ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446376

MOMENTUM-IMBUED LANGEVIN DYNAMICS (MILD) FOR FASTER SAMPLING

Nishanth Shetty[1†], *Manikanta Bandla*[2†], *Nishit Neema*[3†],
Siddarth Asokan[4], *Chandra Sekhar Seelamantula*[5]

[1,5] Department of Electrical Engineering, Indian Institute of Science, Bangalore
[2,3] Department of Computer Science and Automation, Indian Institute of Science, Bangalore
[4] Microsoft Research Lab India (MSRI), Bangalore
{*nishanths, manikantab, nishitneema, css*}*@iisc.ac.in*, *sasokan@microsoft.com*

ABSTRACT

Score-based generative models have emerged as the state-of-the-art in generative modeling. In this paper, we introduce a novel sampling scheme that can be combined with pre-trained score-based diffusion models to speed up sampling by a factor of two to five in terms of the number of function evaluations (NFEs) with a superior Fréchet Inception distance (FID), compared to Annealed Langevin dynamics in noise-conditional score network (NCSN) and improved noise-conditional score network (NCSN++). The proposed sampling algorithm is inspired by momentum-based accelerated gradient descent used in convex optimization techniques. We validate the sampling efficiency of the proposed algorithm in terms of FID on CIFAR-10 and CelebA datasets.

***Index Terms*—** Score-based models, generative AI, deep generative models, diffusion models, momentum.

1. INTRODUCTION

Recently, there has been a surge of interest in generative models within the field of deep learning and artificial intelligence. Generative models are trained to output samples from an unknown, underlying distribution of a dataset, given access to a finite number of samples. Generative models have demonstrated remarkable capabilities in a variety of applications, such as image synthesis, text generation, audio synthesis, music synthesis, image reconstruction and data augmentation.

Song and Ermon [1] introduced *score-based generative models*, wherein the goal is to learn the gradient of the *score*, which is the log-probability density function (p.d.f.) of the data. Since one does not have access to the p.d.f. in closed-form, the score is typically approximated using deep neural networks. Score-based models have achieved state-of-the-art performance on tasks such as image generation [1–6], audio synthesis [7–9], music generation [10] and image restoration [11–13] among others. In score-based models, unlike variational autoencoders (VAEs) [14] and generative adversarial networks (GANs) [15], the sampling process is iterative and is implemented using Langevin dynamics [16]. While these models achieve state-of-the-art Fréchet Inception distance (FID) [17], a major shortcoming is the relatively longer sampling time.

1.1. Related Work

Song et al. [2] improved training and sampling by tuning the noise-scales and improved sampling stability via an exponential moving-average for the weights. Song et al. [3] proposed a unified framework for analyzing diffusion models and score-based models using stochastic differential equations (SDEs). Jolicoeur-Martineau et al. [18] proposed an adaptive stochastic differential equation (SDE) solver as a fast sampler without any drop in the quality of generated images. Their approach reduces computational requirement by a factor of $2\times$ to $5\times$. For a fixed number of function evaluations (NFEs), their sampler results in a lower FID for the generated images. Ma et al. [19] proposed a matrix preconditioning method to accelerate sampling motivated by the intuition to make the rate of curvature similar along all the directions. They employ the Metropolis adjusted Langevin sampling algorithm [20] and achieve a $29\times$ speedup in sampling for high-resolution images. Denoising diffusion implicit models (DDIMs) [21] proposed a generalization involving non-Markovian diffusion processes, achieving a $10\times$ to $50\times$ speedup over Denoising Diffusion Probabilistic Models [4]. Consistency models [22] are a novel class of generative models that support fast one-step and multi-step sampling for generation of high quality samples, and are typically trained by distilling pre-trained diffusion models. They outperform existing non-adversarial generative models on CIFAR-10 [23], ImageNet [24] and LSUN [25] datasets. Karras et al. [26] proposed a unified framework for implementing all score-based generative models by focusing on the design space of generative models.

† denotes equal contribution.

