Discriminator Contrastive Divergence: Semi-Amortized Generative Modeling by Exploring Energy of the Discriminator

Anonymous Author(s)
Affiliation
Address
email

Abstract

Generative Adversarial Networks (GANs) have shown great promise in modeling high dimensional data. The learning objective of GANs usually minimizes some measure discrepancy, *e.g.*, f-divergence (f-GANs) or Integral Probability Metric (Wasserstein GANs). With f-divergence as the objective function, the discriminator essentially estimates the density ratio, and the estimated ratio proves useful in further improving the sample quality of the generator. However, how to leverage the information contained in the discriminator of Wasserstein GANs (WGAN) is less explored. In this paper, we introduce the Discriminator Contrastive Divergence, which is well motivated by the property of WGAN's discriminator and the relationship between WGAN and energy-based model. Compared to standard GANs, where the generator is directly utilized to obtain new samples, our method proposes a semi-amortized generation procedure where the samples are produced with the generator's output as an initial state. Then several steps of Langevin dynamics are conducted using the gradient of the discriminator. We demonstrate the benefits of significantly improved generation on both synthetic data and several real-world image generation benchmarks.

1 Introduction

Generative Adversarial Networks (GANs) [10] proposes a popular way to learn likelihood-free generative models, which have shown promising results on various challenging tasks. Specifically, GANs are learned by finding the equilibrium of a min-max game between a generator and a discriminator or a critic. Assuming the optimal discriminator can be obtained, the generator substantially minimizes some discrepancy between the generated distribution and the target distribution.

Improving training GANs by exploring the discrepancy measure with the excellent property has stimulated fruitful lines of research works and is still an active area. Two well-known discrepancy measures for training GANs are f-divergence and Integral Probability Metric (IPM) [26]. f-divergence is severe for directly minimization due to the intractable integral, f-GANs provide minimization instead of a variational approximation of f-divergence between the generated distribution p_{G_θ} and the target distribution p_{data}. The discriminator in f-GANs serves as a density ratio estimator [36]. The other families of GANs are based on the minimization of an Integral Probability Metric (IPM). According to the definition of IPM, the critic needs to be constrained into a specific function class. When the critic is restricted to be 1-Lipschitz function, the corresponding IPM turns to the Wasserstein-1 distance, which inspires the approaches of Wasserstein GANs (WGANs) [25, 1, 13].

No matter what kind of discrepancy is evaluated and minimized, the discriminator is usually discarded at the end of the training, and only the generator is kept to generate samples. A natural question to ask is whether, and how we can leverage the remaining information in the discriminator to construct a more superior distribution than simply sampling from a generator.

Recent work [2, 35] has shown that a density ratio can be obtained through the output of discriminator, and a more superior distribution can be acquired by conducting rejection sampling or Metropolis-Hastings sampling with the estimated density ratio based on the original GAN [10].

However, the critical limitation of previous methods lies in that they can not be adapted to WGANs, which enjoy superior empirical performance over other variants. How to leverage the information of a WGAN's critic model to improve image generation remains an open problem. In this paper, we do the following to address this:

- We provide a generalized view to unify different families of GANs by investigating the informativeness of the discriminators.
- We propose a semi-amortized generative modeling procedure so-called discriminator contrastive divergence (DCD), which achieves an intermediate between implicit and explicit generation and hence allows a trade-off between generation quality and speed.

Extensive experiments are conducted to demonstrate the efficacy of our proposed method on both synthetic setting and real-world generation scenarios, which achieves state-of-the-art performance on several standard evaluation benchmarks of image generation.

2 Methodology

We first introduce the Fenchel dual of the intractable partition function Z_θ in Eq. 8:

Theorem 1. *[38] With $H(q) = -\int q(x)\log q(x)dx$, the Fenchel dual of log-partition Z_θ is as follows:*

$$A(E_\theta) = \max_{q\in\mathcal{P}} \langle q(x), E_\theta(x)\rangle + H(q), \tag{1}$$

where $\mathcal{P}$ denotes the space of distributions, and $\langle q(x), E_\theta(x)\rangle = \int E_\theta(x)q(x)dx$.

We put the Fenchel dual of $A(E_\theta)$ back into the MLE objective in Eq. 9, we achieve the following min-max game formalization for training energy-based model based on MLE:

$$\min_{q\in\mathcal{P}} \max_{E_\theta\in\mathcal{E}} \underbrace{\mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}}\left[E_\theta(\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{x}\sim q}\left[E_\theta(\boldsymbol{x})\right]}_{\text{WGAN's objective for critic}} - \underbrace{H(q)}_{\text{entropy regularization}}. \tag{2}$$

The Fenchel dual view of MLE training in the energy-based model explicitly illustrates the gap and connection between the WGAN and Energy based model. If we consider the dual distribution q as the generated distribution p_{G_θ}, and the D_ϕ as the energy function E_θ. The duality form for training energy-based models is essentially the WGAN's objective with the entropy of the generator is regularized.

Hence to turn the discriminator in WGAN into an energy function, we may conduct several fine-tuning steps, as illustrated in Eq. 2. Note that maximizing the entropy of the p_{G_θ} is indeed a challenging task, which needs to either use a tractable density generator, *e.g.*, normalizing Flows [7], or maximize the mutual information between the latent variable $\mathcal{Z}$ and the corresponding $G_\theta(Z)$ when the G_θ is a deterministic mapping. However, instead of maximizing the entropy of the generated distribution p_{G_θ} directly, we derive our method based on the following fact:

Proposition 1. *[19] Update the generated distribution p_{G_θ} according to the gradient estimated through Equation. 2, essentially minimized the Kullback–Leibler (KL) divergence between p_{G_θ} and the distribution p_{D_ϕ}, which refers to the distribution implied by using D_ϕ as the energy function, as illustrated in Eq. 8,* i.e. $D_{\mathrm{KL}}(p_{G_\theta}||p_{D_\phi})$.

