Journey to the BAOAB-limit: finding effective MCMC samplers for score-based models

Ajay Jain[*]
UC Berkeley
ajayj@berkeley.edu

Ben Poole[*]
Google Research
pooleb@google.com

Abstract

Diffusion and score-based generative models have achieved remarkable sample quality on difficult image synthesis tasks. Many works have proposed samplers for pretrained diffusion models, including ancestral samplers, SDE and ODE integrators and annealed MCMC approaches. So far, the best sample quality has been achieved with samplers that use time-conditional score functions and move between several noise levels. However, estimating an accurate score function at many noise levels can be challenging and requires an architecture that is more expressive than would be needed for a single noise level. In this work, we explore MCMC sampling algorithms that operate at a single noise level, yet synthesize images with acceptable sample quality. We show that while näive application of Langevin dynamics and a related noise-denoise sampler produces poor samples, methods built on integrators of underdamped Langevin dynamics using splitting methods can perform well. Our samplers also have great diversity, allowing many samples to be generated in a single long-run MCMC chain. Further, by combining MCMC methods with existing multiscale samplers, we begin to approach competitive sample quality without using scores at large noise levels. Find videos and code at https://ajayj.com/journey.

1 Introduction

Given access to samples from a target density $x \sim p(x)$, generative modeling aims to learn a model of the density that can be used to generate new samples. While directly estimating the density of $p(x)$ may be challenging, the task can be simplified by modeling a sequence of smoother distributions. Score-based generative modeling learns the score functions for a sequence of smoothed distributions that go from no noise (matching the data density) to entirely smooth (matching a Gaussian density) [22, 23, 20, 10]. These smoothed distributions are specified via a tractable Gaussian noising process on the clean data: $p(z_t|x) \triangleq \mathcal{N}(\alpha_t x, \sigma_t^2 I)$, where the resulting marginal distribution of the smoothed density at noise level t is given by $p(z_t) = \int p(z_t|x)p(x)dx$. The score function for this smoothed density can be learned from data using denoising score matching [24] so that $s(z_t; t) \approx \nabla \log p(z_t)$.

Our work studies approaches for generating samples from this learned sequence of score functions. Prior work has focused on methods that utilize all noise levels from this sequence to generate samples, annealing from high noise levels to low noise levels over the course of the sampling process. Song and Ermon [22] build an annealed Langevin dynamics approach where an overdamped Langevin sampler is run at each noise level decaying from high noise to low noise. [23] uses these score functions to integrate an SDE that corresponds to the reverse of the forward noising process. Sampling by integrating the reverse SDE requires discretizing time, which introduces error that can be corrected with MCMC. The integration step, or *predictor*, reduces the noise level, and *corrector* steps can be

[*] Equal contribution. Work done in part at Google Research.

NeurIPS 2022 Workshop on Score-Based Methods.