Siddarth Asokan and Nishanth Shetty are supported by Qualcomm Innovation Fellowship 2023. Nishanth Shetty is also supported by the Prime Minister's Research Fellowship.

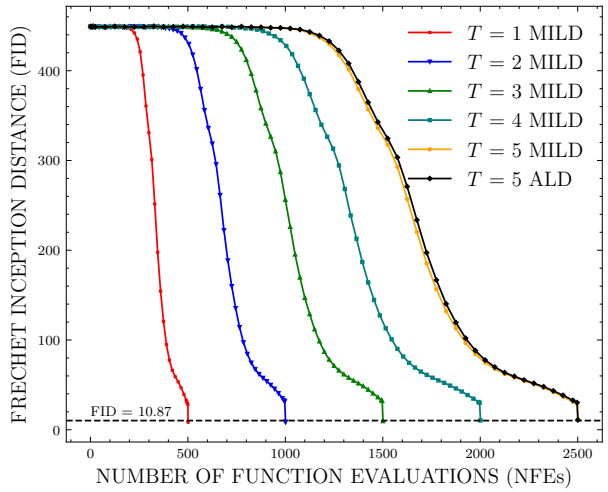

Fig. 1: (Colour online) A comparison of FID versus the number of function evaluations (NFEs) for the proposed MILD vis-à-vis the baseline ALD samplers, for varying number of sampling steps T. MILD sampler with $T = 1$ reduces the NFEs required to achieve a given FID by $2\times$ to $5\times$.

This allows for a separation of training and sampling techniques to optimize score-based generative models. Recent works that tackle the problem of optimizing the sampler for diffusion models [21, 27–31] are motivated by the theory of numerical methods applied to stochastic differential equations (SDEs).

1.2. Our Contributions

In this paper, we focus on score-based generative models, such as NCSN [1] and NCSN++ [2]. We propose a modified sampling algorithm, entitled *momentum-imbued Langevin dynamics (MILD)*, to generate samples from pre-trained score-based generative models such as NCSN [1] and NCSN++ [2]. It can be interpreted as a modification to the unadjusted Langevin algorithm (ULA) [32] for sampling from high-dimensional image distributions with the addition of a momentum update to accelerate sampling when used with the pre-trained NCSN and NCSN++ models. Experimental results show that MILD accelerates sampling for pre-trained score-based generative models by a factor of two to five. Figure 1 gives a flavor of the speed-up obtained using the proposed MILD accelerator. More details will be presented in Sections 2 and 3. MILD is *lightweight* and can be integrated with existing/pre-trained score-based generative models to speed up sampling with improvement in the quality of generated images compared with the baseline ALD. We consider both NCSN and NCSN++ models and present experimental results on CIFAR-10, CelebA, and LSUN-Bedroom datasets. The closest approach to ours is that of Dockhorn et al. [33], who proposed a critically-damped Langevin dynamics (CLD) algorithm, wherein the training and sampling is carried out in a joint extended space involving both position and velocity coordinates. While CLD is efficient in terms of sampling speed and generated sample quality, the approach involves sampling using an SDE integrator and training in a joint extended space. On the other hand, MILD does not require a change in the training loss, and can be used as an add-on to accelerate sampling for pre-trained score-based generative models.

2. SCORE-BASED GENERATIVE MODELLING

Score-based generative models are starkly different from traditional approaches to generative modeling such as the well-established GANs [15] and VAEs [14], where the generation is guided by the minimization of the Jensen-Shannon divergence and the maximization of the evidence lower bound (ELBO), respectively. In contrast, score-based generative models are trained to learn the gradient of the log of the density function directly. This is achieved by minimizing the Fisher divergence $D_{\mathrm{F}}(p\|q)$ between the unknown density $p(\boldsymbol{x})$ and the learnt density $q(\boldsymbol{x})$ [34], defined as follows:

$$D_{\mathrm{F}}(p\|q) = \int p(\boldsymbol{x}) \left\| \nabla_{\boldsymbol{x}} \ln \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} \right\|^2 \mathrm{d}\boldsymbol{x}. \tag{1}$$