To avoid the computation of $H(p_{G_\theta})$, motivated by the monotonic property of MCMC, as illustrated in Eq. 11, we propose Discriminator Contrastive Divergence (DCD), which replaces the gradient-based optimization on $q(p_{G_\theta})$ in Eq. 2 with several steps of MCMC for finetuning the critic in WGAN into an energy function. To be more specific, we use Langevin dynamics[33] which leverages the gradient of the discriminator to conduct sampling:

$$x_k = x_{k-1} - \frac{\epsilon}{2}\nabla_x D_\phi\left(x_{k-1}\right) + \sqrt{\epsilon}\omega, \omega \sim \mathcal{N}(0, \mathcal{I}), \tag{3}$$

where ϵ refers to the step size. The GAN-based approaches are implicitly constrained by the dimension of the latent noise, which is based on a widely applied assumption that the high dimensional data, *e.g.*, images, actually distribute on a relatively low-dimensional manifold. Apart from searching the reasonable point in the data space, we could also find the lower energy part of the latent manifold by conducting Langevin dynamics in the latent space which are more stable in practice, *i.e.*:

$$z_t^l = z_t^{l-1} - \frac{\epsilon}{2}\nabla_z D_\phi\left(G_\theta(z_t)^{l-1}\right) + \sqrt{\epsilon}\omega, \omega \sim \mathcal{N}(0, \mathcal{I}). \tag{4}$$

Ideally, the proposal should be accepted or rejected according to the Metropolis–Hastings algorithm:

$$\alpha := \min\left\{1, \frac{D_\phi\left(x_k\right) q\left(x_{k-1}|x_k\right)}{D_\phi\left(x_{k-1}\right) q\left(x_k|x_{k-1}\right)}\right\}, \tag{5}$$

where q refers to the proposal which is defined as:

$$q\left(x'|x\right) \propto \exp\left(-\frac{1}{4\tau}\left\|x' - x - \tau\nabla\log\pi(x)\right\|_2^2\right). \tag{6}$$

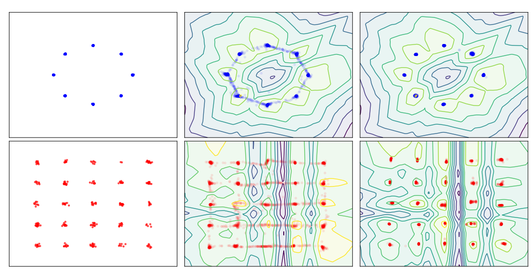
In practice, we find the rejection steps described in Eq. 5 do not boost performance. For simplicity, following [31, 8], we apply Eq. 3 in experiments as an approximate version. The whole tuning procedure is illustrated in Algorithm 1.

After fine-tuning, the discriminator function can be approximated seen as an unnormalized probability function, which implies a unique distribution p_{D_ϕ}. And similar to the p_* implied in the rejection sampling-based method, it is reasonable to assume that p_{D_ϕ} is a superior distribution of p_{G_θ}. Sampling from p_{D_ϕ} can be implemented through the Langevin dynamics, as illustrated in Eq. 3 with p_{G_θ} serves as the initial distribution.

3 Experiments

3.1 Synthetic Density Modeling

Displaying the level sets is a meaningful way to study learned critic. Following the [2, 13], we investigate the impacts of our method on two challenging low-dimensional synthetic settings: twenty-five isotropic Gaussian distributions arranged in a grid and eight Gaussian distributions arranged in a ring (Fig. 1a). For all different settings, both the generator and the discriminator of the WGAN model are implemented as neural networks with four fully connected layers and Relu activations. The Lipschitz constraint is restricted through spectral normalization [25], while the prior is a two-dimensional multivariate Gaussian with a mean of 0 and a standard deviation of 1.



(a) Target (b) SNGAN (c) SNGAN-DCD

Table 1: Density modeling on synthetic distributions. **Top**: 8 Gaussian distribution. **Bottom**: 25 Gaussian distribution. **Left**: Distribution of real data. **Middle**: Distribution defined by the generator of SNGAN. The surface is the level set of the critic. Yellow corresponds to higher value while purple corresponds to lower. **Right**: Distribution defined by the SNGAN-DCD. The surface is the level set of the proposed energy function.

To investigate whether the proposed Discriminator Contrastive Divergence is capable of tuning the distribution induced by the discriminator as desired energy function, *i.e.* p_{D_ϕ}, we visualize both the value surface of the critic and the samples obtained from p_{D_ϕ} with Langevin dynamics. The results are shown in Figure. 1. As can be observed, the original WGAN (Fig. 1b) is strong enough to cover most modes, but there are still some spurious links between two different modes. The enhanced distribution p_{D_ϕ} (Fig. 1c), however, has the ability to reduce spurious links and recovers the modes with underestimated density. More precisely, after the MCMC fine-tuning procedure (Fig. 1c), the gradients of the value surface become more meaningful so that all the regions with high density in data distribution p_{data} are assigned with high D_ϕ value, *i.e.*, lower energy($\exp(-D_\phi)$). By contrast, in the original discriminator (Fig. 1b), the lower energy regions in p_{D_ϕ} are not necessarily consistent with the high-density region of p_{data}.

Model	Inception	FID
CIFAR-10 Unconditional		
PixelCNN [37]	4.60	65.93
PixelIQN [28]	5.29	49.46
EBM [8]	6.02	40.58
WGAN-GP [13]	$7.86 \pm .07$	18.12
MoLM [29]	$7.90 \pm .10$	18.9
SNGAN [25]	$8.22 \pm .05$	21.7
ProgressiveGAN [18]	$8.80 \pm .05$	-
NCSN [31]	$8.87 \pm .12$	25.32
DCGAN w/ DRS [2]	3.073	-
DCGAN w/ MH-GAN [35]	3.379	-
ResNet-SAGAN w/ DOT [32]	$8.50 \pm .12$	19.71
SNGAN-DCD (Pixel)	$8.54 \pm .11$	21.67
SNGAN-DCD (Latent)	$\mathbf{9.11} \pm .04$	$\mathbf{16.24}$
CIFAR-10 Conditional		
EBM [8]	8.30	37.9
SNGAN [25]	$8.43 \pm .09$	15.43
SNGAN-DCD (Pixel)	$8.73 \pm .13$	22.84
SNGAN-DCD (Latent)	$8.81 \pm .11$	15.05
BigGAN [3]	$\mathbf{9.22}$	$\mathbf{14.73}$

Table 2: Inception and FID scores for CIFAR-10.

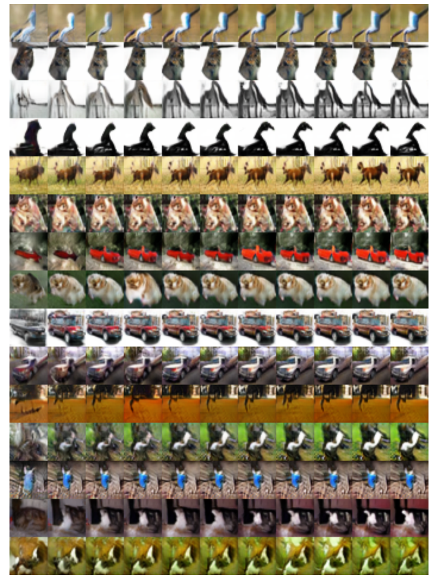


Figure 1: Unconditional CIFAR-10 Langevin dynamics visualization.

