A Proof and Derivations

A.1 Proof of Theorem 2.1

The existence is straightforward, since $\mathrm{FD}(\tilde{p}_d||\tilde{q}_{\theta^*}) = 0 \to \tilde{p}_d = \tilde{q}_{\theta^*}$, we can simply let $q(x) = p_d(x)$, which makes $\int q(x)p(\tilde{x}|x)\,\mathrm{d}x = \int p_d(x)p(\tilde{x}|x)\,\mathrm{d}x = \tilde{p}_d$. To show the uniqueness, we denote density $k(\epsilon) = \mathcal{N}(0, \sigma^2 I)$, so $\tilde{q}_\theta(\tilde{x})$ and $\tilde{p}_d(x)$ can be written as convolutions

$$\tilde{q}_\theta(\tilde{x}) = q * k, \quad \tilde{p}_d(\tilde{x}) = p_d * k, \tag{20}$$

we then have

$$\tilde{p}_d = \tilde{q}_\theta \Leftrightarrow q * k = p_d * k \Leftrightarrow \mathcal{F}(q)\mathcal{F}(k) = \mathcal{F}(p_d)\mathcal{F}(k), \tag{21}$$

where $\mathcal{F}$ denotes the Fourier transform. Since the Fourier transform of a Gaussian is also a Gaussian, so $\mathcal{F}(k) > 0$ everywhere, we have

$$\tilde{p}_d = \tilde{q}_{\theta^*} \Leftrightarrow \mathcal{F}(q)\cancel{\mathcal{F}(k)} = \mathcal{F}(p_d)\cancel{\mathcal{F}(k)} \Leftrightarrow \mathcal{F}(q) = \mathcal{F}(p_d) \Leftrightarrow q = p_d. \tag{22}$$

Therefore, $q = p_d$ is the unique distribution that makes $\tilde{p}_d = \tilde{q}_\theta$. This technique has also been used to construct spread KL divergence (we denote as $\widetilde{\mathrm{KL}}$) [46], which is defined as $\widetilde{\mathrm{KL}}(p_d||q_\theta) \equiv \mathrm{KL}(p_d * k||q_\theta * k)$ where $k(\epsilon) = N(0, \sigma^2 I)$, to train implicit model q_θ. Different from the DSM situation, when $\widetilde{\mathrm{KL}}(p_d||q_\theta) = 0$, the underlying model $q_\theta = p_d$ is directly available, whereas the EBM $\tilde{q}_\theta$ trained by DSM learns to be the noisy distribution $\tilde{q}_\theta = p_d * k$.

A.2 General Conditions Characterising the Existence of the Clean Model

In the previous section, we assume for a flexible neural network parameterized f_θ, the energy-based model $\tilde{q}_\theta(\tilde{x}) = \exp(-f(\tilde{x}))/Z(\theta)$ trained by Equation 5 can recover the target noisy data distribution $\tilde{q}_{\theta^*} = \tilde{p}_d$ so there exists an underlying model q such that $\tilde{q}_{\theta^*} = q * k$ and $q = p_d$. This assumption is commonly used in the literature on score-based methods. For example, in the score-based diffusion models literature [32, 13, 2], for any data $x \in \mathbb{R}^D$, the score function $\nabla_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x})$ is usually parameterized by a neural network $\mathrm{NN}_\theta(\cdot) : \mathbb{R}^D \to \mathbb{R}^D$. However, this parameterization cannot guarantee $\mathrm{NN}_\theta(\tilde{x})$ is a conservative vector field, or in other words, there doesn't exist a distribution $\tilde{q}_\theta(\tilde{x})$ such that $\nabla_{\tilde{x}}\tilde{q}_\theta(\tilde{x}) = \nabla_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x})$ and $\nabla^2_{\tilde{x}} \log \tilde{q}(\tilde{x})$ is symmetric [29, 30]. Therefore, perfect score estimation $\nabla_{\tilde{x}} \log \tilde{p}_d(\tilde{x}) = \nabla_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x})$ is implicitly assumed to allow an EBM interpretation.

