

Flow-based 生成モデルの学習安定化手法の提案

○立浪祐貴 瀧雅人 (立教大学)

Methods for Stabilizing Flow-based Generative Models.

*Y. Tatsunami and M. Taki (Rikkyo University)

Abstract— Nowadays, flow-based generative models have been attracting attention from the perspective that they are capable of not only generation but also density estimation. On the other hand, the models require more computational resources in exchange for its reversibility than other generative models. As a result, environments where there are no sufficient computational resources are forced to train in small mini-batch sizes. The condition induces problems such as divergence of the objective function and outputs which the models generate being outside the expected value range. In our work, we focus on Jacobian determinants and initializations of parameters, of layers that compose the models, especially GLOW, and propose new improvement methods. Additionally, we use some benchmark image datasets to attest that our techniques can contribute to stable learning.

Key Words: Deep generative models, Regularization, Initialization of parameters

1 導入

近年、深層学習による生成モデル、特に敵対的生成ネットワーク (GAN)¹²⁾、変分オートエンコーダー (VAE)¹⁷⁾ は大きな注目を集めている。これらのモデルはサンプルデータを1回の順伝播によって望むデータを生成することができるが、正確な密度推定を行うことはできない。一方で、MADE¹¹⁾、PixelRNN²⁷⁾、PixelCNN²⁶⁾ のような自己回帰モデルは正確な密度推定は可能であるものの、出力次元にわたって連続的にループをする必要があり、サンプリングコストが高い。これらの深層生成モデルのデメリットを克服するモデルに Flow-based 生成モデルの一種である Coupling Flow がある。Coupling Flow は⁶⁾NICE、RealNVP⁷⁾、GLOW¹⁶⁾ などがよく知られている。このネットワークは、Coupling と呼ばれるモジュールを含む、可逆な写像の合成から成る。正確な密度推定と、シンプルな最尤法で学習が可能である。さらに、等方性正規分布などの計算しやすい分布を仮定した潜在空間から学習時に使った写像の逆写像を用いて写像することで、容易にデータをサンプリングすることができる。中でも、GLOW¹⁶⁾ は定量的に良い結果が報告されている上に、高品質な画像を生成することもできる。更に、GLOW は GAN のようにモード崩壊²⁴⁾ に悩まされることはない。

一方で Coupling Flow には可逆であるがゆえに使用するメモリ資源が多いという課題もある。従って、十分なメモリ資源を持たない環境においては、学習時のバッチサイズを小さくせざるを得ない。ところが、我々の研究によると、バッチサイズを小さくした場合には GLOW には不安定さが顕れる。例えば、学習中に損失関数の出力が発散する、生成したデータのある成分が発散する、もしくは想定するレンジ外のデータが生成されるといった問題が浮き彫りになる。

これらの問題を改善するために、理論的な考察から、二つのアプローチをする。最初のアプローチは、写像の可逆性と等長性に注目した。我々は GLOW が非特異であることの証明を知らない。それゆえに、広く知られている非特異であるための十分条件 (ヤコビ行列式が非零) を満たすような学習を考察することには意味がある。さらに言えば、スケールリングが必要とされてい

ない層については、等長性を課すことで、出力の発散を抑止することができる。具体的には、ヤコビ行列式が入力に依存しない場合は対数ヤコビ行列式を0に近づけるように学習すること、そうでない場合は対数ヤコビ行列式が0付近となるように各 flow モジュールを修正することである。これらは GLOW の安定化に貢献する。

もう一つのアプローチは、データ駆動のパラメータ初期化を改善することである。データ駆動のパラメータ初期化を行う Actnorm¹⁶⁾ は、バッチサイズが小さいほどバイアスがかかる初期化方法をしていた。我々の方法ではバイアスがかかりにくい方法でデータの選択を行い、初期化に使用する。

この研究では、これらのアプローチを以って GLOW を改善した新しい安定的な Flow-based 生成モデルを提案する。このモデルは限られたリソースであっても安定的に学習できる非常に実用的な深層生成ニューラルネットワークである。

