Learning and Generalization in Overparameterized Normalizing Flow

Kulin Shah, Amit Deshpande, Navin Goyal
kulin.shah98@gmail.com, amitdesh@microsoft.com, navingo@microsoft.com

Generative Models

- Impressive performance of generative models over last few years
- Capabilities of Generative Models:
  - Generate new samples from the data distribution
  - Estimate probability density of any sample
- Normalizing Flows is one of the few generative models that can do both tasks

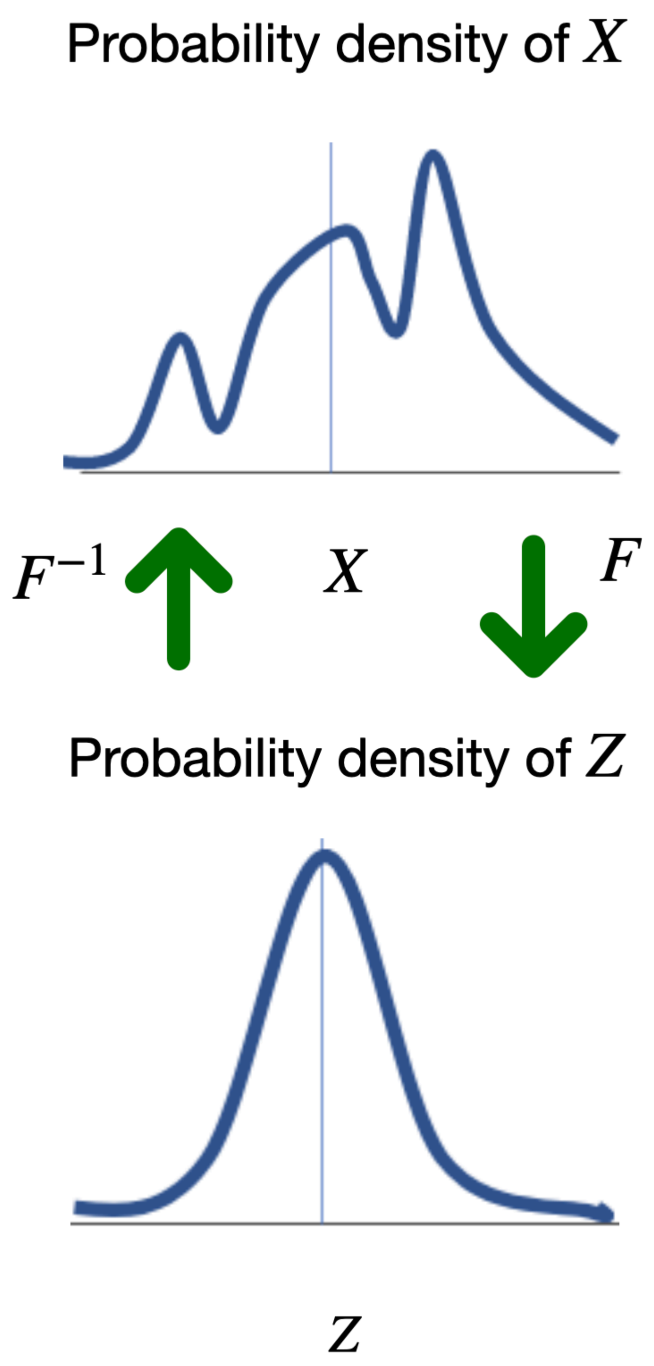
Goal: Theoretical understanding of learning and generalization of (autoregressive) normalizing flows when they are parameterized using a neural network

Normalizing Flows (NFs)

X : target (data) random variable whose distribution we want to learn

Z : base random variable such as standard Gaussian or exponential

For every data distribution, there exists a **monotone** function such that $F(X) = Z$



Learning problem: Given samples of X, learn the function F

Objective: Maximum likelihood on training data. Contains function and its derivative.