3.2 Real-World Image Generation

For quantitative evaluation, we report the inception score [30] and FID [15] scores on CIFAR-10 in Tab. 2. As shown in the Tab. 2, in pixel space, by introducing the proposed DCD algorithm, we achieve a significant improvement of inception score over the SNGAN. The reported inception score is even higher than most values achieved by class-conditional generative models. Our FID score of 21.67 on CIFAR-10 is competitive with other top generative models. When the DCD is conducted in the latent space, we further achieve a 9.11 inception score and a 16.24 FID, which is a new state-of-the-art performance of IS. When combined with label information to perform conditional generation, we further improve the FID to 15.05, which is comparable with current state-of-the-art large-scale trained models [3]. Some visualization of generated examples can be found in Fig 1, which demonstrates that the Markov chain is able to generate more realistic samples, suggesting that the MCMC process is meaningful and effective. Tab. 4 and Tab. 5 shows the performance on STL-10 and ImageNet respectively, which demonstrate that as a generalized method, DCD is not over-fitted to the specific dataset. More experiment details and the generated samples can be found in Appendix. I.

4 Conclusion and Future Work

Based on the density ratio estimation perspective, the discriminator in f-GANs could be adapted to a wide range of applications, *e.g.*, mutual information estimation [17] and bias correction of generative models [12]. However, as another important branch in GANs, the available information in WGANs' discriminator is less explored. In this paper, we narrow down the scope and focus on how to leverage the discriminator of WGANs to further improve the sample quality. We first present a comprehensive theoretical analysis on the informativeness of WGANs' discriminator. Motivated by the theoretical understanding, we investigate the possibility of turning the discriminator of WGANs into an energy function and propose a tuning and sampling procedure named "Discriminator Contrastive Divergence". The final generation process is semi-amortized, where we take the generator as the initial state and then conduct several MCMC steps. Empirical results demonstrate the effectiveness of the proposed method on several tasks. We hope our work can shed some light on a generalized view to a method of connecting different GANs and energy-based models, which will stimulate more exploration into the potential of current deep generative models. One potential direction for future work is to conduct DCD in each layer of the generator. This can be seen as a compromise between the latent and the pixel space, which may lead to further sampling quality improvements.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. Discriminator rejection sampling. *arXiv preprint arXiv:1810.06758*, 2018.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[5] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[8] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *arXiv preprint arXiv:1903.08689*, 2019.

[9] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[11] 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.

[12] Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *arXiv preprint arXiv:1906.09531*, 2019.

[13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017.

[14] 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*, pages 770–778, 2016.

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[16] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2012.

[17] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] Taesup Kim and Yoshua Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[22] Rithesh Kumar, Anirudh Goyal, Aaron Courville, and Yoshua Bengio. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*, 2019.

[23] Yingzhen Li, Richard E Turner, and Qiang Liu. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.

[24] Takeru Miyato and Masanori Koyama. cgans with projection discriminator. *arXiv preprint arXiv:1802.05637*, 2018.

[25] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[26] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[27] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

[28] Georg Ostrovski, Will Dabney, and Rémi Munos. Autoregressive quantile networks for generative modeling. *arXiv preprint arXiv:1806.05575*, 2018.

[29] Suman Ravuri, Shakir Mohamed, Mihaela Rosca, and Oriol Vinyals. Learning implicit generative models with the method of learned moments. *arXiv preprint arXiv:1806.11006*, 2018.

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, pages 11895–11907, 2019.

[32] Akinori Tanaka. Discriminator optimal transport. *arXiv preprint arXiv:1910.06832*, 2019.

[33] Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec): 1235–1260, 2003.

[34] Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844*, 2015.

[35] Ryan Turner, Jane Hung, Yunus Saatci, and Jason Yosinski. Metropolis-hastings generative adversarial networks. *arXiv preprint arXiv:1811.11357*, 2018.

[36] Masatoshi Uehara, Issei Sato, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

[37] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016.

[38] Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

[39] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[40] Zhiming Zhou, Jiadong Liang, Yuxuan Song, Lantao Yu, Hongwei Wang, Weinan Zhang, Yong Yu, and Zhihua Zhang. Lipschitz generative adversarial nets. *arXiv preprint arXiv:1902.05687*, 2019.

A Related Works

Both empirical [1] and theoretical [15] evidence has demonstrated that learning a discriminative model with neural networks is relatively easy, and the neural generative model (sampler) is prone to reach its bottleneck during the optimization. Hence, there is strong motivation to further improve the generated distribution by exploring the remaining information. Two recent advancements are discriminator rejection sampling (DRS) [2] and MH-GANs [35]. DRS conducts rejection sampling on the output of the generator. The vital limitation that lies in the upper bound of D_ϕ is needed to be estimated for computing the rejection probability. MH-GAN sidesteps the above problem by introducing a Metropolis-Hastings sampling procedure with generator acting as the independent proposal; the state transition is estimated with a well-calibrated discriminator. However, the theoretical justification of both the above two methods is based on the fact that the output of discriminator needs to be viewed as an estimation of density ratio $\frac{p_{\text{data}}}{p_{G_\theta}}$. As pointed out by previous work [40], the output of a discriminator in WGAN [1] suffers from the free offset and can not provide the density ratio, which prevents the application of the above methods in WGAN.

Our work is inspired by recent theoretical studies on the property of discriminator in WGANs [13, 40]. [32] proposes discriminator optimal transport (DOT) to leverage the optimal transport plan implied by WGANs' discriminator, which is orthogonal to our method. Besides, turning the discriminator of WGAN into an energy function is closely related to the amortized generation methods in energy-based model (EBM) literature [19, 39, 22] where a separate network is proposed to learn to sample from the partition function in [9]. Recent progress [31, 8] in the area of EBM has shown the feasibility of generating high dimensional data with Langevin dynamics. From the perspective of EBM, our proposed method can be seen as an intermediary between an amortized generative model and an implicit generative model, *i.e.*, a semi-amortized generation method, which allows a trade-off between speed and quality of generation. With a similar spirit, [11] also illustrates the potential connection between neural classifier and energy-based model in supervised and semi-supervised scenarios.

