

# SDS 384-11 PS #5, Spring 2020

Instructor: Purnamrita Sarkar

Solutions by: Anonymous

Due: Friday, May 8, 2020 at 11:59pm

Last modified: Tuesday 12<sup>th</sup> May, 2020 at 9:40am

## 1 Uniform Law of Large Numbers

**Exercise 1.1.** *In class you upper bounded the Rademacher complexity of a function class. Now you will derive a lower bound.*

1. For function classes  $\mathcal{F}$  with function values in  $[0, 1]$ , prove that  $\mathbb{E}[\hat{P}_n - P]_{\mathcal{F}} \geq \frac{\mathcal{R}_{\mathcal{F}}}{2} - \sqrt{\frac{\log 2}{2n}}$ .

*Hint: it may be easier to start from  $\mathcal{R}_{\mathcal{F}}$  and show that  $\mathcal{R}_{\mathcal{F}} \leq 2\mathbb{E}\|\hat{P}_n - P\|_{\mathcal{F}} + \sqrt{\frac{2\log 2}{n}}$ . In order to do this, you would need to add and subtract  $\mathbb{E}f(X)$  and then use triangle inequality.*

*Solution.*

Let  $X = \{X_1, \dots, X_n\}$  be i.i.d samples from a distribution  $P$ , and let  $\varepsilon \in \{-1, 1\}^n$  be a vector of i.i.d Rademacher random variables. Recalling the definition of  $\mathcal{R}_{\mathcal{F}}$ , and applying a triangle inequality, we have that

$$\begin{aligned} \mathcal{R}_{\mathcal{F}} &= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i f(X_i) \right| \\ &= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X_i) + \mathbb{E}f(X_i)) \right| \\ &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X_i)) \right| + \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i \mathbb{E}f(X_i) \right| \\ &= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X_i)) \right| + \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{E}f(X_i) \left| \frac{1}{n} \sum_i \varepsilon_i \right| \quad \text{since } f \geq 0 \text{ and } X_i \text{ i.i.d} \\ &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X_i)) \right| + \mathbb{E} \left| \left\langle \varepsilon, \frac{1}{n} \mathbf{1} \right\rangle \right| \quad \text{since } f \leq 1 \end{aligned}$$

We now proceed by bounding each of these two terms. To bound the first term, we will employ a similar symmetry argument as presented in class. In particular, we take  $X' = \{X'_1, \dots, X'_n\}$

as an i.i.d copy of  $X$ . Then the first term above can be written as follows:

$$\begin{aligned}
& \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X_i)) \right| \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X'_i)) \right| \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X'} \left[ \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - f(X'_i)) \right] \right| \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - f(X'_i)) \right| && \text{by Jensen's and convexity of } |\cdot| \text{ and sup} \\
&= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i f(X_i) - f(X'_i) \right| && \text{by symmetry of } f(X_i) - f(X'_i) \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i f(X_i) - \mathbb{E}f(X_i) + \mathbb{E}f(X'_i) - f(X'_i) \right|
\end{aligned}$$

We may then apply another triangle inequality to the above, recall the definition of  $\|P - \hat{P}_n\|_{\mathcal{F}}$ , and combine our inequalities above to conclude that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_i \varepsilon_i (f(X_i) - \mathbb{E}f(X_i)) \right| \leq 2\mathbb{E}\|P - \hat{P}_n\|_{\mathcal{F}}$$

To conclude, we observe that, by the finite class lemma, taking the set  $A = \{\frac{1}{n}\mathbf{1}\}$ , and thus  $R = \|\frac{1}{n}\mathbf{1}\|_2 = \frac{1}{\sqrt{n}}$  we have

$$\mathbb{E} \left| \left\langle \varepsilon, \frac{1}{n}\mathbf{1} \right\rangle \right| \leq \sqrt{\frac{2 \log 2}{n}}$$

Combining these results and rearranging yields the claimed bound.  $\square$

2. Now prove that  $\|P - \hat{P}_n\|_{\mathcal{F}} \geq \mathbb{E}\|P - \hat{P}_n\|_{\mathcal{F}} - \epsilon$  with probability at least  $1 - \exp(-c n \epsilon^2)$ , for some constant  $c$ .

