Gaussian Complexity, Metric Entropy, and Statistical Learning of Deep Nets

Andrew R. Barron

YALE UNIVERSITY
Department of Statistics and Data Science

Presentation, September 17, 2019

Joint work with Jason Klusowski, Rutgers University

IMA Workshop on Foundations of Data Science

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

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

- 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: $E\tilde{Z}_a^2 \le EZ_a^2$ and

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

since this becomes

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

- Consequent contraction of complexity: In Gaussian ξ case

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

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 conv($\pm A$) be the signed convex hull

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

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

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

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

- Optimized Complexity bound

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

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

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

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
  - $f(x)$ assumed Bounded by a constant B
  - ϵ assumed subGaussian with parameter σ

Statistical Risk

- Statistical risk $E\|\hat{f} - f\|^2 = E(\hat{f}(\underline{X}) - f(\underline{X}))^2$
- Expected squared generalization error on new $\underline{X} \sim P$
- Approximation, complexity trade-off

$$E\|\hat{f} - f\|^2 \;\le\; \|f_\delta - f\|^2 + c\frac{1}{n}\log N(\mathcal{F}, \delta)$$

- the metric entropy $\log N(\mathcal{F}, \delta)$ is the smallest log cardinality of cover such that for all $f \in \mathcal{F}$ there is an approximation f_δ in the cover with $\|f_\delta - f\| \le \delta$.
- The minimax risk corresponds to the optimal approximations, complexity tradeoff,

$$r_n(\mathcal{F}) = \min_{\hat{f}} \max_{f \in \mathcal{F}} E\|\hat{f} - f\|^2 \;\approx\; \min_{\delta} \left\{ \delta^2 + c\frac{1}{n}\log N(\mathcal{F}, \delta) \right\}$$

(Yuhong Yang and A.B. 1998).

Metric Entropy and Gaussian Complexity

Relationship between metric entropy, Gaussian Complexity, and statistical risk

- If $\mathcal{F}$ has Gaussian complexity not more than $\sqrt{n}\, C_{\mathcal{F}}$ then it has the risk bound

$$r_n(\mathcal{F}) \leq (B+\sigma)\frac{C_{\mathcal{F}}}{\sqrt{n}}$$

  and the metric entropy bound

$$\log N(\delta, \mathcal{F}) \leq \frac{16 C_{\mathcal{F}}^2}{\delta^2}$$

- The latter is an instance of a Sudakov inequality relating metric entropy and Gaussian complexity.
- It can be seen as a consequence of the risk bound together with an information theory argument (via the Fano inequality in a manner similar to Yang and B. 1998)

Statistical Risk for the Neural Net class

- Specializing to the class $\mathcal{F}_V$ of networks with variation not more than V our risk bound is

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