2 関連研究

可逆なニューラルネットワーク 本研究は NICE⁶⁾、RealNVP⁷⁾、GLOW¹⁶⁾ に代表される Coupling Flow のアイデアを元に構築されている。NICE⁶⁾ では可逆な Coupling 層が考案された。RealNVP⁷⁾ では、Masked Convolution が提案され、NICE よりも良い性能を示した。GLOW¹⁶⁾ は Masked Convolution を改良した 1×1 LU Convolution と、Actnorm を導入した結果、さらなる性能改善を実現した。本研究では GLOW のアーキテクチャを継承しており、GLOW と比較して論じている。Coupling Flow のアイデアは他にもある。Flow++¹⁴⁾ では uniform dequantization²⁵⁾ に代わる variational dequantization、あるいは CDF や self-attention を利用した Coupling の手法が提案されている。また、Spline Flow^{22, 8, 9)} というアーキテクチャも知られている。Coupling Flow 以外のアイデアでは、Residual Network を可逆化した i-ResNet¹⁾、Residual Flow⁴⁾ や、ODE ベースの Flow の研究^{5, 13)} がある。また、可逆なニューラルネットワークの可逆性について論じた研究^{3, 2)} もある。

ヤコビ行列式に着目した正則化 本研究では、パラ

メータの直交化のためにヤコビ行列式を正則化に利用している。我々の知る限り、本研究のような意図で、ヤコビ行列式を正則化に利用した研究はない。一方、他の意図でヤコビ行列式を正則化に利用してる例はある。例えば、ヤコビ行列式のプロベニウスノルムを正則化に利用している研究^{23, 15)}がある。特に Hoffman らの研究¹⁵⁾はこの正則化がモデルのロバスト化に貢献することも示した。可逆なニューラルネットにおいてもヤコビ行列式のプロベニウスノルムは有用である。例えば、ODE ベースの Flow の正則化の研究¹⁰⁾ではヤコビ行列式のプロベニウスノルムを正則化が使用されている。Mok らの研究²¹⁾ではトポロジー保存変換を意図した選択的ヤコビアン行列式正則化が提案されている。Behrmann らの研究²⁾では可逆なニューラルネットワークのアーキテクチャにおける不正確な勾配が計算される問題や、正規化フローモデルからのサンプリングができない問題が提起されている。その問題への対応策の一つとして、双方向有限差分正則化が提案されている。

3 背景

この節では Coupling Flow の基礎となる Normalizing Flow の定義を始めとする、我々の結果を説明するために必要な背景知識を取り上げる。Normalizing Flow は可逆な写像の合成によって定義される。同じ次元の Z, X の2つの空間が与えられ、 C^1 級でパラメトリックな可逆写像 $f_\theta: X \rightarrow Z$ と事前分布 $p_Z(z)$ が与えられているとする。この条件のもとで $\log p_X(x)$ は以下の通り計算することができる。

$$\log p_X(x) = \log p_Z(f_\theta(x)) + \log \left| \frac{\partial f_\theta(x)}{\partial x} \right| \quad (1)$$

ここで、 $\frac{\partial f_\theta(x)}{\partial x}$ は f_θ の x でのヤコビアンである。このような flow f_θ をたくさん用意してそれらを合成することで、複雑な可逆写像 f を構成できる。事前分布として等方性ガウス分布のような扱いやすい分布を仮定し、負の対数尤度 $-\log p_X(x)$ を目的関数として最小化する。こうして得られたネットワーク f は密度推定に、逆写像 f^{-1} は生成に利用できる。

Normalizing Flow における課題の1つは、計算量が少なく扱いやすいヤコビ行列になるように各 flow f_θ を設計することである。GLOW¹⁶⁾ のアーキテクチャで使用されている可逆なモジュールについての概要は Table 1 に示し、説明は以下に記述する。

3.1 Coupling 層

Coupling 層は NICE⁶⁾, RealNVP⁷⁾ で提案されている。この層は入力 x を x_1, x_2 に分割し、その一方 x_2 にパラメトリックな写像 s, t を作用させる。これを利用して x_1 を変換した結果 z_1 を x_2 と連結し、それを出力 z とする。Additive Coupling は Affine Coupling のスケール部分がない特別な場合である。

3.2 Actnorm 層

Actnorm 層は GLOW¹⁶⁾ で提案されている。スケールリングと並行移動のパラメータを持つ正規化層である。初期のミニバッチデータを使用して、この層への入力に対して、チャンネルごとに精度 σ と要素の平均の反