B Preliminaries

B.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [10] is an implicit generative model that aims to fit an empirical data distribution p_{data} over sample space $\mathcal{X}$. The generative distribution p_{G_θ} is implied by a generated function G_θ, which maps latent variable Z to sample X, *i.e.*, $G_\theta : \mathcal{Z} \to \mathcal{X}$. Typically, the latent variable Z is distributed on a fixed prior distribution $p(z)$. With i.i.d samples available from p_{G_θ} and p_{data}, the GAN typically learns the generative model through a min-max game between a discriminator D_ϕ and a generator G_θ:

$$\min_\theta \max_\phi \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}} \left[r(D_\phi(\boldsymbol{x})) \right] - \mathbb{E}_{\boldsymbol{x}\sim p_{G_\theta}} \left[m(D_\phi(\boldsymbol{x})) \right]. \tag{7}$$

With r and m as the function $r(x) = m(x) = x$ and the $D_\phi(x)$ is constrained as 1-Lipschitz function, the Eq. 7 yields the WGANs objective which essentially minimizes the Wasserstein distance between p_{data} and p_{G_θ}. With $r(x) = x$ and $m(x)$ as the Fenchel conjugate[16] of a convex and lower-semicontinuous function, the objective in Eq. 7 approximately minimize a variational estimation of f-divergence[27] between p_{data} and p_{G_θ}.

B.2 Energy Based Model and MCMC basics

The energy-based model tends to learn an unnormalized probability model implied by an energy function $E_\theta(x)$ to prescribe the ground truth data distribution p_{data}. The corresponding normalized density function is:

$$q_\theta(x) = \frac{e^{-E_\theta(x)}}{Z_\theta}, \quad Z_\theta = \int e^{-E_\theta(x)} \mathrm{d}x, \tag{8}$$

where Z_θ is so-called normalization constant. The objective of training an energy-based model with maximum likelihood estimation is as:

$$\mathcal{L}_{\text{MLE}}(\theta; p) := -\mathbb{E}_{x\sim p_{\text{data}}(x)} \left[\log q_\theta(x) \right]. \tag{9}$$

The estimated gradient with respect to the maximum likelihood estimation objective is as follows:

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\mathrm{MLE}}(\theta; p) &= \nabla_\theta \mathbb{E}_{x\sim p_{\text{data}}(x)}[E_\theta(x)] - \frac{\int e^{-E_\theta(x)} \nabla_\theta E_\theta(x) \mathrm{d}x}{Z_\theta} \\ &= \mathbb{E}_{x\sim p_{\text{data}}(x)}[\nabla_\theta E_\theta(x)] - \mathbb{E}_{x\sim q_\theta(x)}[\nabla_\theta E_\theta(x)]. \end{aligned} \tag{10}$$

The above method for gradient estimation in Equation 10 is called contrastive divergence (CD).

Markov chain Monte Carlo is a powerful framework for drawing samples from a given distribution. An MCMC is specified by a transition kernel $\mathcal{K}(x'|x)$ which corresponds to a unique stationary distribution p. More specifically, MCMC can be viewed as drawing x_0 from the initial distribution q_0 and iteratively get sample x_t at the t-th iteration by applied the transition kernel on the previous step, *i.e.*, $x_t|x_{t-1} \sim \mathcal{K}(x_t|x_{t-1})$. Following [23], we formalized the distribution q_t of x_t as obtained by a fixed point update of form $q_t(x) \leftarrow \mathcal{K}q_{t-1}(x)$, and $\mathcal{K}q_{t-1}(x)$:

$$\mathcal{K}q_{t-1}(x) := \int q_{t-1}(x')\, \mathcal{K}(x|x')\, dx'.$$

As indicated by the standard theory of MCMC, the following monotonic property is satisfied:

$$D_{\mathrm{KL}}(q_t||p) \leq D_{\mathrm{KL}}(q_{t-1}||p). \tag{11}$$

And q_t converges to the stationary distribution p as $t \to \infty$.

B.3 Informativeness of WGAN Discriminator

So far, it is well known that the discriminator D_ϕ in f-GAN is optimized to estimate a statistic related to the density ratio between $\frac{p_{\text{data}}}{p_{G_\theta}}$ [2]. In this section, we seek to investigate the following questions:

- What kind of information is contained in the discriminator of WGANs?
- Why and how can the information be utilized to further improved the quality of generated distribution?

Different from f-GANs, the objective of WGANs is derived from the Integral Probability Metric, and the discriminator can not naturally be derived as an estimated density ratio. Before leveraging the remaining information in the discriminator, the property of the discriminator in WGANs needs to be investigated first. We introduce the primal problem implied by WGANs objective as follows:

Let π denote the joint probability for transportation between P and Q, which satisfies the marginality conditions,

$$\int d\boldsymbol{y}\pi(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x}), \quad \int d\boldsymbol{x}\pi(\boldsymbol{x},\boldsymbol{y}) = q(\boldsymbol{y}) \tag{12}$$

The primal form first-order Wasserstein distance W_1 is defined as:

$$W_1(\mathcal{P}, \mathcal{Q}) = \inf_{\pi \in \Pi(\mathcal{P},\mathcal{Q})} \mathbb{E}_{(x,y)\sim\pi}[\|x-y\|_2]$$

the objective function of the discriminator in Wasserstein GANs is the Kantorovich-Rubinstein duality of Eq. 12, and the optimal discriminator has the following property[13]:

Theorem 2. *Let π^* as the optimal transport plan in Eq. 12 and $x_t = tx + (1-t)y$ with $0 \leq t \leq 1$. With the optimal discriminator D_ϕ as a differentiable function and $\pi^*(x,x) = 0$ for all x, then it holds that:*

$$\mathrm{P}_{(x,y)\sim\pi^*}\left[\nabla_{x_i} D_\phi^*(x_t) = \frac{y-x}{\|y-x\|}\right] = 1$$

Theorem. 2 states that for each sample x in the generated distribution p_{G_θ}, the gradient on the x directly points to a sample y in the p_{data}, where the (x, y) pairs are consistent with the optimal transport plan π^*. All the linear interpolations x_t between x and y satisfy that $\nabla_{x_k} D_\phi^*(x_t) = \frac{y-x}{\|y-x\|}$. It should also be noted that similar results can also be drawn in some variants of WGANs, whose loss functions may have a slight difference with standard WGAN [40]. For example, the SNGAN uses the hinge loss during the optimization of the discriminator, *i.e.*, $r(\cdot)$ and $g(\cdot)$ in Eq. 7 is selected as

$\max(0, -1-u)$ for stabilizing the training procedure. We provide a detailed discussion on several surrogate objectives in Appendix. H.