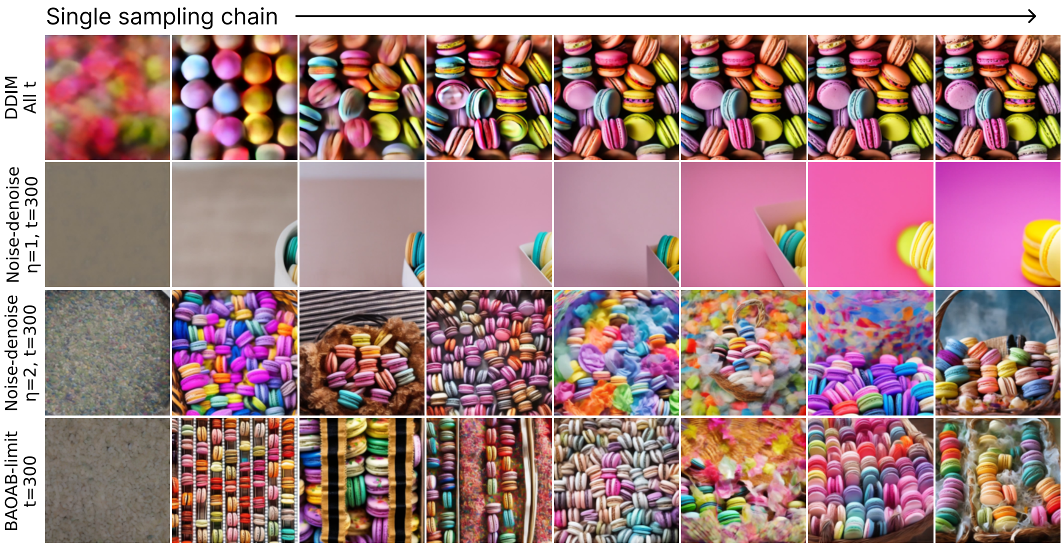

Figure 1: Steps along a single sampling chain for the caption *"a DSLR photo of a large basket of rainbow macarons"* with Stable Diffusion v2. Existing diffusion samplers like DDIM (**top**) generate a single image per sampling trajectory and require scores at many noise levels. Our noise-denoise $\eta = 2$ and BAOAB-limit samplers (**bottom**) freely traverse between modes, showing **greater diversity** within a chain, and **only use a single noise level**.

run to move the sample back towards the marginal for that smaller noise level. In [23], methods are explored that either just run corrector steps like in annealed Langevin dynamics, just run predictor steps to integrate the reverse SDE, or hybrid methods that alternate predictor and corrector steps. Improvements in MCMC could be leveraged in either annealed Langevin dynamics-based sampling approaches or in the corrector step of these hybrid approaches.

We are interested in developing methods that can not only improve the sampling quality in this "annealed" regime where all noise levels are used, but also can act as effective samplers from single or a reduced set of noise levels. Generating from a reduced set of noise levels could lead to two types of improvements. First, training is more efficient if the score function approximator can be trained on a reduced set of noise levels. Past work on text-to-image diffusion models improves quality by training separate models for different noise levels [1, 7], but a subset may suffice. Second, we could improve sampling efficiency by not running over all noise levels. Using annealed samplers with models like [1, 7] is cumbersome since model weights need to be switched across noise levels. If we are able to build effective MCMC samplers for individual noise levels that were able to mix well between modes, then there could be an additional speed improvement from the ability to generate a diverse set of samples from a single chain (as in strided sampling in MCMC).

In this work, we explore MCMC-based methods with momentum for improving sampling in score-based models. Our experiments began when we discovered a simple modification of an existing sampling method that performed rather well, both when annealed and when sampling single noise levels. After an extended literature search over many months, we found this update actually corresponds to an integrator of the underdamped Langevin dynamics (ULD) known in molecular dynamics as BAOAB-limit [13]. We explore this and several other integrators of ULD in the context of score-based models, and find that they are effective at producing samples from a *single* noise level of a pretrained score model, as well as when used in annealing approaches over multiple noise levels. We believe these MCMC methods may be useful in applications where typical diffusion sampling breaks down. In the future, we hope to explore difficult settings like using guided sampling from classifiers or introducing strong conditioning.

2 Sampling Algorithms

2.1 A simple, blurry baseline, and a bug that leads to a surprisingly effective generalization

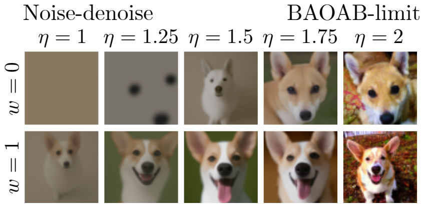

Figure 2: Comparing noise-denoise and our generalized sampler with an annealed noise level. We use the GLIDE filtered text-to-image model [14] conditioned on the caption *"a photograph of a corgi"*. **Left:** Noise-denoise with $\eta = 1$ is oversmoothed, though classifier-free guidance adds some detail. **Right:** More details appear as η approaches 2 (our accidental sampler, called BAOAB-limit).

Generating samples from a single noise level is challenging as we need a way to map from a noisy sample back to something that looks like denoised data. Given a noise process $p(z|x) = \mathcal{N}(\alpha z, \sigma)$, we can use an approximate score to estimate the posterior mean via Tweedie's formula [15, 5]: $\hat{x}(z) \triangleq \mathbb{E}[X|z] = (z + \sigma^2 s(z))/\alpha$. If σ^2 is small and $z \sim p(z)$, then $\hat{x}(z)$ will be a good sample from $p(x)$. As a simple baseline sampler, we considered alternating adding noise to the image and denoising:

$$\begin{aligned} \epsilon^{i+1} \sim \mathcal{N}(0, I), \quad z^{i+1} &= \alpha x^i + \sigma \epsilon^{i+1}, && \text{Noise} \\ x^{i+1} &= \hat{x}(z^{i+1}) && \text{Denoise} \end{aligned} \tag{1}$$