Modules	Maps	Inverse Maps	L.J.D.
Additive Coupling	$x_1 \oplus x_2 = x$ $z_1 = x_1 + t(x_2)$ $z = z_1 \oplus x_2$	$z_1 \oplus z_2 = z$ $x_1 = z_1 - t(z_2)$ $x = x_1 \oplus z_2$	0
Affine Coupling	$x_1 \oplus x_2 = x$ $z = \exp(s(x_2)) \odot x_1 + t(x_2)$ $z = z_1 \oplus x_2$	$z_1 \oplus z_2 = z$ $x = \frac{z_1 - t(z_2)}{\exp(s(z_2))}$ $x = x_1 \oplus z_2$	$\text{sum}(\log s(x_2))$
Actnorm	$z = x \odot \sigma + \tau$	$x = (z - \tau) / \sigma$	$h \cdot w \cdot \text{sum}(\log \sigma)$
1 × 1 LU Convolution	$\forall i, j: z_{i,j} = W x_{i,j}$	$\forall i, j: x_{i,j} = W^{-1} z_{i,j}$	$h \cdot w \cdot \text{sum}(\log u)$

Table 1: The forward and inverse maps and the logarithm Jacobian determinants are described for the 4 types of layers mentioned. Let x be the input to the layer and let z be the output from the layer. x and z are tensors, whose shape is $[h \times w \times c]$, where h is the height of the space, w the width of the space, and c the dimension of the channels. We denote by W the $[c \times c]$ weight matrix, by s, t the any maps, and by σ, τ trainable parameters. \oplus is a direct sum, and \odot is a Hadamard product, i.e., an element-wise product. Note that the results of the logarithm Jacobian determinant are discussed in GLOW¹⁶⁾.

数 τ を計算し、それらをスケールリングと並行移動のパラメータとする。このデータ駆動的初期化方法は学習の安定化に貢献する。

3.3 1 × 1 LU Convolution 層

1 × 1 LU Convolution 層は GLOW¹⁶⁾ で提案されており、NICE⁶⁾, RealNVP⁷⁾ で使用されている固定置換の一般化である。Convolution に使用される重み行列 W をランダムな直交行列で初期化し、重み行列の LU 分解 $W = PLU$ に対して、 L, U のみをパラメータライズする。行列式は $u = \text{diag}(U)$ を用いて容易に計算できるため、LU 分解を使用しない場合と比較すると、行列式の計算量を $\mathcal{O}(n^3)$ から $\mathcal{O}(n)$ に削減できる。

4 手法

この節では、GLOW の学習の安定に有用な正則化や初期化の手法を提案する。1 × 1 LU Convolution 層の改善を図る対数ヤコビ行列式罰則と、Actnorm 層の改善を図る Actnorm の累積的初期化を導入する。

4.1 対数ヤコビ行列式罰則

1 × 1 LU Convolution のように、変換そのものが入力に依存しない、線形変換として表現可能な層は、その対数ヤコビ行列式が固定的である。それがゆえに、対数ヤコビ行列式に静的な制限を与え、その重み行列を直交行列に近づけることができる。すなわち、目的関数を次で定義可能である：

$$J(x) = \log p_X(x) + \lambda \cdot \sum_{1 \times 1 \text{ LU Conv.}} |h \cdot w \cdot \text{sum}(\log |u|)|^2. \quad (2)$$

ここで λ は罰則の大きさを与えるハイパーパラメータである。次元に依存せずに λ を調節するため、5 節の実験では bpd 基準での負の対数尤度に対して、 λ を設定している。この正則化手法は行列の直交化を強制して、GLOW 全体を非特異から乖離しないようにする。さらに、1 × 1 LU Convolution を等長写像に近づける働きをし、急激なスケール変化を抑制する。その結果、生成するデータが想定範囲の値に収まるような学習を実現している。これら理由からも安定化に貢献すると考えられる。実際に我々の手法が安定的な学習にとってよい効果をもたらすことは 5 節で実験的に示す。ま

た、 1×1 LU Convolution に与えられた役割は固定置換と同様のシャッフルの役割であるが、直交行列に近づけたとしてもその役割を果たせること注意する。

4.2 Actnorm の累積的初期化

Actnorm 層はバッチサイズが小さい場合に、悪い挙動を示すことがある。この事実は実際に 5.2 の実験でも示す。このように悪い挙動を示す理由は、最初のミニバッチのデータを元に初期化をしているため、ミニバッチサイズが小さい場合はその初期値が高バイアスとなるためである。単純に最初のいくつかのミニバッチのデータの平均と分散を使って初期パラメータを決定することで、この問題を回避できる。この初期化は少ない計算コストで高速に最適な初期値を設定することができ、後で実験で示すようにミニバッチサイズ 1 の場合であっても、問題なく学習できる。

