RUTGERS
UNIVERSITY | NEW BRUNSWICK

Approximation and Estimation for Deep Learning Networks

Jason M. Klusowski

The Wharton School, Department of Statistics

STAT 991: Topics in Deep Learning

November 20$^{\text{th}}$, 2018

Two Important Questions in Deep Learning

- Approximation error: how well can one approximate a general function of many variables with a neural network?
- Model complexity: how difficult is it to describe a parameterized family of neural networks that have good expressive power?

Data Setting

- Data are of the form $\{(X_i, Y_i)\}_{i=1}^n$, drawn independently from a joint distribution $P_{X,Y}$ with P_X on $[-1, 1]^d$
- Inputs: explanatory variable vectors $X_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})$
- Random design: independent $X_i \sim P$, with P probability measure
- The target function is $f(x) = \mathbb{E}[Y \mid X = x]$, the mean of the conditional distribution $P_{Y|X=x}$, optimal in mean square for the prediction of future Y from corresponding input X
  - Classification: $Y \in \{0, 1\}$ with $f(X) = \mathbb{P}[Y = 1 \mid X]$
- In some cases, assumptions are made on the error of the target function $\epsilon_i = Y_i - f(X_i)$ (i.e. bounded, Gaussian, or sub-Gaussian)

Statistical Risk

- Estimators $\hat{f}(x) = \hat{f}(x, \{(X_i, Y_i)\}_{i=1}^n)$ are formed from the data
- Loss at a target f^* is the $L_2(P_X)$ square error $\|f^* - \hat{f}\|^2$
- Risk is the expected squared error $\mathbb{E}[\|f - \hat{f}\|^2]$
- Minimax risk is best worst case risk

$$R_{n,d}(\mathcal{F}) = \inf_{\hat{f}} \sup_{f^* \in \mathcal{F}} \mathbb{E}[\|f^* - \hat{f}\|^2]$$

Why Are Our Two Questions Important?

- Let $\hat{f}$ be a complexity penalized least squared estimator over a class of candidate functions $\mathcal{F}$ (could be deep neural networks), i.e., $\hat{f}$ is chosen to optimize or approximately optimize

$$\sum_i (Y_i - f(X_i))^2 + \text{pen}(f),$$

  over a collection $\mathcal{F}$ of candidate functions.
  - Nonconvex objective function is computationally difficult to optimize!
  - Gradient descent (back propagation for neural networks)
  - Tensor methods (method of moments)

Upper Bounds

- Complexity penalized least squares estimators [Barron, Birge, and Massart, 1999] satisfy

$$\mathbb{E}[\|f^* - \hat{f}\|^2] \leq \inf_{f \in \mathcal{F}} \left\{ \|f^* - f\|^2 + \frac{\text{complexity}(f)}{n} \right\}.$$

- An important aspect of the above adaptive risk bound is that f^* need not belong to $\mathcal{F}$. The only requirement is that it is well-approximated by certain members of $\mathcal{F}$.
- Right side is an index of resolvability expressing the tradeoff between approximation error and descriptive complexity relative to sample size n.
- Log-cardinality of d-dimensional covers of the dictionary provide a descriptive complexity.

Takeaway

- For penalized least squares estimators, there is a tradeoff between approximation error and model complexity.

Approximation Error

One of the earliest approximation results for single hidden-layer neural networks is [Cybenko, 1989].

- Bounded activation function $\phi(z)$ on $\mathbb{R}$, sigmoidal if $\lim_{z\to\pm\infty}\phi(z) = \pm 1$.
- Work with parameterized family of functions $\mathcal{F}_m$

$$f_M(x) = f(x; W) = \sum_{j_1=1}^{M} w_{j_1}\phi\left(\sum_{j_2=1}^{d} w_{j_1,j_2}x_{j_2}\right).$$

- $W_1[j_1] = w_{j_1}$ outer layer parameters.
- $W_2[j_1, j_2] = w_{j_1,j_2}$ inner layer parameters for single hidden-layer.

Cybenko showed that for sigmoidal ϕ and for large enough M, any continuous function f can be approximated by an f_M with arbitrary accuracy. [Hornik, 1991] later showed f_M enjoys the same approximation properties as long as ϕ is not the constant function.