The above property of discriminator in WGANs can be interpreted as that given a sample x from generated distribution p_{G_θ} we can obtain a corresponding y in data distribution p_{data} by directly conducting gradient decent with the optimal discriminator D^*_ϕ:

$$y = x + w_x * \nabla_x D^*_\phi, \quad w_x \geq 0 \tag{13}$$

It seems to be a simple and appealing solution to improve p_{G_θ} with the guidance of discriminator D_ϕ. However, the following issues exist:

1) there is no theoretical indication on how to set w_x for each sample x in generated distribution. We noticed that a concurrent work [32] introduce a search process called Discriminator Optimal Transport(DOT) by finding the corresponding y^* through the following:

$$y_x = \arg\min_{\boldsymbol{y}} \left\{ \|\boldsymbol{y} - \boldsymbol{x}\|_2 - D^*_\phi(\boldsymbol{y}) \right\} \tag{14}$$

However, it should be noticed that Eq. 14 has a non-unique solution. We further extend the fact into the following theorem:

Theorem 3. *With the π^* and D^*_ϕ as the optimal solutions of the primal problem in Eq. 12 and Kantorovich-Rubinstein duality of Eq. 12, the distribution p_{ot} implied by the generated distribution $p_{G_\theta}$and the discriminator D^*_ϕ is defined as (y_x is defined in Eq. 14):*

$$p_{ot}(\boldsymbol{y}) = \int d\boldsymbol{x} \delta(y - y_x) p_{G_\theta}(\boldsymbol{x})$$

when $p_{data} \neq p_{G_\theta}$, there exists infinite numbers of p_{ot} with p_{data} as a special case.

Theorem 3 provides a theoretical justification for the poor empirical performance of conducting DOT in the sample space, as shown in their paper.

2) Another problem lies in that samples distributed outside the generated distribution (p_{G_θ}) are never explored during training, which results in much adversarial noise during the gradient-based search process, especially when the sample space is high dimensional such as real-world images.

To fix the issues mentioned above in leveraging the information of discriminator in Wasserstein GANs, we propose viewing the discriminator as an energy function. With the discriminator as an energy function, the stationary distribution is unique, and Langevin dynamics can approximately conduct sampling from the stationary distribution. Due to the monotonic property of MCMC, there will not be issues like setting w_x in Eq. 13. Besides, the second issue can also be easily solved by fine-tuning the energy spaces with contrastive divergence. In addition to the benefits illustrated above, if the discriminator is an energy function, the samples from the corresponding energy-based model can be obtained through Langevin dynamics by using the gradients of the discriminator which takes advantage of the property of discriminator as shown in Theorem 2. With all the facts as mentioned above, there is strong motivation to explore further and bridge the gap between discriminator in WGAN and the energy-based model.

B.4 Semi-Amortized Generation with Langevin Dynamics

B.5 Real-World Image Generation

To quantitatively and empirically study the proposed DCD approach, in this section, we conduct experiments on unsupervised real-world image generation with DCD and its related counterparts. On several commonly used image datasets, experiments demonstrate that our proposed DCD algorithm can always achieve better performance on different benchmarks with a significant margin.

B.6 Experimental setup

Baselines. We evaluated the following models as our baselines: we take PixelCNN [37], PixelIQN [28], and MoLM [29] as representatives of other types of generative models. For the energy-based model, we compared the proposed method with EBM [8] and NCSN [31]. For GAN models,

Algorithm 1 Discriminator Contrastive Divergence

1: **Input:** Pretrained generator G_θ, discriminator D_ϕ.
2: Set the step size ϵ, the length of MCMC steps K and the total iterations T.
3: **for** iteration $i = 1, \cdots, T$ **do**
4: Sample a batch of data samples $\{x_t\}_{t=1}^m$ for empirical data distribution p_{data} and $\{z_t\}_{t=1}^m$ for the prior distribution $p(z)$.
5: **for** iteration $l = 1, \cdots, K$ **do**
6: **Pixel Space:** $G_\theta(z_t)^l = G_\theta(z_t)^{l-1} - \frac{\epsilon}{2}\nabla_x D_\phi\left(G_\theta(z_t)^{l-1}\right) + \sqrt{\epsilon}\omega, \omega \sim \mathcal{N}(0, \mathcal{I})$ **or**
7: **Latent Space:** $z_t^l = z_t^{l-1} - \frac{\epsilon}{2}\nabla_z D_\phi\left(G_\theta(z_t)^{l-1}\right) + \sqrt{\epsilon}\omega, \omega \sim \mathcal{N}(0, \mathcal{I})$
8: **end for**
9: Optimized the following objective w.r.t. ϕ:
10: **Pixel Space:** $L = \frac{1}{m}\sum_t (D_\phi(x_t) - D_\phi(G_\theta(z_t)^K))$ **or**
11: **Latent Space:** $L = \frac{1}{m}\sum_t (D_\phi(x_t) - D_\phi(G_\theta(z_t^K)))$
12: **end for**

Model	Inception	FID
CIFAR-10 Unconditional		
PixelCNN [37]	4.60	65.93
PixelIQN [28]	5.29	49.46
EBM [8]	6.02	40.58
WGAN-GP [13]	7.86 ± .07	18.12
MoLM [29]	7.90 ± .10	18.9
SNGAN [25]	8.22 ± .05	21.7
ProgressiveGAN [18]	8.80 ± .05	-
NCSN [31]	8.87 ± .12	25.32
DCGAN w/ DRS [2]	3.073	-
DCGAN w/ MH-GAN [35]	3.379	-
ResNet-SAGAN w/ DOT [32]	8.50 ± .12	19.71
SNGAN-DCD (Pixel)	8.54 ± .11	21.67
SNGAN-DCD (Latent)	**9.11** ± .04	**16.24**
CIFAR-10 Conditional		
EBM [8]	8.30	37.9
SNGAN [25]	8.43 ± .09	15.43
SNGAN-DCD (Pixel)	8.73 ± .13	22.84
SNGAN-DCD (Latent)	8.81 ± .11	15.05
BigGAN [3]	**9.22**	**14.73**

Table 3: Inception and FID scores for CIFAR-10.

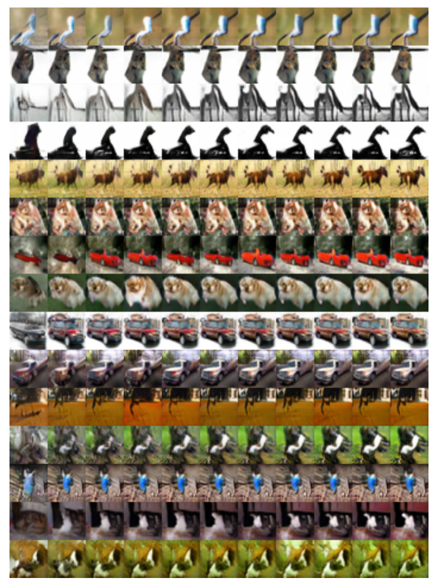


Figure 2: Unconditional CIFAR-10 Langevin dynamics visualization.

we take WGAN-GP [13], Spectral Normalization GAN (SNGAN) [25], and Progressiv eGAN [18] for comparison. We also take the aforementioned DRS [2], DOT [32] and MH-GAN [35] into consideration. The choices of EBM and GANs are due to their close relation to our proposed method, as analyzed in Section 2. We omit other previous GAN methods since as a representative of a state-of-the-art GAN model, SNGAN and Progressive GAN has been shown to rival or outperform several former methods such as the original GAN [10], the energy-based generative adversarial network [39], and the original WGAN with weight clipping [1].