Bansal et al. [2] refer to (1) as näive sampling. We can rewrite the update with Tweedie's formula:

$$\begin{aligned} x^{i+1} &= \hat{x}(z^{i+1}) = (z^{i+1} + \sigma^2 s(z^{i+1}))/\alpha \\ &= (\alpha x^i + \sigma\epsilon^{i+1} + \sigma^2 s(z^{i+1}))/\alpha \\ &= x^i + \sigma/\alpha * (\epsilon^{i+1} - \hat{\epsilon}(z^{i+1})) \end{aligned} \tag{2}$$

where $\hat{\epsilon}(z^{i+1}) = -\sigma\, s(z^{i+1})$ is the prediction of the noise that was added. However, the noise-denoise update produces blurry samples, shown in Fig. 2 (top left) using the GLIDE model [14] and annealing the noise level σ gradually. Fig. 3 (left) shows similarly poor quality with an unconditional CIFAR10 model, both at single noise levels and when annealing. During experiments, we accidentally duplicated update step (2):

```
z_next = alpha * x_prev + sigma * eps                    # add noise to $x^i$
epshat = epshat_predictor(z_next)                        # predict noise with model
x_next = x_prev + sigma / alpha * (eps − epshat)         # set $x^{i+1}$ to $\hat{x}(z^{i+1})$
x_next = x_next + sigma / alpha * (eps − epshat)         # bug! redundant step
```

Surprisingly, this noise-double-denoise update performed remarkably well! Generalizing, we get:

$$z^{i+1} = \alpha x^i + \sigma\epsilon^{i+1}, \qquad x^{i+1} = x^i + \eta\sigma/\alpha(\epsilon^{i+1} - \hat{\epsilon}(z^{i+1})) \qquad \text{Generalized noise-denoise} \tag{3}$$

Fig. 2 shows sampling trajectories when sweeping η in (3). For $\eta = 1$, the noise-denoise method, we collapse to a constant oversmoothed image. Adding classifier-free guidance [9] improves sample quality for short chains (second row), but still is blurry. Running the sampler for longer even further deteriorates quality. Our surprisingly effective but accidental sampler with $\eta = 2$ produces good text-to-image samples (right), even without guidance.

2.2 Understanding the generalized sampler as correlating noise in Langevin dynamics

We hypothesize that noise-denoise is mode-seeking, and modes of likelihood-based models are known to have poor samples in complex distributions. To understand this, we rewrite updates in z-space:

$$\begin{aligned} z^{i+1} &= \alpha\hat{x}(z^i) + \sigma\epsilon^{i+1} && \text{One step of noise-denoise} \\ &= \alpha(z^i + \sigma^2 s(z^i))/\alpha + \sigma\epsilon^{i+1} && \text{Substitute Tweedie's for } \hat{x} \\ &= z^i + \sigma^2 s(z^i) + \sigma\epsilon^{i+1} && \text{Noise-denoise in z-space} \end{aligned} \tag{4}$$

Eq. (4) resembles overdamped Langevin dynamics (OLD), which uses step size κ and the update:

$$z^{i+1} = z^i + \kappa s(z^i) + \sqrt{2\kappa}\epsilon^{i+1}, \quad \epsilon^{i+1} \sim \mathcal{N}(0, I) \qquad \text{Overdamped Langevin} \tag{5}$$

Setting $\kappa = \sigma^2$, OLD follows $z^{i+1} = z^i + \sigma^2 s(z^i) + \underline{\sqrt{2}}\sigma\epsilon^{i+1}$, *i.e.* the same step size as noise-denoise, but with additional noise injected to the iterate after every step. The extra noise avoids the mode collapse present in the noise-denoise sampler and turns the heuristic update into MCMC. Similarly, we can rewrite the $\eta = 2$ setting of the generalized update (3) in z-space:

$$z^{i+1} = z^i + 2\sigma^2 s(z^i) + \sigma\epsilon^{i+1} + \underline{\sigma\epsilon^i}, \quad \epsilon^{i+1} \sim \mathcal{N}(0, I) \tag{6}$$

Unlike overdamped Langevin (5), our update (6) depends on the noise added at the previous iteration, and the scale on the noise is different. It is similar to OLD with $\kappa = 2\sigma^2$, but reusing half of the noise. Empirically, Langevin performs much worse with this step size (Fig. 3B). A heuristic for adapting the step size to score norm from [23] (their Alg. 5) improves OLD samples, but (6) was more robust.