*Solution.*

Take  $X = \{X_1, \dots, X_n\}$  to be  $n$  i.i.d samples from some distribution  $P$ , and take  $X'$  to be  $n$  samples where  $X'_i = X_i$  for every  $i \neq j$ , and  $X'_j$  is another i.i.d sample. Let us denote

$$g(X) = \|\hat{P}_n - P\|_{\mathcal{F}}$$

then we have that  $g(X)$  satisfies the bounded difference property, since  $f \in [0, 1]$ , and since,

by triangle inequality,

$$\begin{aligned}
|g(X) - g(X')| &= \left| \left| \frac{\sum_i f(X_i) - \mathbb{E}f(X_i)}{n} \right| - \left| \frac{\sum_i f(X'_i) - \mathbb{E}f(X'_i)}{n} \right| \right| \\
&\leq \left| \frac{\sum_i f(X_i) - \mathbb{E}f(X_i)}{n} - \frac{\sum_i f(X'_i) - \mathbb{E}f(X'_i)}{n} \right| \\
&= \frac{1}{n} |(f(X_j) - f(X'_j))| \\
&= \frac{1}{n}
\end{aligned}$$

Hence, we may apply the one-sided McDiarmid's inequality to conclude that, for any  $\varepsilon > 0$ ,

$$\begin{aligned}
\mathbb{P} \left( \|\hat{P}_n - P\|_{\mathcal{F}} - \mathbb{E}\|\hat{P}_n - P\|_{\mathcal{F}} < -\varepsilon \right) &= \mathbb{P}(g(X) - \mathbb{E}g(X) < -\varepsilon) \\
&\leq \exp \left( -\frac{2\varepsilon^2}{n \frac{1}{n^2}} \right) \\
&= \exp(-2\varepsilon^2 n)
\end{aligned}$$

Thus, with probability  $1 - \exp(-2\varepsilon^2 n)$ ,  $g(X) \geq \mathbb{E}g(X) - \varepsilon$ , as desired.  $\square$

3. Recall the class of all subsets with finite size in  $[0, 1]$ . Prove that the Rademacher complexity of this class is at least  $\frac{1}{2}$ . What does this imply?

*Solution.*

**Note:** I am a bit uncertain if this question is asking for the Rademacher complexity of the *function class of indicator functions on  $S$*  or of the *set of finite subsets of  $[0, 1]$* . Because of this uncertainty, I will provide a proof of both.

- (a) Assuming the question is asking for the Rademacher complexity of the function class of indicators on  $S$

Let  $\mathcal{F}_S = \{\mathbb{1}_S : S \subset [0, 1], |S| < \infty\}$ . Let  $X_1, \dots, X_n$  be drawn i.i.d from some distribution  $P$  with no atoms. Then, taking  $\hat{S} = \{X_1, \dots, X_n\}$ , we have that

$$\begin{aligned}
\|P - \hat{P}_n\|_{\mathcal{F}_S} &= \sup_{S \subset [0, 1], |S| < \infty} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_S(X_i) - \mathbb{E}\mathbb{1}_S(X_i) \right| \\
&= \sup_{S \subset [0, 1], |S| < \infty} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_S(X_i) - \underbrace{\mathbb{P}(X_i \in S)}_{=0 \forall S \text{ since } P \text{ has no atoms}} \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\hat{S}}(X_i) - \underbrace{\mathbb{P}(X_i \in S)}_{=0 \forall S \text{ since } P \text{ has no atoms}} \right| \\
&= 1
\end{aligned}$$

Now, as we showed in class,  $\mathcal{R}_{\mathcal{F}} \geq \frac{1}{2} \mathbb{E}\|P - \hat{P}_n\|_{\mathcal{F}_S}$ , and thus by the above result,

$$\mathcal{R}_{\mathcal{F}_S} \geq \frac{1}{2}$$

Therefore, we know that  $\mathcal{F}_S$  is *not* a Glivenko-Cantelli class for any  $P$  with no atoms.