Instead of modelling the density function directly, we model the score function of the density $q(\boldsymbol{x})$, which is defined as $s_{\boldsymbol{\theta}}(\boldsymbol{x}) \triangleq \nabla_{\boldsymbol{x}} \ln q(\boldsymbol{x})$. Directly minimizing the divergence in (1) is not possible since one does not have access to the ground-truth score $\nabla_{\boldsymbol{x}} \ln p(\boldsymbol{x})$. This can be circumvented by using an equivalent loss, namely, denoising score matching [35, 36] or sliced score matching [37]. In denoising score matching (DSM), the original dataset is perturbed by noise of various scales and the score-based model is trained to learn the score of the noise-perturbed data distribution for each noise scale. The noise is typically isotropic Gaussian, and there are L levels of noise with decreasing standard deviations $\sigma_1 > \sigma_2 > \ldots > \sigma_L$. The noise perturbed distribution corresponding to the i^{th} noise level is given by the convolution $p_{\sigma_i}(\boldsymbol{x}) = \int p(\boldsymbol{y}) \mathcal{N}(\boldsymbol{x}; \boldsymbol{y}, \sigma_i^2 \mathbb{I}) \mathrm{d}\boldsymbol{y}$, where $p(\boldsymbol{y})$ denotes the underlying true, but unknown, distribution, and $\mathcal{N}(\boldsymbol{x}; \boldsymbol{y}, \sigma_i^2 \mathbb{I})$ is a Gaussian with mean $\boldsymbol{y}$ and standard deviation σ_i in all directions. In NCSN, the score network $s_{\boldsymbol{\theta}}(\boldsymbol{x}, i)$ is trained to estimate the score of the noise-perturbed distribution $\nabla_{\boldsymbol{x}} \ln p_{\sigma_i}(\boldsymbol{x})$ for each $i \in \{1, \ldots, L\}$. In this setting, with $\lambda(i) \triangleq \sigma_i^2$, the loss function becomes a weighted sum of Fisher divergences for all noise-scales given by

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{L} \lambda(i) \mathbb{E}_{p_{\sigma_i}(\boldsymbol{x})} \left[\| s_{\boldsymbol{\theta}}(\boldsymbol{x}, i) - \nabla_{\boldsymbol{x}} \ln p_{\sigma_i}(\boldsymbol{x}) \|_2^2 \right].$$

Once the model is trained, new samples are generated from the model using a modified version of the unadjusted Langevin algorithm [32], referred to as the annealed Langevin dynamics (ALD), since the noise-scale σ_i anneals gradually as i

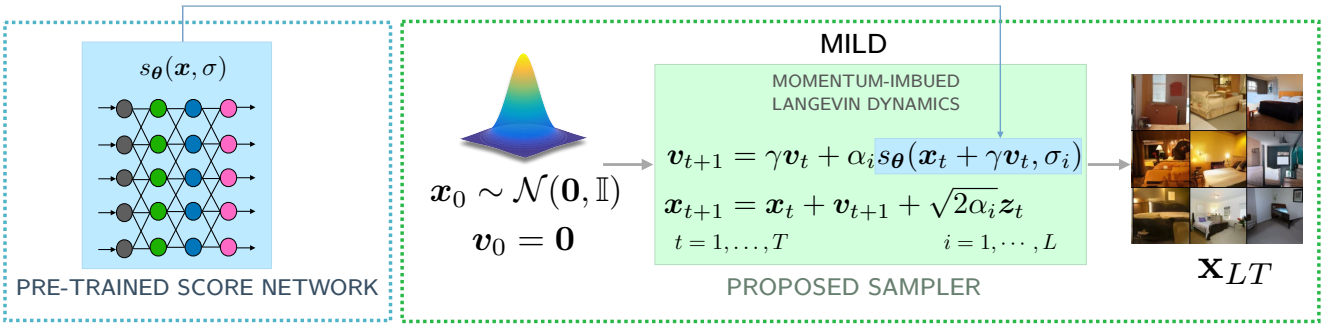

Fig. 2: (Colour online) Schematic of the proposed accelerated Momentum-Imbued Langevin Dynamics (MILD) sampler.

decreases. In NCSN++ [2], practical recommendations are given to improve the quality of the generated images by introducing modifications to the noise-scales, model architectures and employing an exponential moving-average of the neural network parameters.