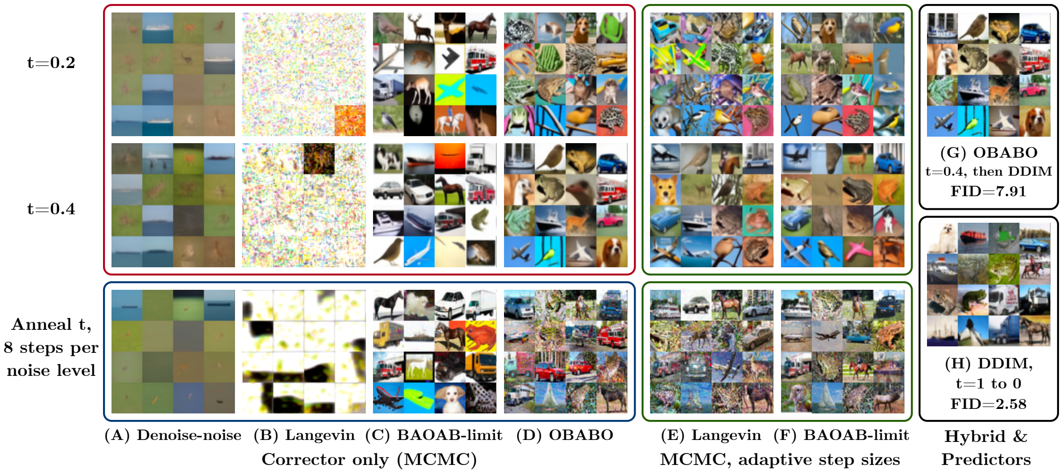

Figure 3: MCMC methods can sample images with reasonable quality, even at a single noise level (top two rows). The stateless sampler, **(A)** denoise-noise, produces overly smooth images, while **(B)** discretized overdamped Langevin MCMC fails to converge. In contrast, splitting methods from MD like the **(C)** infinite friction limit of the BAOAB sampler and **(D)** finite friction OBABO sampler have reasonable quality. **(E-F)** Tuning step sizes to be adaptive based on the norm of the score function [23] improves sample quality. **(G)** Refining single noise level samples with the DDIM predictor significantly boosts quality without using large timestep score estimates, unlike **(H)**.

2.3 Underdamped Langevin, splitting methods, BAOAB-limit and OBABO sampling

We sought to find the exact update (6) in the literature. It is tantalizingly similar to OLD, but the reuse of ϵ resembles a momentum-based sampler. We turned to *underdamped* Langevin dynamics (ULD), an SDE (7) that defines the evolution of position z and momentum p variables:

$$dz = M^{-1}pdt \qquad dp = s(z)dt - \gamma pdt + \sigma M^{1/2}dW, \tag{7}$$

where dW is a stochastic Wiener process (Brownian motion) injecting noise into the momentum term, γ is a friction coefficient, and M is a mass matrix that we set to the identity matrix. Many works propose methods for integrating the Langevin SDE (7). At a high level, integrators follow an Euler-Maruyma discretization scheme, or use splitting methods. Splitting methods offer efficient integrators for many dynamical systems by breaking apart an SDE into parts, exactly solving individual parts of the SDE alone, and using the solution to form partial updates to the parameters. The vector field defined by the ULD SDE can be split into three terms labeled A, B, and O [13, 8, 3]:

$$d\begin{bmatrix} z \\ p \end{bmatrix} = \underbrace{\begin{bmatrix} M^{-1}p \\ 0 \end{bmatrix} dt}_{A} + \underbrace{\begin{bmatrix} 0 \\ s(z) \end{bmatrix} dt}_{B} + \underbrace{\begin{bmatrix} 0 \\ -\gamma pdt + \sigma M^{1/2}dW \end{bmatrix}}_{O} \tag{8}$$

