SHOULD EBMS MODEL THE ENERGY OR THE SCORE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in training unnormalized models through *denoising score matching with Langevin dynamics* (SMLD) and *denoising diffusion probabilistic modeling* (DDPM) has made unnormalized models a competitive model class for generative modeling. Unlike earlier work on energy based models, these recent works construct generative models by directly parameterizing the score function of the model density, rather than the density itself. Such *unconstrained* score models are not guaranteed to output a conservative vector field, meaning they do not correspond to the gradient of any function, unlike *constrained* score models that are obtained through explicitly differentiating a parameterized energy function. Explicit energy based models thus seem to have a theoretical advantage, but empirical evidence currently points to unconstrained score models performing better in practice. Here we compare both methods for modeling the score of the data distribution, finding that constrained score models, i.e. energy based models, can perform just as well as unconstrained models when using a comparable model structure.

1 TRAINING SCORE-BASED GENERATIVE MODELS

SMLD (Song & Ermon, 2019) and DDPM (Sohl-Dickstein et al., 2015; Ho et al., 2020) are generative models that train a sequence of denoising autoencoders on Gaussian-perturbed data at multiple noise scales, producing samples at test time by running processes similar to Langevin dynamics using the denoising autoencoders as vector fields in data space. More specifically, SMLD trains on denoising score matching losses (Vincent, 2011):

$$\sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0}\, \mathbb{E}_{p_t(\mathbf{x}\,|\,\mathbf{x}_0)}\, \sigma_t^2\, \|\mathbf{s}_\theta(\mathbf{x},\mathrm{t}) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x}\,|\,\mathbf{x}_0)\|^2 \tag{1}$$

where $p_t(\mathbf{x}\,|\,\mathbf{x}_0) = \mathcal{N}(\mathbf{x};\mathbf{x}_0,\sigma_t^2\,\mathbf{I})$ adds Gaussian noise with variance σ_t^2 to the raw data $\mathbf{x}_0$ for a total of T noise levels. This loss is equivalent to score matching on the marginals $p_t(\mathbf{x}) = \int \mathcal{N}(\mathbf{x};\mathbf{x}_0,\sigma_t^2)p(\mathbf{x}_0)d\,\mathbf{x}_0$ for some weights γ_t:

$$\sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0}\, \mathbb{E}_{p_t(\mathbf{x}\,|\,\mathbf{x}_0)}\, \gamma_t\, \|\mathbf{s}_\theta(\mathbf{x},\mathrm{t}) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\|^2 \tag{2}$$

and therefore, under ideal conditions, $\mathbf{s}_\theta(\mathbf{x},\mathrm{t})$ learns the true scores $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ for all t. Sampling proceeds by running multiple Langevin dynamics steps for each $t = T, \ldots, 1$ using $\mathbf{s}_\theta(\mathbf{x},\mathrm{t})$ as the gradient, warm-starting the first sample for step t from the final sample from step $t+1$.

DDPM defines a forward process $p(\mathbf{x}_t\,|\,\mathbf{x}_{t-1}) = \mathcal{N}\left(\mathbf{x}_t, \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\,\mathbf{I}\right)$ that gradually adds noise with variances β_t to the data, resulting in noisy scaled data $\tilde{p}_t(\mathbf{x}\,|\,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}; \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0, (1-\bar{\alpha}_t)\,\mathbf{I})$ at timestep t. The model is a reverse process

$$p_\theta(\mathbf{x}_{t-1}\,|\,\mathbf{x}_t = \mathbf{x}) = \mathcal{N}\left(\mathbf{x}_{t-1}; \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x} - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\mathbf{s}_\theta(\mathbf{x},\mathrm{t})\right), \sigma_t^2\,\mathbf{I}\right) \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Training is performed by maximizing the variational lower bound treating the noisy data $\mathbf{x}_t$ as latent variables, which can be shown to be equivalent to optimizing the score matching losses for some weights $\tilde{\gamma}_t$:

$$\sum_{t=1}^{T} \mathbb{E}_{\mathbf{x}_0}\, \mathbb{E}_{\tilde{p}_t(\mathbf{x}\,|\,\mathbf{x}_0)}\, \tilde{\gamma}_t\, \|\mathbf{s}_\theta(\mathbf{x},\mathrm{t}) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\|^2 \tag{4}$$

where above $p_t(\mathbf{x}) = \int \tilde{p}_t(\mathbf{x} \,|\, \mathbf{x}_0)p(\mathbf{x}_0)d\,\mathbf{x}_0$ is the marginal distribution of the forward process. Sampling is defined to be ancestral sampling from $p_\theta(\mathbf{x}_{t-1} \,|\, \mathbf{x}_t)$, which, by examining (3), resembles a warm-started Langevin dynamics on $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ similar to SMLD.