- (b) Assuming the question is asking for the Rademacher complexity of the set of finite subsets of  $[0, 1]$ .

Let  $\mathcal{T} = \{S \subset [0, 1] \mid |S| < \infty\}$ . We wish to provide a lower bound on  $\mathcal{R}(\mathcal{T})$ . Beginning with the definition, we observe that, taking  $\varepsilon$  as a vector of i.i.d Rademacher random variables,

$$\begin{aligned}\mathcal{R}(\mathcal{T}) &= \mathbb{E} \sup_{d < \infty} \sup_{\theta \in [0, 1]^d} \langle \theta, \varepsilon \rangle \\ &\geq \mathbb{E} \sup_{\theta_1 \in [0, 1]} \langle \theta_1, \varepsilon_1 \rangle && \text{by def of sup, taking } \theta_i = 0 \text{ for } i > 1 \\ &= \mathbb{E} \varepsilon_1 \mathbb{1}\{\varepsilon_1 = 1\} \\ &= \mathbb{P}(\varepsilon_1 = 1) \\ &= \frac{1}{2}\end{aligned}$$

as desired.

Now, recalling the definition of Rademacher complexity for function classes, this lower bound implies that any function class with functions which can take values in  $[0, 1]$  for any number of points, the function class is not a Glivenco-Cantelli class.

□

**Exercise 1.2.** In this exercise, we explore the connection between VC dimension and metric entropy. Given a set class  $\mathcal{S}$  with finite VC dimension  $\nu$ , we show that the function class  $\mathcal{F}_{\mathcal{S}} := \mathbb{1}_{\mathcal{S}}, \mathcal{S} \in \mathcal{S}$  of indicator functions has metric entropy at most

$$N(\delta; \mathcal{F}_{\mathcal{S}}, L^1(P)) \leq \left( \frac{K \log(3e/\delta)}{\delta} \right)^{\nu} \quad (1)$$

for a constant  $K$ .

Let  $\{\mathbb{1}_{S_1}, \dots, \mathbb{1}_{S_N}\}$  be a maximal delta packing in the  $L^1(P)$  norm, so that

$$\|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|_1 = \mathbb{E} |\mathbb{1}_{S_i}(X) - \mathbb{1}_{S_j}(X)| > \delta$$

for all  $i \neq j$ . This is an upper bound on the  $\delta$  covering number.

1. Suppose that we generate  $n$  samples  $X_i$  drawn i.i.d from  $P$ . Show that the probability that every  $S_i$  picks out a different subset  $\{X_1, \dots, X_n\}$  is at least  $1 - \binom{N}{2} (1 - \delta)^n$ .

*Solution.*

We observe that, by a union bound, and applying the above definitions,

$$\begin{aligned}
& 1 - \mathbb{P}(\text{every } S_i, i \in [N] \text{ picks different subset of } X_1, \dots, X_n) \\
&= \mathbb{P}(\text{at least two } S_i, S_j, i \neq j \text{ pick same subset}) \\
&= \mathbb{P}\left(\bigcup_{(i,j) \in \binom{[N]}{2}} \{S_i, S_j \text{ pick same subset}\}\right) \\
&\leq \binom{N}{2} \mathbb{P}(S_i, S_j \text{ pick same subset}) \\
&= \binom{N}{2} \mathbb{P}\left(\bigcap_{k=1}^n \mathbb{1}_{S_i}(X_k) = \mathbb{1}_{S_j}(X_k)\right) \\
&= \binom{N}{2} \mathbb{P}(\mathbb{1}_{S_i}(X_k) = \mathbb{1}_{S_j}(X_k))^n \\
&= \binom{N}{2} (1 - \|\mathbb{1}_{S_i} - \mathbb{1}_{S_j}\|_1)^n \\
&\leq \binom{N}{2} (1 - \delta)^n
\end{aligned}$$

Rearranging terms yields the desired inequality.  $\square$

2. Using part (a), show that for  $N \geq 2$  and  $n = \lceil 2 \log(N)/\delta \rceil$ , there exists a set of  $n$  points from which  $\mathcal{S}$  picks out at least  $N$  subsets, and conclude that  $N \leq \left(\frac{3e \log N}{\nu \delta}\right)^\nu$ .