However, the underlying clean model doesn't always exist for imperfect model $\tilde{q}_\theta \neq \tilde{p}_d$. We here provide the sufficient and necessary conditions which guarantee the existence of the underlying clean model.

Theorem A.1 (Necessary and Sufficient conditions for the existence of the underlying clean model.)**.** *For a model $\tilde{q}_\theta$ with the convolutional noise distribution $k(\epsilon) = \mathcal{N}(0, \sigma^2 I)$, there exists an underlying model q such that $q * k = \tilde{q}$ if and only if $\mathcal{F}(\tilde{q}_\theta)/\mathcal{F}(k)$ is positive semi-definite* [7]*. Additionally, the underlying distribution q can be written as*

$$q = \mathcal{F}^{-1}(\mathcal{F}(\tilde{q}_\theta)/\mathcal{F}(k)), \tag{23}$$

where $\mathcal{F}^{-1}$ is the inverse Fourier transform. This theorem is a straightforward corollary of Bochner's Theorem [8]. However, for the energy model $\tilde{q}_\theta(\tilde{x}) \propto \exp(-f_\theta(\tilde{x}))$, it's difficult to design a functioning family of f that satisfies the positive semi-definite condition and have the tractable score function at the same time [9]. We thus leave the design of better energy function parameterizations as a promising future direction.

[7] A continuous function $f : \mathbb{R}^d \to \mathbb{C}$ is positive semi-definite if for all $n \in \mathbb{N}$, all sets of pairwise distinct centers $X = \{x_1, ..., x_N\} \in \mathbb{R}^d$ and all $\alpha \in \mathbb{C}^N$, $\sum_{i=1}^N \sum_{j=1}^N \alpha_i \overline{\alpha_j} f(x_i - x_j) \geq 0$, see [41, Definition 6.1]

[8] Bochner's Theorem [41, Theorem 6.6]: A continuous function $f : \mathbb{R}^d \to \mathbb{C}$ is positive semi-definite if and only if it is the Fourier transform of a finite non-negative Borel measure on $\mathbb{R}^d$.

[9] For example, one can define a noisy energy-based model $\tilde{q}_\theta = \exp(-f_\theta(\tilde{x}))/Z(\theta)$ with $-f_\theta(\tilde{x}) = \int (g_\theta(x) + 1/\sigma^2||\tilde{x} - x||_2^2)\,\mathrm{d}x$, which always allows an underlying clean energy-based model $q_\theta(x) = \exp(-g_\theta(x))/Z(\theta)$ such that $\tilde{q}_\theta(\tilde{x}) = q_\theta(x) * k$ with $k(\epsilon) = \mathcal{N}(0, \sigma^2 I)$. However, the score function $\nabla_{\tilde{x}} \log \tilde{q}(\tilde{x}) = -\nabla_{\tilde{x}} f_\theta(\tilde{x})$ is intractable in this case.

A.3 Proof of Theorem 2.2

Derivation of the Mean Identity

We let $\tilde{q}_\theta(\tilde{x}) = \int k(\tilde{x}|x)q_\theta(x)\,\mathrm{d}\tilde{x}$, where $k(\tilde{x}|x) = \mathcal{N}(0, \sigma^2 I)$, we have

$$\begin{aligned}\nabla_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) = \frac{\nabla_{\tilde{x}}\tilde{q}_\theta(\tilde{x})}{\tilde{q}_\theta(\tilde{x})} &= \frac{\int \nabla_{\tilde{x}} k(\tilde{x}|x)q_\theta(x)\,\mathrm{d}x}{q_\theta(x)} \\ &= -\frac{1}{\sigma^2}\int \left((\tilde{x}-x)\frac{k(\tilde{x}|x)q_\theta(x)}{\tilde{q}_\theta(\tilde{x})}\right)\mathrm{d}x \\ \Longrightarrow \sigma^2 \nabla_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) + \tilde{x} &= \int x \frac{k(\tilde{x}|x)q_\theta(x)}{\tilde{q}_\theta(\tilde{x})}\,\mathrm{d}x = \langle x \rangle_{q_\theta(x|\tilde{x})}\end{aligned}$$