Both methods learn a function $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ modeling the score of of noisy data, either $p_t(\mathbf{x})$ or $\tilde{p}_t(\mathbf{x})$, but successful implementations currently parameterize $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ as an unconstrained neural network, not as the gradient of a parameterized energy function. Our work here investigates whether $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ networks constrained to be the gradient of a parameterized energy function are able to attain sample quality results similar to those of unconstrained networks.

2 Conservative vector fields and higher order gradients

The goal of score-based generative modeling is to approximate the score of the data distribution $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ with the learned score model $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$. Here, we know that $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is a *conservative vector field*, which means that the line integral between two points $\int_{\mathbf{x}_1}^{\mathbf{x}_2} \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) d\,\mathbf{x}$ is independent of the path taken from $\mathbf{x}_1$ to $\mathbf{x}_2$. This property is essential in guaranteeing that MCMC sampling based on $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ indeed samples $\mathbf{x}_1$ and $\mathbf{x}_2$ according to their relative probability under $p_t(\mathbf{x})$. When $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ is parameterized as the gradient of a (smooth) energy function, i.e. $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t}) = -\nabla_{\mathbf{x}} E_\theta(\mathbf{x}, t)$, our model is guaranteed to give a conservative vector field. This is not true for the general case where $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ is modeled by an unconstrained neural network.

A related property that is satisfied by the ground-truth score is that $\partial_i[\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]_j = \partial_j[\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]_i$, i.e. the Jacobian of the score is symmetric. This property is hard to reproduce in feedforward neural nets, as discussed by Saremi (2019).

It makes intuitive sense to add these properties as constraints when constructing our model $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$, by parameterizing it as the gradient of an energy function. However, in personal communication multiple researchers in this area indicated that they did not obtain good results following this approach. In Section 3 we investigate how to build in these constraints in a way that does not hurt results.

3 Constrained and unconstrained score models

Recent work on score-based generative modeling (Song & Ermon, 2019; Ho et al., 2020) models the score as an unconstrained convolutional neural network, i.e. $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t}) = f_\theta(\mathbf{x}, t)$, where $f_\theta(\mathbf{x}, t)$ maps from $\mathbb{R}^d$ to $\mathbb{R}^d$ for a d-dimensional input image $\mathbf{x}$. In contrast, earlier work on energy based model usually modeled the score by taking the derivative of a parameterized energy function, $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t}) = -\nabla_{\mathbf{x}} E_\theta(\mathbf{x}, t)$, with $E_\theta(\mathbf{x}, t)$ a feedforward neural network mapping from $\mathbb{R}^d$ to $\mathbb{R}$. In addition to guaranteeing a conservative vector field $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$, the latter approach allows us to use standard architectures from the image classification literature for specifying $E_\theta(\mathbf{x}, t)$. When $E_\theta(\mathbf{x}, t)$ is chosen as such a classifier, it enables combining image generation and classification in interesting ways (Grathwohl et al., 2019).

Although elegant, specifying the score model by taking the gradient of an image classifier has so far not produced competitive results in image generation. We hypothesize that the reason is that the function $\nabla_{\mathbf{x}} E_\theta(\mathbf{x}, t)$ is severely restricted when $E_\theta(\mathbf{x}, t)$ is a standard feedforward classifier. Standard image classifiers are not designed to preserve all the detailed information in $\mathbf{x}$ needed to accurately model the high dimensional score of the data distribution, but instead use deep stacks of layers with downsampling to arrive at more abstract representations. In contrast, Ho et al. (2020) obtain their best results when specifying $\mathbf{s}_\theta(\mathbf{x}, \mathrm{t})$ as a U-Net (Ronneberger et al., 2015), with many short-cut connections specifically designed to propagate fine details from the inputs $\mathbf{x}$ to the high-dimensional output.

To combine the strengths of the U-net model $f_\theta(\mathbf{x}, t)$ used by Ho et al. (2020) with the guarantees of using an energy function, we propose specifying the energy as follows:

$$E_\theta(\mathbf{x}, t) = \frac{1}{2\sigma(t)} || \,\mathbf{x} - f_\theta(\mathbf{x}, t)||^2, \tag{5}$$

When taking the gradient of the energy function in (5), we get

$$-\mathbf{s}_\theta(\mathbf{x},\mathrm{t}) = \nabla_{\mathbf{x}} E_\theta(\mathbf{x},t) = \frac{1}{\sigma(t)}(\mathbf{x} - f_\theta(\mathbf{x},t)) - \frac{1}{\sigma(t)}(\mathbf{x} - f_\theta(\mathbf{x},t))\nabla_{\mathbf{x}} f_\theta(\mathbf{x},t). \tag{6}$$