Actnorm の累積的初期化のアルゴリズムは Algorithm 1 に従って、Actnorm のスケールパラメータ σ とバイアスパラメータ τ を初期化する。 M は初期化に使うミニバッチ数、 $\{\mathbf{x}_m\}_{m=1}^M$ を初期化に使うミニバッチ、 b はミニバッチインデックス、 c はチャンネルのインデックスを表す。

Algorithm 1 The cumulative initialization of Actnorm

```

1: procedure INITIALIZATION( $\{\mathbf{x}_m\}_{m=1}^M, \sigma, \tau$ )
2:    $\mu \leftarrow \mathbf{0}$ 
3:    $\nu \leftarrow \mathbf{0}$ 
4:   for  $m \in 1, \dots, M$  do
5:      $\mu \leftarrow \mu + \frac{1}{bc} \sum_{b,c} \mathbf{x}_{m;b,c}$ 
6:      $\nu \leftarrow \nu + \frac{1}{bc} \sum_{b,c} \mathbf{x}_{m;b,c}^2$ 
7:   end for
8:    $\sigma \leftarrow \frac{1}{\sqrt{\frac{\nu}{M} - (\frac{\mu}{M})^2 + \epsilon}}$ 
9:    $\tau \leftarrow -\frac{\sigma \odot \mu}{M}$ 
10: end procedure

```

5 実験

我々の実験は GLOW¹⁶⁾ のアーキテクチャを基本とする。このアーキテクチャの構造は Fig. 2 のように、squeeze、step of flow、split から成る。これらの要素を繰り返す数を L 、step of flow を繰り返す数を K と記載している。Fig. 1 は step of flow の構成要素を示している。我々の実験では Coupling 層の種類を実験によって変更する。

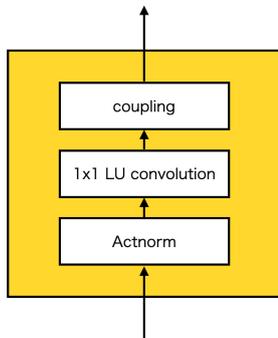


Fig. 1: The step of the flow are composed of Actnorm, 1×1 LU Convolution and coupling.

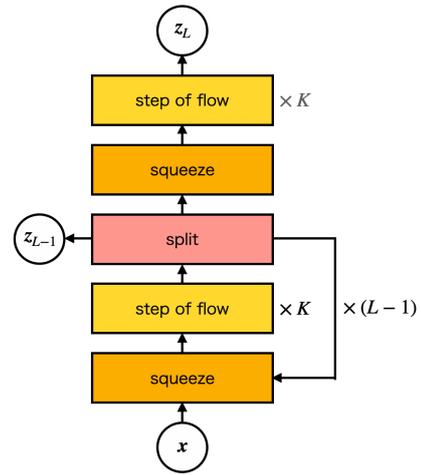


Fig. 2: The overall architecture of GLOW, which achieves complex representations and reduces the amount of computation through the iterative structure consisting of 3 components.

小さいミニバッチサイズであっても安定した学習を行うため、オプティマイザは RAdam¹⁹⁾ を採用し、 $\beta_1 = 0.9$ 、 $\beta_2 = 0.999$ 、 $\epsilon = 10^{-8}$ に設定する。特に記載がない場合は、Actnorm の初期化は 256 個のミニバッチで初期化している。また、学習時にはデータに対して uniform dequantization²⁵⁾ を適用している。

5.1 対数ヤコビ行列式罰則についての定量的評価

1×1 LU Convolution の対数ヤコビ行列式罰則の効果を検証するために、さまざまなデータセットで学習し、テストデータの bpd で定量的に評価する。学習率は 10^{-4} 、ミニバッチサイズ 32、 $K = 32$ 、 $L = 4$ とし、エポック数は最大 200 に設定し、CIFAR-10¹⁸⁾ について検証した。Additive Coupling については対数ヤコビ行列式罰則の有用性を比較検証するため、正則化係数 $\lambda = 10^{-5}$ で正則化した場合と、対数ヤコビ行列式罰則を一切与えない場合を検証した。Affine Coupling については正則化係数 $\lambda = 10^{-5}$ のみ検証した。