*Solution.*

We proceed by the probabilistic method, showing that, for the stated choices of parameters,  $\binom{N}{2} (1 - \delta)^n < 1$ .

We assume without loss of generality that  $0 < \delta < 1$ . Thus, we have that

$$\begin{aligned}
\binom{N}{2} (1 - \delta)^{\lceil 2 \log(N)/\delta \rceil} &\leq \binom{N}{2} (1 - \delta)^{2 \log(N)/\delta} \\
&\stackrel{\text{want}}{<} 1
\end{aligned}$$

Taking log on both sides, it is sufficient to show that

$$\begin{aligned}
\frac{2 \log N}{\delta} \log(1 - \delta) &< -\log \binom{N}{2} \\
\iff \frac{2 \log N}{\delta} &> \frac{\log(N(N-1)/2)}{\log \frac{1}{1-\delta}}
\end{aligned}$$

Now, since  $N \geq 2$ , we have that  $N^2 > \binom{N}{2}$  and thus  $2 \log(N) > \log(N(N-1)/2)$ . Finally, using the well-known inequality  $\log \frac{1}{1-\delta} > \delta$  when  $\delta \in (0, 1)$ , we conclude that the above inequality is true. Therefore, by the probabilistic method, there exists a set of  $n$  points from which  $\mathcal{S}$  picks out at least  $N$  subsets.

Now, by definition of the growth function,  $\Pi_{\mathcal{F}_S}(n) \geq N$ . By Sauer's Lemma, we have the following bound on the growth function:

$$\begin{aligned}
N &\leq \Pi_{\mathcal{F}_S}(n) \\
&\leq \sum_{i=0}^{\nu} \binom{n}{i} \\
&\leq \left(\frac{en}{\nu}\right)^{\nu} && \text{assuming } n \geq \nu \\
&= \left(\frac{e \lceil 2 \log(N)/\delta \rceil}{\nu}\right)^{\nu} \\
&\leq \left(\frac{3e \log(N)}{\nu \delta}\right)^{\nu}
\end{aligned}$$

as desired. □

3. Use part (b) to show that Equation 1 holds with  $K = 3e^2/(e-1)$ . Hint: Note that you have  $\frac{N^{1/\nu}}{\log N} \leq \frac{3e}{\nu \delta}$ . Let  $g(x) = x/\log x$ . We are solving for  $g(m^{1/\nu}) \leq 3e/\delta$ . Prove that  $g(x) \leq y$  implies  $x \leq \frac{e}{e-1} y \log y$ .

*Solution.*

Following the hint, let us suppose that  $\frac{x}{\log x} \leq y$ . Assume that  $y > e$  and  $x > 1$ . Therefore,

$$\begin{aligned}
\frac{e}{e-1} y \log y &\geq \frac{e}{e-1} \frac{x}{\log x} (\log x - \log \log x) \\
&= \frac{e}{e-1} x - \frac{e}{e-1} \frac{x \log \log x}{\log x} \\
&\stackrel{\text{want}}{\geq} x
\end{aligned}$$

Now, the final inequality above is equivalent to

$$\frac{x}{1-e} \geq \frac{e}{e-1} \frac{x \log \log x}{\log x}$$

Now, for  $x \in (1, e)$  the above inequality (and thus the claim) is always true, since  $\log \log x < 0$ . Thus, we may assume that  $x \geq e$ . In this case, the above is equivalent to

$$\log x \geq e \log \log x$$

Now, since this inequality is satisfied for  $x \geq e$ , the claim is established.