Evaluation Metrics. For evaluation, we concentrate on comparing the quality of generated images since it is well known that GAN models cannot perform reliable likelihood estimations [34]. We choose to compare the Inception Scores [30] and Frechet Inception Distances (FID) [15] reached during training iterations, both computed from 50K samples. A high image quality corresponds to high Inception and low FID scores. Specifically, the intuition of IS is that high-quality images should lead to high confidence in classification, while FID aims to measure the computer-vision-specific similarity of generated images to real ones through Frechet distance.

Data. We use CIFAR-10 [21], STL-10 [4] and ImageNet [6], which are all standard datasets widely used in generative literature. STL-10 consists of unlabeled real-world color images, while CIFAR-10 and ImageNet is provided with class labels, which enables us to conduct conditional generation tasks. For STL-10, we also shrink the images into 32×32 as in previous works.

Network Architecture. For all experiment settings, we follow Spectral Normalization GAN (SNGAN) [25] and adopt the same Residual Network (ResNet) [14] structures and hyperparameters, which presently is the state-of-the-art implementation of WGAN. Details can be found in Appendix. G. We take their open-source code and pre-trained model as the base model for the experiments on CIFAR-10 and ImageNet. For STL-10, since there is no pre-trained model available to reproduce the results, we train the SNGAN from scratch and take it as the base model.

B.6.1 Results

For quantitative evaluation, we report the inception score [30] and FID [15] scores on CIFAR-10 in Tab. 3. As shown in the Tab. 3, in pixel space, by introducing the proposed DCD algorithm, we achieve a significant improvement of inception score over the SNGAN. The reported inception score is even higher than most values achieved by class-conditional generative models. Our FID score of 21.67 on CIFAR-10 is competitive with other top generative models. When the DCD is conducted in the latent space, we further achieve a 9.11 inception score and a 16.24 FID, which is a new state-of-the-art performance of IS. When combined with label information to perform conditional generation, we further improve the FID to 15.05, which is comparable with current state-of-the-art large-scale trained models [3]. Some visualization of generated examples can be found in Fig 2, which demonstrates that the Markov chain is able to generate more realistic samples, suggesting that the MCMC process is meaningful and effective. Tab. 4 and Tab. 5 shows the performance on STL-10 and ImageNet respectively, which demonstrate that as a generalized method, DCD is not over-fitted to the specific dataset. More experiment details and the generated samples can be found in Appendix. I.

Model	Inception	FID
SNGAN [25]	$8.90 \pm .12$	18.73
SNGAN-DCD (Pixel)	$9.25 \pm .09$	22.25
SNGAN-DCD (Latent)	$\mathbf{9.33} \pm .04$	**17.68**

Table 4: Inception and FID scores for STL-10

Model	Inception
cGAN	36.23
cGAN w/ DOT [2]	37.29
SNGAN [25]	36.8
SNGAN-DCD	**38.9**

Table 5: Inception scores for ImageNet

C Broader Impact

It should be noted that the semi-amortized generation allows a trade-off between the generation quality and sampling speed, which holds a slower sampling speed than a direct generation with a generator. Hence the proposed method is suitable to the application scenario where the generation quality is given vital importance. Another interesting observation during the experiments is the discriminator contrastive divergence surprisingly reduces the occurrence of adversarial samples during training, so it should be a promising future direction to investigate the relationship between our method and bayesian adversarial learning.

However, negative consequences also exist since advances in generative models may lead to more realistic fake images, which have the capacity to deceive, emotionally distress, and affect public opinions and actions. To mitigate the risks associated with deep generative models, we encourage researchers to understand and avoid the bad influence of using generative models in particular real-world scenarios.

D Proof of Theorem 2

It should be noticed that Theorem. 2 can be generalized to that Lipschitz continuity with l_2-norm (Euclidean Distance) can guarantee that the gradient is directly pointing towards some sample[40]. We introduce the following lemmas, and Theorem. 2 is a special case.

Let (x, y) be such that $y \neq x$, and we define $x_t = x + t \cdot (y - x)$ with $t \in [0, 1]$.

Lemma 1. *If $f(x)$ is k-Lipschitz with respect to $\|.\|_p$ and $f(y) - f(x) = k\|y-x\|_p$, then $f(x_t) = f(x) + t \cdot k\|y-x\|_p$.*

Proof. As we know $f(x)$ is k-Lipschitz, with the property of norms, we have

$$\begin{aligned} f(y) - f(x) &= f(y) - f(x_t) + f(x_t) - f(x) \\ &\leq f(y) - f(x_t) + k\|x_t - x\|_p = f(y) - f(x_t) + t \cdot k\|y-x\|_p \\ &\leq k\|y - x_t\|_p + t \cdot k\|y-x\|_p = k \cdot (1-t)\|y-x\|_p + t \cdot k\|y-x\|_p \\ &= k\|y-x\|_p. \end{aligned} \tag{15}$$

$f(y) - f(x) = k\|y-x\|_p$ implies all the inequalities is equalities. Therefore, $f(x_t) = f(x) + t \cdot k\|y-x\|_p$. □

Lemma 2. *Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. If $f(x_t) = f(x) + t \cdot k\|y-x\|_2$, then $\frac{\partial f(x_t)}{\partial v}$ equals to k.*

Proof.

$$\begin{aligned} \frac{\partial f(x_t)}{\partial v} &= \lim_{h \to 0} \frac{f(x_t + hv) - f(x_t)}{h} = \lim_{h \to 0} \frac{f(x_t + h\frac{y-x}{\|y-x\|_2}) - f(x_t)}{h} \\ &= \lim_{h \to 0} \frac{f(x_{t+\frac{h}{\|y-x\|_2}}) - f(x_t)}{h} = \lim_{h \to 0} \frac{\frac{h}{\|y-x\|_2} \cdot k\|y-x\|_2}{h} = k. \quad \square \end{aligned}$$

Then we derive the formal proof of Theorem 2.