Here, A is an update to the position, B is an update to momentum driven by the score function, and together A and B are Hamilton's equation. O is often referred to as a partial momentum refreshment, labeled O after the Ornstein-Uhlenbeck process. Splitting methods to solve ULD choose a repeated order for solving these parts. Repeatedly alternating ABABA corresponds to leapfrog steps in HMC. There are a variety of methods for integrating Langevin dynamics based on different orderings, e.g. BAOAB, ABOBA. For conditional score sampling, we found that the updates of [8] did not perform well. In Dockhorn et al. [4], we found a reference to splitting methods commonly used in molecular sampling [13], which presents an update based on the BAOAB ordering. BAOAB can be simplified to our setting:

$$p^{i+1} = c(p^i + \sqrt{2\kappa}s(z^i)) + \sqrt{1-c^2}\epsilon^{i+1}, \qquad z^{i+1} = z^i + \kappa s(z^i) + \sqrt{\kappa/2}(p^{i+1} + p^i) \tag{9}$$

where $c = e^{-\gamma\sqrt{2\kappa}}$, κ is a step size, and γ is friction. Taking the limit as $\gamma \to \infty$ so $c = 0$, we arrive at the **BAOAB-limit** update where $p^{i+1} = \epsilon^{i+1}$:

$$z^{i+1} = z^i + \kappa s(z^i) + \sqrt{\kappa/2}(p^{i+1} + p^i)$$

$$\text{Or, } z^{i+1} = z^i + \kappa s(z^i) + \sqrt{\kappa/2}(\epsilon^{i+1} + \epsilon^i) \tag{10}$$

For $\kappa = 2\sigma^2$, this is $z_s = z_t + 2\sigma^2 s(z_t) + \sigma(\epsilon_s + \epsilon_t)$, **exactly the accidental update we found effective in** (6), giving us a proper derivation for the sampler.

OBABO is another ordering with $\kappa = 2\sigma_t(1-a)$ where we find that $a = 0.9$ performs well:

$$v_s = av_t + \kappa(a+1)/2\, s(z_t) + \sqrt{1-a^2}\epsilon_s, z_s = z_t + hp_s$$

2.4 Evaluating BAOAB-limit and other samplers

Text-to-image Fig. 1 compares a sampling chain from DDIM with long-run single noise level chains from our MCMC samplers. We compare noise-denoise (ND) with $\eta = 1$ and $\eta = 2$, and the generalized BAOAB-limit sampler (6). We use a latent diffusion model [16], Stable Diffusion v2, and visualize $\hat{x}$ predictions (Tweedie's denoiser). DDIM produces a single high-quality sample. ND$_{\eta=1}$ produces oversmoothed samples, but ND$_{\eta=1}$ and BAOAB-limit are sharp. They also have great diversity within a single sampling chain. This could allow users to generate hundreds of diverse samples from a single chain, more efficiently than sampling many images parallel with DDIM.

CIFAR-10, single noise level In quantitative experiments, we evaluate the sample quality of images generated by MCMC samplers with an unconditional diffusion model trained on CIFAR10 [17]. Fig. 4 shows the Frechét Inception Distance (FID) achieved by MCMC using about 512 score function evaluations. We compare single noise level MCMC (corrector-only) sampling using the BAOAB-limit sampler with alternatives including discretized OLD with adaptive step sizes and OBABO. To evaluate FID, we map the final MCMC iterate z^i to the clean data space with Tweedie's denoiser using the last score function evaluation. BAOAB-limit has better FID than alternatives at low noise levels, 27.49 at $t = 0.225$ (256 steps), even though nonsmooth energy landscapes typically cause slow mixing.

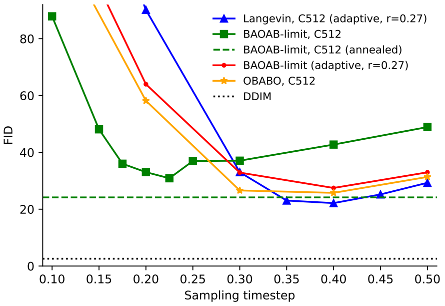

Figure 4: Some integrators like BAOAB-limit can sample reasonable quality images using a score function at only one noise level. Overdamped Langevin with standard step sizes is not shown as it did not converge, but adaptive step sizes allow OLD to operate on intermediate score functions e.g. $s(z_{t=0.4})$. Note that DDIM, a predictor-only method, has better FID since FID is evaluated on clean images, $t = 0$. Results use 512 score evaluations.

Hybrid samplers Still, predictor-only methods like DDIM [21] have the best quality, 2.58 FID with 512 steps, since they do not rely on Tweedie's to produce clean images. Running OBABO at a fixed noise level of $t = 0.4$ for 512 steps has FID 25.76. Subsequently refining with 64 steps of DDIM, annealing $t = 0.4$ to 0 **achieves competitive FID of 7.91**. See Section A.1 for experimental details and more results.