2.1. Connection between Optimization and Sampling

Consider the minimization of a differentiable function f using the gradient-descent update:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \nabla_x f(\boldsymbol{x}_t). \tag{2}$$

Assuming an energy-based model, i.e., $p(\boldsymbol{x}) = \exp(-f(\boldsymbol{x}))/Z$, the Langevin Monte Carlo (LMC) sampling step is given by

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha_t \underbrace{\nabla_x \ln p(\boldsymbol{x}_t)}_{=-\nabla_x f(\boldsymbol{x}_t)} + \sqrt{2\alpha_t}\, \boldsymbol{z}_t \tag{3}$$

Comparing Eqs. 2 and 3, we observe that LMC sampling can be interpreted as a noisy gradient-ascent on the log-likelihood with progressively smaller noise as the iterations progress. Leveraging this link between sampling and optimization, we posit that acceleration schemes used in the optimization literature can be used to advantage in sampling as well. In Fig. 3, the parallels between optimization and sampling are shown through an illustration for a quadratic function f and a log-concave density p. Our acceleration scheme for sampling is inspired by Nesterov's momentum [38]. For convex, β-smooth functions, Nesterov's technique [38] accelerates the convergence rate of standard gradient-descent optimization of convex costs from linear to quadratic. The accelerated sampling updates are given by

$$\boldsymbol{v}_{t+1} = \gamma \boldsymbol{v}_t - \alpha \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t + \gamma \boldsymbol{v}_t), \tag{4}$$

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{v}_{t+1} + \sqrt{2\alpha}\boldsymbol{z}_t, \tag{5}$$

where $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$. We observed through experimentation that keeping the value of the parameter γ small and constant across iterations yielded better FID. More details will be presented in a journal version. Intuitively, momentum provides an advantage over traditional gradient-descent by introducing a moving-average of the past gradients, accelerating and stabilizing the optimization. The resulting momentum-imbued Langevin dynamics algorithm is named MILD (cf. Fig. 2 for a schematic, and Algorithm 1 for the listing).

Algorithm 1: Faster generative sampling with momentum-imbued Langevin dynamics (MILD).

Input: $\{\sigma_i\}_{i=1}^{L}, \varepsilon, T, \gamma$, score-network $s_{\boldsymbol{\theta}}$
Initialize: $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$, $\boldsymbol{v}_0 = \mathbf{0}$
for $i = 1$ *to* L **do**
  $\alpha_i = \dfrac{\varepsilon \cdot \sigma_i^2}{\sigma_L^2}$
  for $t = 0$ *to* $T - 1$ **do**
    Draw $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$
    $\boldsymbol{v}_{t+1} = \gamma \boldsymbol{v}_t + \alpha_i s_{\boldsymbol{\theta}}(\boldsymbol{x}_t + \gamma \boldsymbol{v}_t, \sigma_i)$
    $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \boldsymbol{v}_{t+1} + \sqrt{2\alpha_i}\boldsymbol{z}_t$
  $\boldsymbol{x}_0 \leftarrow \boldsymbol{x}_T$
  $\boldsymbol{v}_0 \leftarrow \boldsymbol{v}_T$
if denoise $\boldsymbol{x_T}$ **then**
  return $\boldsymbol{x_T} + \sigma_T^2 s_\theta(\boldsymbol{x_T}, \sigma_T)$
else
  return $\boldsymbol{x_T}$

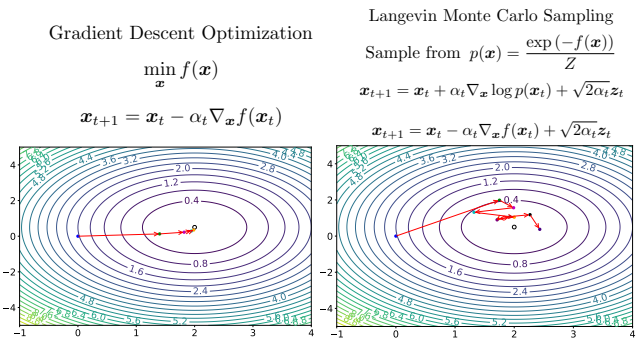

Fig. 3: (Colour online) Comparison between minimization of a convex function f using the gradient descent algorithm and sampling from a log-concave p.d.f. p using Langevin Monte Carlo algorithm.