Proof. Assume $p = 2$, if $f(x)$ is k-Lipschitz with respect to $\|.\|_2$ and $f(x)$ is differentiable at x_t, then $\|\nabla f(x_t)\|_2 \leq k$. Let v be the unit vector $\frac{y-x}{\|y-x\|_2}$. We have

$$k^2 = k\frac{\partial f(x_t)}{\partial v} = k\,\langle v, \nabla f(x_t)\rangle = \langle kv, \nabla f(x_t)\rangle \leq \|kv\|_2\|\nabla f(x_t)\|_2 = k^2. \tag{16}$$

Because the equality holds only when $\nabla f(x_t) = kv = k\frac{y-x}{\|y-x\|_2}$, we have that $\nabla f(x_t) = k\frac{y-x}{\|y-x\|_2}$. □

E Proof of Theorem 3

Theorem. 3 states that following the following procedure as introduced in [32], there is non-unique stationary distribution. The complete procedure is to find the following y for $x \sim P_{G_\theta}$:

$$y^* = \arg\min_x\{\|x-y\|_2 - D(x)\}. \tag{17}$$

To find the corresponding y^*, the following gradient based update is conducted:

$$\{x \leftarrow x - \epsilon\nabla_x\left\{||x-y||_2 - D(x)\right\}. \tag{18}$$

For all the points x_t in the linear interpolation of x and target y^* as defined in the proof of Theorem 2,

$$\nabla_{x_t}\left\{||x_t-y||_2 - D(x_t)\right\} = \frac{y-x}{\|y-x\|_2} - \frac{y-x}{\|y-x\|_2} = 0, \tag{19}$$

which indicates all points in the linear interpolation satisfy the stationary condition.

F Proof of Proposition 1

Proposition. 1 is the direct result of the following Lemma. 3. Following [23], we provide the complete proof as following.

Lemma 3. *[5] Let q and r be two distributions for $\boldsymbol{z}_0$. Let q_t and r_t be the corresponded distributions of state $\boldsymbol{z}_t$ at time t, induced by the transition kernel $\mathcal{K}$. Then $\mathrm{D}_{KL}[q_t||r_t] \geq \mathrm{D}_{KL}[q_{t+1}||r_{t+1}]$ for all $t \geq 0$.*

Proof.

$$
\begin{aligned}
\mathrm{D}_{\mathrm{KL}}[q_t||r_t] &= \mathbb{E}_{q_t}\left[\log\frac{q_t(\boldsymbol{z}_t)}{r_t(\boldsymbol{z}_t)}\right] \\
&= \mathbb{E}_{q_t(\boldsymbol{z}_t)\mathcal{K}(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)}\left[\log\frac{q_t(\boldsymbol{z}_t)\mathcal{K}(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)}{r_t(\boldsymbol{z}_t)\mathcal{K}(\boldsymbol{z}_{t+1}|\boldsymbol{z}_t)}\right] \\
&= \mathbb{E}_{q_{t+1}(\boldsymbol{z}_{t+1})q_{t+1}(\boldsymbol{z}_t|\boldsymbol{z}_{t+1})}\left[\log\frac{q_{t+1}(\boldsymbol{z}_{t+1})q(\boldsymbol{z}_t|\boldsymbol{z}_{t+1})}{r_{t+1}(\boldsymbol{z}_{t+1})r(\boldsymbol{z}_t|\boldsymbol{z}_{t+1})}\right] \\
&= \mathrm{D}_{\mathrm{KL}}[q_{t+1}||r_{t+1}] + \mathbb{E}_{q_{t+1}}\mathrm{D}_{\mathrm{KL}}[q_{t+1}(\boldsymbol{z}_t|\boldsymbol{z}_{t+1})||r_{t+1}(\boldsymbol{z}_t|\boldsymbol{z}_{t+1})].
\end{aligned}
$$

□

G Network architectures

The ResNet architectures for CIFAR-10 and STL-10 datasets are shown in Tab. 6, which are similar to the ones in [13]. For the ImageNet datasets, we follow the ResNet architectures in [25]. The details are shown in Tab. 7.

Table 6: ResNet architectures for CIFAR-10 and STL-10 datasets.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times 256$
ResBlock up 256
ResBlock up 256
ResBlock up 256
BN, ReLU, 3×3 conv, 3 Tanh

(a) Generator

RGB image $x \in \mathbb{R}^{32\times32\times3}$
ResBlock down 128
ResBlock down 128
ResBlock 128
ResBlock 128
ReLU
Global sum pooling
dense $\rightarrow$ 1

(b) Discriminator

Table 7: ResNet architectures of the Generator for ImageNet dataset. As for the model of the *projection discriminator*, we used the same architecture used in [24]. Please see the paper for the details.

$z \in \mathbb{R}^{128} \sim \mathcal{N}(0, I)$
dense, $4 \times 4 \times 1024$
ResBlock up 1024
ResBlock up 512
ResBlock up 256
ResBlock up 128
ResBlock up 64
BN, ReLU, 3×3 conv 3
Tanh

(a) Generator

H Discussions on Objective Functions

Optimization of the standard objective of WGAN, *i.e.* with $r(x) = m(x) = x$ in Eq. 7, are found to be unstable due to the numerical issues and free offset [40, 25]. Instead, several surrogate losses are actually used in practice. For example, the logistic loss($r(x) = m(x) = -\log(1 + e^{-x})$) and hinge loss($r(x) = m(x) = \min(0, x)$) are two widely applied objectives. Such surrogate losses are valid due to that they are actually the lower bounds of the Wasserstein distance between the two distributions of interest. The statement can be easily derived by the fact that $-\log(1 + e^{-x}) \leq x$ and $\min(0, x) \leq x$. A more detailed discussion could also be found in [32].

Note that $\min(0, -1 + x)$ and $-\log(1 + e^{-x})$ are in the function family proposed in [40], and Theorem 4 in [40] guarantees the gradient property of discriminator.

I More Experiment Details

I.1 CIFAR-10

For the meta-parameters in DCD Algorithm 1, when the MCMC process is conducted in the pixel space, we choose $6-8$ as the number of MCMC steps K, and set the step size ϵ as 10 and the standard deviation of the Gaussian noise as 0.01, while for the latent space we set K as 50, ϵ as 0.2 and the deviation as 0.1. Adam optimizer [20] is set with 2×10^{-4} learning rate with $\beta_1 = 0, \beta_2 = 0.9$. We use 5 critic updates per generator update, and a batch size of 64.

I.2 STL-10

We show generated samples of DCD during Langevin dynamics in Fig. 3. We run 150 steps of MCMC steps and plot generated samples for every 10 iterations. The step size is set as 0.05 and the noise is set as $N(0, 0.1)$.