This indicates that the sample quality achievable using scores only estimated for $t \in [0, 0.4]$ can approach quality using all timesteps. Future work could improve the efficiency of diffusion model training by limiting the number of modeled noise levels akin to [1, 7], and using a hybrid sampler like OBABO followed by DDIM.

Acknowledgements

We thank Alex Alemi, Matt Hoffman, Ruiqi Gao, Pieter Abbeel, Jon Barron, Ben Mildenhall, Aditi Jain and Paras Jain for helpful discussions. StabilityAI provided computational resources used to evaluate samplers on CIFAR10. Ajay Jain was supported in part by the BAIR Industrial Consortium.

References

[1] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, K. Kreis, M. Aittala, T. Aila, S. Laine, B. Catanzaro, T. Karras, and M.-Y. Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[2] A. Bansal, E. Borgnia, H.-M. Chu, J. S. Li, H. Kazemi, F. Huang, M. Goldblum, J. Geiping, and T. Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise, 2022. URL `https://arxiv.org/abs/2208.09392`.

[3] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped langevin mcmc: A non-asymptotic analysis. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323. PMLR, 06–09 Jul 2018. URL `https://proceedings.mlr.press/v75/cheng18a.html`.

[4] T. Dockhorn, A. Vahdat, and K. Kreis. Score-based generative modeling with critically-damped langevin diffusion, 2021. URL `https://arxiv.org/abs/2112.07068`.

[5] B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181. URL `https://doi.org/10.1198/jasa.2011.tm11181`. PMID: 22505788.

[6] J. Fass, D. Sivak, G. Crooks, K. Beauchamp, B. Leimkuhler, and J. Chodera. Quantifying configuration-sampling error in langevin simulations of complex molecular systems. *Entropy*, 20(5):318, Apr 2018. ISSN 1099-4300. doi: 10.3390/e20050318. URL `http://dx.doi.org/10.3390/e20050318`.

[7] Z. Feng, Z. Zhang, X. Yu, Y. Fang, L. Li, X. Chen, Y. Lu, J. Liu, W. Yin, S. Feng, Y. Sun, H. Tian, H. Wu, and H. Wang. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts, 2022. URL `https://arxiv.org/abs/2210.15257`.

[8] A. Garriga-Alonso and V. Fortuin. Exact langevin dynamics with stochastic gradients. In *Third Symposium on Advances in Approximate Bayesian Inference*, 2021. URL `https://openreview.net/forum?id=Rprd8aVUYkE`.

[9] J. Ho and T. Salimans. Classifier-free diffusion guidance, 2022. URL `https://arxiv.org/abs/2207.12598`.

[10] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf`.

[11] M. D. Hoffman and P. Sountsov. Tuning-free generalized hamiltonian monte carlo. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7799–7813. PMLR, 28–30 Mar 2022. URL `https://proceedings.mlr.press/v151/hoffman22a.html`.

[12] A. Jolicoeur-Martineau, R. Piché-Taillefer, R. T. d. Combes, and I. Mitliagkas. Adversarial score matching and improved sampling for image generation, 2020. URL `https://arxiv.org/abs/2009.05475`.

[13] B. Leimkuhler and C. Matthews. Rational Construction of Stochastic Numerical Methods for Molecular Sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56, 06 2012. ISSN 1687-1200. doi: 10.1093/amrx/abs010. URL `https://doi.org/10.1093/amrx/abs010`.

[14] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021. URL `https://arxiv.org/abs/2112.10741`.

[15] H. E. Robbins. *An Empirical Bayes Approach to Statistics*, pages 388–394. Springer New York, New York, NY, 1992. ISBN 978-1-4612-0919-5. doi: 10.1007/978-1-4612-0919-5_26. URL `https://doi.org/10.1007/978-1-4612-0919-5_26`.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[17] T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=TIdIXIpzhoI`.

[18] S. Saremi and R. K. Srivastava. Multimeasurement generative models. *ArXiv*, abs/2112.09822, 2022.

[19] J. Serrà, S. Pascual, and J. Pons. On tuning consistent annealed sampling for denoising score matching. *CoRR*, abs/2104.03725, 2021. URL `https://arxiv.org/abs/2104.03725`.

[20] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/sohl-dickstein15.html`.

[21] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=St1giarCHLP`.

[22] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=PxTIG12RRHS`.

