Gaussian Complexity, Metric Entropy & Risk of Deep Nets

Andrew R. Barron

Yale Dept Statistics and Data Science

Joint work with Jason Klusowski

Rutgers Dept Statistics and Biostatistics

NESS Presentation in Celebration of Rick Vitale

May 16, 2019

Target of Investigation

- Deep Nets: $f(x, W)$. Inputs x in $[-1, 1]^d$. Weights W. Rectified linear activation functions. L layers.
- Network Variation V: Sums of weights of network paths.
- Risk bound: Least squares $\hat{f}$. Observations $Y_i = f(X_i) + \epsilon_i$ with (sub-)Gaussian error, sample size n.

$$E[\|\hat{f} - f\|^2] \leq V \left(\frac{L + \log d}{n} \right)^{1/2}$$

- Precursor Work: Neyshabur et al ('15), Golowich et al ('18), Barron & Klusowski ('18) with other complexity controls.
- Gaussian process comparison inequalities: Key to provide the risk bounds in current form.

Geometric width of sets

- Arbitrary set of interest: A_n in R^n. For statistical application

$$A_n = \mathcal{F}_{x^n} = \{(f(x_1), f(x_2), \ldots, f(x_n)) \, : \, f \in \mathcal{F}\}$$

restriction of a class $\mathcal{F}$ of functions to data $x_1, x_2, \ldots, x_n$.

- Half space in direction determined by $\xi = (\xi_1, \xi_2, \ldots, \xi_n)$ with threshold t

$$\{a : \xi \cdot a \leq t\}$$

- Half space supporting A_n in the direction determined by ξ uses the threshold

$$t_n = t_n(\xi, A_n) = \sup_{a \in A_n} \xi \cdot a$$

- Support function $t_n(\xi, A_n)$ is "width" of A_n in direction ξ. The least threshold such that the half space contains A_n.

Probabilistic Geometry Width

- Probabilistic width: for random ξ with distribution μ.
- Mean width: The μ complexity of A_n

$$C_\mu(A_n) = E_\xi \sup_{a \in A_n} \xi \cdot a$$

- Cummulant generating function of the width:

$$C_{\lambda,\mu}(A_n) = \frac{1}{\lambda} \log E[e^{\lambda \sup_{a \in A_n} \xi \cdot a}]$$

- General width: Positive increasing convex g with inverse ψ

$$C_{g,\mu}(A_n) = \psi(E[g(\sup_{a \in A_n} \xi \cdot a])$$

- For Rademacher Complexity: ξ_i indep symmetric Bernoulli
- For Gaussian Complexity: ξ_i independent Gaussian
- Some relationship: Tomczak-Jaegermann ('89). There are positive constants $\underline{c}$, $\overline{c}$ such that for all A_n

$$\underline{c}\, C_{Rad}(A_n) \leq C_{Gaussian}(A_n) \leq \overline{c}\, C_{Rad}(A_n) \log n$$

Random process perspective

- Random process: indexed by a in A_n

$$Z_a = \xi \cdot a = \sum_{i=1}^{n} a_i \xi_i$$

- This Z_a is of course a Gaussian process if ξ is Gaussian

- Isometry: If ξ has identity covariance then

$$E[(Z_a - Z_b)^2] = \|a - b\|^2$$

- Probabilistic width studies the maximum of the process

$$C_\mu(A_n) = E[\sup_{a \in A_n} Z_a]$$

Merit of Gaussian versus Rademacher Complexity

- More general error distributions: sub-Gaussian instead of bounded error
- Stronger link to the metric entropy: via Sudakov and Dudley inequalities. The Sudakov lower bound can also be revealed via statistical risk and information theory analysis using Fano's inequality.
- Analogous contraction properties: Most important for our present purposes.

Gaussian Comparison Inequality

- Let $\tilde{Z}_a$ be Gaussian majorized by Z_a in expectation

$$E[\tilde{Z}_a^2] \leq E[Z_a^2] \qquad *$$

and

$$E[(\tilde{Z}_a - \tilde{Z}_b)^2] \leq E[(Z_a - Z_b)^2]$$

- By Vitale (2000), equation 13, for increasing convex g,

$$E[g(\sup_{a \in A_n} \tilde{Z}_a)] \leq E[g(\sup_{a \in A_n} Z_a)]$$

- Refines Fernique (1975) which worked with

$$E[\sup_{a,b \in A_n} (Z_a - Z_b)]$$

- Refines Slepian (1962) which assumed equality in $*$.
- Avoids a factor of 2.