検証結果については Table 2 に示した。実験結果では、Additive Coupling、Affine Coupling のいずれの場合も、対数ヤコビ行列式罰則を使用しない場合は損失が nan を出力しているが、対数ヤコビ行列式罰則を使用した場合は 200 エポックの学習が問題なく完了している。この事実から、対数ヤコビ行列式罰則の有効性が確認できた。

対数ヤコビ行列式罰則を使用して学習した各モデルで画像生成も行える。その結果は Fig. 3 に示した。

5.2 Actnorm の累積的初期化の検証

Actnorm の累積的初期化の有効性を示すために、従来の Actnorm の初期化手法では学習が難しい状況において、我々の初期化手法と比較検証を実施した。256 個のミニバッチを使用して初期化して Actnorm の累積的初期化を行ったケースと GLOW で提案された Actnorm の初期化を行ったケースとを比較検証を行った。実験には Additive Coupling を使用し、ハイパーパラメータはミニバッチサイズ 1、正則化係数 $\lambda = 10^{-3}$ 、学習率 10^{-5} を使用した。データセットは CelebA²⁰⁾、CIFAR-10¹⁸⁾ を用いて検証を行った。Fig. 4 に示した通り、我々

Coupling	λ	bpd	loss
Additive Coupling	0	nan	nan
Additive Coupling	10^{-5}	3.60	4.30
Affine Coupling	0	nan	nan
Affine Coupling	10^{-5}	3.56	4.27

Table 2: In the experiment in Subsection 5.1, we evaluated the results on the validation dataset every 10 epochs out of 200 epochs. If there was no improvement of bits per dims for two epochs in a row in the evaluation of the validation dataset, the training was terminated. The table shows the best bits per dims of the evaluated values and the corresponding objective values.



(a) Additive Coupling

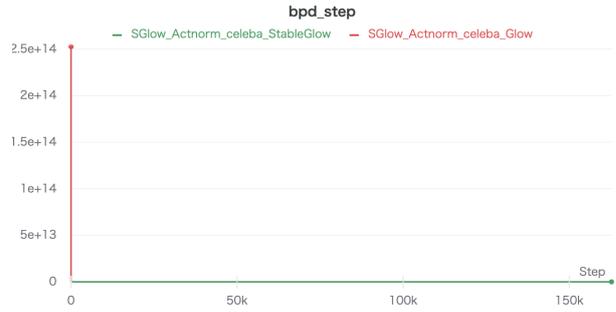


(b) Affine Coupling

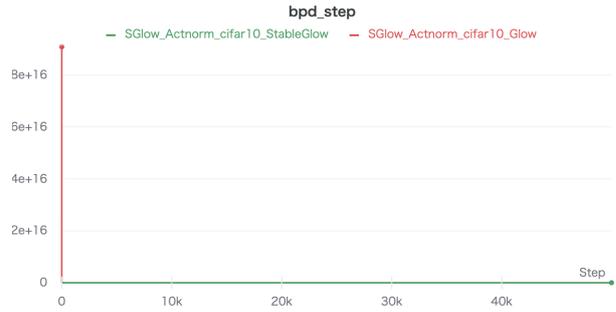
Fig. 3: Example images generated by each model trained with the natural logarithm of the Jacobian determinant penalty, on CIFAR-10.

の初期化方法では損失が発散することなく学習が進むが、GLOW オリジナルの初期化方法では学習の初期段階で損失が発散した。この実験結果は、我々の手法が小さいミニバッチサイズで安定的に GLOW の学習を

行える有効的な手段であることを示している。



(a) The CelebA case



(b) The CIFAR-10 case

Fig. 4: The experiment in Subsection 5.2. **Red**: the results when using the traditional Actnorm initialization, and the bpd diverges on both datasets. **Green**: the results of using the cumulative Actnorm initialization which we proposed, and the training is progressing without divergence.

加えて、CelebA を使用した、ミニバッチサイズ 1、かつ 1 バッチで初期化した場合の実験において、2 回目のイテレーションにおける Additive Coupling 層の出力を観察した。その結果を Fig. 5 に示した。

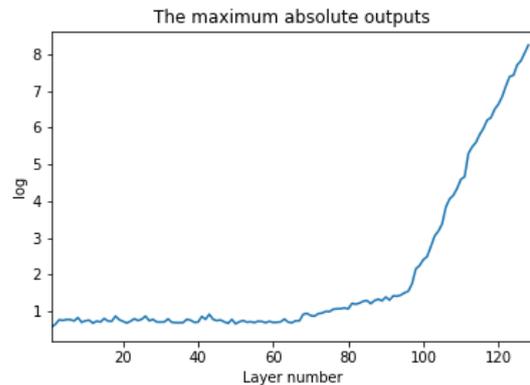


Fig. 5: The figure shows the example of absolute maximum outputs in the initial stage of learning. The horizontal axis is the number of layers of additive couplings in a GLOW, with the layer closest to the input being labeled 1, and the layers sequentially labeled thereafter. The vertical axis is the common logarithm of absolute maximum output of each additive coupling.