Approximation Error

- Cybenko and Hornik results are useful starting points, but do not show quality of approximation as a function of m, the number of terms. So $\|f^* - f_M\|^2$ is small, but what about complexity(f_M)?
- Barron provides an answer:

Theorem (Barron, 1993)

For functions f satisfying $v_f = \int_{\mathbb{R}^d} \|\omega\|_1 |\tilde{f}(\omega)| d\omega < +\infty$, there exists a single hidden-layer neural network f_M such that $\|f - f_M\|^2 \leq v_f^2 / M$.

Approximation Error

- Makovoz gives a near optimal refinement:

Theorem (Makovoz, 1995)

For functions f satisfying $v_{f,1} = \int_{\mathbb{R}^d} \|\omega\|_1 |\tilde{f}(\omega)| d\omega < +\infty$, there exists a single hidden-layer neural network f_M with bounded $\|W_1\|_1$ such that $\|f - f_M\|^2 \leq c v_{f,1}^2 / M^{1+1/d}$. Furthermore, the exponent $1 + 1/d$ cannot be improved beyond $1 + 2/d$.

- Utility of these accuracy bounds is that we now bound can count the number of functions of the form f_M (by further discretizing them).
- Proofs of Makovoz and Barron are based on a powerful probabilistic argument. [Barron and Klusowski, 2018] extend argument to multi-layer networks.

Estimation Error

Theorem (Barron, 1994)

Suppose $\hat{f}$ is a complexity penalized least squares estimator with $\mathrm{pen}(f_M) = \lambda_n \sum_{j_1=1}^{M} |w_{j_1}| = \|W_1\|_1$. Then,

$$\mathbb{E}[\|f^* - \hat{f}\|^2] \leq C v_{f^*,1} \left(\frac{d \log(n/d)}{n} \right)^{1/2}.$$

- Not good in high-dimensional settings where $d \gg n$.
- ℓ^1 penalty $\mathrm{pen}(f_M) = \lambda_n \sum_{j_1=1}^{M} |w_{j_1}| = \lambda_n \|W_1\|_1$ does not regularize internal parameters w_{j_1,j_2}.

Estimation Error

Theorem (Klusowski and Barron, 2018)

For functions f satisfying $v_{f,2} = \int_{\mathbb{R}^d} \|\omega\|_1^2 |\tilde{f}(\omega)| d\omega < +\infty$, there exists a single hidden-layer neural network f_M with ReLU activation function $\phi(z) = \max\{0, z\}$ and bounded $\|W_1\|_1$ and $\|W_2\|_{1,\infty}$ such that $\|f - f_M\|^2 \leq c v_{f,2}^2 / M^{1+2/d}$. Furthermore, the exponent $1 + 2/d$ cannot be improved beyond $1 + 4/d$.

Theorem (Klusowski and Barron, 2018)

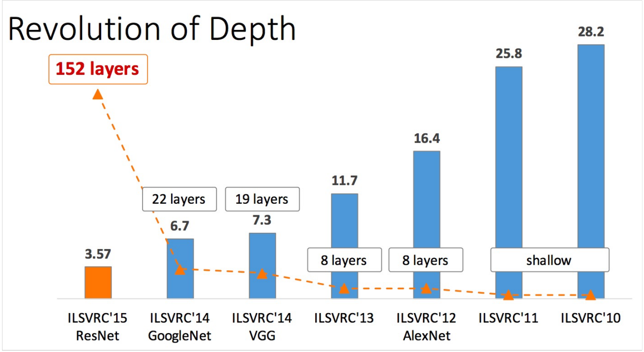
Suppose $\hat{f}$ is a complexity penalized least squares estimator with $\mathrm{pen}(f_M) = \lambda_n \sum_{j_1=1}^{M} \sum_{j_2=1}^{d} |w_{j_1}||w_{j_1,j_2}| = \lambda_n \|W_1 W_2\|_1$. Then,

$$\mathbb{E}[\|f^* - \hat{f}\|^2] \leq C v_{f^*,2} \left(\frac{\log(d)}{n} \right)^{1/2}.$$

Furthermore, the bound is minimax optimal for the collection of functions f with finite $v_{f,2}$.