where we define the model denoising posterior using Bayes rule $q_\theta(x|\tilde{x}) \equiv k(\tilde{x}|x)q_\theta(x)/\tilde{q}_\theta(\tilde{x})$. The second equality is due to the following Gaussian distribution property

$$\nabla_{\tilde{x}} k(\tilde{x}|x) = \frac{1}{\sqrt{2\pi\sigma^2}}\nabla_{\tilde{x}} e^{\frac{-(\tilde{x}-x)^2}{2\sigma^2}} = -\frac{\tilde{x}-x}{\sigma^2}\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(\tilde{x}-x)^2}{2\sigma^2}} = -\frac{\tilde{x}-x}{\sigma^2}k(\tilde{x}|x). \tag{24}$$

Derivations of the Analytical Full Covariance Identity

We have derived the mean identity

$$\mu_q(\tilde{x}) \equiv \langle x \rangle_{q_\theta(x|\tilde{x})} = \sigma^2 \nabla_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) + \tilde{x}. \tag{25}$$

Taking the gradient over x in both side and scaling with σ^2, we have

$$\sigma^2 \nabla_x \mu_q(\tilde{x}) = \sigma^4 \nabla^2_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) + \sigma^2 I. \tag{26}$$

We can also expand the hessian of the $\log \tilde{q}_\theta(\tilde{x})$:

$$\begin{aligned}\nabla^2_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) &= -\frac{1}{\sigma^2}\int \nabla_{\tilde{x}}\left((\tilde{x}-x)\frac{k(\tilde{x}|x)p_\theta(x)}{\tilde{p}_\theta(\tilde{x})}\right)\mathrm{d}x \\ = -\frac{1}{\sigma^2}\int \frac{k(\tilde{x}|x)p_\theta(x)}{\tilde{p}_\theta(\tilde{x})}\,\mathrm{d}x &+ \frac{1}{\sigma^2}\int (\tilde{x}-x)\frac{\nabla_{\tilde{x}} k(\tilde{x}|x)\tilde{p}_\theta(\tilde{x})p_\theta(x) - \nabla \tilde{p}_\theta(\tilde{x})k(\tilde{x}|x)p_\theta(x)}{\tilde{p}^2_\theta(\tilde{x})} \\ \Longrightarrow \sigma^2 \nabla^2_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) + 1 &= \int (\tilde{x}-x)\frac{\nabla_{\tilde{x}} k(\tilde{x}|x)p_\theta(x) - \nabla \log \tilde{q}_\theta(\tilde{x})k(\tilde{x}|x)p_\theta(x)}{\tilde{q}_\theta(\tilde{x})}\,\mathrm{d}x \\ &= \int (\tilde{x}-x)\frac{-\frac{1}{\sigma^2}(\tilde{x}-x)k(\tilde{x}|x)p_\theta(x) + \frac{1}{\sigma^2}(\tilde{x} - \langle x \rangle_{p_\theta(x|\tilde{x})})k(\tilde{x}|x)p_\theta(x)}{\tilde{q}_\theta(\tilde{x})}\,\mathrm{d}x \\ \Longrightarrow \sigma^4 \nabla^2_{\tilde{x}} \log \tilde{q}_\theta(\tilde{x}) + \sigma^2 I &= \int \left(-(\tilde{x}-x)^2 + (\tilde{x}-x)(\tilde{x} - \langle x \rangle_{p_\theta(x|\tilde{x})})\right) p_\theta(x|\tilde{x})\,\mathrm{d}x \\ &= \langle x^2 \rangle_{p_\theta(x|\tilde{x})} - \langle x \rangle^2_{p_\theta(x|\tilde{x})} \equiv \Sigma_q(\tilde{x})\end{aligned}$$