Fig. 4: (Colour online) Samples generated by NCSN++ [2] with MILD acceleration.

Table 1: FID values for samples generated using annealed Langevin dynamics (ALD) and momentum-imbued Langevin dynamics (MILD) for the pre-trained NCSN [1] and NCSN++ [2] models. L denotes the number of noise levels, and T denotes the number of sampling steps at each noise-scale. $L \times T$ is the total number of sampling steps.

NCSN	Dataset	L	T	FID
ALD	CIFAR-10 (32×32)	10	100	**25.32**
MILD **(Ours)**	CIFAR-10 (32×32)	10	100	46.98
		10	50	28.23
		10	40	29.78
		10	30	40.35
		10	20	41.18
ALD	CelebA (32×32)	10	100	75.78
MILD **(Ours)**	CelebA (32×32)	10	100	81.85
		10	50	70.97
		10	40	**70.48**
		10	30	71.14
		10	20	73.25
NCSN++	**Dataset**	L	T	**FID**
ALD	CIFAR-10 (32×32)	232	5	12.9
MILD **(Ours)**	CIFAR-10 (32×32)	232	5	12.97
		232	4	**12.43**
		232	3	12.55
		232	2	13.58
		232	1	15.5
ALD	CelebA (64×64)	500	5	11.1
MILD **(Ours)**	CelebA (64×64)	500	5	10.93
		500	4	9.91
		500	3	9.37
		500	2	**8.59**
		500	1	8.98

3. EXPERIMENTAL RESULTS

MILD can be used with any pre-trained score network. To evaluate the sampling efficiency of MILD, we conducted a comprehensive set of experiments aimed at assessing the speedup obtained with MILD sampling algorithm for two prominent score-based generative models, namely, NCSN and NCSN++, which employ annealed Langevin dynamics (ALD) [1, 2] for sampling. Consistent with the baseline approaches [1, 2], we report FID on CIFAR-10 and CelebA datasets. L is the number of noise levels used during sampling and T is the number of ALD updates done. We ablate on T with MILD and calculate the Fréchet Inception Distance (FID) [17] and the number of function evaluations (NFEs) $L \times T$. Table 1 reports the FID achieved while sampling with ALD and MILD. Figure 1 shows the FID as a function of the NFEs. The results show that MILD achieves FIDs comparable to lower than the baseline ALD algorithm, in two-to-five fold fewer function evaluations. Fig. 4 shows the images generated using the NCSN++ model with MILD acceleration on CIFAR-10 (32×32), CelebA (64×64), and LSUN-bedrooms (128×128) datasets [1].

4. CONCLUSIONS

Incorporating momentum clearly accelerates the sampling speed as well as improves the quality of the generated samples (as measured by FID). MILD achieves lower FID than ALD, which is the sampler used in NCSN and NCSN++ across CIFAR-10 and CelebA datasets with fewer sampling steps. For high-resolution images, MILD generates consistently higher quality samples than ALD, which is also consistent with that of Jolicoeur-Martineau et al. [18]. Obtaining theoretical guarantees on the rate of convergence and exploring alternative acceleration schemes for faster sampling are potential directions for future research in this area.

[1] Source code to implement MILD is accessible at `https://github.com/mani-312/mild`

5. REFERENCES

[1] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Adv. Neural Inf. Process. Syst.*, 2019.

[2] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in *Adv. Neural Inf. Process. Syst.*, 2020.

[3] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Intl. Conf. on Learning Representations*, 2021.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Adv. Neural Inf. Process. Syst.*, 2020.