The first term in this equation, $(\mathbf{x} - f_\theta(\mathbf{x},t))/\sigma(t)$ is equivalent to one of the score models explored by Song & Ermon (2019), and is closely related to the model of Ho et al. (2020) that directly predicts the noise that was used to perturb $\mathbf{x}$. The second term in the equation is new, and this is what guarantees the score model $\mathbf{s}_\theta(\mathbf{x},\mathrm{t})$ to be a conservative vector field. As we show in Section 4, this additional term does not hurt the performance of the generative model.

4 EXPERIMENTS

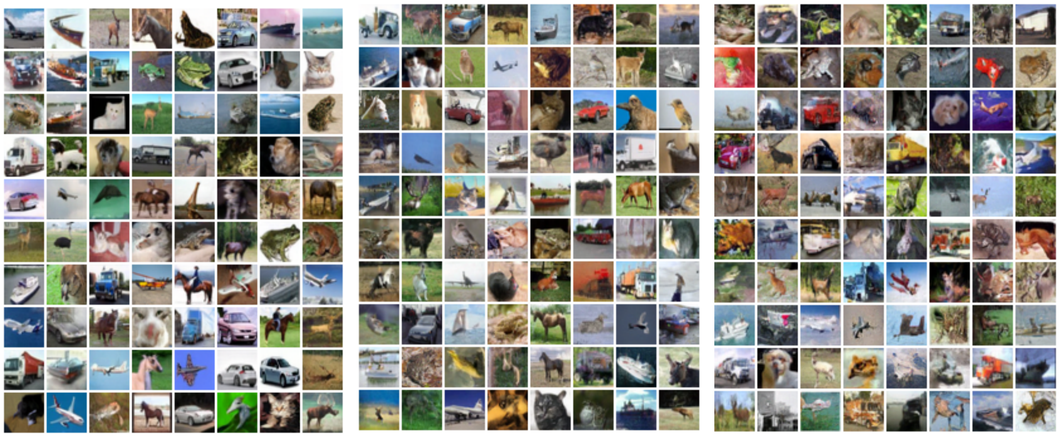
To empirically investigate the questions posed above, we train unconditional generative models on the CIFAR-10 dataset of small images. Here we compare the three different model types discussed in Section 3: A: an unconstrained U-net style model as used successfully by Ho et al. (2020), corresponding to just the first term of (6), B: an energy-based constrained U-net model corresponding to both terms of (6), guaranteed to give a conservative vector field, and C: an energy-based model based on a standard feedforward ResNet (He et al., 2016) as often used for image classification. Otherwise we follow the same setup used by Ho et al. (2020) in their experiments.

4.1 ENERGY BASED MODELS PERFORM ON PAR WITH UNCONSTRAINED SCORE MODELS

As Table 1 shows, energy based models perform on par with unconstrained score models when keeping the model structure the same, while standard feedforward energy models perform less well.

Model type	Conservative?	Short-cuts to inputs?	FID ↓	Inception Score ↑
A: Unconstrained U-net	no	yes	6.5	9.2
B: Energy-based U-net	yes	yes	6.8	9.3
C: Energy-based ResNet	yes	no	21	7.8

Table 1: Results on unconditional CIFAR-10 image modeling after 300k steps of training. The constrained and unconstrained U-net models perform similarly, with one being better on FID and the other being better on Inception Score. The model with the more commonly used feedforward ResNet energy model does not perform well.



(a) Unconstrained U-Net

(b) Energy-based U-Net

(c) Feedforward ResNet

Figure 1: Unconditional CIFAR-10 samples from the three score models we consider.

5 Conclusion

The recent success of score based generative modeling gives rise to some fundamental questions about energy based models: Do we actually need to model an energy function, or is it enough to model its gradient directly? Can unconstrained neural networks learn to approximate the gradient of a function, or does this constraint need to be built in? Are energy based models, i.e. constrained score models, empirically less powerful than unconstrained models? Here we explore some first steps in trying to answer these questions: By comparing constrained and unconstrained score models in a way that minimizes the difference in the resulting model structure we find that constrained score models, i.e. energy based models, can perform just as well as unconstrained models for image generation. This suggests that future work in energy-based models and score-based models can focus on concrete modeling architectures, without getting distracted by differences in the model formalism, enabling the two approaches to build on each other's results.

References

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saeed Saremi. On approximating ∇f with neural networks. *arXiv preprint arXiv:1910.12744*, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2019.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.