7. Approximation: Sampling bounds

Recap

- In the last lecture, we have established 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 $q(\mathbf{w}, b)$.

- Slightly rephrasing, can be written as:

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

  - π: probability of drawing some neuron
  - a: 2nd layer weights
- Note: There are many different ways to decompose!

Today

- We sample the neurons to construct a finite-width network
    - Independently draw m neurons $(\mathbf{w}_i, b_i) \sim \pi$
    - Construct

$$f(\mathbf{x}) = \sum_{i=1}^{m} \frac{1}{m} \cdot a(\mathbf{w}_i, b_i) \cdot \mathbf{1}\{\mathbf{w}_i^\top \mathbf{x} \geq b_i\}$$

- **Claim.**
    - DON'T: Any $f(\cdot)$ will be close to $g(\cdot)$ if m grows (way too pessimistic)
    - DO: There is at least one $f(\cdot)$ that is close to $g(\cdot)$
        - Turns out that how we decompose to π, a matters

Overview

- **Want-to-show:** "There is at least one $f(\cdot)$ that is close to $g(\cdot)$"

- We will show this in three steps

  - If f and g are similar in expectation, there exists at least one f that is close to g
    - random coding
  - If each neuron has a small variance, f is close to its mean in expectation
    - Maurey's empirical method
  - We can make neuron variance small by tuning (π, a)
    - importance sampling

Random coding

Random coding argument

- Roughly, want to show that

"If f and g are similar in expectation, there exists at least one f that is close to g"

Claim.

Let ν be a distribution of functions, from which we can sample. Suppose that we have

$$\mathbb{E}_{f\sim\nu}[\|f-g\|^2] \le \varepsilon$$

Then, there exists at least one $f^* \in \mathrm{supp}(\nu)$ such that

$$\|f^* - g\|^2 \le \varepsilon$$

- **Proof.** Volunteer?

Random coding argument

- **Proof.** By contradiction ✏️

- **Trivia.** Called “random coding” argument, in information theory
  - due to Shannon / Erdös
  - also known as “probabilistic method”

Maurey’s empirical method
(a.k.a. Maurey’s sparsification)

Rough claim

- Roughly, we wanted to show:

"If each neuron has a small variance, f is close to its mean in expectation"

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

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

Rough claim

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

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

- Looks way too complicated?
    - Let’s find out and remove the easiest parts so that we can focus on others.

Rough claim

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

- To show this, we can simply proceed as:

$$\mathbb{E}\left\| X - \frac{1}{m}\sum \mathbf{V}_i \right\|^2 = \mathbb{E}\left\| \frac{1}{m}\sum (X - \mathbf{V}_i) \right\|^2$$

$$= \frac{1}{m^2}\mathbb{E}\left\| \sum (X - \mathbf{V}_i) \right\|^2$$

Rough claim

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

- To show this, we can simply proceed as:

$$\begin{aligned}
\mathbb{E}\left\|X - \frac{1}{m}\sum \mathbf{V}_i\right\|^2 &= \mathbb{E}\left\|\frac{1}{m}\sum (X - \mathbf{V}_i)\right\|^2 \\
&= \frac{1}{m^2}\mathbb{E}\left\|\sum (X - \mathbf{V}_i)\right\|^2 \\
&= \frac{1}{m^2}\mathbb{E}\left(\sum (X - \mathbf{V}_i)^2 + \sum_{i \neq j} \langle X - \mathbf{V}_i, X - \mathbf{V}_j \rangle\right)
\end{aligned}$$

Rough claim

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

- To show this, we can simply proceed as:

$$\begin{aligned}
\mathbb{E}\left\|X - \frac{1}{m}\sum \mathbf{V}_i\right\|^2 &= \mathbb{E}\left\|\frac{1}{m}\sum (X - \mathbf{V}_i)\right\|^2 \\
&= \frac{1}{m^2}\mathbb{E}\left\|\sum (X - \mathbf{V}_i)\right\|^2 \\
&= \frac{1}{m^2}\mathbb{E}\left(\sum (X - \mathbf{V}_i)^2 + \sum_{i \neq j} \langle X - \mathbf{V}_i, X - \mathbf{V}_j\rangle\right) \\
&= \frac{1}{m^2}\mathbb{E}\left(\sum (X - \mathbf{V}_i)^2\right)
\end{aligned}$$

Rough claim

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

- To show this, we can simply proceed as:

$$\mathbb{E}\left\|X - \frac{1}{m}\sum \mathbf{V}_i\right\|^2 = \mathbb{E}\left\|\frac{1}{m}\sum (X - \mathbf{V}_i)\right\|^2$$

$$= \frac{1}{m^2}\mathbb{E}\left\|\sum (X - \mathbf{V}_i)\right\|^2$$

$$= \frac{1}{m^2}\mathbb{E}\left(\sum (X - \mathbf{V}_i)^2 + \sum_{i\neq j}\langle X - \mathbf{V}_i, X - \mathbf{V}_j\rangle\right)$$

$$= \frac{1}{m^2}\mathbb{E}\left(\sum (X - \mathbf{V}_i)^2\right)$$

$$= \frac{1}{m^2}\sum \mathbb{E}(X - \mathbf{V})^2 \qquad = \frac{1}{m}\mathbb{E}(X - \mathbf{V})^2$$

Why the special name?

- 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...

Importance sampling

TBD

- TBD!