Given the claim, the desired result is immediate. Indeed, from the previous problem, we have that

$$\begin{aligned}
\frac{N^{1/\nu}}{\frac{1}{\nu} \log N} &= g(N^{1/\nu}) \\
&\leq \frac{3e}{\delta}
\end{aligned}$$

and thus, by the claim we just proved,

$$\begin{aligned} N^{1/\nu} &\leq \frac{e}{e-1} \frac{3e}{\delta} \log \frac{3e}{\delta} \\ \implies N &\leq \left( \frac{3e^2}{\delta(e-1)} \log \frac{3e}{\delta} \right)^\nu \end{aligned}$$

and thus, Equation 1 holds with  $K = \frac{3e^2}{e-1}$ , as desired.  $\square$

**Exercise 1.3.** We will find the covering number of ellipses in this problem. Given a collection of positive numbers  $\{\mu_j\}_{j=1}^d$ , consider the ellipse

$$\mathcal{E} = \{\theta \in \mathbb{R}^d : \sum_i \theta_i^2 / \mu_i^2 \leq 1\}$$

1. Show that

$$\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) \geq d \log \frac{1}{\epsilon} + \sum_{j=1}^d \log \mu_j$$

*Solution.*

Suppose that  $\{\theta_1, \dots, \theta_N\}$  is an  $\epsilon$ -cover of  $\mathcal{E}$ . Then, by definition,  $\mathcal{E} \subset \cup_{i=1}^N \mathcal{B}_\epsilon(\theta_i)$ , where  $\mathcal{B}_\epsilon(\theta_i) = \{\|\theta - \theta_i\|_2 \leq \epsilon : \theta \in \mathbb{R}^d\}$ . Thus, we have that

$$\begin{aligned} \text{Vol}(\mathcal{E}) &\leq \sum_{i=1}^N \text{Vol}(\mathcal{B}_\epsilon(\theta_i)) \\ &= N \text{Vol}(\mathcal{B}_\epsilon(\mathbf{0})) \end{aligned}$$

Now, let us consider the change of coordinates from points in the ellipsoid to points in the ball. Given coordinates  $\{u_i\}_{i=1}^d$  from the  $\epsilon$ -ball, we may map these coordinates in a one-to-one manner to points  $\{x_i\}_{i=1}^d$  in  $\mathcal{E}$  by the formula:

$$x_i = \frac{\mu_i}{\epsilon} u_i$$

Indeed, since by definition  $\sum_i u_i^2 \leq \epsilon^2$ , and so

$$\begin{aligned} \epsilon^2 &\geq \sum_i u_i^2 = \sum_i \frac{\epsilon^2}{\mu_i^2} x_i^2 \\ \implies \sum_i \frac{x_i^2}{\mu_i^2} &\leq 1 \end{aligned}$$

as desired. Therefore, we may compute the volume of  $\mathcal{E}$  using the change of variable formula

$$\begin{aligned} \text{Vol}(\mathcal{E}) &= \int_{\mathcal{E}} dx_1, \dots, x_n \\ &= \int_{\mathcal{B}_\epsilon(\mathbf{0})} \left| \frac{\partial(x_1, \dots, x_n)}{\partial(u_1, \dots, u_n)} \right| du_1, \dots, u_n \\ &= \int_{\mathcal{B}_\epsilon(\mathbf{0})} \left( \prod_{i=1}^d \frac{\mu_i}{\epsilon} \right) du_1, \dots, u_n \\ &= \left( \prod_{i=1}^d \frac{\mu_i}{\epsilon} \right) \text{Vol}(\mathcal{B}_\epsilon(\mathbf{0})) \end{aligned}$$

Hence,

$$\begin{aligned} \left( \prod_{i=1}^d \frac{\mu_i}{\epsilon} \right) \text{Vol}(B_\epsilon(\mathbf{0})) &= \text{Vol}(\mathcal{E}) \\ &\leq N \text{Vol}(\mathcal{B}_\epsilon(\mathbf{0})) \end{aligned}$$

and thus,

$$\begin{aligned} N &\geq \prod_{i=1}^d \frac{\mu_i}{\epsilon} \\ \implies \log N &\geq d \log \frac{1}{\epsilon} + \sum_{i=1}^d \log \mu_i \end{aligned}$$

as desired.  $\square$

2. Now consider the infinite-dimensional ellipse, specified by the sequence  $\mu_j = j^{-2\beta}$  for some parameter  $\beta > \frac{1}{2}$ . Show that

$$\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) \geq C \left( \frac{1}{\epsilon} \right)^{1/2\beta}$$

where  $\|\theta - \theta'\|_{\ell_2}^2 = \sum_{i=1}^\infty (\theta(i) - \theta'(i))^2$ .