Contraction Inequality

- Let ϕ be a contraction: Lipshitz 1 with $\phi(0) = 0$.
- Compare the processes:

$$\tilde{Z}_a = \sum_i \xi_i \, \phi(a_i) \text{ and } Z_a = \sum_i \xi_i \, a_i$$

- Satisfy the majorization inequalities: $E\tilde{Z}_a^2 \leq EZ_a^2$ and

$$E(\tilde{Z}_a - \tilde{Z}_b)^2 \leq E(Z_a - Z_b)^2$$

since this becomes

$$\sum (\phi(a_i) - \phi(b_i))^2 \leq \sum (a_i - b_i)^2$$

- Consequent contraction of complexity: In Gaussian ξ case

$$E[\sup_{a \in A_n} g(\sum \xi_i \phi(a_i))] \leq E[\sup_{a \in A_n} g(\sum \xi_i a_i)]$$

This Gaussian complexity contraction is an extension (with different proof) of the Rademaker complexity contraction obtained by Ledoux and Talagrand ('91), inequality (4.20).

Network Layer Complexity Comparison

- For arbitrary set A in R^n and a contraction ϕ, let $\phi \circ A$ be

$$\{(\phi(a_i), \phi(a_2), \ldots, \phi(a_n)) \, : \, a \in A$$

- and let $\text{conv}(\pm A)$ be the signed convex hull

$$\{\sum w_j \, \underline{a}_j \, : \, \underline{a}_j \in A \, , \sum |w_j| = 1\}$$

- $A' = \text{conv}(\pm \phi \circ A)$ is the set of values realizable by a layer of a network for given original input values.
- As in Neyshabur et al ('15) and Golowich et al ('18), which was for Rademachers, we have also for Gaussian complexity

$$C(A') \leq 2C(A)$$

and

$$C_\lambda(A') \leq C_\lambda(A) + (\log 2)/\lambda$$

- What happens with multiple layers?

Multilayer networks for given inputs

- Set of input vectors: $A^0 = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_d\}$ each in R^n.
- Set of one layer network outputs: restricted to said inputs

$$A^1 = conv(\pm\phi \circ A^0)$$

- Intermediate layers: preserving unit total weight variation

$$A^\ell = (A^{\ell-1})' = conv(\pm\phi \circ A^{\ell-1})$$

- Set of L layer networks outputs: restricted to said inputs

$$A^L = (((A^0)')' \ldots)'$$

Tracking Complexity through the layers

- Assume each given $x_{i,j}$ has magnitude not exceeding 1
- Initial complexity of signed input set: $C(\pm A^0) \leq C_\lambda(\pm A^0)$.
- A familar bound often attributed to Massart uses a cummulant generating function trick and replaces the supremum by a sum.
- Resulting complexity is not more than

$$C_\lambda(\pm A^0) \leq n\lambda/2 + (1/\lambda)\log(2d)$$

- when optimized over λ yields the complexity bound

$$C(\pm A^0) \leq \sqrt{2n\log(2d)}.$$

Multilayer Complexity

- Intermediate layer complexity: for $A^\ell = conv(\pm\phi \circ A^{\ell-1})$

$$C(A^\ell) \le 2C(A^{\ell-1}) \quad \text{and} \quad C_\lambda(A^\ell) \le C_\lambda(A^{\ell-1}) + (\log 2)/\lambda$$

- Complexity for the class of L layer networks:
- Crude: $C(A^L) \le 2^L C(A^0)$.
- Better: $C(A^L) \le C_\lambda(A^L) \le C_\lambda(A^0) + (L \log 2)/\lambda$

- Optimized Complexity bound

$$C(A^L) \leq \sqrt{2n[L \log 2 + \log 2d]}$$

Follows Golowich et al, but now, thanks to Vitale's comparison inequality it is seen to hold for Gaussian complexity and not just Rademacher.

- Corresponding risk: based on $C(A^L)/n$ equal to

$$\left(\frac{2L \log 2 + 2 \log 2d}{n} \right)^{1/2}$$

Standard Deep Network Formulation

- Deep net function $f(W, x)$, weights W, inputs x in $[-1, 1]^d$,

$$\phi_{out}\Big(\sum_{j_1} w_{j_1}\phi\Big(\sum_{j_2} w_{j_1,j_2}\phi\Big(\sum_{j_3} w_{j_2,j_3}\cdots\phi\Big(\sum_{j_L} w_{j_{L-1},j_L}x_{j_L}\Big)\Big)\Big)\Big),$$

  where ϕ_{out} is any specified Lipschitz(1) function.

