7. Approximation: Sampling bounds (cont’d)

Recap

- Deriving sampling-based approximation bounds for neural networks
  - **Part 1.** GT is an ∞-width two-layer threshold network
  - **Part 2.** Sampling m neurons give you a good approximation

- In part 1, we showed that

Theorem (informal).

Under some conditions, we have

$$g(\mathbf{x}) = \int\int q(\mathbf{w}, b) \cdot \mathbf{1}[\mathbf{w}^\top \mathbf{x} \geq b]\, \mathrm{d}\mathbf{w}\, \mathrm{d}b$$

for some parameter density $q(\mathbf{w}, b)$.

Recap

- For part 2, we have studied a powerful tool:

Lemma (Maurey)

Let $\mathbf{V}$ be a random element in some Hilbert space, supported on the set $\mathcal{S}$, and let $X = \mathbb{E}\mathbf{V}$.

Let $(\mathbf{V}_1, \ldots, \mathbf{V}_m)$ be i.i.d. draws of $\mathbf{V}$. Then, we have

$$\mathbb{E}\left\|X - \frac{1}{m}\sum_{i=1}^{m}\mathbf{V}_i\right\|^2 \leq \frac{\mathrm{Var}(\mathbf{V})}{m} \leq \frac{\mathbb{E}\|\mathbf{V}\|^2}{m} \leq \frac{\sup_{U\in\mathcal{S}}\|U\|^2}{m}$$

Moreover, there exists $U_1, \ldots, U_m \in \mathcal{S}$ such that

$$\left\|X - \frac{1}{m}\sum_{i=1}^{m}U_i\right\|^2 \leq \mathbb{E}\left\|X - \frac{1}{m}\sum_{i=1}^{m}\mathbf{V}_i\right\|^2$$

Why is Maurey great?

- Maurey's method is quite versatile — if we choose the right $\mathbf{V}$, one can show the results like:

Corollary.

Let B_1 be a unit ball in $\mathbb{R}^d$. Consider covering this ball with ℓ_2-norm balls with radius ε.

Let $N(B_1, \|\cdot\|_2, \varepsilon)$ be the covering number, i.e., the minimum number of ℓ_2 balls so that the union of these balls have B_1 as a subset.

Then, we have:

$$\log N(B_1, \|\cdot\|_2, \varepsilon) \leq \min\left\{2d \log\left(1 + \frac{1}{2\varepsilon^2 d}\right), \quad \frac{1}{\varepsilon^2}\log(1 + 2d\varepsilon^2)\right\}$$

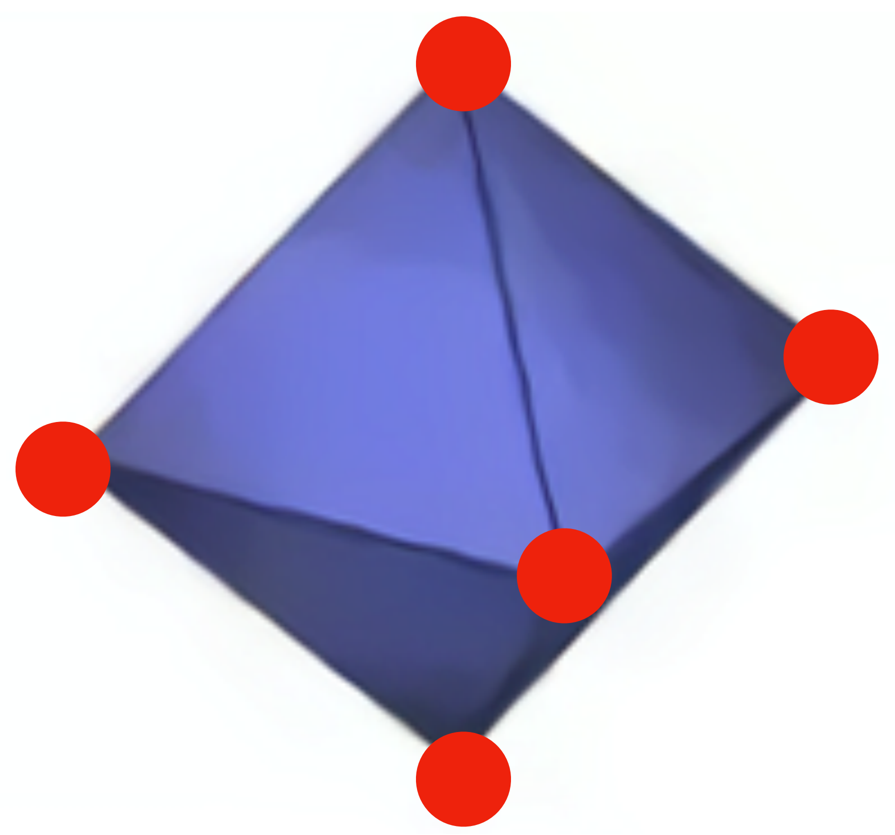
- Note. There should be a wrong term here…

Proof idea

- Select an arbitrary $X \in B_1$
- Define a $d-$dimensional random vector

$$\mathbf{V} = \begin{cases} \text{sgn}(x_i)e_i & \text{w.p. } |x_i| \\ 0 & \text{w.p. } 1 - \|x\|_1 \end{cases}$$

- Then, we know that
  - $\mathbb{E}[\mathbf{V}] = X$
  - $\mathbf{V}$ is supported on the origin & critical points (total 2d+1 points)
    - Call this set $\mathcal{S}$