GLOW の通常の Actnorm の初期化では、2 イテレーション目、つまりパラメータの更新を 1 回のみ行った

状況でも、出力の絶対値の最大が 10^8 程度になっている。それゆえに、bpd も大きくなる。このことは、学習が初期段階ですでにうまくいっていないことを示している。

6 結論及び今後の展望

我々の研究では Flow-Based 生成モデルとしてよく知られている GLOW¹⁶⁾ を改善した。新たに提案した対数ヤコビ行列式罰則による正則化手法、Actnorm の累積的初期化の 2 つの手法を以って、安定的な学習が行えることを示した。更に、我々が提案した手法は小さいミニバッチサイズであっても有用であることに言及し、限られたリソース上での Coupling Flow の学習可能性を明らかにした。今後のこの研究の展望としては、bpd がオリジナルの GLOW に劣らず強力であること、より高解像度の画像であっても、定性的によい画像が生成可能であることを検証しようと試みている。加えて、実施済の実験に対して、他のデータセットでの検証を進める。また、我々の研究で提案している手法の 1 つである対数ヤコビ行列式罰則については、一般の Convolution にも適用できる。そのため、他の CNN の学習の安定化にも効果がある可能性もある。それについては今後の研究の課題としたい。

参考文献

- 1) J. Behrmann, W. Grathwohl, R. T. Chen, D. Duvenaud, and J.-H. Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.
- 2) J. Behrmann, P. Vicol, K.-C. Wang, R. Grosse, and J.-H. Jacobsen. Understanding and mitigating exploding inverses in invertible neural networks. *arXiv preprint arXiv:2006.09347*, 2020.
- 3) J. Behrmann, P. Vicol, K.-C. Wang, R. B. Grosse, and J.-H. Jacobsen. On the invertibility of invertible neural networks. 2019.
- 4) R. T. Chen, J. Behrmann, D. K. Duvenaud, and J.-H. Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, pages 9916–9926, 2019.
- 5) R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.
- 6) L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.
- 7) L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real nvp. 2017.
- 8) C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Cubic-spline flows. *Workshop on Invertible Neural Networks and Normalizing Flows, International Conference on Machine Learning*, 2019.
- 9) C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, pages 7511–7522, 2019.
- 10) C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *International Conference on Machine Learning*, pages 3154–3164. PMLR, 2020.
- 11) M. Germain, K. Gregor, I. Murray, and H. Larochelle. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pages 881–889, 2015.
- 12) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- 13) W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Fjord: Free-form continuous dynamics for scalable reversible generative models. In *The 7th International Conference on Learning Representations (ICLR)*, 2018.
- 14) J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In *International Conference on Machine Learning*, pages 2722–2730, 2019.
- 15) J. Hoffman, D. A. Roberts, and S. Yaida. Robust learning with jacobian regularization. *arXiv preprint arXiv:1908.02729*, 2019.
- 16) D. P. Kingma and P. Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pages 10215–10224, 2018.
- 17) D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *The 2nd International Conference on Learning Representations (ICLR)*, 2013.
- 18) A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- 19) L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2019.
- 20) Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

- 21) T. C. Mok and A. Chung. Fast symmetric diffeomorphic image registration with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4644–4653, 2020.
- 22) T. Müller, B. McWilliams, F. Rousselle, M. Gross, and J. Novák. Neural importance sampling. *ACM Transactions on Graphics (TOG)*, 38(5):1–19, 2019.
- 23) R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *In The 6th International Conference on Learning Representations (ICLR)*, 2018.
- 24) T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.
- 25) B. Uria, I. Murray, and H. Larochelle. Rnade: The real-valued neural autoregressive density-estimator. *Advances in Neural information processing systems*, 26:2175–2183, 2013.
- 26) A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29:4790–4798, 2016.
- 27) A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756. PMLR, 2016.