[5] P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Adv. Neural Inf. Process. Syst.*, 2021.

[6] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *J. Machine Learning Research*, 2022.

[7] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," in *Intl. Conf. on Learning Representations*, 2021.

[8] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," in *Intl. Conf. on Learning Representations*, 2021.

[9] S. Pascual, G. Bhattacharya, C. Yeh, J. Pons, and J. Serrà, "Full-band general audio synthesis with score-based diffusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[10] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021.

[11] B. Kawar, G. Vaksman, and M. Elad, "SNIPS: Solving noisy inverse problems stochastically," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[12] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," in *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

[13] B. Kawar, J. Song, S. Ermon, and M. Elad, "JPEG Artifact Correction using Denoising Diffusion Restoration Models," in *NeurIPS Workshop on Score-Based Methods*, 2022.

[14] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *Intl. Conf. on Learning Representations*, 2014.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Adv. Neural Inf. Process. Syst. 27*. 2014.

[16] G. Parisi, "Correlation functions and computer simulations," *Nuclear Physics B*, vol. 180, no. 3, 1981.

[17] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Adv. Neural Inf. Process. Syst.*, 2017, vol. 30.

[18] A. Jolicoeur-Martineau, K. Li, R. Piché-Taillefer, T. Kachman, and I. Mitliagkas, "Gotta go fast when generating data with score-based models," *arxiv:2105.14080*, 2021.

[19] H. Ma, L. Zhang, X. Zhu, and J. Feng, "Accelerating score-based generative models with preconditioned diffusion sampling," in *17th European Conf. on Comp. Vis.*, 2022.

[20] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami, "Langevin diffusions and the Metropolis-adjusted Langevin algorithm," *Statistics Probability Letters*, 2014.

[21] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Intl. Conf. on Learning Representations*, 2021.

[22] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in *Intl. Conf. on Machine Learning*, 2023.

[23] A. Krizhevsky, "Learning multiple layers of features from tiny images," *Master's thesis, University of Toronto*, 2009.

[24] J. Deng, W. Dong, R. Socher, L. Li, L. Kai, and Li F., "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Intl. Conf. Comput. Vis. Pattern Recognit.*, 2009.

[25] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv:1506.03365*, 2015.

[26] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Adv. Neural Inf. Process. Syst.*, 2022.

[27] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Adv. Neural Inf. Process. Syst.*, 2022.

[28] C. Lu, Y. Zhou, . Bao, J. Chen, C. Li, and J. Zhu, "DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models," *arxiv:2211.01095*, 2023.

[29] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," in *Intl. Conf. on Learning Representations*, 2023.

[30] Q. Zhang, J. Song, and Y. Chen, "Improved order analysis and design of exponential integrator for diffusion models sampling," *arxiv:2308.02157*, 2023.

[31] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *Intl. Conf. on Learning Representations*, 2022.

[32] A. Wibisono, "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem," in *Proceedings of the 31st Conference On Learning Theory*, 2018.

[33] T. Dockhorn, A. Vahdat, and K. Kreis, "Score-Based Generative Modeling with Critically-Damped Langevin Diffusion," in *Intl. Conf. on Learning Representations*, 2022.

[34] S. Lyu, "Interpretation and generalization of score matching," in *Proc. of the* 25^{th} *Conf. on Uncertainty Artif. Intell.*, 2009.

[35] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural Computation*, vol. 23, no. 7, 2011.

[36] A. Hyvärinen, "Estimation of non-normalized statistical models by score matching," *J. Mach. Learn. Res.*, vol. 6, 2005.

[37] Y. Song, S. Garg, J. Shi, and S. Ermon, "Sliced score matching: A scalable approach to density and score estimation," in *Proc. of the* 35^{th} *Conf. on Uncertainty Artif. Intell.*, 2019.

[38] Y. E. Nesterov, "A method of solving a convex programming problem with convergence rate o$\frac{1}{(k^2)}$," in *Doklady Akademii Nauk*. Russian Academy of Sciences, 1983, vol. 269.