- Activation functions are $\pm$positive part, rectified linear units
  $\phi(z) = \quad (z)_+$ for first half of nodes on each layer
  $\phi(z) = -(z)_+$ for the second half.

- Weights $w_{j_{\ell-1},j_\ell}$ may thus be arranged to be non-negative.

- Computation at node j_ℓ on layer ℓ.

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

Composite Weights and their Variation

- 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

$$V = \sum_{j_1, \ldots, j_L} w_{j_1, \ldots, j_L}.$$

Probabilistic Characterization of Deep Nets

- Path weights provide a joint probability distribution

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

- It has a Markov structure

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

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

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

- Iterated expectation representation, interspersed with nonlinearities

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

Interpretation of Variation

Interpretations of the Variation V

- Probabilistic: The total variation of the measure $W = V\,q$ provided by the weight paths.
- Calculus: With one hidden layer, as in B.1991, V extends the notion of bounded variation of a function on an interval (with respect to unit step functions) to functions in R^d (with respect to half spaces). Generalizes to variation of functions with respect to depth $L-1$ subnets.
- Functional Analysis: V is the atomic norm of f with respect to depth $L-1$ subnets.
- Range: For x in $[-1,1]^d$ the range of $f(x,W)$ is in $[-V,V]$.
- Linear Algebra: V is the entry-sum of the product of the weight matrices $W_1 W_2 \cdots W_L$, where $(W_\ell)_{j_{\ell-1},j_\ell} = w_{j_{\ell-1},j_\ell}$.

Sparse Deep Net Approximation

- Approximate the weights q by $\tilde{q}$ from a sparse set.
- Draw sample, size M, independent from distrib $q_{j_1,j_2,\ldots,j_L}$
- 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 $\tilde{q}_{j_\ell,j_{\ell+1}} = K_{j_\ell,j_{\ell+1}}/M$.
- Marginals $\tilde{q}_{j_\ell} = K_{j_\ell}/M$.
- Conditionals $\tilde{q}_{j_{\ell+1}|j_\ell} = K_{j_\ell,j_{\ell+1}}/K_{j_\ell}$ (when $K_{j_\ell} > 0$ and $0/0 = 0$ otherwise).
- Size of set of indices $j_1, j_2, \ldots, j_L$

$$D = d_1 d_2 \cdots d_L = d^L$$

where d is the geometric mean of $d_1, d_2, \ldots, d_L$.

Size of Cover of Deep Nets

- Log Cardinality of set of counts with specified sum M

$$log\binom{M+D-1}{M} \leq M \log(2ed_1 d_2 \cdots d_L/M) \leq ML \log d$$

- At each layer, at most M of the nodes can have positive weight, at most M from first half and at most M from second half. So when $d_\ell \geq 2M$ may replace d_ℓ with $d_\ell^{new} = \min\{d_\ell, 2M\}$ in representation of $f(\tilde{a}, x)$.
- Refined Log Cardinality bound

$$(L-2)M \log(\min\{\bar{d}, 2M\}) \; + \; M \log(4e\, d_{in}),$$

where $\bar{d}$ is the geometric mean of $d_2, d_3, \ldots, d_{L-1}$.

- The bound is independent of d_1.

Accuracy of Deep Net Cover

- Use the $L_2(P)$ norm for any $P = P_X$ on $[-1,1]^d$,

$$\|f(\cdot,q) - f(\cdot,\tilde{q})\|^2 = \int (f(x,q) - f(x,\tilde{q}))^2 P_X(dx).$$

- For each q there is a representor $\tilde{q}$ such that

$$\|f(\cdot,q) - f(\cdot,\tilde{q})\| \leq C_v \frac{L}{M^{1/2}}$$

Also

$$\|f(\cdot,q) - f(\cdot,\tilde{q})\| \leq 2\, C_v^{red} \frac{L}{M^{1/2}}$$

- Variation coefficient

$$C_v = \frac{1}{L} \sum_{\ell=0}^{L-1} \sum_{j_\ell} (V_{j_\ell}^{out} V_{j_\ell}^{in} / V)^{1/2} \leq \overline{V} / V^{1/2}$$

- C_v^{red} is the same but with $V_{j_\ell}^{in,red}$ in place of $V_{j_\ell}^{in}$ with the largest incoming weighted sub-variation via $j_{\ell+1}^*$ removed.