Analogous Results for Deep Networks

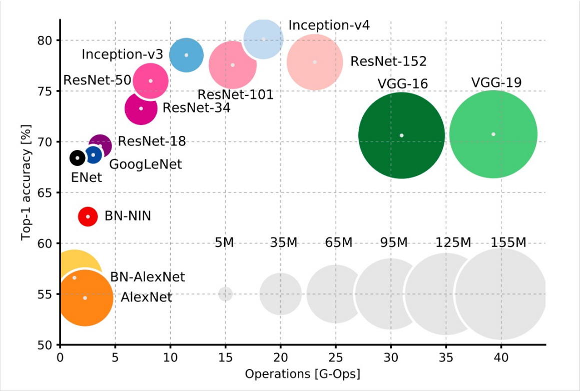
- What are the approximation capabilities of deep networks? What types of flexible high-dimensional function spaces admit sparse representations by deep networks?
- Largely an open question. State-of-the art results are for spaces that are really big (and hence already suffer from curse of dimensionality): e.g., Sobolev-like spaces [Yarotsky, 2017, 2018, Schmidt-Hieber, 2017] with proofs based on local Taylor expansions.
- What are the benefits of depth? Again, only known for large function classes or specific data structures [Telgarsky, 2016].



Source: Kaiming He, Deep Residual Networks (2016)

- Networks can be very deep. Versions of ResNet have 152 layers.

Number of Parameters



Source: Canziani, Culurciello, and Paszke (2017)

- Number of parameters far exceeds sample size. AlexNet uses $\approx$ 60 million parameters with only $\approx$ 1.2 million training samples.

Standard Deep Network Formulation

- Deep net function $f(x; W)$, weights $W = (W_1, W_2, \ldots, W_L)$, inputs x in $[-1, 1]^d$,

$$\phi_{out}(\sum_{j_1}^{d_1} w_{j_1} \phi(\sum_{j_2}^{d_2} w_{j_1,j_2} \phi(\sum_{j_3}^{d_3} w_{j_2,j_3} \cdots \phi(\sum_{j_L}^{d_L} w_{j_{L-1},j_L} x_{j_L}))))$$

- Computation at node j_ℓ on layer ℓ.

$$z_{j_\ell} = \phi(\sum_{j_{\ell+1}} w_{j_\ell,j_{\ell+1}} z_{j_{\ell+1}})$$

- Total number of parameters:

$$\sum_{\ell=1}^{L} d_{\ell-1} d_\ell$$

- Let $\mathcal{F}_L$ denote the class of all depth L networks.

Estimation for Deep ReLU Networks

Theorem (Schmidt-Hieber, 2017)

Suppose $f^* = g_1 \circ g_2 \circ \cdots \circ g_q$, where $g_j : \mathbb{R}^{d'_{j+1}} \mapsto \mathbb{R}^{d'_j}$ and each of the d'_j components of g_j is β_j-smooth and depends only on $t_j \ll d'_{j+1}$ variables. Suppose $L \asymp \log n$ and width $= \max_\ell d_\ell \asymp \max_{j=1,\dots,q} n^{\frac{t_j}{2\beta_j + t_j}} \log n$. If $\hat{f}$ is the least squares estimator over $\mathcal{F}_L$, then

$$\mathbb{E}[\|f^* - \hat{f}\|^2] \leq C \max_{j=1,\dots,q} n^{-\frac{2\beta_j}{2\beta_j + t_j}}.$$

Approximation for Deep ReLU Networks

Theorem (Barron and Klusowski, 2018)

Fix a positive integer M. For any deep ReLU network $f(x; W)$, there exists a sparse approximant $f(x; \widetilde{W})$ (with at most LM nonzero parameters) from a subfamily with log-cardinality at most $cLM \log(\max_\ell d_\ell)$, such that

$$\|f(\cdot; \widetilde{W}) - f(\cdot; W)\|^2 \leq CL^2 \|W_1 W_2 \cdots W_L\|_1^2 / M.$$

- Complexity constant $\|W_1 W_2 \cdots W_L\|_1$ is a norm of a matrix product
- Differs from results involving the product of individual matrix norms of the weight matrices, i.e., $\prod_{\ell=1}^{L} \|W_\ell\|$, [Neyshabur, 2017, Golowich 2017, Bartlett, 2017].