Therefore, we obtain the analytical full covariance identity.

$$\Sigma_q(\tilde{x}) = \sigma^2 \nabla_{\tilde{x}} \mu_q(\tilde{x}). \tag{27}$$

A.4 Proof of Theorem 2.3

Lemma A.2 (KL to Gaussian [2]). *Let $p(x)$ be a distribution with mean μ_p and covariance Σ_p and $q(x) = \mathcal{N}(\mu_q, \Sigma_q)$, denote the differential entropy as $\mathrm{H}(p) \equiv -\int p(x)\log p(x)\,\mathrm{d}x$, we have*

$$\mathrm{KL}(p||q) = \mathrm{KL}(\mathcal{N}(\mu_p, \Sigma_p)||q) + \mathrm{H}(\mathcal{N}(\mu_p, \Sigma_p)) - \mathrm{H}(p) \tag{28}$$

The proof can be found in [2] Lemma 2.

We can then prove Theorem 2.3. Since $p(\tilde{x}|x)p_d(x) = p(x|\tilde{x})\tilde{p}_d(\tilde{x})$, where $\tilde{p}_d(\tilde{x}) = \int p_d(x)p(\tilde{x}|x)\,\mathrm{d}x$, we have

$$\mathrm{KL}(p(\tilde{x}|x)p_d(x)\|q(x|\tilde{x})\tilde{p}_d(\tilde{x})) = \langle \mathrm{KL}(p(x|\tilde{x})||q(x|\tilde{x}))\rangle_{\tilde{p}(\tilde{x})} \tag{29}$$

Assume Gaussian distribution $q(x|\tilde{x}) = \mathcal{N}(\mu_q(\tilde{x}), \Sigma_q(\tilde{x}))$and denote the mean and covariance of the true posterior are $\mu_p(\tilde{x})$ and $\Sigma_p(\tilde{x})$, then the optimal q^* is

$$\begin{aligned} q^* &= \arg\min_q \mathrm{KL}(p(\tilde{x}|x)p_d(x)\|q(x|\tilde{x})\tilde{p}_d(\tilde{x})) & (30) \\ &= \arg\min_q \left\langle \mathrm{KL}(p(x|\tilde{x})||q(x|\tilde{x})) \right\rangle_{\tilde{p}(\tilde{x})} & (31) \\ &= \arg\min_q \left\langle \mathrm{KL}(\mathcal{N}(\mu_p, \Sigma_p)||q(x|\tilde{x})) + \mathrm{H}(\mathcal{N}(\mu_p, \Sigma_p)) - \mathrm{H}(p(x|\tilde{x})) \right\rangle_{\tilde{p}(\tilde{x})} & (32) \\ &= \arg\min_q \left\langle \mathrm{KL}(\mathcal{N}(\mu_p, \Sigma_p)||q(x|\tilde{x})) \right\rangle_{\tilde{p}(\tilde{x})} + const.. & (33) \end{aligned}$$

Therefore, the optimal $q(x|\tilde{x}) = \mathcal{N}(\mu_q(\tilde{x}), \Sigma_q(\tilde{x}))$ under the joint KL has the mean and covariance $\mu_q^*(\tilde{x}) = \mu_p(\tilde{x}), \Sigma_q^*(\tilde{x})) = \Sigma_p(\tilde{x})$.