[24] P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

A Appendix

A.1 Hybrid samplers and acceleration with short-run MCMC

Figure 5 shows the FID score for CIFAR10 samples versus the total number of score functional evaluations for the BAOAB-limit and OBABO z-space samplers, Equations (6, 11). We apply the two splitting methods at a single noise level but vary the number of MCMC steps run from initialization. Each chain is initialized with a draw from a standard normal. The BAOAB-limit sampler converges to the target distribution faster than OBABO, perhaps because momentum takes time to warmup. In fact, a shorter-run chain with 256 steps performs better than 512 steps for BAOAB-limit.

To close the sample quality gap between MCMC samplers run at a single noise level and predictor-only methods, we explore hybrid samplers (dashed lines, Figure 5). The hybrid samplers only require score function evaluations up to a maximum noise level: $t \leq 0.4$ for hybrid OBABO, and $t \leq 0.225$ for hybrid BAOAB-limit. These max timesteps were chosen based on the best single noise level samplers in Figure 4. After running MCMC for $K_{\text{corrector}} \in \{32, 64, 128, 256, 512, 768\}$ iterations at a fixed noise level, we initialize a DDIM sampler with the resulting z_t^i and run it for $K_{\text{predictor}} \in \{8, 16, 32, 64\}$ steps to remove remaining noise. This significantly outperforms the MCMC-only approaches shown with solid lines. The MCMC-only approaches rely on Tweedie's one step denoiser at the end of sampling map from the marginal distribution $p(z_t)$ to $p(x)$, but predictor methods are more accurate. This almost closes the gap to DDIM-only sampling (2.58 FID at $K_{\text{corrector}} = 0, K_{\text{predictor}} = 512$).

The promise of hybrid samplers suggests that the score function may not need to be learned at large noise levels. Training only for a small range of noise levels would be more efficient by permitting a smaller function approximator, and potentially would converge faster.

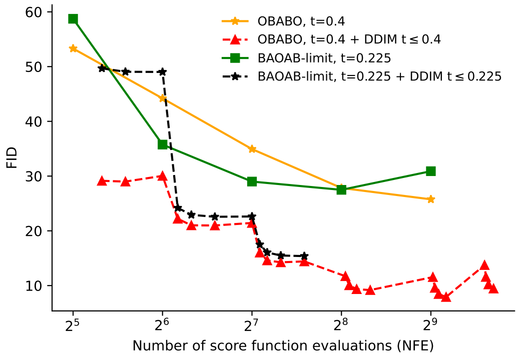

Figure 5: CIFAR10 FID (sample quality) versus number of score function evaluations. The BAOAB-limit sampler converges faster than the momentum-based OBABO method.

A.2 Relationship between DDPM sampler, noise-denoise and BAOAB-limit

The DDPM ancestral sampler from Ho et al. [10] can be expressed in a similar form to the generalized denoise-noise sampler presented in Equation (3):

$$\begin{aligned} z_t &= \alpha_t x_t + \sqrt{c}\sigma_t \epsilon \\ &= x_t + \frac{\sigma_t}{\alpha_t}(\sqrt{c}\epsilon - c\hat{\epsilon}_\theta(z_t)) \\ \text{where } c &= 1 - \frac{\sigma_s^2 \alpha_t^2}{\alpha_s^2 \sigma_t^2} \text{ for DDPM.} \end{aligned} \tag{11}$$

Fig. 6 shows c for the β-linear schedule used by Ho et al. [10], which is less than 1 across most timesteps and increases at the start and end of the sampling trajectory. Interestingly, setting $c = 1$ recovers the noise-denoise update (2). In early experiments, setting c to a small fixed value did not perform well.

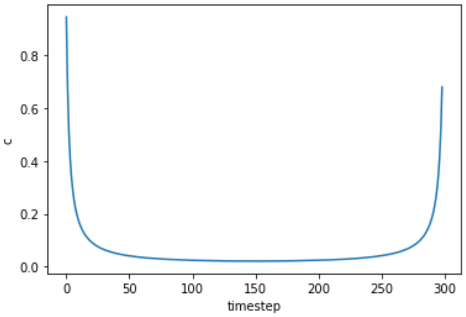

Figure 6: The value of the scaling coefficient c for the DDPM ancestral sampler with a β-linear noise schedule when expressed in a form similar to our noise-denoise sampler.

