k d} + E\left[\left\langle \delta_k x_0 - \delta_k f_k(y_{k-1}),\ \sqrt{\delta_k}\, \hat{W}_k \right\rangle \right] \Bigg)$$

$\hat{W}_k \perp\!\!\!\perp (x_0, y_{k-1})$

$$- d/2.$$

$$= \frac{\delta_k}{2} E\left[\left\| x_0 - f_k(y_{k-1}) \right\|^2 \right]$$

$$= \frac{\delta_k}{2} E\left[\left\| x_0 - E[x_0 | y_{k-1}] \right\|^2 \right]$$

$$+ \frac{\delta_k}{2} E\left[\left\| E[x_0 | y_{k-1}] - f_k(y_{k-1}) \right\|^2 \right]$$

principle of orthogonality: $(x_0 - E[x_0|y_{k-1}]) \perp h(y_{k-1})$ (error vector; any estimate)

for any $h(\cdot)$, because $E[x_0|y_{k-1}]$ is the minimum mean square estimate of x_0 given y_{k-1}.

Denominator term of (5)

So far: $-E\left[\ln q_k(Y_k|Y_{k-1})\right] - \frac{d}{2}\ln(2\pi e\delta_k)$

$= \frac{\delta_k}{2} E\left[\|X_0 - E[X_0|Y_{k-1}]\|^2\right]$ → (6)

$+ \frac{\delta_k}{2} E\left[\| E[X_0|Y_{k-1}] - f_k(Y_{k-1})\|^2\right]$

Conditional differential entropy

Defn: $h(X|Z) = E[\ln p(X|Z)]$

$= \int\int p(x,y)\ln p(x|y)\,dx\,dy$

Numerator term of (5)

Next, recall that given (X_0, Y_{k-1}), Y_k is conditional Gaussian (0 mean, $\delta_k I_d$ covariance).

Thus, $h(Y_k|Y_{k-1}, X_0) = \frac{d}{2}\ln(2\pi e\delta_k)$

We now have:

$E\left[\ln p_k(Y_k|Y_{k-1})\right] + \frac{d}{2}\ln(2\pi e\delta_k)$

$$= -h(Y_k|Y_{k-1}) + h(Y_k|Y_{k-1}, X_0)$$

$I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$

$$= I(X_0; Y_k | Y_{k-1})$$

$$= h(X_0|Y_{k-1}) - h(X_0|Y_{k-1}, Y_k)$$

$Y_{k-1} - Y_k - X_0$ is a M.C from (b), Lemma 2

$$= \left(h(X_0) - h(X_0|Y_{k-1})\right) + \left(h(X_0) - h(X_0|Y_{k-1})\right)$$

$$= I(X_0; Y_k) - I(X_0; Y_{k-1}).$$

In summary, we have:

$$E\left[\ln p_k(Y_k|Y_{k-1})\right] + \frac{d}{2}\ln(2\pi e \delta_k) \rightarrow (6)$$

$$= I(X_0; Y_k) - I(X_0; Y_{k-1}).$$

Now, substitute (5), (6) in (4), telescope and note $I(X_0; Y_0) = 0$ ∎

With these lemmas, we prove Theorem 1. This uses the I-MMSE relationship [5].

Recall note 9a on I-MMSE. Rewriting in integral form:

Suppose $X_0 \sim \mu_X$, $N \sim N(0, I_d)$, $0 \le a < b$:

Then:
$$I(X_0; \sqrt{b}X_0 + N) - I(X_0; \sqrt{a}X_0 + N) = \frac{1}{2}\int_a^b M(s)\,ds$$

where $M(s) = E\left[\left\| X_0 - E[X_0 | \sqrt{s}X_0 + N] \right\|^2\right]$ ↦ (7)

and $M(0) = \mathrm{tr}(\mathrm{Cov}(X_0))$.

Fact: $M(s)$ is monotone non-increasing in s.

Proof of Theorem 1:

Recall from Lemma 2 that:

$$\Delta_n = \sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| X_0 - E[X_0|Y_{k-1}] \|^2\right]$$

$$- I(X_0; Y_n) + \sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| f_k(Y_{k-1}) - E[X_0|Y_{k-1}] \|^2\right]$$

We now show that:

$$\sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| X_0 - E[X_0|Y_{k-1}] \|^2\right] - I(X_0; Y_n)$$

$$\leq \frac{\delta_{max}}{2} \operatorname{tr}\left(\operatorname{Cov}(X_0)\right),$$

which leads to the Theorem statement.

Before we proceed, we note the following property of Mutual Information in our setting:

$$I(X_0; Y_k) = I\left(X_0;\ t_k X_0 + W_k\right)$$

($W_k \mapsto N(0, t_k I_d)$)

(from (3), pg. 3.)

$\sim N(0, I_d)$.

$$= I\left(X_0; \sqrt{t_k}\, X_0 + N\right)$$

(MI is invariant to 1-1 transformation)

Then, we have the following:

$$I(X_0; Y_k) - I(X_0; Y_{k-1}) = \frac{1}{2}\int_{t_{k-1}}^{t_k} M(s)\, ds \quad ,$$

where $t_k = \sum_{i=1}^{k} \delta_i$ $\left(\because t_k - t_{k-1} = \delta_k\right)$, and $M(s)$ is monotone non-increasing. Thus,

$$\frac{\delta_k}{2} M(t_k) \le I(X_0; Y_k) - I(X_0; Y_{k-1}) \le \frac{\delta_k}{2} M(t_{k-1})$$

→ (8)

Comment: The above bound is important; it allows us to control the "error" that is accumulated in each time step.

Here, $M(t_k) = E\left[\left\| X_0 - E[X_0 | Y_k] \right\|^2\right]$.

Thus,

$$\sum_{k=1}^{n} \frac{\delta_k}{2} E\left[\| X_0 - E[X_0|Y_{k-1}] \|^2\right] - I(X_0; Y_n)$$

$$= \sum_{k=1}^{n} \frac{\delta_k}{2} M(t_{k-1}) - I(X_0; Y_n)$$

telescope

$$= \sum_{k=1}^{n} \left(\frac{\delta_k}{2} M(t_{k-1}) + I(X_0; Y_{k-1}) - I(X_0; Y_k) \right)$$

(8)

$$\leq \sum_{k=1}^{n} \frac{\delta_k}{2} \left(M(t_{k-1}) - M(t_k) \right)$$

$$\leq \frac{\delta_{max}}{2} \sum_{k=1}^{n} \left(M(t_{k-1}) - M(t_k) \right)$$

$$= \frac{\delta_{max}}{2} \left(M(0) - M(t_n) \right)$$

$M(0) = tr(Cov(X_0))$

$$\leq \frac{\delta_{max}}{2} tr\left(Cov(X_0)\right)$$

∎

Comments: There are many other results in [1] that are of interest, including

(a) optimization of step sizes (sec 4.1 in [1])
(b) improved rates via moment matching (4.2 in [1])

along with other connections to literature.