Proof idea

- By Maurey, exists some $U_1, \ldots, U_m \in \mathcal{S}$ such that

$$\left\| X - \frac{1}{m} \sum_{i=1}^{m} U_i \right\|_2^2 \leq \frac{\sup_{U \in \mathcal{S}} \|U\|_2^2}{m} = \frac{1}{m}$$

  - We'll set $m = 1/\varepsilon^2$

- Now, examine the number of distinct values that

$$\bar{U} = \frac{1}{m} \sum_{i=1}^{m} U_i$$

  can have, regardless of the choice of X

- Any volunteer? 🙋‍♂️

Proof idea

- WLOG, we can count the number of distinct values for

$$\sum_{i=1}^{m} U_i$$

- Define

$$m_j^+ = \sum_{i=1}^{m} \mathbf{1}\{U_i = +e_j\}, \qquad m_j^- = \sum_{i=1}^{m} \mathbf{1}\{U_i = -e_j\}, \qquad m_0 = \sum_{i=1}^{m} \mathbf{1}\{U_i = \mathbf{0}\}$$

  - Then, this satisfies

$$m_0 + \sum_{j=1}^{d} (m_j^+ + m_j^-) \leq m, \qquad 0 \leq m_0, m_j^+, m_j^- \leq m$$

  - We'll count the number of all m_j^+, m_j^- that satisfies the above
    - This will be an upper bound of the original quantity considered — as we dropped a constraint that either m_j^+ or m_j^- should be zero

Proof idea

$$\sum_{j=1}^{d} (m_j^+ + m_j^-) \leq m, \qquad 0 \leq m_j^+, m_j^- \leq m$$

- This is like placing m identical balls in $2d + 1$ rooms
    - Rooms: $m_0, m_1^+, m_1^-, \cdots$
    - Balls: $U_1, \ldots, U_m$

- The total number of choices is

$$\binom{2d+m}{m} = \binom{2d+m}{2d}$$

- Apply the binomial upper bound $\binom{n}{k} \leq (n \cdot e/k)^k$ to get the results

Importance sampling

Importance sampling

- From Maurey, we have that:
  - There exists some neurons $f_1, \ldots, f_m$ such that:

$$\left\| g - \frac{1}{m} \sum_{i=1}^{m} f_i \right\|^2 \leq \frac{\mathrm{Var(neuron)}}{m}$$

  where Var(neuron) denotes the variance of neuron drawn from g

- **Question.** How do we minimize the variance of neurons, by decomposing

$$q(\mathbf{w}, b) = \pi(\mathbf{w}, b) \cdot a(\mathbf{w}, b)$$

  for the GT density

$$g(\mathbf{x}) = \iint q(\mathbf{w}, b) \cdot \mathbf{1}[\mathbf{w}^\top \mathbf{x} \geq b] \, \mathrm{d}\mathbf{w} \, \mathrm{d}b$$

Importance sampling

- Consider a simplified version of our question — fix x
  - GT can be written as:

$$g = \int \pi(z) \cdot \frac{q(z)}{\pi(z)} \cdot \eta(z)\,\mathrm{d}z$$

    - $z = (\mathbf{w}, b)$ Parameterization of each neuron
    - $\pi(z)$ Sampling probability
    - $q(z)/\pi(z)$ 2nd layer weight
    - $\eta(z)$ 1st layer outputs

Importance sampling

- **Want to solve.**

$$\min_{\pi} \mathrm{Var}_{z \sim \pi}\left(\frac{q(z)}{\pi(z)}\eta(z)\right)$$

  - Any volunteer? 🙋‍♂️

Importance sampling

- **Solution.** Select

$$\pi(z) \propto |q(z)| \cdot \eta(z)$$

- **Proof.** ✏️

Combining the tools: Univariate case

Summing up: Univariate case

- Consider the univariate case
    - We have a GT network

$$g(x) = \int_0^1 g'(b) \cdot \mathbf{1}[x \geq b] \, \mathrm{d}b$$

- We want to:
    - Come up with a good sampling distribution $\pi(b)$
    - Provide a clean bound on the approximation error

Summing up: Univariate case

$$g(x) = \int_0^1 g'(b) \cdot \mathbf{1}[x \geq b] \, \mathrm{d}b$$

- The importance sampling tells us that we should use the sampling distribution:

$$\pi(b) \propto |g'(b)| \cdot \mathbf{1}\{x \geq b\}$$

    - However, there is a term about x
        - Our sampling should be input-independent

- **Solution.** Simply ignore it and use

$$\pi(b) \propto |g'(b)|$$

    - May not be ideal, but good enough

Summing up: Univariate case

- The sampling scheme becomes:
  - Draw $b_1, \ldots, b_m$ using

$$\pi(b) = \frac{|g'(b)|}{\int |g'(b)| \mathrm{d}b} =: \frac{|g'(b)|}{G'}$$

  - The finite-width network will be

$$f(x) = \sum_{i=1}^{m} \frac{G'}{m} \cdot \mathrm{sgn}(g'(b_i)) \cdot \mathbf{1}\{x \geq b_i\}$$

    - The second layer weights are simply ± 1

Summing up: Univariate case

- The variance is upper-bounded by a term proportional to:

$$\frac{(G')^2}{m}$$

- That is, this guarantee accounts for the flatness of the GT function
    - Exercise. Check this

Multivariate case

- Similar logic, but much dirtier
  - Read the textbook for details
  - Similar dependency on:

$$\frac{\left(\int \|\widetilde{\nabla g}\| \mathrm{d}w\right)^2}{m}$$

Next up

- Near-initialization approximation and kernel regime
- Benefits of depth