Our Contributions

- We prove that unconstrained NFs can efficiently learn any reasonable data distribution when the underlying network is overparameterized one-hidden layer neural network.
- We provide theoretical and empirical evidence that for NFs with the one-hidden-layer network, overparameterization hurts the training.

Constrained Normalizing Flows (CNFs)

- It directly represents the learner function by a neural network and imposes constraints to make it monotonic
- Most models are of this type (NAF [1], BNAF [2], etc.)
- We provide theoretical and empirical evidence that overparameterization hurts the training of CNFs
  - We don't know of **any other applications** of neural networks with this trend.

Result: For bounded number of training iterations or for bounded change in weights, highly over-parameterized one-layer CNFs can only learn very small class of probability distributions.

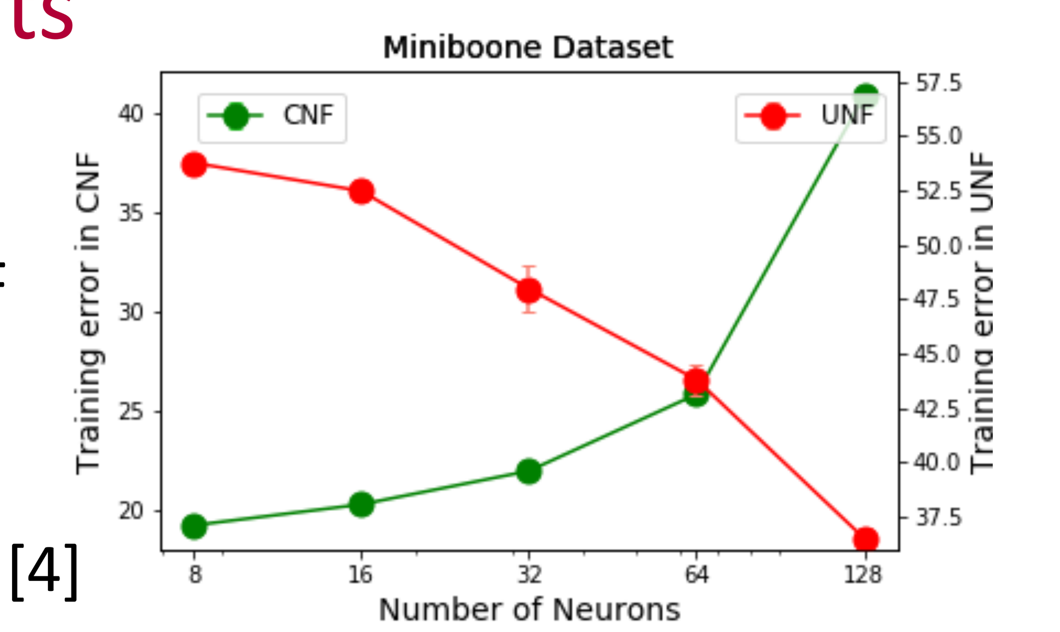
Unconstrained Normalizing Flows (UNFs)

- UNFs represent the derivative of the function using a neural network. [3]
  - Only need positive derivative to get monotonic function
- Function value can be estimated using numerical integration

Result (Informal): Suppose sufficiently overparameterized one-hidden layer UNF learner network is trained with SGD on maximum likelihood. Then, for any data distribution with a "low complexity" monotonic function F such that $F(X) = Z$, UNF will achieve small KL divergence between data distribution and learned distribution.

Experimental Results

Training error after a fixed after a fixed number of iterations increases for CNF and decreases for UNF



Trend:
UNF $\approx$ Supervised learning [4]
UNF $\not\approx$ CNF

References
[1] Neural Autoregressive Flows. Chin-Wei Huang et al. 2018
[2] Block Neural Autoregressive Flow. Nicola De Cao et al. 2019
[3] Unconstrained Monotonic Neural Networks. Antoine Wehenkel et al. 2020
[4] In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. Behnam Neyshabur et al. 2015