A.3 Damping BAOAB-limit sampling for high guidance conditional models

For some diffusion priors like GLIDE, we found that high classifier-free guidance (CFG) led to artifacts for the BAOAB-limit sampler. However, a damping parameter helped remove artifacts in this strongly conditional case. Fig. 7 shows sampling trajectories for an update that reformulates Equation (3) to introduce a damping parameter λ:

$$z^{i+1} = \alpha x^i + \sigma \epsilon^{i+1} \tag{12}$$

$$x^{i+1} = (1+\lambda)\hat{x}(z^{i+1}) - \lambda x^i, \lambda \in [0, 1] \tag{13}$$

The timestep used to determine the noise level σ is annealed over the course of sampling. Setting $\lambda = 0$ recovers the simple denoise-noise update, while $\lambda = 1$ is equivalent to the $\eta = 2$ setting of (3). An intermediate value such as $\lambda = 0.8$ removes all the high-frequency staturated artifacts while preserving sample quality. Early on when the noise level is large, there is significantly diversity within the sampling chain for all settings. Such diverse samplers could be useful for collecting many samples from the same trajectory. This could provide downstream applications multiple options for a generated image without running a large, more expensive batch.

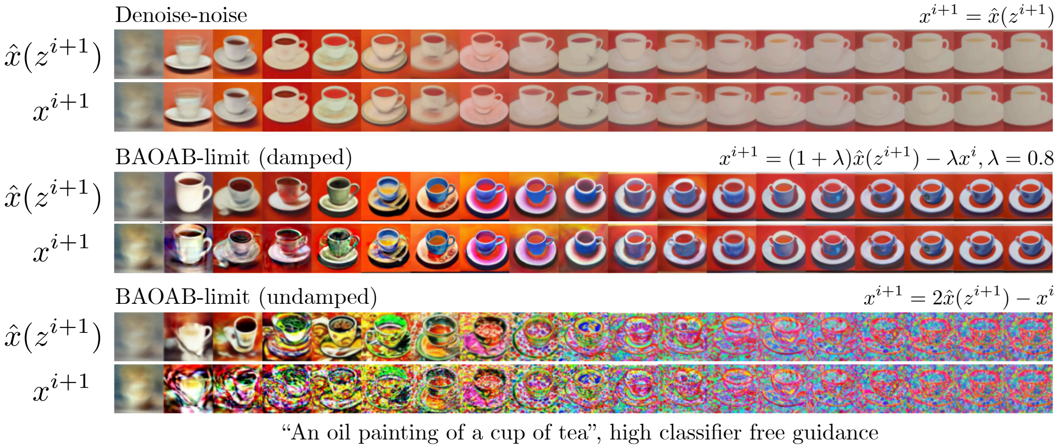
Denoise-noise
$x^{i+1} = \hat{x}(z^{i+1})$
$\hat{x}(z^{i+1})$
x^{i+1}
BAOAB-limit (damped)
$x^{i+1} = (1+\lambda)\hat{x}(z^{i+1}) - \lambda x^i, \lambda = 0.8$
$\hat{x}(z^{i+1})$
x^{i+1}
BAOAB-limit (undamped)
$x^{i+1} = 2\hat{x}(z^{i+1}) - x^i$
$\hat{x}(z^{i+1})$
x^{i+1}
"An oil painting of a cup of tea", high classifier free guidance

Figure 7: The infinite-friction limit BAOAB sampler can be seen as a way of propagating history in the denoise-noise sampler. In this figure, we show a text-conditional sampling trajectory for the GLIDE model using high values of classifier-free guidance. While high CFG can create artifacts, introducing a damping parameter λ mitigates the problem without losing sample quality.

A.4 Related work

There are several methods that have aimed to improve MCMC-based methods for score-based models. Saremi and Srivastava [18] introduce a "walk-jump" method that runs Langevin MCMC jointly on several noise levels and then applies Tweedie's formula to jump to a denoised sample. This method is rather expensive as it requires evaluating several score functions at each step. Jolicoeur-Martineau et al. [12] introduce an improved method for setting step size in annealed Langevin dynamics that we would like to compare against in future work along with Serrà et al. [19]. In the MCMC literature, there has also been recent methods that incorporate partial momentum refreshment to make more progress between Metropolis-Hastings steps [11]. Our work is similar in that integrators of underdamped Langevin dynamics also use partial momentum refreshment but drop the complexity of the Metropolis-Hastings correction (thus introducing bias). However Fass et al. [6] show that rejection rates for Metropolized BAOAB are typically too high, incurring a high cost from momentum flipping when a sample is rejected.