B Connection to Analytical DDPM

Paper [2] considers the constrained variational family $q_\theta(x|\tilde{x}) = \mathcal{N}(\mu_\theta(\tilde{x}), \sigma_q^2 I)$ and derive the optimal σ_q^* as

$$\sigma_q^{*2} = \arg\min_{\sigma_q} \mathrm{KL}(p(\tilde{x}|x)p_d(x)\|q_\theta(x|y))\tilde{p}_d(\tilde{x})) = \frac{1}{d}\left\langle \mathrm{Tr}\left(\mathrm{Cov}_{q(x|\tilde{x})}[x]\right)\right\rangle_{\tilde{p}_d(\tilde{x})}, \tag{34}$$

which can also be rewritten using the score function

$$\sigma_q^{*2} = \sigma^2 - \frac{\sigma^4}{d}\left\langle \|s_{q_\theta}(\tilde{x})\|_2^2 \right\rangle_{\tilde{p}_d(\tilde{x})}. \tag{35}$$

To make a deep connection, we can also plug our analytical full covariance (Equation 11) into Equation 17

$$\begin{aligned} \sigma_q^{*2} &= \sigma^2 + \frac{\sigma^4}{d}\mathrm{Tr}\left\langle \nabla_x^2 \log q_\theta(\tilde{x}) \right\rangle_{\tilde{p}_d(\tilde{x})} \\ &= \sigma^2 - \frac{\sigma^4}{d}\mathrm{Tr}\left\langle s_{q_\theta}(\tilde{x}) s_{q_\theta}(\tilde{x})^T \right\rangle_{\tilde{p}_d(\tilde{x})} = \sigma^2 - \frac{\sigma^4}{d}\left\langle \|s_{q_\theta}(\tilde{x})\|_2^2 \right\rangle_{\tilde{p}_d(\tilde{x})}, & (36) \end{aligned}$$

which recovers Equation 18, where the first equality is due to the well-known Fisher information identity [9].

C Experiments

All the experiments conducted in this paper are run on one single NVDIA GTX 3090.

C.1 Effect of the Single Noise Choice on MNIST

Figure 10 shows the samples generated by our method with the EBM trained with difference $\sigma \in \{0.3, 0.5, 0.8\}$ in the noise distribution $p(\tilde{x}|x)$, we can find the image quality also heavily depends on the choice of the noise scale and $\sigma = 0.5$ achieves the best visual quality, we then use this hyper-parameter in the subsequent comparisons.

C.2 Multi-level Noise Details

For full details on the architecture and noise schedule used in the multi-level noise experiments in Section 5, we refer to Appendix B of [33]. For our multi-level Gibbs sampling procedure, we used 3 Gibbs steps at each noise level and 3 Rademacher samples for each diagonal Hessian computation. Following [33], we used a total of 232 noise levels, distributed according to their proposed geometric schedule, and applied a final denoising step in which the mean of the clean distribution conditioned on the final output of the sampling procedure is returned (the final output of the sampling procedure is a sample from the noised distribution from the noise distribution at the smallest noise level). This denoising step was previously found to improve FID scores [16] significantly.

(a) $\sigma = 0.3$ (b) $\sigma = 0.5$ (c) $\sigma = 0.8$

Figure 10: Sample comparisons with different σ value.

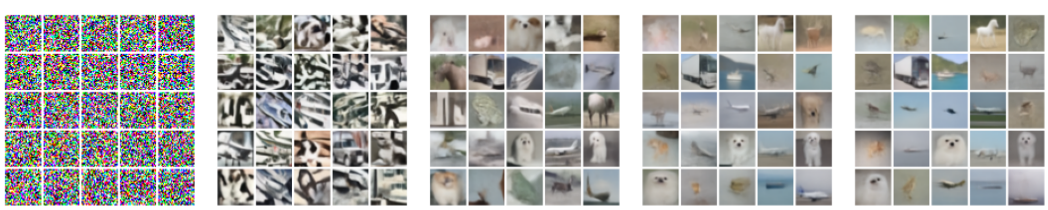

Figure 11: Mode Collapse visualization of 25 Markov chains, we plot the samples every 20 Gibbs steps, we can find less modes are covered if we run the Gibbs sampling for a longer time.