Data Setting

- Data: $(\underline{X}_i, Y_i)$, $i = 1, 2, \ldots, n$
- Inputs: explanatory variable vectors with arbitrary dependence
$$\underline{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})$$
- Domain: Cube $[-1, 1]^d$ in R^d
- Random design: independent $\underline{X}_i \sim P$
- Output: response variable Y_i in R
  - Bounded or subgaussian
- Relationship: $E[Y_i|\underline{X}_i] = f(\underline{X}_i)$ as in:
  - Perfect observation: $Y_i = f(\underline{X}_i)$
  - Noisy observation: $Y_i = f(\underline{X}_i) + \epsilon_i$ with ϵ_i indep

Statistical Risk

- Statistical risk $E\|\hat{f} - f\|^2 = E(\hat{f}(\underline{X}) - f(\underline{X}))^2$
- Expected squared generalization error on new $\underline{X} \sim P$
- Minimax optimal risk, in the class $\mathcal{F}_v$ of functions with composite variation not more than v

$$E\|\hat{f} - f\|^2 \leq \|f_M - f\|^2 + c\frac{1}{n}\log N(\mathcal{F}, \delta_M)$$

with $\log N(\mathcal{F}, \delta)$ the metric entropy of $\mathcal{F}$ at $\delta_M = \|f_M - f\|$

- Achieves ideal approximation, complexity trade-off.

Statistical Risk

- Statistical risk $E\|\hat{f} - f\|^2 = E(\hat{f}(\underline{X}) - f(\underline{X}))^2$
- Expected squared generalization error on new $\underline{X} \sim P$
- Minimax optimal risk, in a class $\mathcal{F}$ of functions

$$E\|\hat{f} - f\|^2 \ \leq \ \|f_M - f\|^2 + c\frac{1}{n}\log N(\mathcal{F}, \delta_M)$$

with $\log N(\mathcal{F}, \delta)$ the metric entropy of $\mathcal{F}$ at $\delta_M = \|f_M - f\|$

- Specializing to the class $\mathcal{F}_v$ of functions with composite variation $V\,C_v^{red}$ not more than v

$$E\|\hat{f} - f\|^2 \ \leq \ \frac{(L\,v)^2}{M} + c\,\frac{LM\log d}{n}$$

- With best M, our risk bound is

$$E\|\hat{f} - f\|^2 \ \leq \ 2v\left(\frac{c\,L^3\log d}{n}\right)^{1/2}$$

Summary

- Formulation of Deep Nets: Positive part activation function provides weight homogeneity, probabilistic interpretation.
- Average Variation $\overline{V}$ and Geometric Mean Variation C_v: Extends notion to multi-layer nets and their sub-nets
- Metric Entropy: Simple log-cardinality bound $LM \log d$
- Approximation: Sample M paths and set weights based on second order counts.
- Approximation bound: From telescoping control of accuracy of each layer. Bounds $C_v L/\sqrt{M}$ and $2C_v^{red} L/\sqrt{M}$.
- Estimation: Risk bound $C_v \left(\frac{L^3 \log d}{n}\right)^{1/2}$. From trade-off of approximation and complexity relative to the sample size n.

The norm of f with respect to a dictionary G

- Minimal ℓ_1 norm on coefficients in approximation of f

$$\|f\|_G = \lim_{\epsilon \longrightarrow 0} \inf \{ \sum_j |c_j| \ : \ \| \sum_j c_j g_{a_j} - f \| \leq \epsilon \}$$

- Also called the variation of f with respect to G (B. 1991)

$$\|f\|_G = V_G(f) = \inf\{V \ : \ f/V \in \mathit{closure}(\mathit{conv}(\pm G))\}$$

- Later also called the atomic norm of f with respect to G

Summary

- Flexible approximation models
  - Subset selection
  - Nonlinearly parameterized bases as with neural nets
  - ℓ_1 control on coefficients of combination
- Accurate approximation with moderate number of terms
  - Proof analogous to random coding
- Information theoretic risk bounds
  - Based on the minimum description length principle
  - Shows accurate estimation, even for very large dimension
- Computational challenges are being addressed by
  - Adaptive annealing
  - Nonlinear power methods