*Solution.*

Let us denote the ellipse truncated to  $d$  dimensions as:

$$\mathcal{E}_d = \{\tilde{\theta} \in \mathbb{R}^d : \theta \in \mathcal{E}, \tilde{\theta}(i) = \theta(i) \forall i \in [d]\}$$

Let  $S = \{\theta_1, \dots, \theta_N\}$  be an  $\epsilon$ -covering of  $\mathcal{E}$ . Define  $S_d$  as the elements of  $S$  truncated to  $d$  dimensions, that is, the set of  $N$  elements  $\tilde{\theta}_i$  such that  $\tilde{\theta}_i(j) = \theta_i(j)$  for  $j \in [d]$ .

Now, we will show that  $S_d$  is an  $\epsilon$ -covering of  $\mathcal{E}_d$ . Indeed, fix any  $\tilde{\theta} \in \mathcal{E}_d$ . By definition, there is some  $\theta$  such that  $\tilde{\theta}(j) = \theta(j)$  for every  $j \in [d]$ . By definition of  $S$ , there exists some  $\theta_i$  satisfying  $\|\theta - \theta_i\|_{\ell_2} \leq \epsilon$ . Therefore,

$$\begin{aligned} \epsilon^2 &\geq \|\theta - \theta_i\|_{\ell_2}^2 \\ &= \sum_{j=1}^d (\theta(j) - \theta_i(j))^2 + \sum_{j=d+1}^\infty (\theta(j) - \theta_i(j))^2 \\ &= \sum_{j=1}^d (\tilde{\theta}(j) - \tilde{\theta}_i(j))^2 + \sum_{j=d+1}^\infty (\theta(j) - \theta_i(j))^2 \\ &\geq \sum_{j=1}^d (\tilde{\theta}(j) - \tilde{\theta}_i(j))^2 + \sum_{j=d+1}^\infty (0 - 0)^2 \\ &= \|\tilde{\theta} - \tilde{\theta}_i\|_2^2 \end{aligned}$$



and thus  $S_d$  is also an  $\epsilon$ -cover of  $\mathcal{E}_d$ . Therefore, we have that

$$\begin{aligned}
\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) &\geq \log N(\epsilon, \mathcal{E}_d, \|\cdot\|_2) \\
&\geq d \log \frac{1}{\epsilon} + \sum_{i=1}^d \log \mu_i && \text{by the previous problem} \\
&\geq d \log \frac{1}{\epsilon} - 2\beta \log d! \\
&\geq d \log \frac{1}{\epsilon} - 2\beta \log(d^{d+1/2} e^{-d+1}) && \text{by Sterling's approximation} \\
&= d \log \frac{1}{\epsilon} - 2\beta d \log d + 2\beta \left( d - 1 + \frac{1}{2} \log d \right)
\end{aligned}$$

Now, choose  $d = \left\lceil \left( \frac{1}{\epsilon} \right)^{1/2\beta} \right\rceil$ . Then the above inequality becomes

$$\begin{aligned}
\log N(\epsilon; \mathcal{E}, \|\cdot\|_2) &\geq d \log \frac{1}{\epsilon} - \underbrace{2\beta d \log \left( \left( \frac{1}{\epsilon} \right)^{1/2\beta} + 1 \right)}_{\leq \frac{1}{2\beta} \log \left( \frac{1}{\epsilon} \right) + \frac{1}{2}} + 2\beta \left( d - 1 + \frac{1}{2} \underbrace{\log d}_{\geq 0} \right) \\
&\geq \beta (d - 2) \\
&\geq C\beta d && \text{for } C < 1 \text{ small enough} \\
&\geq C\beta \left( \frac{1}{\epsilon} \right)^{1/2\beta}
\end{aligned}$$

as desired. □