I.3 ImageNet

We show generated samples of DCD during Langevin dynamics in Fig. 4. We run 1000 Langevin dynamics steps and plot generated samples for every 100 iterations. The initial step size and the Gaussian noise are set as 0.05 and $N(0, 0.1)$ respectively. The step size and standard deviation of Gaussian noise are simultaneously decayed with a factor 0.3 for every 100 iterations.

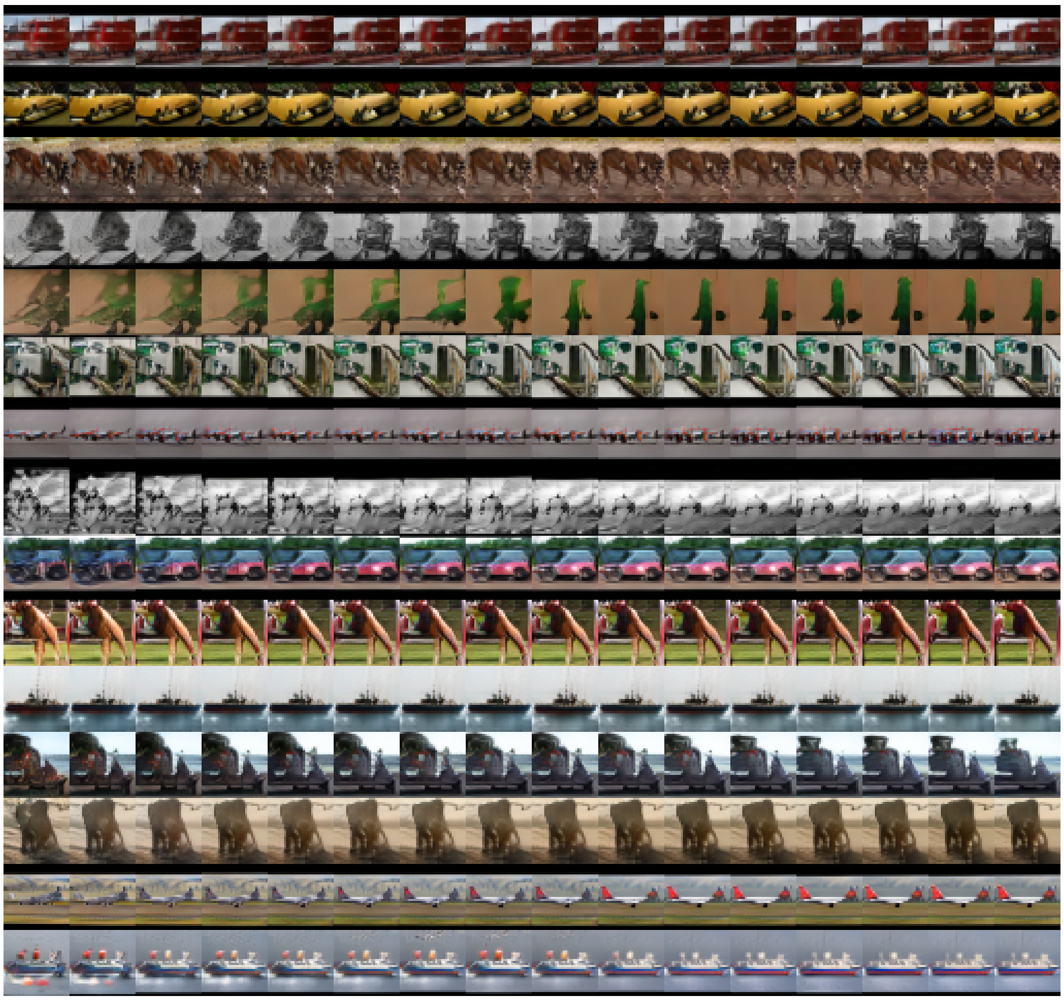


Figure 3: STL-10 Langevin dynamics visualization.

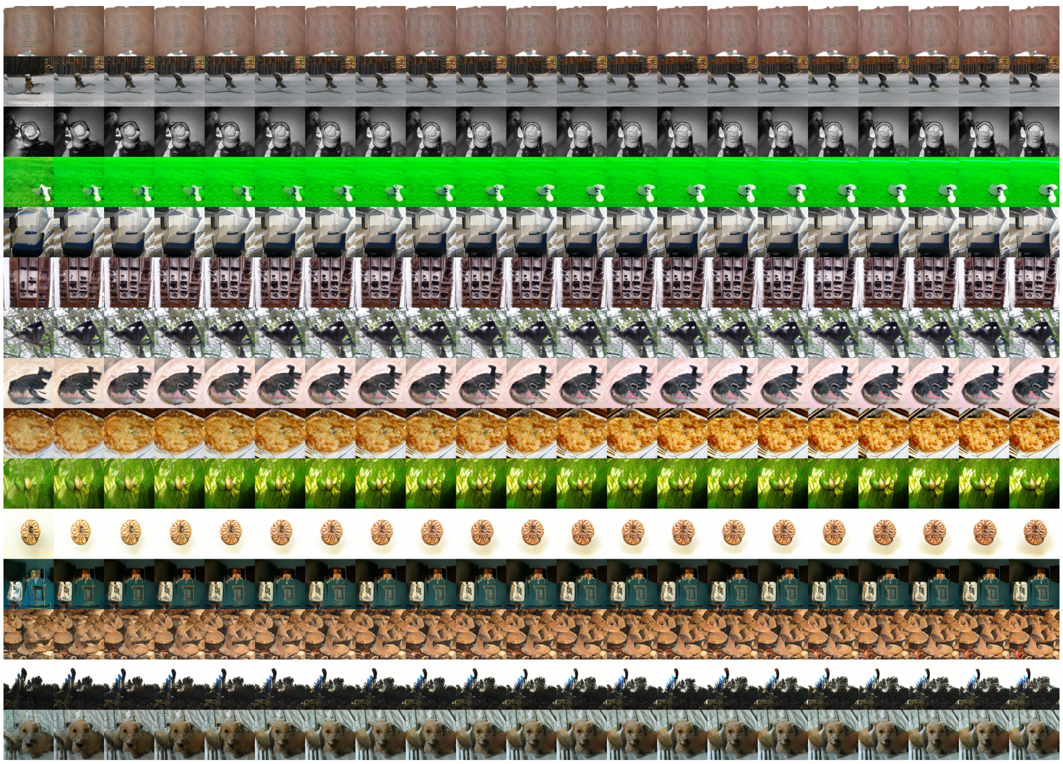


Figure 4: ImageNet Langevin dynamics visualization.