Estimation for Deep ReLU Networks

Theorem (Barron and Klusowski, 2018)

Suppose $\hat{f}$ is a complexity penalized least squares estimator with $\mathrm{pen}(f(\cdot; W)) = \lambda_n \|W_1 W_2 \cdots W_L\|_1$. Suppose $f^* = f(\cdot; W^*)$ is a deep ReLU network with $\|W_1 W_2 \cdots W_L\|_1 \leq V$. Then,

$$\mathbb{E}[\|f^* - \hat{f}\|^2] \leq \inf_M \left\{ CL^2 V^2 / M + \frac{cLM \log(\max_\ell d_\ell)}{n} \right\}$$

$$\leq CV \left(\frac{L^3 \log(\max_\ell d_\ell)}{n} \right)^{1/2}.$$

Proof: Main Idea

- Homogeneity property of positive part. For $w \geq 0$

$$w\phi(z) = \phi(wz).$$

- Implication. May push weights to the innermost layer

$$f(W, x) = \sum_{j_1} \phi\Big(\sum_{j_2} \phi\Big(\sum_{j_3} \cdots \phi\Big(\sum_{j_L} w_{j_1, j_2, \ldots, j_L} x_{j_L} \Big)\Big)\Big).$$

- Composite weights of paths $j_1, j_2, \ldots, j_L$

$$w_{j_1, j_2, \ldots, j_L} = w_{j_1} w_{j_1, j_2} w_{j_2, j_3} \cdots w_{j_{L-1}, j_L}.$$

- Full network variation (related to "path norm" [Neyshabur et. al., 2015])

$$V = \sum_{j_1, \ldots, j_L} w_{j_1, \ldots, j_L} = \| W_1 W_2 \cdots W_L \|_1.$$

Probabilistic Characterization of Deep Nets

- Path weights provide a joint probability distribution

$$a_{j_1,j_2,\ldots,j_L} = \frac{w_{j_1,j_2,\ldots,j_L}}{V}.$$

- It has a Markov structure

$$a_{j_1,j_2,\ldots,j_L} = a_{j_1} a_{j_2|j_1} a_{j_3|j_2} \cdots a_{j_L|j_{L-1}}.$$

- Probability characterization of deep net $f(W,x) = V f(a,x)$

$$f(a,x) = \sum_{j_1} \phi(\sum_{j_2} \phi(\sum_{j_3} \cdots \phi(\sum_{j_L} a_{j_1,j_2,\ldots,j_L} x_{j_L}))).$$

- Iterated expectation representation, interspersed with nonlinearities

$$\sum_{j_1} a_{j_1} \phi(\sum_{j_2} a_{j_2|j_1} \phi(\sum_{j_3} a_{j_3|j_2} \cdots \phi(\sum_{j_L} a_{j_L|j_{L-1}} x_{j_L}))).$$

Sparse Deep Net Approximation

- Approximate the weights a by $\tilde{a}$ from a sparse set.
- Draw sample, size M, independent from distrib. $a_{j_1,j_2,\ldots,j_L}$
- Let $K_{j_1,j_2,\ldots,j_L}$ be the counts of $(j_1, j_2, \ldots, j_L)$, usually zero.
- Let $K_{j_\ell,j_{\ell+1}}$ be the marginal counts.
- Let $\tilde{a}$ be the Markov distribution on $(j_1, j_2, \ldots, j_L)$, consistent with the pairwise marginals $\tilde{a}_{j_\ell,j_{\ell+1}} = K_{j_\ell,j_{\ell+1}}/M$.
- Marginals $\tilde{a}_{j_\ell} = K_{j_\ell}/M$.
- Conditionals $\tilde{a}_{j_{\ell+1}|j_\ell} = K_{j_\ell,j_{\ell+1}}/K_{j_\ell}$ (when $K_{j_\ell} > 0$ and $0/0 = 0$ otherwise).